Statistical Decision Rules in Econometrics

Keisuke Hirano
Department of Economics
Pennsylvania State University
hirano@psu.edu

Jack R. Porter
Department of Economics
University of Wisconsin
jrporter@ssc.wisc.edu

December 20, 2018

Abstract

Statistical decision rules map data into actions. Point estimators, inference procedures, and forecasting methods can be viewed as statistical decision rules. However, other types of rules are possible, such as rules for assigning individuals to treatments based on covariates, and methods for designing auctions. We discuss heuristics for constructing statistical decision rules, and survey results that characterize the properties of various classes of decision rules. Particular attention is paid to developing large-sample approximations to the distributions and associated risk properties of statistical decision rules.

Keywords: statistical decision theory; treatment assignment rules; limit experiments; risk.

1 Introduction

A statistic can be defined simply as a quantity computed from a sample of data. Thus any statistical procedure is a mapping from the sample space of the statistical "experiment" into some space of "actions." Standard procedures such as parameter estimators, hypothesis tests, and confidence intervals, can be viewed in this way, but so can other types of procedures, such as forecasting rules, allocation decisions based on past data, and methods for designing experiments. If we can also specify a criterion with which to evaluate such decision rules, then we can draw upon a rich set of results from statistical decision theory to analyze, compare, and choose among feasible rules.

Decision theoretic concepts and methods permeate statistics and econometrics with varying degrees of formality, and we can only touch on some aspects of their use in this chapter. Our goal is to survey some recent work in econometrics that adopts this framework to study different types of empirical problems. We review some classical problems, such as point estimation and confidence interval construction, but our goal is to highlight a richer set of decision problems that arise naturally in economic applications and can be handled within the statistical decision theory framework.

We primarily focus on analyzing rules from a frequentist perspective, calculating their *ex ante* expected performance, where the expectation is with respect to the sampling distribution of the data. In some cases, it is possible to characterize the finite sample properties of decision rules and obtain exact optimality results. However, in many economic applications, the richness of the hypothesized set of probability distributions, the complexity of the action space, or the nature of the evaluation criterion, make exact analysis infeasible. We consider how large sample approximations can be used to obtain approximate measures of performance and, in some cases, approximate optimality results for statistical decision rules. Exact and large sample analyses are often complementary. Finite sample results for carefully chosen special cases of the general decision problem can both provide intuition and often form the technical basis for large sample theory.

The next section reviews the general statistical decision theory framework, which can be used to study many applications of statistical treatment rules in economics. We set up the basic components of the framework, including data, action, and evaluation criterion, and review some standard notions of optimality. We also introduce some of the large sample theory that will be used in the remainder of the chapter. The remaining sections then consider various applications of the general approach. Section 3 considers "point" decision problems, where the action space is a subset of a Euclidean space. The leading case is point estimation in parametric models, but we also want to extend classical results on point estimation to handle other decision problems, for example the problem of choosing a reserve price for a first-price auction. Section 4 considers treatment assignment, where the action is a treatment protocol that specifies how to allocate a treatment among individuals in some population. This problem has gained renewed interest in economics and other fields in recent years. Section 5 considers other applications, including problems involving nonregular functionals, partially identified models, and ex-

perimental design, where the decision-theoretic perspective can provide useful insights.

2 General Setup and Evaluation of Decision Rules

2.1 Setup

Our setup broadly follows the frequentist statistical decision theory approach of Wald (1950). Many texts, such as Ferguson (1967) and Berger (1993), cover statistical decision theory in detail.[1] We will consider a *statistical model* to be a collection of probability measures $\mathscr{M}$ on some measurable space $\mathscr{Z}$. We will usually parametrize the probability measures $P \in \mathscr{M}$ as P_θ, where the parameter θ lies in the parameter space Θ:

$$\mathscr{M} = \{P_\theta : \theta \in \Theta\}$$

In nonparametric and semiparametric settings, the set of possible probability measures $\mathscr{M}$ (and hence Θ) is infinite-dimensional. In this case it is sometimes more convenient to drop the parametric notation and simply identify θ with P.

The decision maker observes a random variable $Z \sim P_\theta$, with support in $\mathscr{Z}$, and chooses an action based on the observation. The action space $\mathscr{A}$ is a measurable space. A (nonrandomized) statistical decision rule is a measurable mapping $\delta : \mathscr{Z} \to \mathscr{A}$. A randomized decision rule is a Markov kernel with source $\mathscr{Z}$ and target $\mathscr{A}$. In practice, we can view a randomized decision rule as a function $\delta(Z, U)$, where U is a random variable independent of Z.[2]

We evaluate decision rules based on a welfare function $W : \Theta \times \mathscr{A} \to \mathbb{R}$, where we interpret higher values of welfare as more desirable. The expected welfare of a (nonrandomized) rule δ under distribution P_θ is

$$E_\theta\left[W(\theta, \delta(Z))\right] = \int W(\theta, \delta(z)) dP_\theta(z).$$

(For a randomized rule $\delta(Z, U)$, we further integrate W over the appropriate distribution of U.) Note that this quantity depends on the value of θ, which is not known. Therefore, to evaluate decision rules we will need to aggregate the expected welfare over possible values of θ in some way as we will discuss further below. In some cases it is convenient to work instead with a loss function $L(\theta, a)$, where smaller values of loss are preferred. The *risk* of a decision rule is its expected loss:

$$R(\theta, \delta(Z)) = \int L(\theta, \delta(z)) dP_\theta(z).$$

Example 1 *Suppose that after observing $Z \sim P_\theta$, it is desired to produce a point estimate of θ. Then the*

[1] See also Strasser (1985), Le Cam (1986), Le Cam and Yang (2000), and Liese and Miescke (2008).

[2] The additional randomness induced by U is often undesirable from the standpoint of minimizing risk, but we need the additional generality to characterize some procedures used in practice. For example, Section 3.5 discusses some shrinkage estimators that are asymptotically equivalent to randomized estimators.

action space can be taken as $\mathscr{A} = \Theta$, and we can denote a point estimator as a function $\hat{\theta} : \mathscr{Z} \to \Theta$. Suppose θ is scalar and we evaluate the performance of the point estimator by squared error loss $L(\theta, a) = (\theta - a)^2$. Then its frequentist risk is

$$R(\theta, \hat{\theta}) = \int L(\theta, \hat{\theta}(z)) dP_\theta(z) = \int \left(\theta - \hat{\theta}(z)\right)^2 dP_\theta(z).$$

This fits into our framework by setting $W(\theta, a) = -L(\theta, a)$. We will discuss point estimation and other point decision problems, where the action space is a subset of a Euclidean space, in more detail in Section 3.

