INVERSE PROBLEMS AS STATISTICS

STEVEN N. EVANS AND PHILIP B. STARK

ABSTRACT. What mathematicians, scientists, engineers, and statisticians mean by "inverse problem" differs. For a statistician, an inverse problem is an inference or estimation problem. The data are finite in number and contain errors, as they do in classical estimation or inference problems, and the unknown typically is infinite-dimensional, as it is in nonparametric regression. The additional complication in an inverse problem is that the data are only indirectly related to the unknown. Canonical abstract formulations of statistical estimation problems subsume this complication by allowing probability distributions to be indexed in more-or-less arbitrary ways by parameters, which can be infinite-dimensional. Standard statistical concepts, questions, and considerations such as bias, variance, mean-squared error, identifiability, consistency, efficiency, and various forms of optimality, apply to inverse problems. This article discusses inverse problems as statistical estimation and inference problems, and points to the literature for a variety of techniques and results. It shows how statistical measures of performance apply to techniques used in practical inverse problems, such as regularization, maximum penalized likelihood, Bayes estimation, and the Backus-Gilbert method. The article generalizes results of Backus and Gilbert characterizing parameters in inverse problems that can be estimated with finite bias. It establishes general conditions under which parameters in inverse problems can be estimated consistently.

CONTENTS

Date: 31 August 2001; updated 25 February 2002. Technical Report 609, Department of Statistics, U.C. Berkeley.

Key words and phrases. Inverse problems, statistics, decision theory, ill-posed problem, regularization, inference, constraints, Backus-Gilbert, consistency.

1. Introduction

This paper casts inverse problems as statistical estimation and inference problems. It was written to introduce some standard statistical ideas and approaches to the Inverse Problems community. It is mostly expository, but Section 2.4 contains new results concerning consistent estimation in linear inverse problems.

In *forward problems* in statistics, one has a class of possible descriptions of the world, and a forward operator that maps each description into a probability measure for the observables. The data consist of a realization of (*i.e.*, a sample from) the probability measure. The probability measure tells the whole story: it captures any stochastic variability in the "truth," contamination by measurement error, systematic error, *etc.* We refer to each possible description of the world as a *model*. Applied mathematicians generally write forward problems as a composition of steps: (a) transforming the correct description of the world into ideal, noise-free, infinite-dimensional data ("physics"), (b) censoring the ideal data to retain only a finite list of numbers, because we can only measure, record, and compute with such lists, and (c) possibly corrupting the list with deterministic measurement error. This sequential procedure is equivalent to a single-step procedure in which the corruption (c) is on a par with the physics (a), and the mapping yields only the actual observables, incorporating the censoring (b). The probability distribution of the observables is degenerate if the observational error is deterministic. Hence, the statistical framework for forward problems is at least as general as that of applied mathematics: Forward problems of applied mathematics are instances of statistical forward problems.

Typically, the class of models is indexed by a set Θ with some special structure. For example, Θ could be a convex subset of a separable Banach space $\mathcal{T}$. For convenience, we refer to Θ as the class of possible models, and to the model with index $\theta \in \Theta$ as the model θ. The forward mapping is then $\theta \mapsto \mathbb{P}_\theta$, the mapping from (the index of) the model to a probability distribution for the observables. The index θ generally has a physical significance that gives the forward mapping $\theta \mapsto \mathbb{P}_\theta$ reasonable analytic properties, *e.g.*, continuity. The class of possible models is denoted $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. The forward problem is *linear* if $\mathbb{P}_\theta$ is the probability distribution of $K\theta + \epsilon$, where K is a linear operator and ϵ is a random variable whose distribution does not depend on θ. A *parameter* of a model θ is the value $g(\theta)$ at θ of a function g defined on Θ; the function g could be the identity—and often is.

In *inverse problems*, we observe data X drawn from the probability distribution $\mathbb{P}_\theta$ for some unknown $\theta \in \Theta$; we want to use X and the knowledge that $\theta \in \Theta$ to learn about θ, for example, to estimate a parameter $g(\theta)$. In essence, the goal of inverse problems is to invert partially the forward operator. We shall always assume that Θ contains at least two points; otherwise, there is no problem to solve—there is only one possible value of $g(\theta)$, and data are superfluous. The differences between how applied mathematicians and how statisticians view inverse problems center on the number of observations, whether the observations are contaminated, the nature of such contamination, and the questions whose answers are interesting. For example, to a statistician, the number of data is finite—although the behavior of the problem as the number of data grows is investigated frequently—and the data contain errors that are modeled at least in part as stochastic. Bias, variance, identifiability, consistency, and similar notions figure prominently; emphasis is on estimation and inference. Applied mathematicians often are more interested in existence, uniqueness, and construction of a solution consistent with an infinite number of ideal noise-free data, and stability of the solution when the data are contaminated by a deterministic disturbance.

The two viewpoints are related. For example, identifiability—distinct models θ yield distinct probability distributions $\mathbb{P}_\theta$ for the observables—is similar to uniqueness—the forward operator maps at most one model into the observed data. Consistency (the parameter can be estimated with arbitrary accuracy as the number of data grows) is related to stability of a recovery algorithm (small changes in the data produce small changes in the recovered model), because consistency essentially requires that arbitrarily small changes in the model produce detectable changes in the data. There are also quantitative connections between the two points of view. For example, statistical measures of the difficulty estimating a linear functional of an element of a Hilbert space from observations of linear functionals of the element contaminated by Gaussian errors can be calculated by scaling the error bound given by the theory of optimal recovery of a linear functional from linear data corrupted maliciously [19]. Many of the tools used to study inverse problems in statistics and applied mathematics are the same, too: functional analysis, convex analysis, optimization theory, nonsmooth analysis, approximation theory, harmonic analysis, and measure theory.

This paper is organized as follows. Sections 2.1, 2.2 and 2.3 summarize a standard abstract statistical framework that subsumes many inverse problems. Section 2.4 studies the possibility of *consistent* estimates in inverse problems, and gives necessary conditions and sufficient conditions involving the class Θ of possible models, the forward operator, and the observational errors. Section 2.5 introduces a simple example and shows how the ideas developed previously in Section 2 apply to this example. Section 3.1 introduces some ideas and notation from statistical decision theory. Section 3.2 applies these ideas to estimation, and presents some common loss functions used to compare estimators and to define what it means for an estimator to be optimal. Section 3.3 does the same thing for confidence sets. Section 4 examines some estimators and confidence sets used in statistics and inverse problems, including the Backus-Gilbert method, Bayes estimation, maximum likelihood and some of its variants involving penalization and regularization (such as stochastic inversion and the method of sieves), shrinkage estimators (including wavelet and wavelet-vaguelette shrinkage), and strict bounds. Appendix A presents some results from probability theory used elsewhere in the paper, and Appendix B is a brief refresher on measure-theoretic probability as used in the paper.

2. Identifiability and Consistency in Inverse Problems

We re-state a canonical inverse problem as a statistical inference problem. Let Θ be a non-empty subset of a separable Banach space $\mathcal{T}$. The set Θ represents possible "theories" about the state of nature—the competing models θ that might spawn the observations X. We allow X to take values in any separable metric space $\mathcal{X}$, but in our examples, $\mathcal{X} = \mathbb{R}^n$ and $X = \{X_j\}_{j=1}^n$. Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a collection of probability measures defined on a common σ-algebra $\mathcal{F}$ on $\mathcal{X}$. If theory $\theta \in \Theta$ is true, the probability distribution of the data X is $\mathbb{P}_\theta$; that is, $X \sim \mathbb{P}_\theta$. Each possible value of $\theta \in \Theta$ induces a probability distribution for X; many different values of θ might yield the same probability distribution for X. In the language of applied mathematics, this is the non-uniqueness problem; in the language of statistics, it is non-identifiability of θ. We discuss identifiability in more detail below.

One of the theories $\theta \in \Theta$ is true—in fact, $X \sim \mathbb{P}_\theta$. We do not know the value of θ, only that it is an element of Θ. We wish to learn something about θ from X. The quadruple $(\Theta, \mathcal{P}, \mathcal{X}, \mathcal{F})$ is a *statistical experiment indexed by* Θ; see [44]. It is our mathematical model for inverse problems in the most general setting.

Parameters are features of θ we might wish to learn about. For our purposes, a parameter is the value at θ of a continuous mapping $g : \Theta \to \mathcal{G}$, where $\mathcal{G}$ is a separable metric space. (Restricting attention to continuous mappings is not always necessary or desirable; see, *e.g.*, [17].) We insist further that $g(\Theta)$ (and thus Θ) contain at least two points—if g were constant on Θ we would know $g(\theta)$ perfectly before observing any data. If we sought to estimate θ in its entirety, we might take $\mathcal{G} = \mathcal{T}$. We might instead be interested in a low-dimensional projection of θ, the norm of θ, or some other function or functional.

We try to distinguish between characteristics of the statistical experiment, and characteristics of methods used to draw inferences in the statistical experiment. One of the most fundamental statistical properties a parameter is identifiability.

A parameter $g(\theta)$ is *identifiable* if for all θ_1, $\theta_2 \in \Theta$,

$$\{g(\theta_1) \neq g(\theta_2)\} \Longrightarrow \{\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}\}. \tag{2.1}$$

That is, a parameter is identifiable if a change in the parameter always is accompanied by a change in the probability distribution of the data. In most inverse problems, $g(\theta) = \theta$ is not identifiable: this is essentially the problem of nonuniqueness. Moreover, in most linear inverse problems (defined below) most linear functionals of θ are not identifiable; see Theorem 2.6. We present some results on indentifiability of parameters in linear inverse problems in Section 2.3

2.1. **Estimators.** We have said that we seek to learn about $g(\theta)$ from X, but we have said nothing about the tools we can use. This section characterizes the kind of tools we shall consider, and gives minimal conditions they must satisfy.

A *(randomized) decision rule*

$$\begin{aligned} \delta : \mathcal{X} &\rightarrow \mathcal{M}_1(\mathcal{A}) \\ x &\mapsto \delta(x)(\cdot), \end{aligned}$$

is a measurable mapping from $\mathcal{X}$ to the collection $\mathcal{M}_1(\mathcal{A})$ of probability distributions on a separable metric space $\mathcal{A}$ of *actions*, where the probability distributions are defined on a sub-σ-algebra of the Borel σ-algebra on $\mathcal{A}$. (A mapping δ from $\mathcal{X}$ to the collection of measures on $\mathcal{A}$ is measurable if for every $\theta \in \Theta$, $\delta(x)(A)$ is a $\mathbb{P}_\theta$-measurable function of x for every Borel set $A \subseteq \mathcal{A}$.)

A *non-randomized decision rule* is a randomized decision rule that, to each $x \in \mathcal{X}$, assigns a unit point mass at a value $a = a(x) \in \mathcal{A}$. It is useful (often in proofs and occasionally in practice) to consider randomized decision rules; in effect, they are convex combinations of non-randomized rules. For ease of notation, a non-randomized decision rule will often be written as an $\mathcal{A}$-valued function rather than a $\mathcal{M}_1(\mathcal{A})$-valued one.

An (non-randomized) *estimator* of a parameter $g(\theta)$ is a (non-randomized) decision rule for which the space $\mathcal{A}$ of possible actions is the space $\mathcal{G}$ of possible parameter values. A common notation for an estimator of a parameter $g(\theta)$ is $\hat{g}$. In keeping with our convention above, a non-randomized estimator will often be written as an $\mathcal{G}$-valued function rather than a $\mathcal{M}_1(\mathcal{G})$-valued one.

To see how a randomized estimator might arise, consider the following example. We wish to estimate the probability p that a given coin lands heads when it is tossed. We know *a priori* that either $p = 1/3$ or $p = 2/3$. We toss the coin 10 times and observe X, the number of times the coin lands heads. A reasonable estimator $\hat{p}(X)$ might be defined as follows: let W be a random variable that equals 0 with probability 1/2 and equals 1 with probability 1/2, independent of X. Define

$$\hat{p}(X) = \begin{cases} 1/3, & X < 5, \\ 1/3, & X = 5 \text{ and } W = 0, \\ 2/3, & X = 5 \text{ and } W = 1, \\ 2/3, & X > 5. \end{cases} \tag{2.2}$$

This estimator only returns possible values for p, but in effect tosses a fair coin to decide which of the two possible values to use as the estimate when the data do not favor either. See [47] §5.1 for more motivation for randomized estimators.

2.1.1. *Mean distance error and bias.* There are many common measures of the performance of estimators; we shall see several in Section 3, but two of the simplest are mean distance error and bias. For simplicity, we restrict attention to non-randomized estimators. Let $d_{\mathcal{G}}(\cdot,\cdot)$ denote the metric on $\mathcal{G}$. The *mean distance error* at θ of the estimator $\hat{g}$ of the parameter $g(\theta)$,

$$\mathrm{MDE}_\theta(\hat{g}, g) = \mathbb{E}_\theta[d(\hat{g}, g(\theta))], \tag{2.3}$$

is the expected value of the distance between the estimator and the parameter when the model is θ. Because the space $\mathcal{G}$ of parameter values is a metric space and the metric takes values in $\mathbb{R}^+$, the mean distance error is always well defined. When the metric derives from a norm, the mean distance error is called the mean norm error (MNE). When $\mathcal{G}$ is a Hilbert space with norm $\|\cdot\|$, the mean squared error (MSE) is

$$\mathrm{MSE}_\theta(\hat{g}, g) = \mathbb{E}_\theta\left[\|\hat{g} - g(\theta)\|^2\right]. \tag{2.4}$$

When $\mathcal{G}$ is a Banach space, we define the *bias* at θ of $\hat{g}$ to be

$$\mathbf{bias}_\theta(\hat{g}, g) = \mathbb{E}_\theta[\hat{g} - g(\theta)] \tag{2.5}$$

when the expectation is well-defined. If $\mathbf{bias}_\theta(\hat{g}, g) = 0$, we say that $\hat{g}$ *is unbiased at* θ (for g). If $\hat{g}$ is unbiased at θ for g for every $\theta \in \Theta$, we say $\hat{g}$ is *unbiased* (for g).

Remark 2.1. If, in an inverse problem, there is some estimator $\hat{g}$ that is unbiased for g (in which case g is said to be *U-estimable*), then g is certainly identifiable.

Let $\bar{g}_\theta \equiv \mathbb{E}_\theta[\hat{g}]$. Then $\mathbf{bias}_\theta(g) = \bar{g}_\theta - g(\theta)$. When $\mathcal{G}$ is Hilbertian, we define the variance of the estimator $\hat{g}$ to be

$$\mathbf{Var}_\theta(g) \equiv \mathbb{E}_\theta[\|\hat{g} - \bar{g}_\theta\|^2]. \tag{2.6}$$

We can use the projection theorem to decompose the mean squared error into a sum of two terms, the variance of $\hat{g}$ and the square of the norm of the bias of $\hat{g}$. That is,

$$\begin{aligned} \mathbb{E}_\theta[\|\hat{g} - g(\theta)\|^2] &= \mathbb{E}_\theta[\|\hat{g} - \bar{g}_\theta\|^2] + \|\bar{g}_\theta - g(\theta)\|^2 \\ &= \mathbf{Var}_\theta(\hat{g}) + \|\mathbf{bias}_\theta(\hat{g})\|^2. \end{aligned} \tag{2.7}$$

Mean distance error and mean squared error are examples of *risk functions*, discussed in more detail in Section 3.

2.1.2. *Sequences of estimators and consistency.* Consistency has to do with the behavior of sequences of estimators for a sequence of inverse problems. While it is possible to consider a very general situation in which there is a different parameter in each of the inverse problems, we will restrict attention the most natural situation: the n^{th} inverse problem in our sequence has its own data space $\mathcal{X}_n$, but all the problems have the same index space Θ and the same parameter g. Moreover, the sequence of problems is nested in the sense that $\mathcal{X}_m$ is a Cartesian factor of $\mathcal{X}_n$ for $m \le n$ (for example, $\mathcal{X}_m = \mathbb{R}^m$ and $\mathcal{X}_n = \mathbb{R}^n$) and the probability measure $\mathbb{P}_{\theta,m}$ governing the data for the m^{th} problem is the $\mathcal{X}_m$ marginal of the probability measure $\mathbb{P}_{\theta,n}$ on $\mathcal{X}_n$ governing the data for the n^{th} problem. Thus, the different problems in the sequence differ only in how many data are available. The sequence of estimators is consistent if for any parameter value the estimated value of the parameter converges to the true value, in a sense made precise below, as more data are used.

In real problems, the number of data is finite and frequently fixed. However, it is often possible, at least notionally, to embed a particular problem within an hypothetical hierarchical sequence in which more experiments of a similar type are conducted or more measurements are made. It is then comforting to know that if one could collect more data, the parameter could be estimated with arbitrary precision.

Notation 2.2. We write $d_{\mathcal{G}}$ for the metric on $\mathcal{G}$.

Definition 2.3. Given a nested sequence of forward problems $\{\{\mathbb{P}_{\theta,n} : \theta \in \Theta\}\}_{n=1}^{\infty}$ and a parameter $g : \Theta \to \mathcal{G}$, a sequence of non-randomized estimators $\{\hat{g}_n\}_{n=1}^{\infty}$, $g_n : \mathcal{X}_n \to \mathcal{G}$ is *consistent* (for g) if for every $\theta \in \Theta$ and every neighborhood U of $g(\theta) \in \mathcal{G}$,

$$\lim_{n\to\infty} \mathbb{P}_{\theta,n}\{\hat{g} \notin U\} = 0. \tag{2.8}$$

A parameter g is *consistently estimable* if there exists a sequence of estimators that is consistent for g. If g is the identity mapping and $g(\theta) = \theta$ is consistently estimable in some topology on Θ (not necessarily the norm topology inherited from $\mathcal{T}$), we say that the model is consistently estimable.

2.2. **The Linear Forward Problem.** Linear forward problems are a special subclass of the general forward problems defined above. *Linearity* refers to linear structure on the set Θ of possible models, the set $\mathcal{X}$ of possible data, and the set $\mathcal{G}$ of parameter values. In linear forward problems, the forward operator also possesses a type of linearity, which we clarify after introducing some notation.

Notation 2.4. Let $\mathcal{T}$ be a separable Banach space. Then $\mathcal{T}^*$ (resp. $\mathcal{T}^{**}$) denotes the normed dual (resp. normed second dual) of $\mathcal{T}$, and the pairing between $\mathcal{T}^*$ and $\mathcal{T}$ (resp. between $\mathcal{T}^{**}$ and $\mathcal{T}^*$) is denoted by $\langle\cdot,\cdot\rangle : \mathcal{T}^* \times \mathcal{T} \to \mathbb{R}$ (resp. $\langle\langle\cdot,\cdot\rangle\rangle : \mathcal{T}^{**} \times \mathcal{T}^* \to \mathbb{R}$). The norms on $\mathcal{T}$, $\mathcal{T}^*$ and $\mathcal{T}^{**}$ are denoted $\|\cdot\|$, $\|\cdot\|_*$, and $\|\cdot\|_{**}$, respectively.

Definition 2.5. A forward problem is *linear* if

(1) Θ is a subset of a separable Banach space $\mathcal{T}$

(2) For some fixed sequence $\{\kappa_j\}_{j=1}^n$ of elements of $\mathcal{T}^*$, the datum $X = \{X_j\}_{j=1}^n$, where

$$X_j = \langle \kappa_j, \theta\rangle + \epsilon_j, \quad \theta \in \Theta, \tag{2.9}$$

and $\epsilon = \{\epsilon_j\}_{j=1}^n$ is a vector of stochastic errors whose probability distribution does not depend on θ. (Thus $\mathcal{X} = \mathbb{R}^n$.)

The functionals $\{\kappa_j\}_{j=1}^n$ are the "representers" or "data kernels" of the linear forward problem. The distribution $\mathbb{P}_\theta$ of the Introduction is the probability distribution of X, and $\mathcal{P}$ is the set of all such distributions as θ ranges over Θ. Typically, $\dim(\Theta) = \infty$; at the very least, $n < \dim(\Theta)$, so estimating θ is an underdetermined problem.

Define

$$\begin{aligned} K : \Theta &\to \mathbb{R}^n \\ \theta &\mapsto \{\langle \kappa_j, \theta \rangle\}_{j=1}^n . \end{aligned} \tag{2.10}$$

We often abbreviate Equation (2.9) by

$$X = K\theta + \epsilon, \quad \theta \in \Theta. \tag{2.11}$$

Using data $X = K\theta + \epsilon$ and the knowledge that $\theta \in \Theta$ to estimate or draw inferences about a parameter $g(\theta)$ is a *linear inverse problem*. In linear inverse problems, the probability distribution of the data X depends on the model θ only through $K\theta$, so if there are two points θ_1, $\theta_2 \in \Theta$ such that $K\theta_1 = K\theta_2$ but $g(\theta_1) \neq g(\theta_2)$, then $g(\theta)$ is not identifiable. We proceed to study some conditions on K, Θ, and g that control whether $K\theta$ determines $g(\theta)$ on Θ.