In practice only some elements of θ, or some function of θ, may be relevant for welfare. In this case, it may be convenient to modify the notation. Let $\kappa : \Theta \to \mathscr{K} \subset \mathbb{R}^d$ be the decision-relevant quantity, and suppose the welfare function given κ and an action a is given by $\widetilde{W}(\kappa, a)$. The problem can be rewritten as above by defining a welfare function W that depends on θ only through $\kappa(\theta)$: $W(\theta, a) = \widetilde{W}(\kappa(\theta), a)$. The expected welfare of rule δ under θ is

$$E_\theta \left[W(\theta, \delta(Z))\right] = \int \widetilde{W}(\kappa(\theta), \delta(z)) dP_\theta(z).$$

In general, the expected welfare may continue to depend on the entire vector θ, through the distribution of Z.

Later, to develop large sample results we will work with a sequence of statistical models:

$$\mathscr{M}_n = \{P_\theta^{(n)} : \theta \in \Theta\},$$

with $Z^{(n)} \sim P_\theta^{(n)}$ supported on $\mathscr{Z}^{(n)}$. Here n usually represents the sample size of the data. A leading case is when we observe an i.i.d. sample $Z^{(n)} = Z^n := (Z_1, \ldots, Z_n)$ where $Z_i \sim P_\theta$. Then $P_\theta^{(n)} = P_\theta^n$, the n-fold product measure of P_θ. We can then modify the notation for decision rules and welfare accordingly.

2.2 Counterfactuals in the Wald Framework

To work within the Wald framework, we need to specify the parameter space Θ and associated observation model $Z \sim P_\theta$, the action space $\mathscr{A}$, and the welfare function $W(\theta, a)$ or loss function $L(\theta, a)$. The framework is quite general and can handle decision problems involving counterfactual outcomes, and more generally structural models with latent variables.

Suppose there are possible treatments or policies $t \in \mathscr{T}$ that are interpreted as interventions that could affect an outcome of interest $Y \in \mathscr{Y}$. There is some primitive utility function $u : \mathscr{Y} \to \mathbb{R}$ that evaluates final outcomes. If Y has a distribution F, then expected utility is $\int u(y) dF(y)$. More generally, there could be some welfare functional $W(F)$ that evaluates different possible distributions of Y.

Since Y is thought to be affected by the action, it is convenient to posit the existence of latent variables

$Y(t)$ for $t \in \mathcal{T}$. These *potential outcomes* reflect counterfactual scenarios under different interventions. The different potential outcomes have marginal distributions $F_{Y(t)}$ in the target population of the intervention.

We need to specify the form of the observed data $Z \sim P \in \mathscr{P}$, where P is related in some way to the distributions $F_{Y(t)}$. Structural models in econometrics typically involve some latent variables which are invariant to interventions. In causal (potential outcome) models the latent variables are the potential outcomes themselves, but other types of structural models may specify more primitive objects from which potential outcomes may be derived. Suppose that the structural model specifies latent variables

$$S \sim G_\theta, \quad \theta \in \Theta,$$

and there is some mapping from θ to the relevant potential outcome distributions $\{F_{Y(t)} : t \in \mathcal{T}\}$. The latent variables are linked to the observed data Z by some observation function $Z = \text{Obs}_\theta(S)$, so

$$Z \sim P_\theta := G_\theta \circ \text{Obs}_\theta^{-1}.$$

We have allowed the observation function Obs to depend on θ, to accomodate cases like Example 3 below. In principle we could also allow the observation process to be noisy, by extending the setup to allow for an additional source of randomness.

In the classical econometric terminology, the distribution P_θ is the *reduced form* distribution of the data under the structural parameter θ. In general, different values of θ may lead to the same observable distribution P_θ, in which case θ is not point identified, see Section 5.2. The set of possible reduced form distributions $\{P_\theta : \theta \in \Theta\}$ can always be reparametrized as $\{P_\gamma : \gamma \in \Gamma\}$ such that the reduced form parameter γ is point identified. (In the sequel, we will use θ as the general symbol for the parameter of the model, which could be a structural or reduced form parameter depending on the context.)

Finally, we need to specify the action space $\mathscr{A}$ and the welfare function $W(\theta, a)$. In the simplest case, the action space corresponds to the set of treatments: $\mathscr{A} = \mathcal{T}$, and the welfare function (in the expected utility case) is

$$W(\theta, a) = \int u(y) dF_{Y(a)}(y).$$

The action space could be more complicated, for example allowing for randomization over treatments (in which case $\mathscr{A}$ is a set of distributions over $\mathcal{T}$), or conditional or dynamic treatments (in which case $\mathscr{A}$ is an appropriate set of functions from some set into $\mathcal{T}$). The welfare function would then need to be defined appropriately relative to $\mathscr{A}$.

Example 2 *(Simple Randomized Experiment) Suppose that there are two possible treatments, one of which we will assign to the entire target population. Then we can take* $\mathscr{A} = \mathcal{T} = \{0, 1\}$. *There is an outcome of interest* Y, *and we define potential outcomes* $Y(0)$ *and* $Y(1)$. *We have a random sample of size n from the target population and randomly assign the treatment to individuals. The latent variables are* $S = (S_1, \ldots, S_n)$,

with

$$S_i = (T_i, Y_i(0), Y_i(1)).$$

Random assignment implies that T_i is independent of $(Y_i(0), Y_i(1))$. Then $S \sim G_\theta$, where θ indexes the joint distribution of all the latent variables. We observe the treatment T_i and the potential outcome corresponding to the treatment assigned to i, $Y_i = Y_i(T_i)$. This defines the observation function Obs, which takes

$$S_i = (T_i, Y_i(0), Y_i(1)) \mapsto (T_i, Y_i(T_i)) = Z_i.$$

Then P_θ is the joint distribution of $Z = (Z_1, \ldots, Z_n)$.

For the treatment assignment problem, the relevant components of θ are the marginal distributions $F_{Y(0)}$ and $F_{Y(1)}$. By randomization of treatment, we have

$$Y_i | T_i = 1 \sim F_{Y(1)}, \quad Y_i | T_i = 0 \sim F_{Y(0)}.$$

In this case, the decision relevant components of θ are point-identified from Z.

The simple binary decision problem in Example 2 can be extended to the case where individuals have observable characteristics $X \in \mathscr{X}$, and we allow the assignment to depend on X. A treatment assignment rule is a mapping $a : \mathscr{X} \to \{0,1\}$, interpreted as the decision whether or not to assign some treatment or policy to individuals with characteristics $x \in \mathscr{X}$. Let $\mathscr{A}$ be a set of possible mappings, which could be the set of all possible mappings $2^{\mathscr{X}}$ or some subset of $2^{\mathscr{X}}$. Hence the action space is a subset of a functional space. Then the statistical treatment rule $\delta : \mathscr{Z} \to \mathscr{A}$ selects a conditional treatment assignment rule based on data Z. We will discuss this problem in detail in Section 4.

The following example illustrates how the framework can handle decision problems involving a structural econometric model of individual behavior.