2.3. **Identifiability of Linear Parameters in Linear Inverse Problems.** Consider a linear forward problem with $\#(\Theta) \geq 2$. Let $\{g_i\}_{i=1}^m$ be a collection of (not necessarily bounded) functionals that are linear on Θ: for a_1, $a_2 \in \mathbb{R}$ and θ_1, $\theta_2 \in \Theta$, $g(a_1\theta_1 + a_2\theta_2) = a_1 g(\theta_1) + a_2 g(\theta_2)$ when $a_1\theta_1 + a_2\theta_2 \in \Theta$. This subsection addresses estimating the *linear parameter vector*

$$g(\theta) \equiv \{g_i(\theta)\}_{i=1}^m \tag{2.12}$$

from data $X = K\theta + \epsilon$.

An example of such a problem is that of estimating a finite collection of spherical harmonic coefficients of Earth's geomagnetic field at the core-mantle boundary from satellite measurements of the field (neglecting sources in the atmosphere, crust, and mantle). In that case, $\mathcal{T}$ is a weighted ℓ_2 space, and Θ is a ball in $\mathcal{T}$. See, *e.g.*, [10, 36]. Linearized travel-time tomography can be cast this way as well; in that problem, $\mathcal{T}$ might be the space of functions of bounded variation, and Θ might be the positive cone in $\mathcal{T}$. Similarly, inverting instances of Abel's equation that arise in seismology and helioseismology can be written in this form, taking $\mathcal{T}$ to be functions of bounded variation, and Θ to be a hyperrectangle (see, *e.g.*, [54]).

Linear functions composed of linear combinations of the data kernels in linear forward problems play a special role in the estimation of parameters. Let Λ be an $m \times n$ matrix with real elements λ_{ij}. We define

$$\begin{aligned} \Lambda \cdot K : \mathcal{T} &\to \mathbb{R}^m \\ t &\mapsto \left(\sum_{j=1}^n \lambda_{1j} \langle \kappa_j, t \rangle, \sum_{j=1}^n \lambda_{2j} \langle \kappa_j, t \rangle, \cdots, \sum_{j=1}^n \lambda_{mj} \langle \kappa_j, t \rangle \right). \end{aligned} \tag{2.13}$$

The following necessary condition for a real-valued parameter to be identifiable extends a theorem of Backus and Gilbert [5]. It addresses parameters that are somewhat more general than the linear parameters just described. Note that a vector-valued parameter is identifiable if and only if each of its components is identifiable, so it suffices to consider real-valued parameters. Recall from Lemma A.4 that if Y is a random n-vector and $a \in \mathbb{R}$, $a \neq 0$, then the probability distribution of Y differs from that of $a + Y$; thus in a linear inverse problem, $K\theta_1 \neq K\theta_2$ iff $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$. It follows that a parameter g is identifiable iff $g(\theta_1) \neq g(\theta_2)$ implies $K\theta_1 \neq K\theta_2$ whenever θ_1, $\theta_2 \in \Theta$.

Theorem 2.6. *Let $g : \Theta \to \mathbb{R}$ be an identifiable real-valued parameter. Suppose there exists a non-empty symmetric convex set $\bar{\Theta} \subseteq \mathcal{T}$ such that:*

i) $\bar{\Theta} \subseteq \Theta$,

ii) $g(-\theta) = -g(\theta)$, $\theta \in \bar{\Theta}$,

iii) $g(a_1\theta_1 + a_2\theta_2) = a_1 g(\theta_1) + a_2 g(\theta_2)$, $\theta_1, \theta_2 \in \bar{\Theta}$, $a_1, a_2 \geq 0$, $a_1 + a_2 = 1$, *and*

iv) $\sup_{\theta \in \bar{\Theta}} |g(\theta)| < \infty$.

Then there is a $1 \times n$ matrix Λ such that the restriction of g to $\bar{\Theta}$ is the restriction of $\Lambda \cdot K$ to $\bar{\Theta}$.

Proof. We may suppose without loss of generality that $\bar{\Theta} = \Theta$. By replacing $\mathcal{T}$ by the closed subspace of $\mathcal{T}$ spanned by Θ, we may further suppose without loss of generality that the closed subspace spanned by Θ is all of $\mathcal{T}$. Then g is the restriction to Θ of a continuous linear functional on $\mathcal{T}$, which we will also denote by g.

Suppose that no such matrix Λ exists. Then g is not a linear combination of the functions $\kappa_1, \ldots, \kappa_n$. A continuity argument shows that there must exist a finite set $T \subset \Theta$ such that g restricted to T is not a linear combination of the functions $\kappa_1, \ldots, \kappa_n$ restricted to T. That is, the finite-dimensional vector $\{\langle g, \theta \rangle\}_{\theta \in T}$ is not a linear combination of the vectors $\{\langle \kappa_i, \theta \rangle\}_{\theta \in T}$, $1 \leq i \leq n$. By standard finite-dimensional linear algebra, there exist constants $\{a_\theta\}_{\theta \in T}$ such that $\sum_{\theta \in T} a_\theta \langle g, \theta \rangle \neq 0$ and $\sum_{\theta \in T} a_\theta \langle \kappa_i, \theta \rangle = 0$, $1 \leq i \leq n$. Furthermore, by replacing $\{a_\theta\}_{\theta \in T}$ by $\{\gamma a_\theta\}_{\theta \in T}$ for some γ sufficiently small, we may suppose that $\sum_{\theta \in T} a_\theta \theta \in \Theta$.

Now observe that

$$g\left(\sum_{\theta \in T} a_\theta \theta \in \Theta\right) = \sum_{\theta \in T} a_\theta \langle g, \theta \rangle \neq 0 = g(0), \tag{2.14}$$

whereas

$$K\left(\sum_{\theta \in T} a_\theta \theta \in \Theta\right) = \left(\sum_{\theta \in T} a_\theta \langle \kappa_i, \theta \rangle\right)_{i=1}^{n} = 0 = K(0), \tag{2.15}$$

contradicting identifiability. $\square$

Theorem 2.6 generalizes somewhat; for example: Suppose there exist $\theta_0 \in \Theta$, a symmetric convex set $\bar{\Theta} \subseteq \mathcal{T}$, a constant $c \in \mathbb{R}$, and a mapping $\bar{g} : \bar{\Theta} \to \mathbb{R}$ such that:

i) $\theta_0 + \bar{\Theta} \subseteq \Theta$
ii) $g(\theta_0 + \bar{\theta}) = c + \bar{g}(\bar{\theta})$, $\bar{\theta} \in \bar{\Theta}$
iii) $\bar{g}(-\bar{\theta}) = -\bar{g}(\bar{\theta})$, $\bar{\theta} \in \bar{\Theta}$,
iv) $\bar{g}(a_1\bar{\theta}_1 + a_2\bar{\theta}_2) = a_1\bar{g}(\bar{\theta}_1) + a_2\bar{g}(\bar{\theta}_2)$, $\bar{\theta}_1, \bar{\theta}_2 \in \bar{\Theta}$, $a_1, a_2 \geq 0$, $a_1 + a_2 = 1$, and
v) $\sup_{\bar{\theta} \in \bar{\Theta}} |\bar{g}(\bar{\theta})| < \infty$.

Then there is a $1 \times n$ matrix Λ such that the restriction of g to $\bar{\Theta} + \theta_0$ is the restriction of $\Lambda \cdot K(\cdot - \theta_0) + c$ to $\bar{\Theta} + \theta_0$.

Theorem 2.6 gives a necessary condition for identifiability. Here is a corresponding sufficient condition.

Theorem 2.7. *Suppose that $g = \{g_i\}_{i=1}^m$ is an $\mathbb{R}^m$-valued parameter that can be written as the restriction to Θ of $\Lambda \cdot K$ for some $m \times n$ matrix Λ. Then g is identifiable. Moreover, if $\mathbb{E}[\epsilon] = 0$, then the statistic $\Lambda \cdot X$ is an unbiased estimator of g. If, in addition, ϵ has covariance matrix $\Sigma = \mathbb{E}[\epsilon\epsilon^T]$, then the covariance matrix of $\Lambda \cdot X$ is $\Lambda \cdot \Sigma \cdot \Lambda^T$ under any $\mathbb{P}_\theta$.*

Proof. The identifiability of g is immediate from Lemma A.4.

Suppose that $\mathbb{E}[\epsilon] = 0$. We compute:

$$\begin{aligned}
\mathbb{E}_\theta[\Lambda \cdot X] &= \mathbb{E}_\theta[\Lambda \cdot K\theta + \Lambda \cdot \epsilon] \\
&= \mathbb{E}_\theta[\Lambda \cdot K\theta] + \mathbb{E}_\theta[\Lambda \cdot \epsilon] \\
&= \Lambda \cdot K\theta + \Lambda \cdot \mathbb{E}_\theta[\epsilon] \\
&= \Lambda \cdot K\theta \\
&= g(\theta), \qquad (2.16)
\end{aligned}$$

so $\Lambda \cdot X$ is an unbiased estimator of $g(\theta)$. Suppose, in addition, that ϵ has covariance matrix Σ. We compute:

$$\begin{aligned}
\mathbf{Cov}_\theta(\Lambda \cdot X) &= \mathbb{E}\left[\Lambda \cdot \epsilon \cdot \epsilon^T \cdot \Lambda^T\right] \\
&= \Lambda \cdot \Sigma \cdot \Lambda^T. \qquad (2.17)
\end{aligned}$$

□

Corollary 2.8 (The fundamental theorem of Backus and Gilbert). *Let $\mathcal{T}$ be a Hilbert space; let $\Theta = \mathcal{T}$; let $g \in \mathcal{T} = \mathcal{T}^*$ be a linear parameter; and let $\{\kappa_j\}_{j=1}^n \subseteq \mathcal{T}^*$. The parameter $g(\theta)$ is identifiable iff $g = \Lambda \cdot K$ for some $1 \times n$ matrix Λ. In that case, if $\mathbb{E}[\epsilon] = 0$, then $\hat{g} = \Lambda \cdot X$ is unbiased for g. If, in addition, ϵ has covariance matrix $\Sigma = \mathbb{E}[\epsilon\epsilon^T]$, then the MSE of $\hat{g}$ is $\Lambda \cdot \Sigma \cdot \Lambda^T$.*

2.4. **Consistency in Linear Inverse Problems.** This subsection derives, in a fairly general setting, necessary and sufficient conditions for the model θ in a linear inverse problem to be consistently estimable.

We assume in this subsection that the observational error ϵ of Equation 2.11 is an n-vector of independent and identically distributed real-valued random variables with common distribution μ. No moment conditions on μ are required for the results here.

Whether the entire model can be estimated consistently depends on the space of models considered, the prior constraints Θ on the model within that space, the functionals K that are observed, and the probability distribution of the observational errors. Our results will be framed in terms of a suitable definition of the "size" of the set of probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$.

Definition 2.9. Let μ_a, $a \in \mathbb{R}$, denote the push-forward of μ under the map $x \mapsto x + a$. That is, $\mu_a(B) = \mu(B - a)$. Define a metric δ on $\mathbb{R}$ by letting $\delta(a, b)$ be the Hellinger distance

$$\begin{aligned} 0 \leq \delta(a,b) &\equiv \left\{\frac{1}{2}\int(\sqrt{d\mu_a} - \sqrt{d\mu_b})^2\right\}^{\frac{1}{2}} \\ &= \left\{\frac{1}{2}\int\left(\sqrt{\frac{d\mu_a}{d(\mu_a+\mu_b)}} - \sqrt{\frac{d\mu_b}{d(\mu_a+\mu_b)}}\right)^2 d(\mu_a+\mu_b)\right\}^{\frac{1}{2}} \\ &\leq 1 \end{aligned} \tag{2.18}$$

between the measures μ_a and μ_b. The metric $\delta(a, b)$ is translation invariant (it depends on a and b only through $|a - b|$).

Definition 2.10. Given $\epsilon > 0$, an ϵ*–net* for a metric space (S, ρ) is a subset $R \subseteq S$ such that for each $s \in S$, $\rho(r, s) < \epsilon$ for some $r \in R$. The metric space (S, ρ) is said to be *totally bounded* if it has a finite ϵ–net for each $\epsilon > 0$. Compactness always implies total boundedness, and the converse implication holds for complete metric spaces (but not in general).

Notation 2.11. Given a strictly positive sequence of constants $\{C_n\}_{n=1}^{\infty}$ define pseudo–metrics d_n, $n \in \mathbb{N}$, on $\mathcal{T}$ by setting

$$d_n(x', x'') = \left\{\frac{1}{C_n}\sum_{i=1}^{n}\delta^2(\langle\kappa_i, x'\rangle, \langle\kappa_i, x''\rangle)\right\}^{\frac{1}{2}}. \tag{2.19}$$

Theorem 2.12. *Suppose that* $\lim_n C_n = \infty$*, that there is a countable collection of subsets* $\Theta_1 \subseteq \Theta_2 \ldots \subseteq \Theta$ *such that* $\Theta = \bigcup_h \Theta_h$*, and* d_n *converges uniformly on each set* $\Theta_h \times \Theta_h$ *to a metric* d *on* Θ*. Suppose further that each set* Θ_h *is totally bounded with respect to* d*. Then the model is consistently estimable in the* d*–topology.*

Proof. Suppose to begin with that d_n converges uniformly to a metric d on Θ and that Θ is totally bounded with respect to d. For $k \in \mathbb{N}$, let $\{\theta_{k,1}, \ldots, \theta_{k,K_k}\}$ be a finite 2^{-k}–net for Θ equipped with d.

By a result of Birgé (see Prop 3, §16.4 of [44]), there exist numbers $a > 0$ and $b > 0$ such that for each $n \in \mathbb{N}$ and pair $(k, \ell'), (k, \ell'')$ we have a $\{0, 1\}$–valued function $\psi_{n,k,\ell',\ell''}$ on $\mathbb{R}^n$ with the property that

$$\begin{aligned}\inf\{\mathbb{P}_\theta\{\psi_{n,k,\ell',\ell''}(X_1,\dots,X_n)=1\} : d_n(\theta,\theta_{k,\ell'}) \le a d_n(\theta_{k,\ell'},\theta_{k,\ell''})\}\\ \ge 1-\exp(-bC_n d_n^2(\theta_{k,\ell'},\theta_{k,\ell''}))\end{aligned} \tag{2.20}$$

and

$$\begin{aligned}\sup\{\mathbb{P}_\theta\{\psi_{n,k,\ell',\ell''}(X_1,\dots,X_n)=1\} : d_n(\theta,\theta_{k,\ell''}) \le a d_n(\theta_{k,\ell'},\theta_{k,\ell''})\}\\ \le \exp(-bC_n d_n^2(\theta_{k,\ell'},\theta_{k,\ell''})).\end{aligned} \tag{2.21}$$

For each k choose $N_k \in \mathbb{N}$ such that such that if $n \ge N_k$, then

$$d_n(\theta_{k,\ell'},\theta_{k,\ell''}) \ge \frac{1}{2} d(\theta_{k,\ell'},\theta_{k,\ell''}), \quad 1 \le \ell' \ne \ell'' \le K_k. \tag{2.22}$$

For $1 \le \ell' \le K_k$ write

$$L_k(\ell') = \{\ell'' : d(\theta_{k,\ell'},\theta_{k,\ell''}) \ge a^{-1}2^{-(k-2)}\}. \tag{2.23}$$

Set

$$\chi_{k,n,\ell'} = \prod_{\ell''\in L_k(\ell')} \psi_{n,k,\ell',\ell''}(X_1,\dots X_n), \tag{2.24}$$

where the product is defined to 1 if $L_k(\ell') = \emptyset$.

By construction, if $n \ge N_k$ and $d_n(\theta,\theta_{k,\ell'}) < 2^{-(k-1)}$, then

$$d_n(\theta,\theta_{k,\ell'}) < 2^{-(k-1)} \le \frac{a}{2} d(\theta_{k,\ell'},\theta_{k,\ell''}) \le a d_n(\theta_{k,\ell'},\theta_{k,\ell''}), \quad \ell'' \in L_k(\ell'), \tag{2.25}$$

and hence

$$\begin{aligned}\mathbb{P}_\theta\{\chi_{k,n,\ell'}=1\} &\ge 1-\sum_{\ell''\in L_k(\ell')} \exp(-bC_n d_n^2(\theta_{k,\ell'},\theta_{k,\ell''}))\\ &\ge 1-\sum_{\ell''\in L_k(\ell')} \exp(-bC_n 2^{-2} d^2(\theta_{k,\ell'},\theta_{k,\ell''}))\\ &\ge 1-K_k\exp(-bC_n a^{-2}2^{-2(k-1)}).\end{aligned} \tag{2.26}$$

Moreover, for any $\ell'' \in L_k(\ell')$ we have

$$\begin{aligned}\mathbb{P}_\theta\{\chi_{k,n,\ell''}=1\} &\le \exp(-bC_n d_n^2(\theta_{k,\ell'},\theta_{k,\ell''}))\\ &\le \exp(-bC_n 2^{-2} d^2(\theta_{k,\ell'},\theta_{k,\ell''}))\\ &\le \exp(-bC_n a^{-2}2^{-2(k-1)}).\end{aligned} \tag{2.27}$$

If $\{1 \le \ell \le K_k : \chi_{k,n,\ell} = 1\}$ is empty, let $\hat{\theta}_{k,n}$ be an arbitrary point $\theta_0 \in \Theta$. Otherwise, set $\hat{\theta}_{k,n} = \theta_{k,p(k,n)}$, where $p(k,n) = \min\{1 \le \ell \le K_k : \chi_{k,n,\ell} = 1\}$. Consider $\theta \in \Theta$ and choose $\theta_{k,\ell'}$ such that $d(\theta, \theta_{k,\ell'}) < 2^{-k}$. By the above

$$
\begin{aligned}
\mathbb{P}_\theta\{d(\hat{\theta}_{k,n}, \theta) > a^{-1}2^{-(k-2)} + 2^{-k}\} \le \mathbb{P}_\theta &\left(\{\chi_{k,n,\ell'} = 0\} \cup \bigcup_{\ell'' \in L_k(\ell')} \{\chi_{k,n,\ell''} = 1\}\right) \\
&+ \mathbf{1}\{d_n(\theta, \theta_{k,\ell'}) \ge 2^{-(k-1)}\} \\
&\le 2K_k \exp(-bC_n a^{-2} 2^{-2(k-1)}) \\
&+ \mathbf{1}\{d_n(\theta, \theta_{k,\ell'}) - d(\theta, \theta_{k,\ell'}) > 2^{-(k-1)} - 2^{-k}\}.
\end{aligned}
\tag{2.28}
$$

Now define $1 = n_1 < n_2 < \cdots \in \mathbb{N}$ inductively by

$$
\begin{aligned}
n_{k+1} = \min\{n > n_k : 2K_{k+1} &\exp(-bC_n a^{-2} 2^{-2k}) \le 2^{-(k+1)}, \\
&\sup_{\theta', \theta''} |d_n(\theta', \theta'') - d(\theta', \theta'')| \le 2^{-k} - 2^{-(k+1)}\},
\end{aligned}
\tag{2.29}
$$

and set

$$
\hat{\theta}_n = \hat{\theta}_{k,n}, \quad n_k \le n < n_{k+1}. \tag{2.30}
$$

It is clear that for each $\eta > 0$

$$
\lim_n \sup_{\theta \in \Theta} \mathbb{P}_\theta\{d(\hat{\theta}_n, \theta) > \eta\} = 0. \tag{2.31}
$$

Now consider the case of general Θ satisfying the conditions of the theorem. Let $\{\hat{\theta}_n^h\}_{n=1}^\infty$ denote the sequence of estimators constructed above with Θ_h playing the role of Θ. Define $1 = m_1 < m_2 < \ldots \in \mathbb{N}$ inductively by

$$
m_{k+1} = \min\{m > m_k : \sup_{\theta \in \Theta_{k+2}} \mathbb{P}_\theta\{d(\hat{\theta}_p^{k+2}, \theta) > 2^{-(k+1)}\} < 2^{-(k+1)}, \ \forall p \ge m\} \tag{2.32}
$$

and set

$$
\hat{\theta}_n = \hat{\theta}_m^{k+1}, \quad m_k \le m < m_{k+1}. \tag{2.33}
$$

It is clear that for all $\theta \in \Theta$, the sequence $\{\hat{\theta}_n\}_{n=1}^\infty$ converges to θ in the d–topology in $\mathbb{P}_\theta$ probability. $\square$

Example 2.13. Suppose that μ is the standard normal distribution, that $\sup_i \|\kappa_i\|_* = 1$ and that, for some $\xi \in \mathcal{T}$ with $\|\xi\| = 1$, Θ is the one–dimensional set $\{t\xi : 0 \le t \le 1\}$. It is not hard to show that $\delta(a,b) = \{1 - \exp(-|a-b|^2/2)\}^{\frac{1}{2}}$ and hence $\alpha|a-b| \le \delta(a,b) \le \beta|a-b|$ for suitable constants $0 < \alpha \le \beta < \infty$ when $|a|, |b| \le 1$. Note for $0 \le s, t \le 1$ that

$$\sum_{i=1}^{n} \delta^2(\langle \kappa_i, s\xi\rangle, \langle \kappa_i, t\xi\rangle) \le \sum_{i=1}^{n} \delta^2(0, \langle \kappa_i, \xi\rangle) \tag{2.34}$$

and that

$$\alpha^2|s-t|^2 \sum_{i=1}^{n} |\langle \kappa_i, \xi\rangle|^2 \le \sum_{i=1}^{n} \delta^2(\langle \kappa_i, s\xi\rangle, \langle \kappa_i, t\xi\rangle) \le \beta^2|s-t|^2 \sum_{i=1}^{n} |\langle \kappa_i, \xi\rangle|^2. \tag{2.35}$$

It follows that if we set

$$C_n = \left\{ \sum_{i=1}^{n} \delta^2(0, \langle \kappa_i, \xi\rangle) \right\}^{\frac{1}{2}}, \tag{2.36}$$

then

$$\lim_n C_n = \infty \text{ if and only if } \sum_{i=1}^{\infty} |\langle \kappa_i, \xi\rangle|^2 = \infty. \tag{2.37}$$

Assume that $\lim_n C_n = \infty$. Then the pseudo–metrics d_n certainly converge uniformly on $\Theta \times \Theta$ to a metric that is equivalent to the metric induced on Θ by the norm and hence, in particular, Θ is compact in the d–topology.

Consequently, a sufficient condition for the model to be consistently estimable in the norm topology is that $\sum_{i=1}^{\infty} |\langle \kappa_i, \xi\rangle|^2 = \infty$. Conversely, if the model θ is consistently estimable, then certainly the probability measures $\mathbb{P}_0$ and $\mathbb{P}_\xi$ are mutually singular. Applying Kakutani's dichotomy A.2 and the calculations above, this will be the case if and only if $\sum_{i=1}^{\infty} |\langle \kappa_i \xi\rangle|^2 = \infty$ (alternatively, one could appeal to the Cameron-Martin theorem on equivalence of shifted Gaussian measures – a result which is itself a consequence of Kakutani's dichotomy). The sufficient condition given by Theorem 2.12 is therefore also necessary in this case.

By similar arguments, if we take μ to be the uniform distribution on $[-1, 1]$, then the model is consistently estimable in the norm topology if and only if $\sum_{i=1}^{\infty} |\langle \kappa_i, \xi\rangle| = \infty$.

Example 2.14. Consider the one–dimensional problem of Example 2.13 with an arbitrary absolutely continuous error distribution μ. We can extend the observation in Example 2.13 that consistent estimation is "easier" for uniform errors than it is for normal errors, by showing that consistent estimation for normal errors is, in fact, the "hardest" among all absolutely continuous error distributions for this problem.

To see this, write f for the density of μ. From Parseval's theorem we have

$$\begin{aligned}\delta^2(a,b) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |\exp(iaz) - \exp(ibz)|^2 |\widehat{\sqrt{f}}(z)|^2\,dz \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} |1 - \exp(i(b-a)z)|^2 |\widehat{\sqrt{f}}(z)|^2\,dz,\end{aligned} \tag{2.38}$$

where $\widehat{\sqrt{f}}$ is the Fourier transform of $\sqrt{f}$.

Note that

$$\delta^2(a,b) \geq \frac{1}{2\pi}\int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} |1 - \exp(i(b-a)z)|^2 |\widehat{\sqrt{f}}(z)|^2\,dz \geq c|a-b|^2 \tag{2.39}$$

for some constant $c > 0$, because for any $\eta > 0$

$$\begin{aligned}\int_{-\eta}^{\eta} |\widehat{\sqrt{f}}(z)|^2\,dz &\geq \int_{-\infty}^{\infty} |\widehat{\sqrt{f}}(z)|^2 \left(1 - \frac{z}{\eta}\right)_+^2 dz \\ &\geq \left|\int_{-\infty}^{\infty} \widehat{\sqrt{f}}(z)\left(1 - \frac{z}{\eta}\right)_+ dz\right|^2 \\ &= \left(2\pi \int_{-\infty}^{\infty} \sqrt{f}(x)\frac{1}{\pi}\frac{1-\cos\eta x}{\eta x^2}\,dx\right)^2 \\ &> 0\end{aligned} \tag{2.40}$$

by the Cauchy-Schwarz inequality and Parseval's identity. Therefore, if consistent estimation is possible in the one–dimensional model under a normal error distribution for some sequence of functionals $\{\kappa_i\}_{i=1}^{\infty}$, then it is possible for any other absolutely continuous error distribution.

Example 2.15. With Examples 2.13 and 2.14 in hand, it is natural to ask for which absolutely continuous error distributions μ is consistent estimation in the one–dimensional model of those examples as "hard" as it is in the normal case. That is, when do we get an upper bound on $\delta^2(a,b)$ corresponding to the lower bound (2.39)?

Again write f for the density of μ. It follows from the dominated convergence theorem that we will get a corresponding upper bound

$$\delta^2(a,b) \leq C|a-b|^2 \tag{2.41}$$

for some constant $C < \infty$ if

$$\int_{-\infty}^{\infty} |z|^2 |\widehat{\sqrt{f}}(z)|^2 \, dz < \infty. \tag{2.42}$$

This is equivalent to

$$\int_{-\infty}^{\infty} |\sqrt{f}'(x)|^2 \, dx = \int_{-\infty}^{\infty} \frac{[f'(x)]^2}{2f(x)} \, dz = \int_{-\infty}^{\infty} \left[\frac{d}{dx} \log f(x)\right]^2 f(x) \, dx < \infty; \tag{2.43}$$

that is, that the *Fisher information* for the shift family $\{\mu_a : a \in \mathbb{R}\}$ is finite.

Example 2.16. It it is clear from Examples 2.13, 2.14 and 2.15 that the existence of a consistent estimator in the absolutely continuous case is intimately related to the lack of smoothness of the density. In essence, for a given sequence of functionals $\{\kappa_i\}_{i=1}^{\infty}$ the estimation problem is "easier" when the density f of the error distribution μ is rougher. This runs counter to the naive intuition that the tail behaviour of the probability measure μ should be the determining feature for whether or not it is possible to construct consistent estimators. We stress this point with the following interesting class of error distributions.

Consider a function $g : \mathbb{R} \to \mathbb{R}$ with the properties:

- g is non-negative and bounded,
- g is even (that is, $g(z) = g(-z)$),
- the restriction of g to the positive half-line is convex and decreasing,
- for some constant $0 < \alpha < 2$ there exist constants $0 < c', c'' < \infty$ such that

$$c'|z|^{-\alpha} \leq \int_{z}^{\infty} g^2(w) \, dw \leq c''|z|^{-\alpha}, \quad |z| \geq 1, \tag{2.44}$$

- $\frac{1}{2\pi} \int_{-\infty}^{\infty} g^2(z) \, dz = 1$.

It follows from Parseval's theorem and Polya's criterion for a function to be a Fourier transform of a positive measure (see Theorem A.3) that g is the Fourier transform of a non-negative function that is the square root of an even probability density. Denote this density by f and let $\mu(dx) = f(x) \, dx$. We have

$$\begin{aligned} \delta^2(0, a) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |1 - \exp(iaz)|^2 g^2(z) \, dz \\ &= \frac{1}{\pi} \left\{ \int_0^1 |1 - \exp(iaz)|^2 g^2(z) \, dz + \int_1^{\infty} |1 - \exp(iaz)|^2 g^2(z) \, dz \right\}. \end{aligned} \tag{2.45}$$

The first integral in the rightmost member is clearly bounded above and below by constant multiples of $|a|^2 \wedge 1$, whilst an integration by parts and a linear change of variable shows that the second integral is bounded above and below by constant multiples of $|a|^\alpha \wedge 1$. Therefore, in the setting of Example 2.13, the one–dimensional model is consistently estimable if and only if $\sum_i |\langle \kappa_i, \xi \rangle|^\alpha = \infty$.

The condition

$$c'|z|^{-\alpha} \le \int_z^\infty g^2(w)\, dw \le c''|z|^{-\alpha}, \quad |z| \ge 1, \tag{2.46}$$

constrains the degree of smoothness of f. For example, it implies that

$$\int_{-\infty}^\infty \int_{-\infty}^\infty \left\{ \frac{|\sqrt{f}(x) - \sqrt{f}(y)|}{|x-y|^\gamma} \right\}^2 dx\, dy < \infty \tag{2.47}$$

if and only if $\gamma < (1+\alpha)/2$ (cf. Example 1.4.1 of [34]), so that, in some sense, f become rougher as α decreases.

One might suspect from the uniform versus normal example that it is the compact support of the uniform distribution that makes consistent estimation easier. However, the densities with Fourier transforms that satisfy Polya's criterion are never compactly supported, and yet when $\alpha < 1$ consistent estimation may be possible under the class of error distributions constructed above when it is not possible under the uniform distribution.

Corollary 2.17. *Suppose that $C_n = n$, that $\Theta = \bigcup_h \Theta_h$ for a countable collection of sets $\Theta_1 \subseteq \Theta_2 \subseteq \ldots \subseteq \Theta$ that are relatively compact in the norm topology, and that d_n converges pointwise on each set $\Theta_h \times \Theta_h$ to a metric d on Θ. Suppose further that μ is absolutely continuous and that $\sup_i \|\kappa_i\|_* < \infty$. Then the model is consistently estimable in the d–topology.*

Proof. We first show that the convergence of d_n to d is uniform on each Θ_h. Observe that

$$d_n(x', x'') \le \sup\{\delta(0,t) : |t| \le \sup_i \|\kappa_i\|_* \|x' - x''\|\}. \tag{2.48}$$

Note also that

$$|d_n(y', y'') - d_n(z', z'')| = |d_n(0, y'-y'') - d_n(0, z'-z'')| \le d_n(y'-y'', z'-z''). \tag{2.49}$$

Because μ is absolutely continuous,

$$\lim_{t \to 0} \delta(0,t) = 0. \tag{2.50}$$

The sequence of functions $\{d_n\}_{n=1}^{\infty}$ is thus equicontinuous on $\Theta_h \times \Theta_h$ in the metric induced by the norm, and uniform convergence follows from Ascoli's theorem.

To complete checking the conditions of Theorem 2.12, we need only show that each Θ_h is totally bounded with respect to d. Because Θ_h is relatively compact in the norm topology, it is totally bounded with respect to the metric induced by the norm. Total boundedness with respect to d now follows from (2.48) and (2.50). □

Example 2.18. Let $\mathcal{T} = C([0,1])$, the Banach space of continuous functions on $[0,1]$ equipped with the supremum norm. For $0 < \alpha \leq 1$, let Θ be the collection of functions satisfying a Hölder condition of order α. That is, Θ is the collection of functions $x \in C([0,1])$ such that

$$\sup\{|x(s) - x(t)|/|s-t|^{\alpha},\ 0 \leq s \neq t \leq 1\} < \infty. \tag{2.51}$$

By the Arzela–Ascoli theorem, the set Θ is the union of a countable collection of sets that are compact in the norm topology. Let μ be any absolutely continuous probability measure and take $C_n = n$. Fix an irrational number ζ and let κ_i be the functional given by evaluation at the fractional part of $i\zeta$. That is, $\langle \kappa_i, x\rangle = x(i\zeta - \lfloor i\zeta \rfloor)$. By the Kronecker–Weyl equidistribution theorem,

$$\lim_n d_n(x', x'') = \left\{ \int_0^1 \delta^2(x'(t), x''(t))\, dt \right\}^{\frac{1}{2}} = d(x', x''). \tag{2.52}$$

for each pair $x', x'' \in C([0,1])$. It follows from Corollary 2.17 that the model is consistently estimable in the d–topology.

Notation 2.19. Define pseudo–metrics D_n, $n \in \mathbb{N}$, on $\mathcal{T}^{**}$ by setting

$$D_n(X', X'') = \left\{ \frac{1}{C_n} \sum_{i=1}^{n} \delta^2(\langle\langle x', \kappa_i\rangle\rangle, \langle\langle x'', \kappa_i\rangle\rangle) \right\}^{\frac{1}{2}}. \tag{2.53}$$

Corollary 2.20. *Suppose that* $\lim_n C_n = \infty$*, that* D_n *converges uniformly on each set of the form* $\{(x', x'') \in \mathcal{T}^{**} : \|x'\|_{**} \leq h, \|x''\|_{**} \leq h\}$ *to a metric* D *on* $\mathcal{T}^{**}$ *that is compatible with the weak* topology on* $\mathcal{T}^{**}$ *(as the dual of* $\mathcal{T}^*$*). Then the model is consistently estimable in the weak topology for any* $\Theta \subseteq \mathcal{T}$*.*

Proof. Recall that there is a canonical isometric embedding $\iota : \mathcal{T} \hookrightarrow \mathcal{T}^{**}$. This embedding is continuous from $\mathcal{T}$ equipped with the weak topology into $\mathcal{T}^{**}$ equipped with the weak* topology (as the dual of $\mathcal{T}^*$). The image of $\{x \in \mathcal{T} : \|x\| \leq h\}$ is weak* dense in $\{x \in \mathcal{T}^{**} : \|x\|_{**} \leq h\}$. It is clear that $D_n(\iota x', \iota x'') = d_n(x', x'')$, so that d_n converges uniformly to a metric d on sets of the form $\{(x', x'') \in \mathcal{T} : \|x'\| \leq h, \|x''\| \leq h\}$, and $D(\iota x', \iota x'') = d(x', x'')$. The assumption that the metric D is compatible with the weak* topology on $\mathcal{T}^{**}$ implies that the metric d is compatible with the weak topology on $\mathcal{T}$.

By the Banach–Alaoglu theorem, $\{x \in \mathcal{T}^{**} : \|x\|_{**} \leq h\}$ equipped with the weak* topology is compact. Because the metric D is compatible with weak* topology, $\{x \in \mathcal{T}^{**} : \|x\|_{**} \leq h\}$ is totally bounded with respect to D. Hence $\{x \in \mathcal{T} : \|x\| \leq h\}$ and, *a fortiori*, $\Theta \cap \{x \in \mathcal{T} : \|x\| \leq h\}$ is totally bounded with respect to the metric d for any $\Theta \subseteq \mathcal{T}$.

The result now follows from Theorem 2.12 and the fact that the metric d is compatible with the weak topology. $\square$

Example 2.21. Let $\mathcal{T}$ and $\Theta \subseteq \mathcal{T}$ be aribitrary. Suppose that μ is absolutely continuous, and hence that the metric δ is compatible with the usual topology on $\mathbb{R}$. Take $C_n = n$. Let $\lambda_1, \lambda_2, \ldots$ be a sequence of linear functionals with closed linear span $\mathcal{T}^*$ (so that $\mathcal{T}^*$ is necessarily separable in the norm topology). Set

$$p_{j,n} = \frac{1}{n}\#\{1 \leq i \leq n : \kappa_i = \lambda_j\}. \tag{2.54}$$

Assume that $\lim_n p_{j,n} = p_j > 0$ exists for all j and $\sum_j p_j = 1$. (For example, if the κ_i were chosen independently at random with the probability that $\kappa_i = \lambda_j$ being p_j, then this property would hold almost surely.) Then

$$D_n(x', x'') = \left\{ \sum_j p_{j,n} \delta^2(\langle\langle x', \lambda_j\rangle\rangle, \langle\langle x'', \lambda_j\rangle\rangle) \right\}^{\frac{1}{2}} \tag{2.55}$$

and

$$D(x', x'') = \left\{ \sum_j p_j \delta^2(\langle x', \lambda_j\rangle, \langle\langle x'', \lambda_j\rangle\rangle) \right\}^{\frac{1}{2}}. \tag{2.56}$$

It is immediate that D is compatible with the weak* topology. Moreover, given $\eta > 0$ choose J such that

$$\sum_{j>J} p_j \leq \frac{\eta}{2} \quad \text{and} \quad \sum_{j \leq J} |p_{j,n} - p_j| \leq \frac{\eta}{2}, \tag{2.57}$$

then

$$|d_n^2(x', x'') - d^2(x', x'')| \leq \eta, \quad x', x'' \in \mathcal{T}, \tag{2.58}$$

and hence d_n certainly converges uniformly to d on norm bounded subsets of $\mathcal{T} \times \mathcal{T}$. Corollary 2.20 applies, and the model is consistently estimable in the weak topology.

2.5. **An Example.** Let $L_2[0,1]$ denote the Hilbert space of Lebesgue square-integrable real-valued functions on the interval $[0,1]$. Let $(f|g)$ denote the inner product of the functions f and g:

$$(f|g) = \int_0^1 f(t)g(t)\,dt. \tag{2.59}$$

Let $\{\Delta_j\}_{j=1}^n$ be a fixed collection of closed, disjoint sub-intervals of $[0,1]$, each of strictly positive length, and such that there exists an open set $\Delta_0 \subset [0,1]$ for which

$$\Delta_0 \cap \left\{ \bigcup_{j=1}^n \Delta_j \right\} = \emptyset. \tag{2.60}$$

For $f \in L_2[0,1]$, define the continuous linear functionals κ_j by

$$\langle \kappa_j, f \rangle = \int_{\Delta_j} f(t)\,dt = (f|1_{\Delta_j}). \tag{2.61}$$

The functions $\{1_{\Delta_j}\}_{j=1}^n \subset L_2[0,1]$ are called "representers" or "data kernels" in much of the inverse problems literature. Note that in this example $\{\kappa_j\}_{j=1}^n$ is a linearly independent subset of $\mathcal{T} = L_2[0,1]$. We observe data $X = \{X_j\}_{j=1}^n$, with

$$X_j = \langle \kappa_j, f \rangle + \epsilon_j,\ j = 1, \ldots, n, \tag{2.62}$$

where, in this subsection, the noise terms $\{\epsilon_j\}_{j=1}^n$ are i.i.d. normal random variables with zero mean and variance σ^2 (we write $\{\epsilon_j\}_{j=1}^n$ i.i.d. $N(0,\sigma^2)$). We abbreviate Equation (2.62) by

$$X = Kf + \epsilon. \tag{2.63}$$

Consider estimating the pair of values $(g_1(f),\, g_a(f))$ from these data, where g_1 and g_a are given by

$$g_1(f) = \int_0^1 f(t)\,dt, \tag{2.64}$$

and

$$g_a(f) = \int_0^1 \sum_{j=1}^n a_j 1_{\Delta_j}(t) f(t)\, dt, \tag{2.65}$$

with $\{a_j\}_{j=1}^n \in \mathbb{R}^n$. Both g_1 and g_a are bounded linear functionals on $L_2[0,1]$.

To translate this problem into the running notation, identify $\Theta = L_2[0,1]$, $\theta = f$, $\mathcal{X} = \mathbb{R}^n$, and $\mathcal{P}$ to be the location family of n-dimensional normal distributions on $\mathbb{R}^n$ with independent components each of which has variance σ^2:

$$\mathbb{P}_\theta(B) = \int_B \prod_{j=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_j - (\theta|\Delta_j))^2}{2\sigma^2} \right\} \right) d^n x \tag{2.66}$$

for all Borel sets $B \subseteq \mathbb{R}^n$. The parameter space $\mathcal{G} = \mathbb{R}^2$, and the mapping g is given by

$$\begin{aligned} g : \Theta &\rightarrow \mathbb{R}^2 \\ \theta &\mapsto \left((\theta|1), \left(\theta \middle| \sum_{j=1}^n a_j 1_{\Delta_j} \right) \right). \end{aligned} \tag{2.67}$$

In this model, θ is not identifiable. Neither is $g(\theta)$, because its first component $(\theta|1)$ can be perturbed arbitrarily without changing the distribution of X (just change θ on Δ_0—this is a special case of Theorem 2.6). The second component of $g(\theta)$, $(\theta|g_a)$, is identifiable. The estimator

$$\hat{g}_a(X) = \sum_{j=1}^n a_j X_j = a \cdot X \tag{2.68}$$

is unbiased for $(\theta|g_a)$; there is no unbiased estimator of g.

Let $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^n$. Suppose we estimate $g(\theta)$ by the linear estimator

$$\hat{g}(X) = (\mathbf{1} \cdot X, a \cdot X). \tag{2.69}$$

The bias of $\hat{g}$ is

$$\mathbb{E}_\theta [\hat{g}(X) - g(\theta)] = \left(\int_0^1 (1_{\bigcup_{j=1}^n \Delta_j}(t) - 1)\theta(t)\, dt, 0 \right). \tag{2.70}$$

Let $b_1(\theta) = \int_0^1 (1 - 1_{\bigcup_{j=1}^n \Delta_j}(t))\theta(t)\, dt$. The mean squared error of $\hat{g}(X)$ is

$$\mathbb{E}_\theta [\|\hat{g}(X) - g(\theta)\|^2] = b_1^2(\theta) + \sigma^2(n + \|a\|^2). \tag{2.71}$$

The bias and the MSE of $\hat{g}(X)$ are unbounded over Θ.