Example 3 *(Empirical Auction Design) Suppose there are m bidders for a single object. The latent variables of the structural model are the bidders' private valuations for the object $S = (v_1, \ldots, v_m)$, drawn jointly from a distribution G_θ:*

$$S = (v_1, \ldots, v_m) \sim G_\theta.$$

For example, these valuations could be drawn i.i.d. from a common distribution, corresponding to the independent private values case, or they could arise from an affiliated private values model as in Li, Perrigne, and Vuong (2002).

Given the rules of the auction and an equilibrium notion, bidders will choose bids $b_1, \ldots, b_m$ based on their individual valuations. Depending on the application, the observed data could consist of all the bids, or just the winning bid, or some other function of the bid vector. In any case, there will be an observation equation

$$Z = Obs_\theta(S),$$

where $Obs_\theta(\cdot)$ captures the mapping from the valuation draws to the observed data. Since bids could depend on the value of θ, which is unknown to the econometrician but typically assumed to be common knowledge among the bidders, the observation function may depend on θ.

Suppose we wish to use the past data to redesign the auction for future instances. For example, we could impose a reserve price, or choose among different auction formats. Let t represent an auction design, and let $\mathcal{T}$ be the set of possible auction designs under consideration. Let $Y(t)$ be the revenue generated by an auction of type t, when bidders' valuations are drawn from G_θ. If we measure welfare by the expected revenue of the auction, then the welfare function would be

$$W(\theta, t) = \int y dF_{Y(t)}(y).$$

In Example 5 below, we further specialize this example to explore large-sample approximations for empirical auction analysis and design.

2.3 Classes of Statistical Decision Rules

Having set up the basic framework for a statistical decision problem, the decision-maker's problem is to choose a statistical decision rule from the set of possible rules. While this choice will be dictated by the specific problem at hand, there are some general classes of rules that are useful in many problems. We discuss a few of them next.

ML Plug-in Rules: Suppose that $\Theta \subset \mathbb{R}^k$ and the probability measures P_θ admit densities p_θ with respect to some measure. Let $\hat{\theta}_{ML} = \hat{\theta}_{ML}(Z)$ be the maximum likelihood estimator of θ:

$$\hat{\theta}_{ML}(Z) = \arg\max_{\theta \in \Theta} p_\theta(Z).$$

A maximum likelihood (ML) plug-in rule chooses an action that maximizes welfare taking θ at the estimated value:

$$\delta_{ML}(z) = \arg\max_{a \in \mathcal{A}} W(\hat{\theta}(z), a).$$

More generally, for any estimator $\hat{\theta}$ we could define a plug-in rule analogously.

Bayes Rules: Let π be a (prior) probability measure over Θ. (We can also allow π to be a general, not necessarily unitary, measure, if the following expressions remain well defined.) Let the *Bayes welfare* for rule δ be

$$\overline{W}(\pi, \delta) = \int E_\theta \left[W(\theta, \delta(Z)) \right] d\pi(\theta) = \int \left[\int W(\theta, \delta(z)) p_\theta(z) dz \right] d\pi(\theta),$$

where $p_\theta(z)$ is the likelihood function. The Bayes welfare averages the expected welfare, which depends on θ, with respect to the prior measure π. A Bayes rule $\delta_B(z)$ maximizes the Bayes welfare

$$\delta_B(z) = \arg\max_{\delta} \overline{W}(\pi, \delta).$$

When the order of integration can be switched, Bayes welfare can be expressed as

$$\overline{W}(\pi,\delta) = \int \left[\int W(\theta,\delta(z)) p_\theta(z) d\pi(\theta) \right] dz.$$

In this case, the Bayes rule is simply the action that maximizes the inner integral in the Bayes welfare expression for each z:

$$\delta_B(z) = \arg\max_a \int W(\theta,a) p_\theta(z) d\pi(\theta).$$

The Bayesian posterior distribution of θ given z is proportional to the product of the likelihood and prior,

$$d\pi(\theta|z) \propto p_\theta(z) d\pi(\theta),$$

so the Bayes rule maximizes posterior expected welfare:

$$\delta_B(z) = \arg\max_a E[W(\theta,a) \mid z] := \int W(\theta,a) d\pi(\theta|z).$$

Empirical Welfare Maximization: Suppose that $Z^n = (Z_1, Z_2, \ldots, Z_n)$ where Z_i are i.i.d. $P \in \mathscr{M}$. Here we identify θ with P so that the welfare function can be written as $W(P,a)$. A natural nonparametric estimator of P is the empirical distribution $\mathbb{P}_n$ which puts probability $\frac{1}{n}$ on each observed value of Z_i. The empirical welfare maximizer solves:

$$\delta_{EW}(z) = \arg\max_{a \in \mathscr{A}_n} W(\mathbb{P}_n, a),$$

where $\mathscr{A}_n$ is a class of actions which could depend on sample size. This can be viewed as a type of plug-in rule.

An alternative formulation is useful in the case where W has the form of an expectation. If

$$W(P,a) = \int u(z,a) dP(z),$$

for some function $u(\cdot,\cdot)$, then

$$W(\mathbb{P}_n, a) = \frac{1}{n} \sum_{i=1}^n u(Z_i, a),$$

and the empirical welfare maximizer simply maximizes $W(\mathbb{P}_n, a)$ with respect to a.

2.4 Evaluating Statistical Decision Rules

For a given decision rule δ, its expected welfare $E_\theta[W(\theta,\delta(Z))]$ is a function of $\theta \in \Theta$. Typically, we cannot learn θ perfectly from the available data, so it is not possible to construct a rule that achieves the best possible expected welfare uniformly over θ. A rule δ is *admissible* if no other rule does as well for all

possible values of θ and strictly improves upon it for some θ: there does not exist $\tilde{\delta}$ with

$$E_\theta[W(\theta,\delta(Z))] \leq E_\theta[W(\theta,\tilde{\delta}(Z))] \quad \forall\theta\in\Theta, \text{ with}$$

$$E_\theta[W(\theta,\delta(Z))] < E_\theta[W(\theta,\tilde{\delta}(Z))] \quad \text{for some } \theta.$$

In practice, many rules may be admissible. To make finer comparisons among rules, one approach is to aggregate their expected welfare over $\theta\in\Theta$ in some way, for example by averaging with respect to some measure on Θ or considering some notion of worst-case performance.

2.4.1 Bayes Welfare

Recall that for a prior distribution π over Θ, the Bayes welfare of a rule δ is

$$\overline{W}(\pi,\delta) = \int E_\theta\left[W(\theta,\delta(Z))\right] d\pi(\theta),$$

and for given π the corresponding Bayes decision rule maximizes $\overline{W}(\pi,\delta)$. Thus Bayes decision rules automatically maximize average expected welfare. The Bayes welfare criterion can be motivated axiomatically following Anscombe and Aumann (1963) and Savage (1972).