Now suppose that in addition to the stochastic errors $\{\epsilon_j\}_{j=1}^n$, each datum contains also a systematic error τ_j:

$$X_j = \langle \kappa_j, f \rangle + \epsilon_j + \tau_j,\ j = 1, \ldots, n. \tag{2.72}$$

Assume we know that τ satisfies

$$\tau \in \mathbf{T} = \{\tau' \in \mathbb{R}^n : |\tau'_j| \le t_j < \infty\,, j = 1, \cdots, n\} \subset \mathbb{R}^n. \tag{2.73}$$

We can embed this case in the framework we have developed by appending to θ the n-vector $\tau = \{\tau_j\}_{j=1}^n \in \mathbb{R}^n$. These additional parameters, which affect the probability distribution but are not the subject of the estimation problem, are called *nuisance parameters*. The model space Θ is now $L_2[0,1] \times \mathbf{T}$, which we endow with the following norm. If $\theta = (f, \gamma)$ with $f \in L_2[0,1]$ and $\gamma \in \mathbf{T}$,

$$\|\theta\|^2 = \|f\|^2 + \|\gamma\|^2, \tag{2.74}$$

where $\|\gamma\|$ is the ordinary Euclidean norm of $\gamma \in \mathbb{R}^n$. For $\theta = (f, \gamma)$ and $\rho = (g, \delta)$ this corresponds to the inner product

$$(\theta|\rho) = \int_0^1 f(t) g(t)\, dt + \gamma \cdot \delta. \tag{2.75}$$

Let $\mathbf{1_j}$ be the n-vector that is zero in every component but j, for which it is 1. The probability distribution $\mathbb{P}_\theta$ on $\mathbb{R}^n$ is

$$\mathbb{P}_\theta(B) = \int_B \prod_{j=1}^n \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_j - (\theta|(1_{\Delta_j}, \mathbf{1_j})))^{\mathbf{2}}}{2\sigma^2} \right\} \right) d^n x \tag{2.76}$$

for all Borel sets $B \subseteq \mathbb{R}^n$. Now

$$g(\theta) = \left((\theta|(1,0)), \left(\theta \middle| (\sum_{j=1}^n a_j 1_{\Delta_j}, 0) \right) \right). \tag{2.77}$$

Neither component of $g(\theta)$ is identifiable once we have systematic errors. However, if we use the same estimator $\hat{g}(X)$ as before, its bias is

$$\mathbf{bias}_\theta(\hat{g}) = \mathbb{E}_\theta[\hat{g}(X) - g(\theta)] = (b_1(\theta) + 1 \cdot \tau, a \cdot \tau)\,. \tag{2.78}$$

The first component of the bias is still unbounded for $\theta \in \Theta$, but the second is bounded:

$$\max_{\tau' \in \mathbf{T}} |a \cdot \tau'| = b(a, \mathbf{T}) < \infty. \tag{2.79}$$

Hölder's inequality gives a crude bound on the bias:

$$|a \cdot \tau| \leq b(a, \mathbf{T}) \leq \|a\|_1 \|\tau\|_\infty \leq \|a\|_1 \|t\|_\infty. \tag{2.80}$$

The mean squared error of $\hat{g}(X)$ is

$$\mathbb{E}_\theta[\|\hat{g}(X) - g(\theta)\|^2] = ((b_1(\theta) + 1 \cdot \tau)^2 + (a \cdot \tau)^2 + \sigma^2(n + \|a\|^2). \tag{2.81}$$

3. Statistical Decision Theory

3.1. **General Framework.** This section presents a framework for comparing estimators and confidence sets in inverse problems: statistical decision theory [64]. Statistical decision theory can be developed in quite abstract settings (Le Cam [44] frames it in the context of mappings from an arbitrary set to an L-space); here we insist that the model set Θ is a subset of a separable Banach space, and that the observation is an element of a separable Banach space. References more accessible than [44] include Lehmann [46], Lehmann and Casella [47], and Berger [11] (for a Bayesian perspective).

Decision theory frames statistical estimation and inference as a two-player game, Nature versus statistician. Nature picks $\theta \in \Theta$; the value of θ is hidden from the statistician; data X will be generated from $\mathbb{P}_\theta$. Before X is drawn, the statistician chooses a strategy δ for guessing some feature of θ from X. The data are generated; the statistician applies the rule; and the statistician pays a *loss* $\ell(\theta, \delta)$ that depends on his guess $\delta(X)$ and the true value of θ. We shall give a more precise mathematical statement of the game after introducing new terminology.

The game has the following ingredients:

(1) a collection $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability distributions on a separable Banach space $\mathcal{X}$, where Θ is a known subset of a separable Banach space $\mathcal{T}$. The elements of $\mathcal{P}$ are the strategies available to Nature.

(2) a fixed collection $\mathcal{D}$ of randomized decision rules mapping $\mathcal{X}$ into probability distributions onto a space $\mathcal{A}$ of actions. The elements of $\mathcal{D}$ are the strategies available to the statistician.

(3) a loss function $\ell : \Theta \times \mathcal{A} \to \mathbb{R}^+$. The statistician pays $\ell(\theta, a)$ if Nature selects θ and the statistician takes action a.

If the statistician uses the randomized rule $\delta \in \mathcal{D}$ to choose his action on the basis of the data $X \sim \mathbb{P}_\theta$, in repeated play, his expected loss is the *risk at $\theta \in \Theta$ of the decision rule $\delta \in \mathcal{D}$*:

$$r(\theta, \delta) \equiv \mathbb{E}_\theta \left[\int_{\mathcal{A}} \ell(\theta, a)\, \delta(X)(da) \right]. \tag{3.1}$$

When δ is non-randomized, we can think of δ as taking values in $\mathcal{A}$ rather than in the collection of probability measures on $\mathcal{A}$; then

$$r(\theta, \delta) \equiv \mathbb{E}_\theta [\ell(\theta, \delta(X))]. \tag{3.2}$$

The statistician seeks to make $r(\theta, \delta)$ "small" by choosing $\delta \in \mathcal{D}$ cleverly.

Because θ is unknown, different senses of small compete, and lead to different strategies for selecting the decision rule δ. The two most common strategies for picking an "optimal" decision rule are *minimax* and *Bayes* (see Definitions 3.1 and 3.2 below for precise definitions). *Minimax* decision rules minimize over the statistician's choice of decision functions $\delta \in \mathcal{D}$ the maximum risk over Nature's possible choices of the parameter θ. This hedges against the possibility that Nature plays the game aggressively, picking the value of θ that maximizes the statistician's guaranteed loss. *Bayes* decision rules minimize over decision functions a weighted (by a prior probability distribution π) average risk over nature's possible choices of the parameter θ. This treats Nature as if it draws θ at random from Θ according to the prior distribution π. There are connections between these two notions that we do not explore here. For example, a sufficient condition for a decision rule to be minimax is that it be Bayes with respect to some prior probability distribution and that the resulting average risk coincides with the maximum risk.

The choice of a loss function $\ell(\theta, a)$ is essentially arbitrary. Context dictates appropriate choices, but most of the worked examples in decision theory use loss functions chosen for analytic convenience rather than for scientific relevance. The most common loss function for point estimates of parameters in Euclidean spaces is squared error: $\ell(\theta, a) = \|a - g(\theta)\|^2$. For estimating a real-valued functional $g(\theta)$, common choices include absolute error

$$\ell(\theta, a) = |a - g(\theta)|, \tag{3.3}$$

and zero-one loss depending on the distance between a and θ:

$$\ell(\theta, a) = \begin{cases} 0, & |a - g(\theta)| \le c \\ 1, & \text{otherwise.} \end{cases} \tag{3.4}$$

For set-valued actions $S \subseteq \mathcal{G}$, loss functions typically combine coverage of the parameter—$\mathbb{P}_\theta\{S \ni g(\theta)\}$—and a measure of the size of the set S.[1] For example, we might take $\ell(\theta, S) = 1_{S \not\ni \theta} + \lambda|S|$, with $\lambda > 0$, where $|S|$ is the diameter of S if Θ is a subset of a metric space, or the Lebesgue measure of S, if $\Theta \subseteq \mathbb{R}^n$. Another possibility is to combine coverage of the parameter and distance from the parameter to the closest point in the set. We shall restrict attention to loss functions that are finite for all $a \in \mathcal{A}$ and all $\theta \in \Theta$.

When $\mathcal{A}$ is a convex subset of a separable Banach space, it is sometimes helpful to require the loss ℓ to be convex in its second argument: for every $\theta \in \Theta$, for all $\gamma \in [0, 1]$, and for all a_1, a_2 in $\mathcal{A}$,

$$\ell(\theta, \gamma a_1 + (1-\gamma)a_2) \leq \gamma\ell(\theta, a_1) + (1-\gamma)\ell(\theta, a_2). \tag{3.5}$$

This holds, for example, if we seek to estimate a parameter $g(\theta)$ using an estimator δ that takes values in $\mathcal{G}$, and the loss is the norm of the error of the estimate: $\ell(\theta, a) = \|g(\theta) - a\|$.

Definition 3.1. The *maximum risk of $\delta \in \mathcal{D}$ over Θ* is

$$\rho(\delta) \equiv \sup_{\theta \in \Theta} r(\theta, \delta). \tag{3.6}$$

The *minimax risk* is

$$\rho^* = \rho^*(\mathcal{D}) = \inf_{\delta \in \mathcal{D}} \rho(\delta). \tag{3.7}$$

If a decision rule $\delta^* \in \mathcal{D}$ has risk $\rho(\delta^*) = \rho^*$ then δ^* is a *minimax decision rule.*

Definition 3.2. If π is a probability measure on Θ, the *posterior risk of δ for prior π* is

$$\rho_\pi(\delta) = \int_{\mathcal{T}} r(\theta, \delta)\pi(d\theta). \tag{3.8}$$

The smallest posterior risk is the Bayes risk:

$$\rho^*_\pi = \inf_{\delta \in \mathcal{D}} \rho_\pi(\delta). \tag{3.9}$$

If a decision rule attains the Bayes risk (if $\rho_\pi(\delta^*) = \rho^*_\pi$), it is a *Bayes decision for prior π.*

[1]Recall that we require the space $\mathcal{A}$ of possible actions to be a subset of a separable metric space. It is usually possible to represent set-valued actions as elements of such a space — for example, by identifying sets with their indicator functions and working in a suitable function space or working with closed sets and using the Hausdorff distance on closed subsets of metric space.

Although the statistician may not be able to find the minimax or Bayes decision rule, he should at least discard a decision rule if he has another rule that performs better whatever be $\theta \in \Theta$:

Definition 3.3. A decision rule δ is *admissible* for loss ℓ if there is no other decision rule δ' such that

$$r(\theta, \delta') \leq r(\theta, \delta), \ \forall \theta \in \Theta \tag{3.10}$$

and $r(\theta, \delta') < r(\theta, \delta)$ for at least one $\theta \in \Theta$. If such a δ' exists, it is said to *dominate* δ. If δ is not admissible it is *inadmissible*.

Example 3.4. Consider estimating an m-vector g of linearly independent linear functionals in a linear inverse problem, as described in Theorem 2.7; suppose that the errors are Gaussian, and that the conditions of that theorem hold: $g = \Lambda \cdot \kappa$ for some $m \times n$ matrix Λ. Although the Backus-Gilbert estimator $\Lambda \cdot X$ is unbiased for g, Section 4.5 shows that if $m \geq 3$ and the variance-covariance matrix Σ of the data errors has full rank, then $\Lambda \cdot X$ is inadmissible for mean squared error.

However, if $m < 3$, the Backus-Gilbert estimator is minimax for mean squared error, and can be characterized as the limit of a sequence of Bayes estimators for prior probability distributions that are increasingly "flat" on $\mathbb{R}^m$.

Determining whether or not a decision rule is minimax or Bayes (or is even admissible) is essentially an optimization problem. To make this optimization problem as simple as possible, it is useful to make an *a priori* reduction in the range of possible decisions that need to be considered. A useful tool for performing this reduction is the notion of sufficiency.

Definition 3.5. A *statistic* is a measurable mapping from the data space $\mathcal{X}$ into some other measurable space. A statistic T is *sufficient for* $\mathcal{P}$ if there is a version of the conditional distribution under $\mathbb{P}_\theta$ of the data X given $T(X)$ that does not depend on $\theta \in \Theta$. It is trivially true that X is sufficient for $\mathcal{P}$.

For convex loss functions, the following result shows that nothing is lost in restricting attention to estimators that are functions of a sufficient statistic.

Theorem 3.6 (Rao-Blackwell Theorem (see [47] Th. 1.7.8)). *Let X have probability distribution $\mathbb{P}_\theta \in \mathcal{P} = \{\mathbb{P}_{\theta'} : \theta' \in \Theta\}$, and let T be sufficient for $\mathcal{P}$. Let $\hat{g}$ be an estimator of the parameter $g(\theta)$, and let the loss $\ell(\theta, a)$ be strictly convex in a. Suppose that $\hat{g}(X)$ is integrable for all $\mathbb{P}_\theta$,*

$$r(\theta, \delta) = \mathbb{E}_\theta[\ell(\theta, \hat{g}(X)] < \infty, \tag{3.11}$$

and

$$\bar{g}(X) = \mathbb{E}_\theta[\hat{g}(X) \mid T(X)] \tag{3.12}$$

(because $T(X)$ is sufficient for θ, the conditional expectation on the right–hand side does not depend on θ). Then

$$r(\theta, \bar{g}) < r(\theta, \hat{g}) \tag{3.13}$$

unless $\hat{g}(X) = \bar{g}(X)$, $\mathbb{P}_\theta$ almost surely, for all $\theta \in \Theta$.

Remark 3.7. To be completely rigorous, the statement of Theorem 3.6 needs a further condition. Conditional expectations are defined only up to sets of probability zero, so the definition of $\bar{g}(X)$ contains a "hidden" null set that could depend on θ. One way of overcoming this unwanted difficulty is to require that the family $\mathcal{P}$ of probability measures is dominated by some fixed probability measure ν, that is, for every $\theta \in \Theta$, $\mathbb{P}_\theta$ is absolutely continuous with respect to ν.

We will see more telling uses of Theorem 3.6 later, but here is a simple example of how it can be used to justify rigorously the intuitively obvious. Consider the set-up of Section 2.5, but change things so that $\Delta_1 = \Delta_2 = \ldots = \Delta_n$ (instead of $\Delta_1, \Delta_2, \ldots, \Delta_n$ being pairwise disjoint). A simple (unbiased) estimator for $g(\theta) = (\theta | 1_{\Delta_1})$ is X_1. However $X_2, \ldots, X_n$ are also unbiased estimators of g and it seems that we should be able to get a better estimator of g by incorporating the information about g contained in these estimators. To this end, note that the statistic $\mathbf{1} \cdot X = X_1 + X_2 + \cdots + X_n$ is sufficient (the conditional distribution of X given $\mathbf{1} \cdot X = x$ under any $\mathbb{P}_\theta$ is normal with expectation x/n and covariance matrix that doesn't involve θ). Therefore, for loss functions that are strictly convex,

$$\bar{g}(X) = \mathbb{E}[X_1 \mid \mathbf{1} \cdot X] = (X_1 + X_2 + \cdots + X_n)/n \tag{3.14}$$

dominates X_1 as an estimator of g: averaging the data gives a better estimate than using a single datum alone.

3.2. **Estimates as Decisions.** We now specialize to the case of estimating a parameter $g(\theta)$ where $g : \Theta \to \mathcal{G}$, with $\mathcal{G}$ a Banach space with norm $\|\cdot\|$. Take the action space $\mathcal{A}$ to be $\mathcal{G}$ as well, and consider the set $\mathcal{D}$ of decision rules δ that are $\mathbb{P}$-measurable mappings from $\mathcal{X}$ to $\mathcal{G}$. A standard choice of $\ell(\theta, a)$ is then $\|g(\theta) - a\|$, which is convex. Then $r(\theta, \delta)$ is the average error in the estimator, measured in the norm of $\mathcal{G}$—the MNE. A less common choice is $\ell(\theta, a) = 1_{g(\theta) \notin B_c(a)}$, where $B_c(a) = \{\eta \in \mathcal{G} : \|\eta - a\| \leq c\}$. When $\mathcal{G}$ is a Euclidean space, the most common loss function is $\|a - g(\theta)\|^2$. When $\mathcal{G} = \mathbb{R}$ (estimating a single real parameter), common loss functions are $\ell(\theta, a) = |g(\theta) - a|^p$ and $\ell(\theta, a) = 1_{|g(\theta) - a| > c}$.

3.3. **Confidence Sets as Decisions.** In elementary statistics, a confidence set is a mapping from possible data values to sets of parameter values (subsets of $\mathcal{G}$). One can think of this as a mapping from possible data values to functions on $\mathcal{G}$ that take the values 0 (the point is not in the set) and 1 (the point is in the set). Many results in decision theory depend on the assumption that the loss is convex in the action. We can make the set of actions convex by allowing the mapping from $\mathcal{G}$ to take more values—by considering confidence sets to be mappings from possible data values to functions on $\mathcal{G}$ that take values in $[0, 1]$, corresponding to a probability of membership.

A *confidence set for the parameter* $g(\theta) \in \mathcal{G}$ is a decision rule δ whose space of actions $\mathcal{A}$ is a collection of (measurable) functions from the space $\mathcal{G}$ of possible parameter values to the interval $[0, 1]$. Such a decision rule δ can be converted into another rule δ' whose space of actions $\mathcal{A}'$ is a collection of measurable functions from $\mathcal{G}$ into the two points $\{0, 1\}$ (that is, the collection of measurable subsets of $\mathcal{G}$) by defining the probability measure $\delta'(x)$ to be the push-forward of the product measure $\delta \otimes \lambda$ under the map $(a, u) \mapsto \mathbf{1}\{u \leq a(\cdot)\}$, where λ is Lebesgue measure on $[0, 1]$; that is,

$$\int_{\mathcal{A}'} F(a')\, \delta'(x)(da') = \int_{\mathcal{A}} \int_0^1 F(\mathbf{1}\{u \leq a(\cdot)\})\, du\, \delta(x)(da), \tag{3.15}$$

for F a bounded measurable function on $\mathcal{A}'$.

The *coverage probability* (at θ) of a confidence set δ for the parameter g is

$$\gamma(\theta, \delta) = \mathbb{E}_\theta \left[\int_{\mathcal{A}} a(g(\theta))\, \delta(X)(da) \right]. \tag{3.16}$$

A $1-\alpha$ *confidence set* for g is a confidence set that has coverage probability at least $1-\alpha$, whatever be $\theta \in \Theta$. The *expected measure* of δ with respect to the measure μ on $\mathcal{G}$ is

$$\nu(\theta, \delta) = \mathbb{E}_\theta \left[\int_{\mathcal{A}} \int_{\mathcal{G}} a(\zeta)\, \mu(d\zeta)\, \delta(X)(da) \right]. \tag{3.17}$$

For randomized confidence sets, the coverage probability and the expected measure are linear functionals of the decision rule, so they are convex, and we could consider a composite convex risk function

$$r(\theta, \delta) = -\gamma(\theta, \delta) + c\nu(\theta, \delta) \tag{3.18}$$

for some $c \in \mathbb{R}^+$, to optimize a tradeoff between coverage probability and expected measure. This is an example of how constructing confidence sets might be posed as a problem within the decision-theoretic framework we have described.

Suppose we restrict attention to sets $\mathcal{A}$ of actions that map $\mathcal{G}$ measurably into $\{0,1\}$, which thus can be thought of as measurable subsets Γ of $\mathcal{G}$. With this restriction, we might choose $\mathcal{A}$ to consist of sets with a given maximum diameter $|\Gamma| = \sup_{\eta,\zeta \in \Gamma} \|\eta - \zeta\|$, for example. In that case, a reasonable loss function is $\ell(\theta, \Gamma) = 1_{g(\theta) \notin \Gamma}$. Then $r(\theta, \delta)$ is the non-coverage probability: the probability under $\mathbb{P}_\theta$ that the set δ does not include $g(\theta)$. One can seek small confidence sets with a given maximum non-coverage probability whatever be $\theta \in \Theta$ by taking $\mathcal{A}$ to be a collection of Borel subsets of $\mathcal{G}$ with a given maximum diameter, and varying that diameter until the supremum of $r(\theta, \delta)$ over Θ is the target non-coverage probability α.

4. Estimation

There are a variety of "recipes" for constructing estimators; perhaps the most common in statistics are maximum likelihood (MLE) and Bayes, both of which have asymptotically optimal properties under certain restrictive assumptions. However, both can be inconsistent, even when the dimension of the model is finite; see below. In the inverse problems literature, regularization (especially regularized least-squares, which is related to maximum penalized likelihood), using a truncated basis expansion (which is related to the method of sieves) and the Backus-Gilbert method are among the most common procedures. In this section, $\theta \in \Theta \subseteq \mathcal{T}$, a separable Banach space, $K : \Theta \to \mathbb{R}^n$ is linear, and $X = K\theta + \epsilon$, where ϵ is usually a vector of i.i.d. zero-mean Gaussian errors.

We shall use interpolation (nonparametric regression) on the unit interval as an example throughout this section: $\mathcal{T}$ will be some class of functions, for example, a Sobolev space of functions $\theta : [0,1] \to \mathbb{R}$, $\langle \kappa_j, \theta \rangle = \theta(t_j)$, $\{t_j\}_{j=1}^n \subset [0,1]$, and $X = K\theta + \epsilon$, where the components of ϵ are i.i.d. $N(0,1)$. In statistical nomenclature, this problem is an instance of *nonparametric regression*. Let f^k denote the kth derivative of the function f. For integer $m \geq 1$, let W_m denote the Sobolev space of functions on $[0,1]$ that are absolutely continuous and have absolutely continuous derivatives up to order $m-1$, and whose m^{th} derivative is in $L_2[0,1]$, with the norm

$$\|f\|^2 = \sum_{k=0}^{m-1} |f^k(0)|^2 + \int_0^1 |f^m|^2 d\mu. \tag{4.1}$$

The corresponding inner product is

$$(f,g) = \sum_{k=0}^{m-1} f^k(0) g^k(0) + \int_0^1 f^m g^m d\mu. \tag{4.2}$$

A Hilbert space of functions of position, in which the point-evaluation functional $f \to f(t_0)$ is continuous for every t_0 in the domain of f, is a reproducing kernel Hilbert space [1]. By the Riesz representation theorem, $f \to f(t_0)$ is thus the inner product $\langle \kappa_{t_0}, f \rangle$ of f with some other fixed element κ_{t_0} of the space. Reproducing kernel Hilbert spaces are at the heart of the theory of splines; see Wahba [63] for a statistical treatment. Finite-dimensional Hilbert spaces of functions are reproducing kernel Hilbert spaces, as are spaces of sufficiently smooth functions, such as bandlimited functions. In most infinite-dimensional inverse problems, the elements of Θ are not smooth enough for Θ to be a reproducing kernel Hilbert space.

The space W_m is a reproducing kernel Hilbert space. In particular, for $m = 2$, with

$$\kappa_j(t) = \begin{cases} 1 + tt_j + \frac{t^2 t_j}{2} - \frac{t^3}{6}, & t \leq t_j \\ 1 + tt_j + \frac{tt_j^2}{2} - \frac{t_j^3}{6}, & t > t_j, \end{cases} \tag{4.3}$$

we have $\kappa_j \in W_2$ and $\langle \kappa_j, f \rangle = f(t_j)$ for all $f \in W_2$. Moreover, if $t_i \neq t_j$, $1 \leq i \neq j \leq n$, then the n point evaluators for the points $\{t_j\}_{j=1}^n$ are linearly independent.

4.1. **Backus-Gilbert Estimation.** In a seminal series of papers [5, 2, 3, 4] George Backus and Freeman Gilbert developed a rigorous basis for linear inverse theory in Geophysics. Here, "a linear inverse problem" is an inverse problem in which the data are linearly related to the unknown but for additive noise, the unknown is an element of a linear vector space with constraints but the data, and the estimators are linear in the data. In statistical terms, Backus and Gilbert showed that the only linear functionals in an unconstrained linear inverse problem (meaning $\Theta = \mathcal{T}$) that are identifiable and estimable with bounded bias are linear combinations of the measurement functionals. That is, if the functional $g : \Theta \to \mathbb{R}$ is linear, then $g(\theta)$ is identifiable iff $g = \gamma \cdot K$ for some $\gamma \in \mathbb{R}^n$. In that case, as shown in Corollary 2.8, $\gamma \cdot X$ is an unbiased estimate of $g(\theta)$, and if Σ is the covariance matrix of ϵ, the variance of the estimate is

$$\mathbb{E}[(\gamma \cdot X - \gamma \cdot K\theta)^2] = \mathbb{E}[(\gamma \cdot \epsilon)^2] = \gamma \cdot \Sigma \cdot \gamma. \tag{4.4}$$

Backus and Gilbert focused on the case where Θ is a Hilbert space of functions of position $r \in \mathcal{D} \subseteq \mathbb{R}^n$, and developed a measure of "nearness" of g to the point-evaluator $\theta \mapsto \theta(r_0)$, which typically is not a member of Θ. Backus and Gilbert developed a framework for trading off between the variance of the estimate and the nearness of the linear functional estimated to $\theta(r_0)$; the latter is called the "resolution" or "resolving width."

In the interpolation problem stated above for W_2, the point evaluator is a bounded linear functional, but only linear combinations of $\{\kappa_j\}$ are estimable using the Backus-Gilbert formalism. In particular, if there is no measurement at the point t_0, Backus-Gilbert theory tells us that $f(t_0)$ is not estimable with bounded bias.

When there are additional constraints on the unknown θ, for example, if Θ is a norm-bounded ellipsoid in a Hilbert space $\mathcal{T}$, more is possible than Backus-Gilbert theory would suggest; see, *e.g.*, [8, 9, 10, 54] and the sections below. In particular, many more linear functionals can be estimated with bounded bias. Moreover, Backus-Gilbert estimates generally are not optimal for two-norm loss when three or more linear functionals are estimated; see Section 4.5 below.

4.2. **Maximum Likelihood Estimation (MLE) and its Variants.** Suppose that the family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability distributions on a measurable space $\mathcal{X}$ is dominated by a common σ-finite measure μ. Let $p_\theta(x)$, $x \in \mathcal{X}$ denote the density of $\mathbb{P}_\theta$ with respect to μ. For fixed $x \in \mathcal{X}$, the function

$$\begin{aligned} \mathcal{L} = \mathcal{L}_x : \Theta &\to \mathbb{R}^+ \\ \theta &\mapsto p_\theta(x) \end{aligned} \tag{4.5}$$

is called the *likelihood function*. If the value $X = x$ is observed, $\mathcal{L}(\theta'|X = x) = \mathcal{L}(\theta')$ is *the likelihood of* θ' *given* $X = x$. Note that, despite the suggestive notation, $\mathcal{L}(\cdot|X = x)$ is neither a conditional probability nor a probability density.

The basic idea behind the maximum likelihood method is to estimate θ by the value $\theta' \in \Theta$ for which the likelihood function $\mathcal{L}(\theta'|X = x)$ for the observation $X = x$ is largest:

$$\hat{\theta}_{\mathrm{ML}}(x) \equiv \arg\max_{\theta' \in \Theta} \mathcal{L}(\theta'|X = x) \tag{4.6}$$

when a unique maximizer exists. The spirit of the approach is that the maximizing value of θ' is "most likely" to be the correct one. More generally, we would estimate the parameter $g(\theta)$ by $\hat{g}_{\mathrm{ML}}(X) = g(\hat{\theta}_{\mathrm{ML}}(X))$. In smooth finite-dimensional problems, maximum likelihood has some nice asymptotic properties; see Lehmann and Casella [47]. In many problems, however, it runs into trouble.

Here are some technical issues. First, in order to define the likelihood function, we need the set of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ to be dominated by a common σ-finite measure μ. (All such dominating μ lead to the same estimator.) Second, we need the likelihood to attain its maximum (this can be overcome by maximizing the likelihood approximately; see inequalities 4.23 and 4.26). Third, we need the maximizer to be unique, or else we need a rule for choosing among maximizers. Fourth, we need $\arg\max_{\theta' \in \Theta} \mathcal{L}(\theta'|X)$ to be a measurable function of X; this requires additional assumptions. Even when these assumptions are met, maximum likelihood can have pathological properties, including being inconsistent even when a consistent estimator exists. See Le Cam [45] and examples in Lehmann and Casella [47]. For an example where the maximum likelihood estimate is inadmissible for a finite-dimensional parameter with Gaussian errors, see Section 4.5 below.

One problem maximum likelihood faces even in quite regular inverse problems is the existence of infinitely many maximizers. For example, in the interpolation problem with Gaussian errors, the likelihood function is

$$\mathcal{L}(\theta|X = x) = \prod_{j=1}^{n} (2\pi)^{-1/2} \exp\left(-\frac{(x_j - \theta(t_j))^2}{2}\right). \tag{4.7}$$

This attains its maximum, $(2\pi)^{-n/2}$, for every function θ that passes through all the data points $\{(t_j, x_j)\}_{j=1}^{n}$.

It is common to minimize the negative of the logarithm of the likelihood function, instead of maximizing the likelihood function. The likelihood function is nonnegative, so its logarithm is defined; the logarithm is monotonic, so a value of θ' that maximizes the log-likelihood also maximizes the likelihood. When the data errors are independent, the likelihood function factors, so taking the logarithm yields a sum of terms, one for each datum. Furthermore, if the errors are Gaussian, the logarithm inverts the exponentiation in the Gaussian density. For example, the log-likelihood in the interpolation example is

$$\ell(\theta'|X = x) = \log \mathcal{L}(\theta'|X = x) = c_1 - c_2 \sum_{j=1}^{n} (x_j - \theta'(t_j))^2, \tag{4.8}$$

where c_1 and c_2 are positive constants. Minimizing the negative of the log-likelihood in this case leads to minimizing the sum of squares of the discrepancies between the model predictions $\{\theta(t_j)\}_{j=1}^n$ and the data $\{x_j\}_{j=1}^n$: least squares.

One can modify the problem by incorporating a strictly concave penalty term to obtain a problem with a unique maximum. *Maximum penalized likelihood* subtracts a nonnegative penalty term $J(\theta)$ from the likelihood function (or the log of the likelihood function) before maximizing it. The penalty functional (or *regularization functional*) J is typically the square of a Hilbertian norm or seminorm. Maximum penalized likelihood is a form of regularization. Indeed, many regularization schemes can be viewed as maximum penalized likelihood estimators for different choices of the penalty functional. Including the penalty can stabilize the numerical problem; it need not result in an estimator with good statistical properties. See [51] for a treatment of quadratic regularization in the context of geophysical inverse theory; see [50] for a statistical perspective on quadratic regularization in inverse problems, and [59] for a recent tutorial. In the interpolation problem with Gaussian errors, including a positive definite quadratic penalty leads to linear estimates of θ.

For example, in the interpolation problem in W_2, we might choose $\hat{\theta}$ to be

$$\hat{\theta}_\lambda \equiv \arg \min_{\theta' \in \mathcal{T}} \left\{ \sum_{j=1}^{n} (x_j - \theta'(t_j))^2 + \lambda \|\theta'\|^2 \right\} \tag{4.9}$$

for some $\lambda > 0$ (standard choices for λ are discussed below); then $J(\theta) = \lambda\|\theta\|^2$. (That the minimizer exists in this problem is a consequence of the projection theorem, which allows us to conclude that the minimizer is finite-dimensional; *vide infra*.) This optimization problem has a unique solution (a linear combination of $\{\kappa_j\}_{j=1}^n$, which is a cubic spline). The constant λ is called the *regularization parameter*.

Next, we shall find the solution to the optimization problem 4.9 to characterize $\hat{\theta}_\lambda$.

Lemma 4.1 (The Projection Theorem). *Let $\mathcal{T}$ be a Hilbert space, let $\{\kappa_j\}_{j=1}^n \subset \mathcal{T}$ be a linearly independent set, let $M = \mathrm{span}\{\kappa_j\}_{j=1}^n$, and let $x \in \mathbb{R}^n$. Then*

$$\arg\min_{\theta' \in \mathcal{T}} \{\|\theta'\| : \langle \kappa_j, \theta' \rangle = x_j, \quad j = 1, \cdots, n\} \in M. \tag{4.10}$$

The minimum is attained, and by a unique element of M.

Remark 4.2. For a proof, see, *e.g.* [48], §3.10, Thm. 2. Similarly, if M is a closed subspace of a Hilbert space $\mathcal{T}$ and θ is an arbitrary element of $\mathcal{T}$,

$$\min_{m \in M} \|\theta - m\| \tag{4.11}$$

is attained by an element m_0 of M, and $\theta - m_0 \in M^\perp$. See [48], §3.3, Thm. 2.

Remark 4.3. It follows from Lemma 4.1 that if $D \subseteq \mathbb{R}^n$ is closed, and θ_0 solves

$$\theta_0 = \arg\min\{\|\theta'\| : K\theta' \in D\}, \tag{4.12}$$

then $\theta_0 \in M = \mathrm{span}\{\kappa_j\}_{j=1}^n$.

Let Γ be the $n \times n$ Gram matrix with elements $\Gamma_{ij} = \langle \kappa_i, \kappa_j \rangle$, and for $\gamma \in \mathbb{R}^n$, let $\gamma \cdot \mathbf{k} = \sum_{j=1}^n \gamma_j \kappa_j$. That the matrix Γ is positive definite follows from the linear independence of $\{\kappa_j\}_{j=1}^n$. The penalized maximum likelihood estimate is of the form $\gamma \cdot \mathbf{k}$. For such a model, we have

$$\begin{aligned} \|\gamma \cdot \mathbf{k}\|^2 &= (\gamma \cdot \mathbf{k}, \gamma \cdot \mathbf{k}) \\ &= \gamma \cdot \Gamma \cdot \gamma. \end{aligned} \tag{4.13}$$

The vector of noise-free data predictions of such a model is

$$K(\gamma \cdot \mathbf{k}) = \gamma \cdot \Gamma, \tag{4.14}$$

and

$$\begin{aligned} \|x - K(\gamma \cdot \mathbf{k})\|^2 &= (x - \gamma \cdot \Gamma) \cdot (x - \gamma \cdot \Gamma) \\ &= \|x\|^2 - 2\gamma \cdot \Gamma \cdot x + \gamma \cdot \Gamma \cdot \Gamma \cdot \gamma. \end{aligned} \tag{4.15}$$

Thus the penalized log-likelihood of γ is (ignoring additive terms that do not involve the parameter γ)

$$\ell(\gamma|x) = \frac{1}{2}\left(-2\gamma \cdot \Gamma \cdot x + \gamma \cdot \Gamma \cdot \Gamma \cdot \gamma + \lambda\gamma \cdot \Gamma \cdot \gamma\right). \tag{4.16}$$

This is a convex quadratic functional of γ, so its minimum is attained at a stationary point $\tilde{\gamma}$:

$$\ell'(\tilde{\gamma}|x) = -\Gamma \cdot x + \Gamma \cdot \Gamma \cdot \tilde{\gamma} + \lambda\Gamma \cdot \tilde{\gamma} = 0. \tag{4.17}$$

Solving for $\tilde{\gamma}$ gives the familiar expression

$$\tilde{\gamma} = (\Gamma + \lambda I)^{-1} x, \tag{4.18}$$

where I is the identity matrix (recall that Γ is positive definite, so $\Gamma + \lambda I$ is too, and hence $\Gamma + \lambda I$ is invertible). Thus the maximum penalized likelihood estimate of θ for regularization parameter $\lambda \geq 0$ is

$$\hat{\theta}_\lambda(X) = \left((\Gamma + \lambda I)^{-1} X\right) \cdot \mathbf{k}. \tag{4.19}$$

Let $\theta_M = \gamma_0 \cdot \mathbf{k}$, so $\gamma_0 = \Gamma^{-1}K\theta$, $\theta_M = \Gamma^{-1}K\theta \cdot \mathbf{k}$, and $\theta_{M^\perp} = \theta - \theta_M$. The bias of the penalized estimator is

$$\begin{aligned} \mathbf{bias}_\theta(\hat{\theta}_\lambda) &= \mathbb{E}_\theta\left[\theta - \left((\Gamma + \lambda I)^{-1} X\right) \cdot \mathbf{k}\right] \\ &= \theta - \mathbb{E}_\theta\left[(\Gamma + \lambda I)^{-1}(K\theta + \epsilon)\right] \cdot \mathbf{k} \\ &= \theta - \left((\Gamma + \lambda I)^{-1} K\theta\right) \cdot \mathbf{k} \\ &= \theta_{M^\perp} + \left(\Gamma^{-1} - (\Gamma + \lambda I)^{-1}\right) K\theta \cdot \mathbf{k}. \end{aligned} \tag{4.20}$$

The first term is the part of the model θ that is in the null space of the forward mapping: the statistic $\hat{\theta}_\lambda$ estimates $\theta_{M^\perp}$ by 0. As a result, the bias is unbounded as θ ranges over Θ. If a prior bound on $\|\theta\|$ were available, that would limit the bias. In some inverse problems, such as estimating the magnetic field at the Earth's core-mantle boundary from satellite observations [9], physically motivated *a priori* bounds on quadratic functionals are available, but such cases seem to be rare. The variance of $\hat{\theta}_\lambda$ is

$$\begin{aligned}
\mathbf{Var}_\theta(\hat{\theta}_\lambda)) \equiv \mathbb{E}_\theta \|\hat{\theta}_\lambda - \mathbb{E}_\theta \hat{\theta}_\lambda\|^2 &= \mathbb{E}_\theta \| \left((\Gamma + \lambda I)^{-1} X \right) \cdot \mathbf{k} - \left((\Gamma + \lambda I)^{-1} K\theta \right) \cdot \mathbf{k} \|^2 \\
&= \mathbb{E}_\theta \| \left((\Gamma + \lambda I)^{-1} \epsilon \right) \cdot \mathbf{k} \|^2 \\
&= \mathbb{E}_\theta \left[\left((\Gamma + \lambda I)^{-1} \epsilon \right) \cdot \Gamma \cdot \left((\Gamma + \lambda I)^{-1} \epsilon \right) \right] \\
&= \mathbb{E}_\theta \left[\epsilon \cdot (\Gamma + \lambda I)^{-1} \cdot \Gamma \cdot (\Gamma + \lambda I)^{-1} \cdot \epsilon \right] \\
&= \operatorname{tr} \left((\Gamma + \lambda I)^{-1} \cdot \Gamma \cdot (\Gamma + \lambda I)^{-1} \right). \qquad (4.21)
\end{aligned}$$

The mean squared error of $\hat{\theta}_\lambda$ is

$$\mathrm{MSE}_\theta(\hat{\theta}_\lambda) = \|\mathbf{bias}_\theta(\hat{\theta}_\lambda)\|^2 + \mathbf{Var}_\theta(\hat{\theta}_\lambda). \qquad (4.22)$$

An approximate maximum penalized likelihood estimator $\hat{\theta}_\epsilon$ is one that nearly maximizes the penalized likelihood, in the sense that

$$\mathcal{L}(\hat{\theta}_\epsilon | X = x) - \lambda J(\hat{\theta}_\epsilon) \geq \sup_{\theta \in \Theta} \mathcal{L}(\theta | X = x) - \lambda J(\theta) - \epsilon. \qquad (4.23)$$

This notion is fruitful when one has a sequence of estimation problems with increasing numbers of data. Then, with appropriate conditions on the models, if ϵ and λ are driven to zero at the right rates as the number of data grows, such a sequence of approximate maximum penalized likelihood estimators can be consistent and efficient in various senses; see, *e.g.*, [53].

Note that the likelihood function can be replaced by other functions of the parameter and the data that then can be maximized to construct an estimator; estimators that solve general optimization problems are called *M-estimators*. For example, using least-squares to fit a linear model to data with non-Gaussian errors is a form of M-estimation. For results concerning the consistency of least-squares estimators in nonparametric regression and inverse problems for not-necessarily-Gaussian data errors, see [27, 49, 61, 65].

4.2.1. *Choosing the Regularization Parameter λ.* From the point of view of stability, any strictly positive value of λ suffices; the variance decreases and the bias tends to increase as λ increases. If λ is chosen *a priori*, the regularized estimate is linear in the data. If the data are also used to select λ adaptively, the estimator is nonlinear in the data.

A common way to choose λ adaptively in geophysical inverse problems leads to "Occam's inversion," named after William of Occam by Constable *et al.* [15]. Occam's Razor demands that when choosing among competing hypotheses that explain the data adequately, one pick the simplest. The quantitative prescription in [15] is to pick the model that attains the smallest value of the regularization functional among models that predict the data within a normalized sum of squared errors equal to one:

$$\frac{1}{n}\sum_{j=1}^{n}(X_j - \langle \kappa_j, \theta\rangle)^2/\sigma_j^2 = 1. \tag{4.24}$$

For independent Gaussian errors, the expectation of the left hand side for the true value of θ is unity; the value on the right can be tuned more finely to be a quantile of the distribution of the sum of squares of the errors.