2.4.2 Maxmin and Minmax Regret

The maxmin expected welfare of a decision rule δ is

$$\inf_{\theta\in\Theta} E_\theta\left[W(\theta,\delta)\right].$$

A maxmin rule maximizes this quantity.[3] See Gilboa and Schmeidler (1989) for an axiomatic treatment of maxmin expected welfare.

An alternative criterion is minmax regret. Consider the action $a^*(\theta)$ which maximizes welfare under θ:

$$a^*(\theta) \in \arg\max_a W(\theta,a)$$

with corresponding risk $W^*(\theta) = \max_a W(\theta,a)$.

Welfare loss regret is defined as the difference between the welfare of this ideal action and the welfare of the given action: $W^*(\theta) - W(\theta,a)$, and a minmax regret rule minimizes:

$$\sup_{\theta\in\Theta} E_\theta\left[W^*(\theta) - W(\theta,a)\right].$$

[3] More generally, we could consider the worst case Bayes welfare with respect to a set of prior distributions over Θ, leading to what is called the Γ-minmax criterion.

The minmax regret criterion seems to have been first proposed by Savage (1951), in a discussion of the minmax criterion used by Wald (1950). Axioms for minmax-regret were proposed by Milnor (1954); see also Hayashi (2008) and Stoye (2011).

2.4.3 Optimality under Restrictions on Decision Rules

In some applications, we may wish to restrict the class of decision rules to satisfy some constraints. For example, there may be practical or institutional restrictions on the kinds of allocation rules allowed in the treatment assignment problems introduced following Example 2 above and further considered in Section 4. A more familiar example is the classical Neyman-Pearson approach to inference, where one imposes a constraint that the decision rule controls size (and possibly satisfies some other restrictions), and then seeks to find the "best" rule (according to some criterion) within this class. Similarly, in point estimation one could restrict attention to unbiased estimators and then seek a minmax or average risk minimizing rule. One could also consider rules that *approximately* satisfy some constraint, as in Müller and Wang (2017). Sharp results are typically only available in very simple models, though they can often be extended via the local asymptotic approximation approach discussed next.

2.5 Bounds and Large Sample Approximations for Decision Rules

A statistical decision rule δ is a Markov transition from the data space $\mathscr{Z}$ to the action space $\mathscr{A}$. Thus it takes each the probability distribution $\{P_\theta\}$ over $\mathscr{Z}$ into a probability distribution over $\mathscr{A}$. In principle, given a rule δ we could work out its distribution under every θ, and then compute its expected welfare $E_\theta[W(\theta,\delta(Z))]$ as a function of θ. However, in practice this may be computationally infeasible, especially if we want to evaluate and compare a large set of possible decision rules.

Nevertheless, it may be possible to obtain useful approximations to the distributions and expected welfare properties of decision rules, that faciliate comparisons of rules. This is especially the case for minmax regret expected welfare, because the minmax regret criterion recenters expected welfare conveniently for bounding and limiting arguments. In some cases it is possible to obtain a simple characterization of rules that are approximately optimal with respect to a specific criterion.

2.5.1 Welfare Bounds via Concentration Inequalities

It is often difficult to explicitly calculate worst-case expected welfare, but in some cases it is possible to obtain finite-sample bounds on minmax regret expected welfare. Recent applications of this approach in economics include Manski (2004) and Kitagawa and Tetenov (2018).

To give a flavor of these arguments, suppose that Z_i are i.i.d. P and welfare has the form $W(P,a) = E_P[u(Z,a)]$ for some function u. The empirical welfare based on a sample of size n is $W(\mathbb{P}_n,a) = \frac{1}{n}\sum_{i=1}^n u(Z_i,a)$.

The empirical welfare maximizer is

$$\delta_{ew} = \arg\max_{a \in \mathscr{A}_n} W(\mathbb{P}_n, a),$$

but the true welfare of this rule is $W(P, \delta_{ew})$. Since P is not known, we do not know the true welfare of the rule.

One possibility is to obtain bounds on the regret

$$W(P, a^*) - W(P, \delta_{ew}),$$

where a^* solves $\max_{a \in \mathscr{A}} W(P, a)$. Here $\mathscr{A}$ could be equal to $\mathscr{A}_n$, the same class of possible rules considered in the construction of the EWM rule, in which case the welfare of δ_{ew} is compared to the welfare of the best possible rule in the same class. Or $\mathscr{A}$ could be the class of all possible rules, so that the welfare of the EWM rule is compared to the best possible welfare among all rules.

Useful finite-sample bounds on the regret can sometimes be obtained from concentration inequalities for empirical processes. The key step is to bound the empirical process

$$W(\mathbb{P}_n, a) - W(P, a) = \frac{1}{n} \sum_{i=1}^{n} u(Z_i, a) - E_P[u(Z_i, a)],$$

uniformly over a. Under some conditions, one can guarantee that this quantity is small with high probability. See Bousquet, Boucheron, and Lugosi (2004) for a survey of techniques for obtaining these types of bounds.

This approach yields worst-case bounds on regret which typically depend on sample size n. Embedding the fixed-sample decision problem in a sequence where sample size n increases, one then obtains a rate of convergence of minmax regret towards zero. A decision rule is (minmax regret) rate-optimal if its minmax regret expected welfare achieves the best possible rate of convergence.

2.5.2 Large Sample Approximations via Local Asymptotics

Large sample distributional approximations play an important role in statistics and econometrics, especially in constructing inference procedures. Here, we show some ways to use large sample theory to obtain approximate characterizations of the welfare properties of decision rules and comparisons between decision rules. We use the framework developed by Le Cam (1972, 1986). Our notation and key results are adapted from van der Vaart (1991a) and van der Vaart (1998). Other useful treatments include Ibragimov and Hasminskii (1981) and Le Cam and Yang (2000). In this section we focus on the case with i.i.d. data from a smooth parametric model to illustrate the main concepts, but local asymptotic approximations can also be used in some nonstandard parametric models, problems with dependent data, and settings where the parameter space is infinite-dimensional.

Consider a random sample $Z^n = (Z_1, \ldots, Z_n)$, where the Z_i are i.i.d. with distribution P_θ for $\theta \in \Theta \subset \mathbb{R}^k$. So Z^n is distributed P_θ^n. We will take approximations as the sample size n goes to infinity. If the model is identified in the usual sense, typically there will exist point estimators $\hat{\theta}$ of θ that are consistent in the sense that $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$. Then a plug-in decision rule based on $\hat{\theta}$ will typically have expected welfare converging to the infeasible optimal welfare that would obtain if θ were known. Of course, this type of crude approximation does not capture the welfare loss due to estimation error in $\hat{\theta}$, nor does it lead to useful comparisons between, say, plug-in rules based on different consistent estimators of θ.