Genovese and Stark [36] give necessary conditions for such estimators to be consistent, and sufficient conditions, but not necessary and sufficient conditions. O'Sullivan [50] presents regularization of inverse problems from a statistical point of view, and gives various senses in which regularized estimates are optimal among linear estimates, along with a discussion of methods for selecting λ.

Cross-validation and *generalized cross-validation* are popular methods in the statistical literature for choosing the regularization parameter; see Wahba [63] for the nonparametric regression case and O'Sullivan [50] for applications to inverse problems and further references; see [59] for an accessible summary. Cross validation is a method for choosing λ adaptively: one forms n data sets of size $n-1$, omitting each datum in turn. Let $X_{(j)}$ denote the data set with the jth datum deleted, and let $\hat{\theta}_{\lambda,(j)}$ denote the regularized estimate based on the data set $X_{(j)}$ using λ as the value of the regularization parameter. The predictive error for regularization parameter λ is

$$\mathrm{PE}(\lambda) = \sum_{j=1}^{n}(X_j - \langle \kappa_j, \hat{\theta}_{\lambda,(j)}\rangle)^2/\sigma_j^2. \tag{4.25}$$

One selects λ to minimize the predictive error. The dependence of λ on the data makes the resulting estimator of θ nonlinear. Predicting omitted data and recovering the underlying model θ are not the same. See Wahba [63] for more more details about cross validation and its generalizations, and for the connection between the theory of splines and Bayesian nonparametric regression.

4.2.2. *Regularization as a Method for Inference.* If the misfit tolerance in Occam's inversion is chosen appropriately (for example, the quantile $\chi^2_{n,1-\alpha}/n$ in the case of Gaussian errors), the minimal value of $J(\theta)$ can be interpreted as the lower endpoint of a $1-\alpha$ one-sided confidence interval for J applied to the true model: the set of models that agree adequately with the data is a confidence set $\mathcal{D}$ for the model, as described in the development leading to Equation (4.57). Typically, the upper endpoint of a confidence interval for J is infinite; see [17] for examples of functionals for which only one-sided confidence intervals can be constructed.

4.2.3. *The Method of Sieves.* Adding a penalty to the likelihood function—regularization—is not the only way to construct an estimator as the unique maximizer in some optimization problem. An alternative way to modify maximum likelihood to arrive at a problem with a unique maximizer is to limit the dimension of the model to obtain an over-determined problem; the *method of sieves* implements this approach. See Shen [53] for a recent study of MLE, penalized MLE, and the method of sieves.

Let $\{\Theta_k\}_{k=1}^{\infty}$ be a sequence of subsets of the Banach space $\mathcal{T}$ that contains Θ. Let $d_k(\theta) = \inf_{\gamma\in\Theta_k}\|\gamma-\theta\|$. If for all $\theta\in\Theta$, $\lim_{k\to\infty} d_k(\theta) = 0$, $\{\Theta_k\}_{k=1}^{\infty}$ is a *sieve*. The collection of subspaces spanned by the first $k \ge 1$ elements of an ordered basis is a typical sieve. The idea of the method of sieves is to maximize the likelihood approximately within the approximating set Θ_n: find $\hat{\theta}_n \in \Theta_n$ such that

$$L(\hat{\theta}_n|X=x) \ge \sup_{\theta\in\Theta_n} L(\theta|X=x) - \epsilon, \tag{4.26}$$

where n is the number of data.

In a sequence of estimation problems with increasing numbers of data n, if the sieve is chosen appropriately and ϵ is driven to zero at the right rate as n grows, this method can have desirable properties such as consistency.

The method of sieves is quite close to a common numerical approach to inverse problems: approximate Θ by a finite-dimensional subspace and perform least-squares collocation within the subspace. Unfortunately, as mentioned above, the choice of the approximating space matters—it controls the bias/variance tradeoff—and good results for sieve estimators and other regularizing methods depend critically on assumptions about θ that tend to be unverifiable in applications.

4.2.4. *Singular Value Truncation and Weighting.* Singular value truncation is another way to choose an approximating subspace. For a review of the applied mathematics perspective on singular value truncation and related regularization methods, see [13, 14, 12, 62]. We assume in this section that $\mathcal{T}$ is a Hilbert space, and that the data errors $\{\epsilon_j\}_{j=1}^n$ are independent and identically distributed with mean zero and variance σ^2. The linear operator $K : \mathcal{T} \to \mathbb{R}^n$ is compact; it has an infinite-dimensional null-space. Because the functionals $\{\kappa_j\}_{j=1}^n$ are linearly independent *ex hypothesi*, the orthogonal complement of the null space of K is n-dimensional. Let $K^* : \mathbb{R}^n \to \mathcal{T}$ be the adjoint operator to K. There is a collection of n triples $\{(\nu_j, x_j, \lambda_j)\}_{j=1}^n$ with $\nu_j \in \mathcal{T}$, $x_j \in \mathcal{X}$ and $\lambda_j \in \mathbb{R}^+$, such that

$$K\nu_j = \lambda_j x_j \text{ and} \tag{4.27}$$

$$K^* x_j = \lambda_j \nu_j. \tag{4.28}$$

The functions $\{\nu_j\}_{j=1}^n$ can be chosen to be orthonormal in $\mathcal{T}$, and the vectors $\{x_j\}_{j=1}^n$ can be chosen to be orthonormal in $\mathcal{X}$; we assume both conditions hold. In this problem, the singular values $\{\lambda_j\}$ are strictly positive (a consequence of the linear independence of $\{\kappa_j\}_{j=1}^n$). We assume that the singular values are ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. The triples $\{(\nu_j, x_j, \lambda_j)\}_{j=1}^n$ comprise the *singular value decomposition* of the operator K.

By virtue of the projection theorem (4.1), a model ν_* of minimum norm that satisfies the data X exactly can be written as a linear combination of the singular functions $\{\nu_j\}_{j=1}^n$; namely,

$$\hat{\theta}_{MN} = \nu_*(X) = \sum_{j=1}^n \lambda_j^{-1}(x_j \cdot X)\nu_j. \tag{4.29}$$

To verify that $\hat{\theta}_{MN}$ satisfies the data, calculate

$$\begin{aligned} K\hat{\theta}_{MN} &= \sum_{j=1}^n \lambda_j \lambda_j^{-1}(x_j \cdot X)x_j \\ &= X, \end{aligned} \tag{4.30}$$

because $\{x_j\}_{j=1}^n$ is an orthonormal basis for $\mathbb{R}^n$. Because of the linear independence, $\hat{\theta}_{MN}$ is the only linear combination of the singular functions $\{\nu_j\}_{j=1}^n$ that fits the data X exactly, hence it is the function of minimum norm that fits the data exactly.

Let us calculate the bias and variance of the estimator $\hat{\theta}_{MN}(X)$. To begin, we decompose θ into a component $\theta_{||}$ in the span of $\{\nu_j\}_{j=1}^n$ and a component $\theta_\perp$ in the null space of K. It is easily seen that $\mathbb{E}_\theta \hat{\theta}_{MN} = \theta_{||}$, so

$$\begin{aligned} \mathbf{bias}_\theta(\hat{\theta}_{MN}) &= \mathbb{E}_\theta \hat{\theta}_{MN}(X) - \theta \\ &= \theta_\perp. \end{aligned} \tag{4.31}$$

The variance of $\hat{\theta}_{MN}$ is

$$\begin{aligned} \mathbf{Var}_\theta \hat{\theta}_{MN} &= \mathbb{E}_\theta \left\| \sum_{j=1}^{n} \lambda_j^{-1} (x_j \cdot \epsilon) \nu_j \right\|^2 \\ &= \sigma^2 \sum_{j=1}^{n} \lambda_j^{-2}. \end{aligned} \tag{4.32}$$

The components associated with small singular values λ_j contribute substantially to the variance, because the corresponding components of noise in the data are multiplied by λ_j^{-1}. The idea behind singular value truncation is to reconstruct θ using only those singular functions whose singular values exceed some threshold t; that is, to estimate θ by

$$\hat{\theta}_{SVT} = \sum_{j=1}^{m} \lambda_j^{-1} (x_j \cdot X) \nu_j, \tag{4.33}$$

where $m = \max\{k : \lambda_k > t\}$. This mollifies the noise magnification, at the expense of increasing bias: the bias increases in norm by an amount equal to the norm of the projection of θ onto the subspace spanned by $\{\nu_j\}_{j=m+1}^n$. The variance decreases by $\sigma^2 \sum_{j=m+1}^n \lambda_j^{-2}$. If one has adequate prior information about θ (to control the increment to the bias) this bias-variance tradeoff can be exploited to construct an estimator with smaller mean squared error than the minimum norm estimator has.

Singular value truncation can be embedded in a family of estimators that re-weight the singular functions in the reconstruction: regularization using the norm is another. (Maximum entropy [41, 21] also can be viewed as a nonlinear regularization method.) The general form is

$$\hat{\theta}_w = \sum_{j=1}^{n} w(\lambda_j)(x_j \cdot X) \nu_j. \tag{4.34}$$

Singular value truncation corresponds to

$$w(u) = \begin{cases} u^{-1}, & u \geq \lambda_{j_0}, \\ 0, & \text{otherwise.} \end{cases} \tag{4.35}$$

Regularization using the norm as the regularization functional, with regularization parameter λ, corresponds to

$$w(u) = \frac{u}{u^2 + \lambda}. \tag{4.36}$$

See [50, 62] for discussions of regularization as a statistically optimal linear estimator.

Penalization, sieves, and singular-value truncation work essentially by forcing the estimator to lie in a compact subset of Θ, but allowing that subset to grow in a controlled way as $n \to \infty$. These methods tend to produce an estimate with norm smaller than that of the maximum likelihood estimate: they can be viewed as shrinkage estimators (see Section 4.5). Consistency and optimality results for penalized maximum likelihood and the method of sieves depend on assumptions about θ that control the bias of the estimator; for example, if θ is a function, some smoothness is required. The choices of $\epsilon = \epsilon(n)$ and (Θ_n) or $J(\theta)$ and $\lambda = \lambda(n)$ matter, and optimal rules tend to depend on assumptions about θ that cannot be verified empirically. Moreover, to our knowledge, the finite-sample behavior of approximate penalized likelihood and the method of sieves are not guaranteed to be good, even when all the technical conditions are met—the theorems are asymptotic as $n \to \infty$.

4.3. **Bayes Estimation.** The description here is based in part on Lehmann and Casella [47], chapter 4; see Hartigan [38] and Berger [11] for more complete and technical expositions of the Bayesian approach, Gelman *et al.* [35] for Bayesian data analysis; and Le Cam [44] for a more general and rigorous development. See Tarantola [58] for a Bayesian treatment of finite-dimensional geophysical inverse problems.

One of the fundamental tenets of Bayesian inference is that uncertainty always can be represented as a probability distribution; in particular, the Bayesian approach treats the model θ as the outcome of a random experiment. According to I.J. Good [37]

> ... the essential defining property of a Bayesian is that *he regards it as meaningful to talk about the probability $P(H|E)$ of a hypothesis H, given evidence E.*

Whether one adheres to a Bayesian view, estimators that arise from the Bayesian approach can have attractive frequentist properties: the proof of the pudding is in the eating.

The Bayesian paradigm requires a little extra structure:

- The parameter set Θ must be measurable. The probability measure $\mathbb{P}_\theta$ is now thought of as a conditional probability distribution for the data, given that the randomly chosen model has the value θ.
- We assume that the probabilities $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ are dominated by a common σ-finite measure μ; the density of $\mathbb{P}_\theta$ with respect to μ is denoted $p_\theta(x)$ as in Section 4.2. Recall that $p_\theta(x)$ is the likelihood $\mathcal{L}(\theta|x)$. We shall assume that $p_\theta(x)$ is jointly measurable with respect to θ and x.
- Before any data are collected, a Bayesian has a *prior probability distribution* π for the unknown model θ—indeed, a Bayesian can estimate θ without data.

In addition to the probability distributions $\mathbb{P}_\theta$ on $\mathcal{X}$, we now have a probability distribution on another space, Θ; let $\mathbb{E}_\pi$ denote expectation with respect to π. From the prior distribution and the distribution $\mathbb{P}_\theta$ of the data X given θ, we can find the joint distribution of θ and X. The *marginal distribution of* X or the *predictive distribution of* X is a mixture of the distributions $\mathbb{P}_\theta$ according to π. When $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is dominated by μ, as we have assumed, the density with respect to μ of the predictive distribution is

$$m(x) = \int p_\theta(x)\, \pi(d\theta). \tag{4.37}$$

Observing data allows us to update the prior probability distribution using $p_\theta(x)$ (the density of the observation given θ) and Bayes' rule to find the *posterior distribution of* θ *given* $X = x$:

$$\pi(d\theta|X = x) = \frac{p_\theta(x)\ \pi(d\theta)}{m(x)}. \tag{4.38}$$

(Note that $m(x)$ can vanish; this happens with probability 0. It is not necessary that $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ be dominated for the posterior distribution to exist, but it makes the formula simple.)

At this point, some Bayesians are done: the fundamental objects in doctrinaire Bayesian analysis are probability distributions, and (given the prior) the posterior distribution $\pi(d\theta|X = x)$ is all there is to know about θ.

Computing the posterior distribution is much like computing the Gibbs distribution in statistical mechanics, and the difficulties in computing the denominator of the posterior distribution are akin to those encountered in computing the partition function—both are integrals over high-dimensional or infinite-dimensional sets. A considerable amount of computational machinery has been developed to evaluate such integrals or to compute the posterior without calculating the integral explicitly (Markov-chain Monte-Carlo, Gibbs sampling, *etc.*); see Gelman *et al.* [35] for examples and references, and Tenorio *et al.* [60] for an example in an inverse problem in microwave cosmology.

Tarantola [58] presents a Bayesian approach to inverse problems, but truncates the problems to finite dimensions *ab initio*; moreover, the prior probability distributions and data error distributions he treats computationally are limited primarily to Gaussians. Many of the difficulties of the Bayesian approach vanish if all the distributions are Gaussian and the dimension of the model is finite. See below for a few of those problems and references. Backus [7] gives a Bayesian treatment of an infinite-dimensional inverse problem in geomagnetism; Backus [10] develops a framework for inference with quadratic constraints that defers the frequentist/Bayesian decision by showing that, for his method, the mathematics does not depend on the interpretation.

Some Bayesians and most frequentists are interested in estimating a parameter $g(\theta)$; the posterior distribution of θ induces a probability distribution for $g(\theta)$. One can also use the posterior distribution to find point estimates that minimize the posterior risk with respect to some loss function. For example, the mean of the posterior distribution of $g(\theta)$ minimizes the posterior mean squared error. Given a loss function $\ell(\zeta, a) : \mathcal{G} \times \mathcal{G} \to \mathbb{R}^+$, recall that the risk at θ of the estimator $\delta : \mathcal{X} \to \mathcal{G}$ of the value at θ of the parameter g is

$$r(\theta, \delta) = \mathbb{E}_\theta \ell(g(\theta), \delta(X)). \tag{4.39}$$

The *average risk of δ for prior π on* Θ is

$$\rho_\pi(\delta) = \int_\Theta r(g(\theta), \delta)\pi(d\theta). \tag{4.40}$$

As noted in Section 3, an estimator δ^*_π that minimizes the average risk for prior π is called a *Bayes estimator*; its risk is the *Bayes risk for prior* π. Under suitable conditions, δ^*_π exists; for example, if there exists an estimator with finite risk, and if a minimizer $\delta_\pi(x)$ of

$$\mathbb{E}_\pi(\ell(g(\theta), \delta(x)|X = x) \tag{4.41}$$

exists for almost all x and depends measurably on x, then δ_π is a Bayes estimator for prior π and loss ℓ.

Under appropriate conditions, every admissible estimator is either a Bayes estimator for some prior, or a limit of such estimators. See Le Cam [44] for details. The maximum risk over $\theta \in \Theta$

of a Bayes estimator can be quite large. Whatever be the probability measure π on Θ

$$\sup_{\theta \in \Theta} r(g(\theta), \delta) \geq \int_\Theta r(g(\theta), \delta) \pi(d\theta) : \tag{4.42}$$

the average risk lower-bounds the maximum risk. Because the Bayes estimator minimizes the right hand side, the maximum risk of any estimator δ is bounded below by the Bayes risk. In particular, the *minimax risk*

$$\inf_\delta \sup_{\theta \in \Theta} r(g(\theta), \delta) \tag{4.43}$$

(see Section 4.4) is bounded below by the Bayes risk for any prior π. In many circumstances, the minimax risk is equal to the maximum Bayes risk over all priors π on Θ; see Le Cam [44] for precise theorems. Both the Bayes risk and the Bayes estimator depend on the prior π; comparing the Bayes risk to the minimax risk is a way to quantify the sensitivity of the risk to the particular prior chosen.

The Bayesian analogue of a frequentist confidence region is called a *credible region* or *credible set*. A level $1 - \alpha$ credible region for the parameter $g(\theta)$ is a set $S(X) \subset \mathcal{G}$ that contains $g(\theta)$ with posterior probability at least $1 - \alpha$:

$$\pi\{g(\theta) \in S(x) | X = x\} \geq 1 - \alpha \text{ for all } x \in \mathcal{X}. \tag{4.44}$$

There is not a unique S with this property; a criterion often used to obtain a unique S is to take S to be a level set of the posterior distribution of $g(\theta)$. Another is to introduce a loss function associated with a measure of the "size" or "volume" of the confidence set (as we discussed in a frequentist context), and to find the region that minimizes that loss or the risk subject to the posterior coverage constraint.

The frequentist coverage probability of a $1 - \alpha$ Bayesian credible region is typically less than $1 - \alpha$. For example, we have seen the following misleading procedure used in applications: Define the posterior distribution for θ to be proportional to the likelihood function, implicitly defining a uniform prior for the parameter θ. Find a $1 - \alpha$ credible region for θ from this posterior. Finally, report, incorrectly, that the credible region is a $1 - \alpha$ confidence region.

4.3.1. *Stochastic Inversion.* Stochastic inversion (*e.g.*, [32]) is a form of quadratic regularization equivalent to Bayesian estimation for mean squared error loss under a Gaussian prior probability distribution for the coefficients in some finite-dimensional approximation of the model. The distinction is philosophical: in stochastic inversion, the model coefficients are viewed as being drawn at random from a population with known parameters, while in Bayesian estimation, the model coefficients are viewed as unknowns with prior probability distributions with known parameters. See [7] for more discussion and a comparison of Bayesian estimation and stochastic inversion in a linear inverse problem in geomagnetism with quadratic prior information.

4.3.2. *Difficulties and Pathologies in Bayes Estimation.* In both the Bayesian and the frequentist approaches, the choice of Θ captures some *a priori* information about appropriate models for the data. For example, if θ represents a physical quantity, such as an absorption coefficient, which must be in the interval $[0, 1]$, then it makes sense to take $\Theta = [0, 1]$ even though we may have a class of models with an analytic form for $\mathbb{P}_\theta$ (for instance, $X \sim N(\theta, 1)$) that is defined for all $\theta \in \mathbb{R}$.

A frequentist can stop here if there is no more prior information, but a Bayesian must proceed to choose a prior π. In such situations many Bayesians try to choose a "non-informative" prior that contains no additional information about θ by making all points in Θ "equally likely." For example, if $\Theta = [0, 1]$, then it is common to take the prior probability distribution π to be the $U[0, 1]$ distribution. Although this notion seems appealing at first glance, the Bayes risk for a non-informative prior can be much smaller than the Bayes risk for other priors, indicating that estimation is easier for the non-informative prior and hence that the non-informative prior does contain useful information about θ.

In the same vein, Backus [6, 8, 7] points out the difficulty of capturing "hard" constraints, such as $\|\theta\| \le 1$, using prior probability distributions in high- or infinite-dimensional spaces. Consider, for example, the case Θ is the subset $\{\theta : \|\theta\| \le 1\}$ of an infinite-dimensional separable Hilbert space $\mathcal{T}$ (such constraints arise in geomagnetism, for example). Then Θ is rotationally invariant, so absent other information, it is reasonable to insist that π be rotationally invariant in $\mathcal{T}$ as well. Backus shows that any rotationally invariant prior on $\mathcal{T}$ assigns probability one to the event $\{\|\theta\| = 0\}$: it is not possible to capture the constraint $\|\theta\| \le 1$ as a prior probability distribution without injecting additional information (imposing a preference for directions in

$\mathcal{T}$). Freedman [33] gives a complete characterization of infinite-dimensional probability distributions that are rotationally invariant in all finite-dimensional subspaces: they are mixtures of independent zero-mean Gaussian random variables.