Thus, we wish to obtain finer approximations that capture the role of parameter uncertainy in the performance of a decision rule. Suppose that the parametric model $\{P_\theta\}$ satisfies a standard smoothness condition around a centering value $\theta_0 \in \Theta$:

Assumption 1 *(a) Differentiability in quadratic mean: there exists a function $s : \mathcal{Z} \to \mathbb{R}^m$, the score function, such that*

$$\int \left[dP_{\theta_0+h}^{1/2}(z) - dP_{\theta_0}^{1/2}(z) - \frac{1}{2} h' \cdot s(z) dP_{\theta_0}^{1/2}(z) \right]^2 = o(\|h\|^2) \quad \text{as } h \to 0;$$

(b) The Fisher information matrix $J_0 = E_{\theta_0}[ss']$ is nonsingular.

Assumption 1 is a sufficient condition for the model to be locally asymptotically normal (LAN) at θ_0 (van der Vaart, 1998). Given this assumption, it will be useful to adopt the usual local parametrization around a point θ_0,

$$\theta_{n,h} = \theta_0 + \frac{h}{\sqrt{n}}.$$

Differentiability in quadratic mean ensures that the normalized log-likelihood function of the model is approximately quadratic in a local neighborhood of θ_0. This turns out to be a key property that leads to the local asymptotic normality property. More generally, the limiting distributions of natural estimators and the limiting properties of the decision problem will depend crucially on the local behavior of the likelihood function of the model. In "non-regular" models the likelihood may have a different limiting form, but this form may be sufficiently tractable to lead to useful risk approximations and comparisons of decision rules.

In general, we fix $\theta_0 \in \Theta$ and consider sequences of local alternatives $\theta_0 + \phi_n h$, where $h \in \mathbb{R}^k$ and $\phi_n \to 0$ is a normalizing sequence of matrices. In regular parametric models satisfying Assumption 1, the appropriate norming is $\phi_n = \frac{1}{\sqrt{n}} I_k$ where I_k is the k-dimensional identity matrix, and we would approximate the distribution of a root-n consistent estimator $\hat{\theta}$ by taking the distributional limit of

$$\phi_n^{-1}\left(\hat{\theta} - (\theta_0 + \phi_n h)\right) = \sqrt{n}\left(\hat{\theta} - \theta_0\right) - h$$

under the sequence of measures $P_{\theta_0 + h/\sqrt{n}}^n$. However, in other applications a different norming sequence may be appropriate.

Consider the sequence of statistical models $\mathscr{E}_n = \{P^n_{\theta_0+\phi_n h} : h \in \mathbb{R}^k\}$. Their likelihood ratio processes are defined as

$$\Lambda_{n,h_0} = \left(\frac{dP^n_{\theta_0+\phi_n h}}{dP^n_{\theta_0+\phi_n h_0}} \right)_{h\in\mathbb{R}^k}. \tag{1}$$

When working with Λ_{n,h_0}, we take its distribution under the localized probability measure $P^n_{\theta_0+\phi_n h_0}$. The sequence of experiments $\mathscr{E}_n$ are said to converge weakly to the experiment $\mathscr{E}$, supported on some measurable space $\{\mathscr{X},\mathscr{B}\}$, with probability measures $\{F_h : h \in \mathbb{R}^k\}$, if the likelihood ratio processes in (1) converge (in the sense of finite-dimensional weak convergence) to

$$\left(\frac{dF_h}{dF_{h_0}} \right).$$

In other words, the likelihood ratio processes of the model are approximated by the likelihood ratio processes of the model $\{F_h : h \in \mathbb{R}^k\}$.

For example, under the regularity conditions in Assumption 1, it can be shown that

$$\log\left[\frac{dP^n_{\theta_0+\phi_n h}}{dP^n_{\theta_0}} \right] = h'\Delta_n - \frac{1}{2}h'J_0 h + o_p(1),$$

where J_0 is the Fisher information matrix defined in Assumption 1(b), and $\Delta_n \rightsquigarrow N(0, J_0^{-1})$. From this we can show that, for each $h_0 \in \mathbb{R}^k$ and every finite subset $I \subset \mathbb{R}^k$, the vectors

$$\left(\frac{dP^n_{\theta_0+h/\sqrt{n}}}{dP^n_{\theta_0+h_0/\sqrt{n}}} \right)_{h\in I} \overset{\theta_0+h_0/\sqrt{n}}{\rightsquigarrow} \left(\exp\left[(h-h_0)'\Delta - \frac{1}{2}(h-h_0)'J_0(h-h_0) \right] \right),$$

where $\Delta \sim N\left(h_0, J_0^{-1}\right)$. The limit on the right is the log likelihood ratio associated with the experiment of observing a single draw from the shifted normal distribution $N\left(h, J_0^{-1}\right)$, where J_0 is known and h is the parameter of interest.

The fact that the likelihood ratio process of the model of interest, after suitable normalization, converges to the likelihood ratio of the simple shifted normal model, suggests that the normal model serves as a kind of canonical model, characterizing the limit distributions of feasible decision rules. This intuition is made precise in asymptotic representation theorems, which state that, for any sequence of decision rules δ_n that have limiting distributions under the local parameter sequence $\theta_0 + h/\sqrt{n}$, there exists a (possibly randomized) decision rule based on a shifted normal $N(h, J_0^{-1})$ whose exact distributions under every h correspond to the limit distributions of δ_n. This enables us to characterize the limiting distributions, and associated welfare and risk properties, of decision rules, and in some cases leads to tractable welfare- and risk-optimality results. We will examine specific applications of asymptotic representation theorems in later sections.

Other limits besides the shifted normal experiment are also possible, and we examine some examples below. If the set of probability distributions under consideration is infinite-dimensional, corresponding

to a nonparametric or semiparametric model, local asymptotic approximations are also possible by an appropriate construction of the local parameter space; see Bickel, Klaasen, Ritov, and Wellner (1993), van der Vaart (1991a), and Section 4 below. (See also Müller (2011) for an alternative approach to local asymptotic efficiency without parametric restrictions.)

2.5.3 Approximating Risk and Expected Welfare

Recall that we evaluate decision rules by their expected welfare functions, or their risk functions. The expected welfare of a rule δ_n is

$$E_{\theta_0+\phi_n h}\left[W(\theta_0+\phi_n h, \delta_n)\right],$$

where the expectation is with respect to the distribution of δ_n under $\theta_0+\phi_n h$. For a given δ_n, we can view this as a function of the local parameter h.

Taking $n \to \infty$, we could take the limit of the above expression and compare different decision rules by their limiting expected welfare or risk functions. However, not all choices for the utility or loss functions will lead to useful limits. In later sections we will discuss some conditions that ensure that the limiting expected welfare and risk functions are nondegenerate and lead to meaningful comparisons of decision rules; in some cases these conditions may be quite stringent.

A distinct but closely related approach is to work directly with the approximate distribution of the decision rule. Under local parametrizations of the type described above, decision rules δ_n will often have limit distributions after some normalization:

$$r_n(\delta_n - c_n) \rightsquigarrow Q_h,$$

where Q_h are the limiting laws under different values of the local parameter h, and the norming r_n and centering c_n may depend on the underlying model, the nature of the decision problem, and the specific decision rule. We regard this as an approximation to the finite-sample distribution of δ_n:

$$\delta_n \overset{a}{\sim} c_n + Q_h/r_n.$$

With this approximation, we can then obtain heuristic approximations to the expected welfare or risk of δ_n, given a choice for the utility or loss function. We will discuss the relationship between these two approximation strategies in some of the applications below.

3 Point Decisions

In this section, we consider decision problems where the action space $\mathscr{A}$ is a subset of a finite-dimensional Euclidean space. This applies to point estimation in parametric and semiparametric models, but also

includes other decision problems. For example, data from auctions can be used to learn about bidder preferences, and in turn to choose the reserve price in future auctions to increase expected revenue to the seller. This can be viewed as a point decision problem where the action is the future value of the reserve price. Another class of point decision problems are statistical portfolio choice problems, where the goal is to choose an allocation vector a. Each element of the vector a is the fraction of wealth that is to be allocated to a particular asset. Data are used to learn about the joint probability distribution of asset returns, but the ultimate goal is to choose the future allocation.

For point estimation and related problems, there is a well developed theory for standard loss functions. We review some key aspects of this theory, emphasizing the use of invariance arguments. We then consider asymptotic theory for point decisions. Under standard loss functions, the finite-sample theory carries over nicely, leading to local asymptotic optimality results.[4] However, we also wish to consider other loss functions which may be less tractable but are natural in economic applications. We may also consider procedures that involve some preliminary model selection or model averaging. For these more complicated problems, it is much more difficult to obtain optimality results, even in large samples. But tools are available to simplify the characterization of the welfare properties of specific rules and to compare rules.

3.1 Shift Equivariance in Point Estimation

To introduce some key ideas, we first consider a very simple point estimation problem. Suppose we observe $Z \sim N(\theta, 1)$ (where the variance is known). A point estimator for θ is a decision rule $\delta : \mathbb{R} \to \mathbb{R}$ with the interpretation that $\delta(z)$ is the estimate of θ when $Z = z$. Consider squared error loss $L(\theta, a) = (a-\theta)^2$. Then the risk of the estimator δ is

$$R(\theta, \delta) = E_\theta \left[(\delta(Z) - \theta)^2 \right] = \int (\delta(z) - \theta)^2 \phi(z - \theta) dz$$

where $\phi(\cdot)$ is the standard normal density. As we noted above, this fits into our framework with welfare W equal to the negative of loss L. Expected welfare then equals the negative of risk.

Consider the point estimator $\delta(Z) = Z$. This is the maximum likelihood estimator (and hence the ML plug-in rule); it is also the Bayes decision rule under a diffuse prior; and it is the empirical welfare maximizer. This rule has a number of optimality properties. It minimizes Bayes risk with respect to the diffuse prior. It can also be shown to be minmax (and minmax regret), and its risk is constant over the parameter space.

These basic finite-sample results can be generalized in a number of ways. One useful extension is to settings that preserve the "shift" structure of the simple normal model:

Parametric Shift Model *Let $Z \sim P_\theta$, where the parameter $\theta \in \mathbb{R}^k$, and the distribution P_θ of Z under θ has*

[4] Some of the results in this section are drawn from the working paper Hirano and Porter (2003).

a density $p(z-\theta)$ for some fixed function p, with support contained in $\mathbb{R}^k$. We can interpret this model as saying that $Z = \theta + V$, where V has density p independent of θ.

For the parametric shift model, let the action space $\mathscr{A}$ be $\mathbb{R}^k$, corresponding to point estimation or point forecasting. (We could also extend the following analysis for the case where interest centers on a linear function of θ.) Suppose that the loss function has the form $L(a-\theta)$, that is, it only depends on the parameter and action through their difference $a-\theta$. We assume L has the following properties:

Assumption 2 *The loss function $L(a-\theta)$ satisfies:*

1. $L(a-\theta) \geq 0$*;*
2. $L(0) = 0$*;*
3. *The sets $\{w \in \mathbb{R}^k : L(w) \leq \tau\}$ are compact for all $\tau \geq 0$.*

This covers a number of popular loss functions besides squared error loss. We could work with absolute error loss $L(a-\theta) = |a-\theta|$, or asymmetric loss functions such as "check function" loss:

$$L(a-\theta) = \begin{cases} c(a-\theta) & \text{if } a-\theta \geq 0 \\ -(1-c)(a-\theta) & \text{if } a-\theta < 0 \end{cases}$$

for some $c \in (0,1)$. Another asymmetric loss function, used in forecasting and other applications, is the linex (linear-exponential) loss (Varian (1974); see also Zellner (1986) and Christoffersen and Diebold (1997)):

$$L(a-\theta) = \exp(c(a-\theta)) - c(a-\theta) - 1, \quad c \neq 0. \tag{2}$$

This loss function is approximately linear on one side of zero and approximately exponential on the other side, with the degree of asymmetry controlled by c.

In this extension of the normal mean problem, the statistical model, the action space, and the loss function are all compatible with vector addition in $\mathbb{R}^k$. More precisely, we can consider the translation group in $\mathbb{R}^k$ as defining groups of transformations over Θ and $\mathscr{R}^k$, and the loss function is symmetric with respect to these groups . Eaton (1989) provides an extensive discussion of statistical methods for models with a group structure.

It will be useful to consider decision rules that are also symmetric with respect to the translation group. These are the *equivariant* estimators. An equivariant nonrandomized estimator has the following property: for all $z, g \in \mathbb{R}^k$,

$$\delta(z) + g = \delta(z+g).$$

Letting $g = -z$, it follows that $\delta(z) = \delta(0) + z$, so any such rule must have the form $\delta(Z) = Z + s$, where s is some constant.

We can extend the concept of equivariance to randomized estimators. A randomized estimator is *equivariant in law* if the distribution of $\delta - \theta$ under P_θ is invariant to θ. By the convolution theorem (Hájek (1970), Beran (1995)), any equivariant-in-law rule can be written as $\delta = Z + S$, where S is a random variable with a fixed distribution independent of Z. Let μ_S denote the distribution of S. Then the risk of an equivariant-in-law estimator can be written as

$$R(\theta, \delta) = E_\theta[L(\delta - \theta)] = \int \int L(z + s - \theta) p(z - \theta) dz d\mu_S(s).$$

Setting $v = z - \theta$, we have

$$R(\theta, \delta) = \int \int L(v + s) p(v) dv d\mu_S(s),$$

which does not depend on θ. Since equivariant-in-law rules have constant risk, they can be ordered and we can seek to minimize risk across this class. Typically this minimum will be attained by a nonrandomized rule $\delta = Z + s$, for which the risk is

$$R(\theta, \delta) = \int L(v + s) p(v) dv.$$

Suppose that the expression above has a unique minimum at $s^* \in \mathbb{R}^k$. Then $\delta^* = Z + s^*$ will be a best equivariant-in-law rule. Its risk does not depend on θ. By the Hunt-Stein theorem (see, for example, Wesler (1959)), this estimator is also minmax. It is also minmax regret, due to the form of the loss function.