A rejoinder to such remarks about the difficulty of choosing a suitable prior is that if there are sufficiently many data, the prior does not matter; the estimator will be close to the truth no matter what—the estimator is *frequentist consistent.* Formally, a Bayesian estimator is frequentist consistent for $g(\theta)$ if for each $\theta \in \Theta$ and each neighborhood τ of $g(\theta)$, the posterior probability that $g(\theta) \in \tau$ given X converges to one almost surely $\mathbb{P}_\theta$ for all $\theta \in \Theta$ as $n \to \infty$. Under appropriate conditions Bayes estimators in smooth, finite-dimensional problems are frequentist consistent. However, Diaconis and Freedman [16] show that in nonparametric regression, a canonical inverse problem, whether a Bayes estimator is frequentist consistent depends on the prior. In particular, a hierarchical mixture of uniform priors on nested finite-dimensional spaces leads to inconsistency in the problem they study.

4.4. **Minimax Estimation.** The minimax risk ρ^* is the smallest worst-case risk over $\theta \in \Theta$, over a class $\mathcal{D}$ of decisions:

$$\rho^* = \rho^*(\Theta, \mathcal{P}, \ell, \mathcal{D}) \equiv \inf_{\delta \in \mathcal{D}} \sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \ell(\theta', \delta(X)). \tag{4.45}$$

If there exists an estimator $\delta^* \in \mathcal{D}$ that has maximum risk ρ^*, it is a *minimax estimator.*

The class of estimators $\mathcal{D}$ might be all measurable functions of X, or we might limit the complexity of the estimator, for example, by considering only linear, affine (inhomogeneous linear), or quadratic estimators.

This approach is very appealing to many frequentists, and concrete results for inverse problems are possible. For example, Donoho [19] studies minimax estimation of a linear functional g of an element of a convex subset Θ of $\mathcal{T} = \ell_2$ from linear data $K\theta$ contaminated by additive i.i.d. zero-mean Gaussian errors. Donoho considers three loss functions: squared error, absolute error, and the length of a fixed-length confidence interval with a specified minimum coverage probability. He shows for all these measures that the risk of the minimax affine estimator is within a fraction of the risk of the minimax measurable estimator, and shows how to construct the minimax affine estimator. He shows that the risk of the hardest one-dimensional subproblem is equal to the risk in the original infinite-dimensional problem, which allows him to use results about estimating the mean γ of a normal distribution subject to the constraint that $\gamma \in [-\tau, \tau]$ to find the risk in the original problem. The fundamental entity in the development is the *modulus of continuity of* g:

$$\omega(\epsilon; g, K, \Theta) \equiv \sup_{\theta_1, \theta_2 \in \Theta} \{|g(\theta_1) - g(\theta_2)| : \|K\theta_1 - K\theta_2\| \leq \epsilon\}. \tag{4.46}$$

This quantity also arises in the theory of optimal recovery, and Donoho establishes an equivalence between statistical estimation in Gaussian noise and optimal recovery in deterministic noise, through a re-calibration of the noise level. The results do not seem to translate to non-Gaussian noise, nor to estimating the whole object θ; however, Donoho [18] addresses minimax estimation of a function $\theta = \theta(t)$ in nonparametric regression and inverse problems, with L_∞ loss

$$\ell(\theta, g) = \sup_t |\theta(t) - g(t)|, \tag{4.47}$$

and Donoho and Nussbaum [26] study minimax estimation of a quadratic functional in a similar setting.

Donoho's [19] approach has been applied to inverse problems in geomagnetism [55] and microwave cosmology [60].

4.5. **Shrinkage Estimation.** In contrast with the problem of estimating a single linear functional of a model θ in a convex set Θ, where, as noted above, affine estimators are nearly optimal, when one seeks to estimate three or more linear functionals, nonlinear estimators can reap large benefits in mean squared error even when there is no prior constraint on θ, by exploiting a bias/variance tradeoff.

Suppose that X has a d-variate normal distribution with independent coordinates: $X \sim N(\theta, I)$, $d \geq 3$. Charles Stein proved in 1956 the surprising result that for squared-error loss, the maximum likelihood estimator of θ, namely X, is not admissible for θ [57]. He showed that estimators that "shrink" the observations nonlinearly towards the origin dominate the sample mean; in particular,

$$\delta_{\mathrm{S}}(x) = \left(1 - \frac{\alpha}{\beta + \|x\|^2}\right) x \tag{4.48}$$

has uniformly smaller mean squared error than $\delta(x) = x$ when α is sufficiently small and β is sufficiently large. There is no consensus definition of *shrinkage estimate*, but the general flavor of the term is that the estimate derives from a simpler estimate such as maximum likelihood by moving the result towards some distinguished set, such as a subspace or the origin.

Stein's original result has been refined and extended in a variety of ways. James and Stein [40] showed that

$$\delta_{\mathrm{JS}}(x) = \left(1 - \frac{\alpha}{\|x\|^2}\right) x \tag{4.49}$$

with $0 < \alpha \leq 2(d-2)$ suffices; $\alpha = d-2$ is optimal in this family. For further generalizations (*e.g.*, to distributions other than the normal) and references, see [31]. The variable

$$\left(1 - \frac{\alpha}{\|x\|^2}\right) \tag{4.50}$$

is called the *shrinkage factor*. The James-Stein estimator has the slightly unsavory feature that the shrinkage factor can be negative, yielding an estimator that both shrinks and reflects the data. Indeed, the *James-Stein positive part estimator*

$$\delta^{+}_{\mathrm{JS}}(x) = \left(1 - \frac{d-2}{\|x\|^2}\right)^{+} x \tag{4.51}$$

dominates the James-Stein estimator. The positive-part estimator is not minimax, but it is hard to improve upon [47].

Stein's result has implications for Backus-Gilbert theory: if there are three or more data, and one seeks to estimate three or more linear functionals of the model, a shrinkage estimator sometimes can do better in mean-squared error than the Backus-Gilbert unbiased estimates. The following theorem is relevant:

Theorem 4.4. *Lehmann and Casella* [47], *Theorem 5.7.* *Let $X \sim N(\theta, \Sigma)$ in dimension $d \geq 3$. Let $\mathrm{tr}(\Sigma)$ be the trace of Σ and let $\lambda_{\max}(\Sigma)$ be the largest eigenvalue of Σ. For squared error loss $\ell(\theta, a) = \|a - \theta\|^2$, the estimator*

$$\delta(x) = \left(1 - \frac{c(\|x\|^2)}{\|x\|^2}\right) x \tag{4.52}$$

is minimax provided

(1) $0 \leq c(\|x\|^2) \leq 2\mathrm{tr}(\Sigma)/\lambda_{\max}(\Sigma) - 4$, and

(2) $c(\cdot)$ is nondecreasing.

Because the risk of X is constant, it follows that this estimator dominates X in mean squared error; moreover, taking its positive part improves it further.

The following argument might provide intuition into the condition on the trace versus the maximum eigenvalue: suppose that there is one coordinate with very high variance compared with the rest. Then the risk is driven by that coordinate, and the problem is essentially one-dimensional—but shrinkage does not help in one dimension (in the absence of *a priori* constraints on θ). Results about shrinkage are generally derived by positing an ansatz, then verifying that the resulting estimator dominates another, rather than giving general insight into the form an estimator must have to dominate.

As noted above in Section 4.2.4, singular value truncation and other regularization methods can be viewed as shrinkage estimators. If the shrinkage does not depend on the data (for example, in regularized least squares using a fixed value of the regularization parameter, or in singular value truncation retaining a fixed number of singular functions), the shrinkage estimator is linear.

4.6. **Wavelet and Wavelet-Vaguelette Shrinkage.** Recall from Section 4.2.4 that singular value truncation and singluar value weighting can be nearly minimax for mean squared error. The situations in which those estimators work well are those in which the prior information $\theta \in \Theta$ essentially ensures that the coefficients in an expansion of θ as a linear combination of the singular functions become small quickly with increasing index, so that down-weighting those components in the reconstruction (to control the variance of the estimator) does not incur large bias. When that is not the case, singular value truncation or weighting can be far from optimal.

Donoho, Johnstone, and co-authors have shown that in some estimation problems like non-parametric regression where the regression function is in a ball in a Besov or Triebel space and the errors are Gaussian, the following estimator can attain the asymptotic minimax risk for mean squared error: project the data onto an orthonormal basis of compactly supported wavelets; shrink the resulting empirical wavelet coefficients nonlinearly, one component at a time; reconstruct the function from the shrunk coefficients [20, 22, 25, 42, 23, 24]. Moreover, no linear method can attain the minimax rate in some of the problems in which wavelet shrinkage is asymptotically minimax for mean squared error. The keys to the success of wavelet shrinkage seem to be that expressing the prior information $\theta \in \Theta$ is easy in the basis—the representation of the object is *sparse* in the basis—and that the basis is unconditional. (An unconditional basis $\{t_j\}_{j=1}^{\infty}$ for a separable Banach space $\mathcal{T}$ is one for which if $\sum_{j=1}^{\infty} \beta_j t_j$ converges in $\mathcal{T}$ and $|\alpha_j| \leq 1$ for all j, then $\sum_{j=1}^{\infty} \alpha_j \beta_j t_j$ converges in $\mathcal{T}$.)

Donoho [20] extends this approach to inverse problems by introducing the the wavelet-vaguelette decomposition (WVD), which is analogous to a singular value decomposition. In contrast to the singular value decomposition, a wide variety of prior constraints can be represented as rapid decay of coefficients in the WVD. Donoho [20] finds the WVD for some homogeneous linear transformations, including integration, fractional integration (*e.g.*, the Abel transform), and the Radon transformation; and he shows that the WVD exists for some inhomogeneous linear operators, such as one-dimensional convolution. For such operators, if the noise is Gaussian and the unknown is known to lie in a ball in a Besov space, finding the empirical WVD of the data; shrinking the coefficients nonlinearly, one at a time; then reconstructing from the shrunk coefficients, is asymptotically minimax for mean squared error loss. See [59] for an accessible overview. There do not seem to have been many applications of this approach to real data yet, but see, *e.g.*, [43], for a simulation study of wavelet-vaguelette shrinkage in tomography.

4.7. **Strict Bounds.** Let $\{g_\gamma\}_{\gamma \in \Gamma}$ be an arbitrary collection of real-valued parameters. Consider a collection $\{\mathcal{I}_\gamma\}_{\gamma \in \Gamma}$ of non-randomized confidence intervals for $\{g_\gamma(\theta)\}_{\gamma \in \Gamma}$. The collection $\{\mathcal{I}_\gamma\}_{\gamma \in \Gamma}$ of confidence intervals has *simultaneous confidence level* $1-\alpha$ if

$$\mathbb{P}_\theta \left(\bigcap_{\gamma \in \Gamma} \{\mathcal{I}_\gamma \ni g_\gamma(\theta)\} \right) \geq 1 - \alpha \tag{4.53}$$

whatever be $\theta \in \Theta$.

There are many ways to construct simultaneous confidence intervals for a collection of parameters; we examine one, based on choosing a fixed set $D \subseteq \mathbb{R}^n$ that has probability $1-\alpha$ of containing the noise ϵ, transforming D into a confidence set $\mathcal{D} \subseteq \Theta$ for the model θ, and then finding the ranges of values the parameters $\{g_\gamma\}_{\gamma \in \Gamma}$ can take on $\mathcal{D}$. This yields conservative confidence intervals for the parameters. Unfortunately, the intervals can be much longer than necessary to attain their nominal confidence level, and can lengthen when the number of data increases—see [36]. This seems to stem from choosing D generically, rather than tailoring D

to the forward mapping K, the particular functionals $\{g_\gamma\}_{\gamma\in\Gamma}$ of interest, and the geometry of the set Θ, which can reduce the expected lengths of the confidence intervals.

Consider a linear inverse problem with data

$$X = K\theta + \epsilon. \tag{4.54}$$

Let $\mathbb{P}$ be the probability distribution of ϵ on $\mathbb{R}^n$, and recall that $\mathbb{P}_\theta$ is the probability distribution of X. Suppose D satisfies $\mathbb{P}\{\epsilon \in D\} \geq 1-\alpha$. Then

$$\begin{aligned} 1-\alpha &\leq \mathbb{P}\{K\theta + \epsilon \in D + K\theta\} \\ &= \mathbb{P}_\theta\{X \in D + K\theta\} \\ &= \mathbb{P}_\theta\{X - D \ni K\theta\} \\ &\leq \mathbb{P}_\theta\{K^{-1}(X-D) \ni \theta\}. \end{aligned} \tag{4.55}$$

(The last step produces an inequality because K may not be one-to-one.) Therefore, the random set

$$\begin{aligned} \mathcal{D} &= \mathcal{D}(X) = K^{-1}(X-D) \\ &= \{\theta' \in \Theta : K\theta' = X - d \text{ for some } d \in D\}, \end{aligned} \tag{4.56}$$

the set of all models in Θ whose noise-free data image is in the set $X - D$, is a $1-\alpha$ confidence region for the model θ.

The collection of intervals

$$\mathcal{I}_g(X) = \left[\inf_{\theta'\in\mathcal{D}} g(\theta'), \sup_{\theta'\in\mathcal{D}} g(\theta')\right] \tag{4.57}$$

has simultaneous $1-\alpha$ coverage probability for all functionals g: the intervals all cover $g(\theta)$ whenever $\epsilon \in D$, which occurs with probability at least $1-\alpha$. The optimization problems to find $\mathcal{I}_g$ often can be solved exactly, depending on how the set D is defined. This approach is sometimes called "strict bounds." See [54] for examples in seismology, gravimetry, and helioseismology; [39] for examples in probability density estimation for earthquake aftershocks.

How can we construct a set D with probability $1-\alpha$ of containing ϵ? If the joint distribution of the stochastic errors ϵ is known, constructing D is straightforward: ellipsoids and hyperrectangular sets are common choices. For example, if ϵ is zero-mean multivariate Gaussian with covariance matrix Σ, one might take

$$D = \{x : \ x \cdot \Sigma^{-1} \cdot x \ \le \ \chi^2_{n,1-\alpha}\} \tag{4.58}$$

If the joint distribution of ϵ is not known, but we are given intervals $\mathcal{I}_j$ such that $\mathbb{P}\{\epsilon_j \in \mathcal{I}_j\} \ge 1 - \alpha_j$ individually, we can use the Boole-Bonferroni inequality (Lemma A.1) to conclude that

$$\mathbb{P}\left\{\epsilon \in \prod_j \mathcal{I}_j\right\} \ge 1 - \sum_j \alpha_j. \tag{4.59}$$

Either of these approaches leads to optimization problems for each $\mathcal{I}_g$ that can be solved by quadratic programming when D is defined by (4.58) or linear programming when D is defined by (4.59). Some other choices of D also yield optimization problems that can be solved exactly; when the optimization problems cannot be solved exactly, sometimes it is possible to approximate the optimization problems conservatively using conjugate duality. See [54] for examples and techniques, and an application to an inverse problem in seismology. See [39] for an application to a problem in seismicity.

The strict bounds approach is fairly general. For example, in many cases no metric or topology on Θ is needed, limiting the assumptions one needs to make [54]. It is straightforward to incorporate systematic errors into strict bounds; [54] gives an example in helioseismology that includes systematic uncertainty in the functionals measured, as well as the stochastic uncertainty in the measurements.

4.8. **Confidence Set Inference.** Backus [9] gives a method he calls "confidence set inference" for constructing a conservative confidence set for a linear functional g of a model θ in a separable Hilbert space $\mathcal{T}$ using a prior constraint on a quadratic functional of the model: $\Theta = \{\theta \in \mathcal{T} : Q(\theta, \theta) \le 1\}$. The method decomposes the parameter $g(\theta)$ into a part controlled by the prior constraint, and a part controlled by the data. Consider decomposing g into a component g_M in the span of $\{\kappa_j\}_{j=1}^n$ and a component $g_{M^\perp}$ orthogonal to the span of $\{\kappa_j\}_{j=1}^n$. If $\sup_{\theta' \in \Theta} |g_{M^\perp}(\theta')| < \infty$, $g(\theta)$ can be estimated with bounded bias. When the noise is Gaussian, the problems to which the method can be applied are a subset of those covered by [19], whose approach generally leads to shorter confidence intervals. See [55] for a comparison of the methods on a problem in geomagnetism also treated by [9].

5. Conclusions

The statistical view of inverse problems differs from the applied mathematics view, but the two are related. For example, *identifiability* is related to *uniqueness*, and *consistency* is related to *stability* and the theory of optimal recovery. The statistical viewpoint is more encompassing: forward and inverse problems of applied mathematics and physical science are special cases of statistical forward and inverse problems.

Whether a parameter can be estimated well depends on the prior constraints on the unknown model, the nature of the forward problem, the probability law of the observational error, and the definition of the parameter. In particular, whether the entire model can be estimated consistently depends crucially on details of the distribution of the observational error and the extent to which the observations measure the same features of the model—or almost the same features—over and over with sufficiently independent errors. *Ceteris paribus*, estimating the model is easier when the probability distribution of the error is rougher.

Describing inverse problems in statistical language permits a unified view of standard inversion techniques, and provides reasonable criteria for choosing among them. For example, Backus-Gilbert theory concerns estimating linear functionals of an element of a Hilbert space from linear observations with additive noise. The theory characterizes those linear functionals that can be estimated without bias, and hence are identifiable. This paper extends Backus-Gilbert theory to a wider class of problems and of parameters to give necessary conditions and sufficient conditions for parameters in inverse problems to be estimable without bias from linear observations contaminated by additive noise. (Generally, the set of functions that can be estimated without restrictive prior information is rather meager.) Moreover, the statistical viewpoint shows that in estimating a collection of linear functionals from the same data, such as a set of weighted averages of the model, the vector of Backus-Gilbert estimates sometimes can be improved using *shrinkage*, which introduces bias deliberately.

Regularized methods, such as penalized maximum likelihood, Tichonov regularization, singular value truncation and weighting, shrinkage estimation, and the method of sieves, exploit a "bias-variance tradeoff" to reduce some measures of risk, such as mean squared error or the size of confidence sets. Tuning the tradeoff so that the resulting estimator in fact performs well depends on *a priori* information about the unknown model that is not available in every problem. Similarly, Bayes estimates can be thought of as regularized estimates that rely upon prior information—typically stronger than the constraints required to justify other regularized estimates—to achieve their performance advantage.

Viewing inverse problems as statistical decisions helps clarify how statistical tools can be brought to bear on inverse problems, and suggests approaches for developing new methods that behave well in applications.

Acknowledgments

This paper was written in part while the second author was on appointment as Miller Research Professor in the Miller Institute for Basic Research in Science, and received support from NSF grants DMS-97-09320 and DMS-98-72979, and NASA grants NAG5-3941 and NRA-96-09-OSS-034SOHO. The first author was supported by NSF grants DMS-97-03845 DMS-00-71468. We thank D.A. Freedman for countless helpful discussions, and J. Goldstein, M. Hansen, and L. Tenorio for noting errors in an earlier draft and for helpful comments. The treatment of loss functions for confidence sets evolved in joint work in progress with B. Hansen. A shorter, preliminary version of the paper without the original results appeared in *Surveys on Solution Methods for Inverse Problems* [56].

Appendix A. Sundry Useful Results from Probability

Lemma A.1 (Boole's and Bonferroni's inequalities). *For any countable collection of events* $\{F_i\} \subseteq \mathcal{F}$,

$$\mathbb{P}\left(\bigcup_i F_i\right) \leq \sum_i \mathbb{P}(F_i) \tag{A.1}$$

and hence

$$\mathbb{P}\left(\bigcap_i F_i\right) \geq 1 - \sum_i [1 - \mathbb{P}(F_i)]. \tag{A.2}$$

These classical inequalities are immediate from the countable additivity of probability measures and de Morgan's laws.

Theorem A.2 (Kakutani's Dichotomy). *Let* μ *and* ν *be two probability measures on the infinite sequence space* $\mathbb{R}^{\mathbb{N}}$ *such that* $\mu = \mu_1 \otimes \mu_2 \otimes \cdots$ *and* $\nu = \nu_1 \otimes \nu_2 \otimes \cdots$ *for two sequences* $\{\mu_i\}_{i=1}^{\infty}$

and $\{\nu_i\}_{i=1}^{\infty}$ *of probability measures on* $\mathbb{R}$*. That is, if* $X = \{X_i\}_{i=1}^{\infty}$ *is a random sequence with probability distribution* μ *(resp.* ν*), then the* X_i *are independent random variables and* X_j *has probability distribution* μ_j *(resp.* ν_j*). Suppose for each* i *that* μ_i *is absolutely continuous with respect to* ν_i *with Radon–Nikodym derivative* $d\mu_i/d\nu_i$*. Then either* μ *is absolutely continuous with respect to* ν *or* μ *and* ν *are mutually singular depending on whether*

$$\prod_{i=1}^{\infty} \int \sqrt{\frac{d\mu_i}{d\nu_i}}\, d\nu_i \tag{A.3}$$

is positive or zero.

See *e.g.* Theorem 4.3.5 of [30].