The best equivariant estimator is also the Bayes estimator under a flat prior. Recall that the Bayes rule $\delta_B(z)$ minimizes the Bayes risk

$$\overline{R}(\pi, \delta) = \int R(\theta, \delta) d\pi(\theta),$$

where π is the prior distribution. If π is Lebesgue measure on $\mathbb{R}^k$, then the Bayes risk equals

$$\int R(\theta, \delta) d\theta = \int \int L(\delta(z) - \theta) p(z - \theta) dz d\theta.$$

Rearranging the order of integration, it is enough to choose $\delta_B(z)$ for each z to minimize

$$\int L(\delta(z) - \theta) p(z - \theta) d\theta.$$

Setting $v = z - \theta$, we can write the minimand as

$$\int L(\delta(z) + v - z) p(v) dv,$$

and setting $\alpha(z) = \delta(z) - z$,

$$\int L(v + \alpha(z)) p(v) dv.$$

By our previous assumption, this is minimized by setting $\alpha(z) = s^*$ for all z, implying that the Bayes rule is $\delta(z) = z + s^*$. Hence the Bayes rule for a flat prior is best equivariant.

Let us also consider the maximum likelihood estimator. In the shift model, the MLE solves

$$\hat{\theta}(z) = \arg\max_{\theta \in \mathbb{R}^k} p(z-\theta).$$

Suppose that the density $p(v)$ has a unique maximum v^*. Then it is clear that

$$\hat{\theta}(z) = z - v^*.$$

So the MLE is equivariant, and by shifting the MLE we obtain the best equivariant estimator:

$$\delta^*(z) = \hat{\theta}(z) + (s^* + v^*).$$

While we have seen that some natural estimators, such as the MLE and the Bayes estimator under a flat prior, are equivariant, it may not always be desirable to restrict attention to equivariant estimators. We will return to this issue in discussing model choice and shrinkage procedures below.

3.2 Asymptotics for Point Estimators

The exact results surveyed in the previous subsection are powerful but require a very special structure among the parametric model, the action space, and the loss function. However, the shift structure, or at least some aspects of it, often hold approximately in more complicated problems, a point first shown by Le Cam in a number of influential papers including Le Cam (1970, 1972). Here we consider some ways to use Le Cam's local asymptotic theory to simplify the analysis of point decision rules.

As in Subsection 2.5.2, consider a random sample $Z^n = (Z_1, \ldots, Z_n)$, where the Z_i are i.i.d. with distribution P_θ for $\theta \in \Theta \subset \mathbb{R}^k$. Fix $\theta_0 \in \Theta$ and suppose there is a local parametrization $\theta_0 + \phi_n h$ for $h \in \mathbb{R}^k$ and some norming sequence $\phi_n \to 0$. The norming sequence is intended to make the parameter $\theta = \theta_0 + \phi_n h$ difficult to distinguish from θ_0 asymptotically, in the sense of convergence of likelihood ratios below. Typically, there will exist estimators $\hat{\theta}$ such that $\phi_n^{-1}(\hat{\theta} - \theta)$ has nondegenerate limiting distributions under local sequences of parameter values.

Suppose the likelihood ratio processes satisfy convergence to a shift experiment: for every finite subset $I \subset \mathbb{R}^k$ and every $h_0 \in \mathbb{R}^k$,

$$\Lambda_{n,h_0} = \left(\frac{dP^n_{\theta_0 + \phi_n h}}{dP^n_{\theta_0 + \phi_n h_0}} \right)_{h \in I} \overset{\theta_0 + \phi_n h_0}{\rightsquigarrow} \left(\frac{f_{\theta_0}(X - h)}{f_{\theta_0}(X - h_0)} \right)_{h \in I}, \tag{3}$$

where $f_{\theta_0}(\cdot)$ is a probability density function, and X has density $f_{\theta_0}(x - h_0)$. The limit is the likelihood ratio process of the experiment of observing $X = h + V$, where V has density $f_{\theta_0}(v)$.

Example 4 *Recall that if the differentiability in quadratic mean condition, Assumption 1, holds, then*

the model is locally asymptotically normal (LAN). In this case, $\phi_n = n^{-1/2}$ and f_{θ_0} can be taken to be a multivariate normal density:

$$f_{\theta_0}(\cdot) = dN(\cdot \mid 0, J_0^{-1}),$$

where J_0 is the Fisher information matrix at θ_0.

A closely related case occurs when J_0 is random. Such cases are called local asymptotic mixed normal (LAMN), and arise in some time series applications and other settings. (See Davies (1985), Jeganathan (1982), and Jeganathan (1995), among others.) LAMN models are shift experiments conditional on J_0. We will not pursue this case explicitly below, but many of the results for shift experiments generalize to the case of conditional shift experiments.

Example 5 *Structural econometric models with parameter-dependent support are non-regular, but can have interesting and tractable limit distributions. To illustrate this we consider a simplified version of the parametric auction model from Paarsch (1992).*

Consider a first-price procurement auction with m bidders under the independent private values paradigm. Bidders are symmetric with cost $c \sim Exp(\theta)$, where $Exp(\theta)$ denotes an exponential distribution with mean θ. In the symmetric equilibrium, bidders bid

$$b = c + \frac{\theta}{m-1}$$

Note that the support of the distribution of bids depends on the parameter θ.

We have a random sample of n auctions (each with m bidders). Suppose we observe the winning (lowest) bid $\underline{b}_i$ for each auction. Then for $i = 1, \ldots, n$,

$$\underline{b}_i \sim Exp\left(\frac{\theta}{m}\right) + \frac{\theta}{m-1}.$$

Although the distribution of winning bids is shifted by $\theta/(m-1)$, its shape also depends on the parameter, so this is not a shift experiment. However, the structure of the model simplifies under local asymptotics.

Fix a centering value θ_0. The appropriate norming turns out to be $\phi_n = n^{-1}$. Then it can be shown that the experiments converge to a shifted exponential experiment with

$$f_{\theta_0}(\cdot) = dExp\left(\cdot \,\middle|\, \frac{m-1}{m}\theta_0\right). \tag{4}$$

See Hirano and Porter (2003) for further details.

A consequence of the limiting shift experiment is an asymptotic representation theorem for estimators (and other point decisions). The following result is due to van der Vaart (1991a).

Theorem 1 *Suppose that the sequence of likelihood ratio processes converges to the likelihood ratio process of a shift experiment as in Equation (3). Let $\hat{\theta}_n$ be a sequence of estimators that have limits under the local parameter sequences $\theta_0 + \phi_n h$:*

$$\phi_n^{-1}(\hat{\theta}_n - \theta_0) \overset{\theta_0+\phi_n h}{\rightsquigarrow} L_h.$$

Then there exists a (possibly randomized) estimator $T(X,U)$ such that

$$T(X,U) \sim L_h$$

when $X = h + V$, where V has density $f_{\theta_0}(v)$, and U has a standard uniform distribution independent of X.