Theorem A.3 (Polya's Criterion). *Let* $h : \mathbb{R} \to \mathbb{R}$ *be a function with the properties:*

- h *is non-negative and bounded,*
- h *is even (that is,* $h(z) = h(-z)$*),*
- *the restriction of* h *to the positive half-line is convex and decreasing.*

Then there is a finite measure ν *such that*

$$h(z) = \int_{-\infty}^{\infty} \exp(izx)\, \nu(dx), \quad z \in \mathbb{R}. \tag{A.4}$$

See *e.g.* Theorem 2.3.10 of [30].

Lemma A.4 (Identifiability of Shifts). *Let* Y *be a random* n*-vector. Given a constant* $a \in \mathbb{R}^n$*,* $a \neq 0$*, the probability distribution of* Y *differs from that of* $a + Y$*.*

Proof. This is most easily seen using Fourier methods. Note that $\mathbb{E}[\exp(iz \cdot (a+Y))] = \exp(iz \cdot a)\mathbb{E}[\exp(iz \cdot Y)]$ for $z \in \mathbb{R}^n$. The function $z \mapsto \mathbb{E}[\exp(iz \cdot Y)]$ is continuous and takes the value 1 at 0, and hence is non-zero in a neighborhood of 0. Hence $\mathbb{E}[\exp(iz \cdot (a+Y))] \neq \mathbb{E}[\exp(iz \cdot Y)]$ for some $z \in \mathbb{R}^n$, and the result follows from Fourier uniqueness. □

Remark A.5. The result covers the case in which $\mathbb{E}[Y]$ is not defined. When $\mathbb{E}[Y]$ is defined, the result is trivial.

Appendix B. Bits of Measure-Theoretic Probability for Statistics

This section presents a few background ideas and some notation from measure-theoretic probability. The intended audience is applied mathematicians who need a terse refresher of measure-theoretic probability to read the rest of the paper. For more, see [28, 30].

For a given set Ω, a *σ-algebra* is a collection $\mathcal{F}$ of subsets of Ω such that

(1) $\Omega \in \mathcal{F}$
(2) If $F_1, F_2, \dots \in \mathcal{F}$, then $\bigcup_j F_j \in \mathcal{F}$, and
(3) If $F \in \mathcal{F}$ then $F^c \in \mathcal{F}$.

There is a smallest σ-algebra containing any collection of sets, the σ-algebra generated by the collection. The σ-algebra generated by the open sets of a topology is called the *Borel σ-algebra* corresponding to the topology. A *measurable space* is an ordered pair $(\Omega, \mathcal{F})$ where Ω is a set and $\mathcal{F}$ is a σ-algebra $\mathcal{F}$ of subsets of Ω. The elements of $\mathcal{F}$ are called *measurable sets*.

A *measure* is non-negative extended-real-valued function $m : \mathcal{F} \to \mathbb{R}^+ \cup \infty$ such that for every countable collection $\{F_j\}$ of elements of $\mathcal{F}$ such that $F_j \cap F_k = \emptyset$ whenever $j \neq k$,

$$m\left(\bigcup_{j=1}^{\infty} F_j\right) = \sum_{j=1}^{\infty} m(F_j). \tag{B.1}$$

A measure m on a σ-algebra $\mathcal{F}$ of subsets of Ω is *finite* if $m(\Omega) < \infty$. A measure m on a σ-algebra $\mathcal{F}$ of subsets of Ω is σ-finite if there exists a countable collection of sets $\{F_j\} \subset \mathcal{F}$ such that $\Omega = \bigcup_j F_j$ and $m(F_j) < \infty$ for all j. A probability distribution $\mathbb{P}$ on a σ-algebra $\mathcal{F}$ of subsets of a set Ω is a finite measure with total mass one (*i.e.*, $\mathbb{P}(\Omega) = 1$). We refer to $(\Omega, \mathcal{F}, \mathbb{P})$ as a *probability triple*. The elements of $\mathcal{F}$ are called *events*. If a statement is true on Ω except on some set F for which $\mathbb{P}(F) = 0$, the statement is said to hold *almost surely* (a.s., or a.s.($\mathbb{P}$)).

For any subset F of Ω, the *indicator function of the set F* is

$$\begin{aligned} 1_F : \Omega &\to \{0,1\} \\ \omega &\mapsto \begin{cases} 1, & \omega \in F \\ 0, & \omega \notin F. \end{cases} \end{aligned} \tag{B.2}$$

Note that $1_{FG} = 1_F 1_G$.

The function 1_F is a special case of a random variable. A *random variable* X is a mapping from the set Ω of a measurable space $(\Omega, \mathcal{F})$ into a separable Banach space $\mathcal{X}$ such that the inverse-image under X of Borel sets of $\mathcal{X}$ are in $\mathcal{F}$.

The *σ-algebra generated by the random variable X, $\sigma(X)$,* is the smallest σ-algebra $\mathcal{G}$ such that X is a random variable on $(\Omega, \mathcal{G})$.

A random variable X is *integrable* if

$$\int_\Omega \|X(\omega)\| d\mathbb{P}(\omega) < \infty. \tag{B.3}$$

If X is integrable, one can define the *expected value of X*

$$\mathbb{E}X = \int_\Omega X(\omega) d\mathbb{P}(\omega) \tag{B.4}$$

as a limit of integrals of suitably defined "simple functions" that converge to X, where convergence is in a metric defined on functions from a measure space to a Banach space; see Dunford and Schwartz [29] volume I, Chapter III for details. The metric Dunford and Schwartz use is

$$d(X, Y) = \inf_{\alpha > 0} \arctan\left(\alpha + \mathbb{P}\{\omega : \|X(\omega) - Y(\omega)\| > \alpha\}\right). \tag{B.5}$$

For any event $F \in \mathcal{F}$,

$$\mathbb{P}(F) \equiv \mathbb{E}1_F \equiv \int_F d\mathbb{P} \equiv \int_\Omega 1_F d\mathbb{P}. \tag{B.6}$$

If there is more than one probability measure under consideration, for example a family of measures $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, we write

$$\mathbb{E}_{\mathbb{P}_\theta} X \equiv \mathbb{E}_\theta X = \int_\Omega X(\omega) d\mathbb{P}_\theta. \tag{B.7}$$

If X is real-valued and integrable, the variance of X, $\mathbf{Var}(X)$ is

$$\mathbf{Var}(X) \equiv \mathbb{E}[X - \mathbb{E}X]^2. \tag{B.8}$$

In a slight abuse of notation, for a random variable X taking values in a Hilbert space, we define

$$\mathbf{Var}(X) \equiv \mathbb{E}\|X - \mathbb{E}X\|^2. \tag{B.9}$$

Events F, $G \subseteq \mathcal{F}$ are *independent* if $\mathbb{P}(FG) = \mathbb{P}(F)\mathbb{P}(G)$. Two σ-algebras $\mathcal{F}$ and $\mathcal{G}$ are *independent* if every $F \in \mathcal{F}$ is independent of every $G \in \mathcal{G}$. If $\sigma(X)$ is independent of $\mathcal{F}$, we say

X is independent of $\mathcal{F}$. We say that the random variables X and Y are independent if $\sigma(X)$ is independent of $\sigma(Y)$.

The *conditional probability of the event F given the event G* is $\mathbb{P}(F|G) \equiv \mathbb{P}(FG)/\mathbb{P}(G)$ for $\mathbb{P}(G) \neq 0$. *Bayes' Rule* is

$$\mathbb{P}(F|G) = \frac{\mathbb{P}(G|F)\mathbb{P}(F)}{\mathbb{P}(G)} \tag{B.10}$$

for $\mathbb{P}(F)$, $\mathbb{P}(G) > 0$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple, let X be a measurable, integrable random variable, and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. The *conditional expectation of X given $\mathcal{G}$*, $\mathbb{E}[X|\mathcal{G}]$, is any $\mathcal{G}$-measurable function Y such that for every $G \in \mathcal{G}$, $\mathbb{E}[1_G Y] = \mathbb{E}[1_G X]$. If Y is a $\mathbb{P}$-measurable random variable, $\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)]$. If X is $\mathcal{G}$-measurable, it follows that $\mathbb{E}[X|\mathcal{G}] = X$. If X is independent of $\mathcal{G}$, it follows that $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$. If X is $\mathcal{G}$-measurable, $\mathbb{E}[XY|\mathcal{G}] = X\mathbb{E}[Y|\mathcal{G}]$. The *conditional probability of $B \in \mathcal{F}$ given the sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$* is $\mathbb{P}[B|\mathcal{G}] = \mathbb{E}[1_B|\mathcal{G}]$.

A measurable random variable X taking values in a separable Banach space $\mathcal{X}$ induces a probability distribution on the Borel σ-algebra of $\mathcal{X}$ through $\mathbb{P}(B) = \mathbb{P}\{X \in B\}$ for all Borel sets $B \subseteq \mathcal{X}$. If m is a Borel measure on $\mathcal{X}$ such that $\mathbb{P}\{X \in B\} = m(B)$ for all Borel sets B, we write $X \sim m$. If two random variables have the same probability distribution (if they take values in the same space $\mathcal{X}$, and $\mathbb{P}\{X \in B\} = \mathbb{P}\{Y \in B\}$ for all Borel sets B), we write $X \sim Y$ and say *X and Y are identically distributed.* If in addition X and Y are defined on the same probability triple and are independent, then they are *independent and identically distributed*, abbreviated i.i.d.

A measure μ is *dominated by the measure ν*, or *absolutely continuous with respect to ν*, if μ and ν are defined on the same σ-algebra $\mathcal{F}$, and if $\nu(F) = 0$ implies $\mu(F) = 0$, $F \in \mathcal{F}$. If μ and ν are σ-finite and μ is dominated by ν, then the Radon-Nikodym Theorem ([52], pp. 129ff) states that μ has a unique ν-integrable *density* $p : \Omega \to \mathbb{R}^+$ with respect to ν, such that for all $F \in \mathcal{F}$,

$$\mu(F) = \int_F p(\omega) d\nu(\omega). \tag{B.11}$$

If μ is absolutely continuous with respect to ν and ν is absolutely continuous with respect to μ, μ and ν are said to be *equivalent measures*. A set $F \in \mathcal{F}$ for which $\mu(F) = 0$ is called a *null set* of μ. The measures μ and ν are *mutually singular*, written $\mu \perp \nu$, if they are defined on the same σ-algebra $\mathcal{F}$ and there exists a set $F \in \mathcal{F}$ such that $\mu(F) = \mu(\Omega)$ and $\nu(F) = 0$; that is, if μ is concentrated on a null set of ν, and *vice versa*. The Lebesgue decomposition theorem says that given two σ-finite measures defined on the same σ-algebra, one measure can be decomposed into a part that is mutually singular with respect to the other, and a part that is absolutely continuous with respect to the other: given μ, ν, defined on $\mathcal{F}$,

$$\mu = \mu_{||} + \mu_{\perp} \tag{B.12}$$

where $\mu_{||}$ is absolutely continuous with respect to ν and $\mu_{\perp} \perp \nu$. The density of $\mu_{||}$ with respect to ν is called the *Radon-Nikodym derivative of μ with respect to ν.*

If S is a bounded Borel subset of $\mathbb{R}^n$, then $U(S)$ is the uniform probability distribution on S, defined by

$$U(B) = \mu(BS)/\mu(S) \tag{B.13}$$

for all Borel sets B, where μ is Lebesgue measure. The special case $n = 1$, $S = [0, 1]$, is written $U[0, 1]$, the uniform distribution on the closed interval $[0, 1]$, which is defined by

$$\mathbb{P}(B) = \int_B d\mu, \tag{B.14}$$

where μ is Lebesgue measure, for all Borel subsets B of $[0, 1]$.

The *covariance matrix* of an $\mathbb{R}^n$-valued random variable X is

$$\mathbf{Cov}(X) = \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^T)\right], \tag{B.15}$$

provided $\mathbb{E}\|X\|^2$ is finite. The d-variate normal distribution with mean $\theta \in \mathbb{R}^d$ and $d \times d$ symmetric covariance matrix Σ is denoted $N(\theta, \Sigma)$; if Σ is positive-definite, the density of $N(\theta, \Sigma)$ with respect to Lebesgue measure on $\mathbb{R}^d$ is

$$\phi(x) = |\Sigma^{1/2}|^{-1}(2\pi)^{-d/2} \exp\left(-\frac{1}{2}(x - \theta) \cdot \Sigma \cdot (x - \theta)\right), \tag{B.16}$$

where $|\Sigma|$ is the determinant of Σ.

If X is a real-valued random variable, the *cumulative distribution function of X* is

$$\begin{aligned} F_X : \mathbb{R} &\to [0, 1] \\ x &\to \mathbb{P}\{X \leq x\}. \end{aligned} \tag{B.17}$$

The α quantile of a real-valued random variable X is

$$X_\alpha = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\}. \tag{B.18}$$

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

[2] G. Backus. Inference from inadequate and inaccurate data, I. *Proc. Natl. Acad. Sci.*, 65:1–7, 1970.

[3] G. Backus. Inference from inadequate and inaccurate data, II. *Proc. Natl. Acad. Sci.*, 65:281–287, 1970.

[4] G. Backus. Inference from inadequate and inaccurate data, III. *Proc. Natl. Acad. Sci.*, 67:282–289, 1970.

[5] G. Backus and F. Gilbert. The resolving power of gross Earth data. *Geophys. J. R. astron. Soc.*, 16:169–205, 1968.

[6] G.E. Backus. Isotropic probability measures in infinite-dimensional spaces. *Proc. Natl. Acad. Sci.*, 84:8755–8757, 1987.

[7] G.E. Backus. Bayesian inference in geomagnetism. *Geophys. J.*, 92:125–142, 1988.

[8] G.E. Backus. Comparing hard and soft prior bounds in geophysical inverse problems. *Geophys. J.*, 94:249–261, 1988.

[9] G.E. Backus. Confidence set inference with a prior quadratic bound. *Geophys. J.*, 97:119–150, 1989.

[10] G.E. Backus. Trimming and procrastination as inversion techniques. *Phys. Earth Planet. Inter.*, 98:101–142, 1996.

[11] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 2nd edition, 1985.

[12] M. Bertero. Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, 75:1–120, 1989.

[13] M. Bertero, C. De Mol, and E.R. Pike. Linear inverse problems with discrete data: I. General formulation and singular system analysis. *Inverse Problems*, 1:301–330, 1985.

[14] M. Bertero, C. De Mol, and E.R. Pike. Linear inverse problems with discrete data: II. Stability and regularisation. *Inverse Problems*, 4:573–594, 1988.

[15] S.C. Constable, R.L. Parker, and C.G. Constable. Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data. *Geophysics*, 52:289–300, 1987.

[16] P.W. Diaconis and D. Freedman. Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli*, 4:411–444, 1998.

[17] D.L. Donoho. One-sided inference about functionals of a density. *Ann. Stat.*, 16:1390–1420, 1988.

[18] D.L. Donoho. Exact asymptotic minimax risk for sup norm loss via optimal recovery. *Probab. Theory and Rel. Fields*, 99:145–170, 1994.

[19] D.L. Donoho. Statistical estimation and optimal recovery. *Ann. Stat.*, 22:238–270, 1994.

[20] D.L. Donoho. Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harm. Anal.*, 2:101–126, 1995.

[21] D.L. Donoho, I. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B.*, 54:41–81, 1992.

[22] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[23] D.L. Donoho and I.M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Stat.*, 26:879–921, 1998.

[24] D.L. Donoho and I.M. Johnstone. Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica*, 9:1, 1999.

[25] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Stat. Soc., Ser. B*, 57:301–369, 1995.

[26] D.L. Donoho and M. Nussbaum. Minimax quadratic estimation of a quadratic functional. *J. Complexity*, 6:290–323, 1990.

[27] H. Drygas. Weak and strong consistency of the least squares estimators in regression models. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 34(2):119–127, 1976.

[28] R.M. Dudley. *Real Analysis and Probability.* Wadsworth & Brooks/Cole, Pacific Grove, CA, 1989.

[29] N. Dunford and J.T. Schwartz. *Linear Operators.* John Wiley and Sons, New York, 1988.

[30] R. Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[31] S.N. Evans and P.B. Stark. Shrinkage estimators, Skorokhod's problem, and stochastic integration by parts. *Ann. Stat.*, 24:809–815, 1996.

[32] J. Franklin. Well-posed stochastic extensions of ill-posed linear problems. *J. Math. Analysis Applic.*, 31:682–716, 1970.

[33] D.A. Freedman. Invariants under mixing which generalize de Finetti's Theorem: continuous time parameter. *Ann. Math. Stat.*, 34:1194–1216, 1963.

[34] M. Fukushima, Y. Ōshima, and M. Takeda. *Dirichlet forms and symmetric Markov processes.* Walter de Gruyter & Co., Berlin, 1994.

[35] A. Gelman, J. Carlin, H. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, London, 1995.

[36] C.R. Genovese and P.B. Stark. Data reduction and statistical consistency of ℓ_p misfit norms in linear inverse problems. *Phys. Earth Planet. Inter.*, 98:143–162, 1996.

[37] I.J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* MIT Press, Cambridge, MA, 1965.

[38] J.A. Hartigan. *Bayes Theory.* Springer-Verlag, New York, 1983.

[39] N.W. Hengartner and P.B. Stark. Finite-sample confidence envelopes for shape-restricted densities. *Ann. Stat.*, 23:525–550, 1995.

[40] W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symposium on Mathematical Statiatics and Probability*, volume 1, pages 361–380, Berkeley, 1961. Univ. California Press.

[41] E.T. Jaynes. *Papers on Probability, Statistics, and Statistical Physics.* Synthese Library, 1983.

[42] I.M. Johnstone and B.W. Silverman. Wavelet threshold estimators for data with correlated noise. *J. Roy. Stat. Soc., Ser. B*, 59:319–351, 1997.

[43] E.D. Kolaczyk. A wavelet shrinkage approach to tomographic image reconstruction. *J. Am. Stat. Assoc.*, 91:1079–1090, 1996.

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

[45] L. Le Cam. Maximum likelihood: an introduction. *Intl. Stat. Rev.*, 58:153–171, 1990.

[46] E.L. Lehmann. *Testing Statistical Hypotheses.* John Wiley and Sons, New York, 2nd edition, 1986.

[47] E.L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer-Verlag, New York, 2nd edition, 1998.

[48] D.G. Luenberger. *Optimization by Vector Space Methods.* John Wiley and Sons, Inc., New York, 1969.

[49] H. Massam. Consistent directions for least-squares estimates. *Canad. J. Statist.*, 15(1):87–90, 1987.

[50] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1:502–518, 1986.

[51] R.L. Parker. *Geophysical Inverse Theory.* Princeton University Press, Princeton, NJ, 1994.

[52] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, New York, 1974.

[53] X. Shen. On methods of sieves and penalization. *Ann. Stat.*, 25:2555–2591, 1997.

[54] P.B. Stark. Inference in infinite-dimensional inverse problems: Discretization and duality. *J. Geophys. Res.*, 97:14,055–14,082, 1992.

[55] P.B. Stark. Minimax confidence intervals in geomagnetism. *Geophys. J. Intl.*, 108:329–338, 1992.

[56] P.B. Stark. Inverse problems as statistics. In D. Colton, H.W. Engl, A.K. Louis, J.R. Mclaughlin, and W. Rundell, editors, *Surveys on Solution Methods for Inverse Problems*, pages 253–275. Springer-Verlag, New York, 2000.

[57] C. Stein. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, Berkeley, 1956. Univ. California Press.

[58] A. Tarantola. *Inverse Problem Theory: methods for data fitting and model parameter estimation.* Elsevier Science Publishing Co., 1987.

[59] L. Tenorio. Statistical regularization of inverse problems. *SIAM Review*, 43(2):347–366, 2001.

[60] L. Tenorio, P.B. Stark, and C.H. Lineweaver. Bigger uncertainties and the Big Bang. *Inverse Problems*, 15:329–341, 1999.

[61] S. van de Geer and M. Wegkamp. Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.*, 24(6):2513–2523, 1996.

[62] A.C.M. van Rooij and F.H. Ruymgaart. *On Inverse Estimation*, pages 579–613. 1999.

[63] G. Wahba. *Spline Models for Observational Data.* Soc. for Industrial and Appl. Math., Philadelphia, PA, 1990.

[64] A. Wald. *Statistical Decision Functions.* John Wiley and Sons, New York, 1950.

[65] Chien-fu Wu. Characterizing the consistent directors of least squares estimates. *Ann. Statist.*, 8(4):789–801, 1980.

E-mail address: evans@stat.berkeley.edu

Department of Statistics and Department of Mathematics, University of California, Berkeley, CA 94720-3860 USA

E-mail address: stark@stat.berkeley.edu

Department of Statistics, Space Sciences Laboratory, and Center for Theoretical Astrophysics, University of California, Berkeley, CA 94720-3860 USA