This result states that any estimator possessing limits under the local parameter sequences can be represented asymptotically as some estimator in the simple shift model $X = h + V$.

Moreover, under additional regularity conditions, the form of the matching limit experiment estimator T can be deduced from the nature of the original estimator. In particular, suppose that $\hat{\theta}_n$ is the maximum likelihood estimator. Then, under conditions for an argmax theorem (see, e.g., Theorem 3.2.2 in van der Vaart and Wellner (1996)), its limiting representation T is the maximum likelihood estimator in the limit experiment:

$$T(X) = \arg\max_h f_{\theta_0}(X - h).$$

Here T does not depend the auxiliary random term U that appears in the asymptotic representation theorem. Since T is shift-equivariant in the limit experiment, $\hat{\theta}_n$ is locally asymptotically shift-equivariant, i.e. regular. We will discuss regular estimators in more detail below.

3.3 Limits for Risk Functions

We want to use large sample approximations to study the approximate risk properties of point estimators. To do this we need to examine the limiting properties of risk functions.

Suppose that the loss function for point estimation depends only on the difference $\hat{\theta} - \theta$. Then the risk of the estimator is the expected loss

$$R(\theta, \hat{\theta}_n) = E_\theta \left[L(\hat{\theta}_n - \theta) \right].$$

Since the point estimator will typically satisfy $\hat{\theta}_n \xrightarrow{p} \theta$ for all θ, the risk function will typically converge to 0 as $n \to \infty$. This type of limit is too crude to make useful comparisons between point estimators. Thus we will consider rescaling the risk so that its limit is nondegenerate. This is most straightforward when the loss function is homogeneous.

3.3.1 Homogeneous Loss Function

Suppose $L(\cdot)$ is homogenous of degree j. For example, squared error loss $L(w) = w^2$ is homogenous of degree 2 while absolute error loss $L(w) = |w|$ is homogenous of degree 1. Then it is natural to scale up the risk by the factor ϕ_n^{-j}, so that

$$\begin{aligned}\phi_n^{-j} R(\theta, \hat{\theta}_n) &= E\left[\phi_n^{-j} L(\hat{\theta}_n - \theta)\right] \\ &= E\left[L\left(\phi_n^{-1}(\hat{\theta}_n - \theta)\right)\right].\end{aligned}$$

For example, in LAN models, $\phi_n = n^{-1/2}$, and $n^{j/2} R(\theta, \hat{\theta}_n) = E\left[L\left(\sqrt{n}(\hat{\theta}_n - \theta)\right)\right]$. For estimators such that $\sqrt{n}(\hat{\theta}_n - \theta)$ has a nondegenerate limiting distribution, the limit of the rescaled risk expression will typically be nonzero.

In Theorem 1, the limiting behavior of point estimators is considered under local parametrizations $\theta_0 + \phi_n h$. This suggests that we should consider the limiting risk as a function of the local parameter h. Suppose the estimator satisfies the conditions of Theorem 1. Then, under mild additional conditions,

$$\begin{aligned}\phi_n^{-j} R(\theta_0 + \phi_n h, \hat{\theta}_n) &= E_{\theta_0 + \phi_n h}\left[L\left(\phi_n^{-1}\left[\hat{\theta}_n - (\theta_0 + \phi_n h)\right]\right)\right] \\ &= E_{\theta_0 + \phi_n h}\left[L\left(\phi_n^{-1}(\hat{\theta}_n - \theta_0) - h\right)\right] \\ &\to E_h\left[L\left(T(X, U) - h\right)\right] \\ &=: R_\infty(h, T).\end{aligned}$$

In this form, limiting risk captures finer variation in estimator behavior than limits taken pointwise in θ. Local asymptotic risk often approximates the finite sample risk of the estimator well. This finite-sample behavior is also important for shrinkage-type estimators; see Section 3.5.

Having transformed the estimation problem to the local parameter space, and having normalized the loss by a factor based on the localization and the degree of homogeneity of the loss function, we can now compare estimators by the normalized limiting risks. This amounts to considering estimators in the limiting version of the problem, using the same loss function as in the original problem.

Thus we can connect this limiting problem to the finite sample theory in Section 3.1. Theorem 1 gives an asymptotic characterization of all estimators that possess limit distributions under the local parameter sequences $\theta_0 + \phi_n h$. An important subclass of these estimators are the regular estimators. An estimator $\hat{\theta}_n$ is *regular* (at θ_0) if the limit distributions of $\phi_n^{-1}\left[\hat{\theta}_n - (\theta_0 + \phi_n h)\right]$ are the same for all h. Equivalently, regular estimators satisfy

$$\phi_n^{-1}(\hat{\theta}_n - \theta_0) \overset{\theta_0 + \phi_n h}{\rightsquigarrow} L_0 + h,$$

where L_0 is a fixed law that does not depend on h. By Theorem 1, for each such regular estimator $\hat{\theta}_n$, there exists an estimator $T(X, U)$ with the distribution $L_0 + h$ in the limit experiment. Such estimators are equivariant in law, and in Section 3.1 we saw that equivariant-in-law estimators are easy to analyze.

In particular, if the loss function satisfies Assumption 2, the risk of $T(X,U)$ is constant in h, making risk calculations and comparisons relatively straightforward. Under some conditions, there is a unique best equivariant estimator, and this estimator is also minmax among all estimators. The risk of the best equivariant estimator in the limiting version of the problem can then serve as a local asymptotic risk bound for the original problem.

The limiting shift experiment given in (3) involves a random variable X with density $f_{\theta_0}(x-h)$. Following the argument in Section 3.1, if $f_{\theta_0}(\cdot)$ has a unique maximum, then the maximum likelihood estimator $T_{ML} = T_{ML}(X)$ can be written as

$$T_{ML} = X - v^*_{\theta_0},$$

where

$$v^*_{\theta_0} = \arg\max_v f_{\theta_0}(v).$$

Let $s^*_{\theta_0}$ be the minimizer of $\int L(v+s) f_{\theta_0}(v)\,dv$. Then the shifted MLE $T_{ML} + v^*_{\theta_0} + s^*_{\theta_0}$ is the best equivariant estimator in the limit experiment.

We can then seek to construct an estimator sequence in the original problem that matches the optimally shifted MLE in the limit. As noted above, under regularity conditions, if $\hat{\theta}_{ML}$ is the MLE in the original problem, then $\phi_n^{-1}(\hat{\theta}_{ML} - \theta_0) \overset{\theta_0+\phi_n h}{\rightsquigarrow} T_{ML}(X)$. Hence, to achieve asymptotic optimality, we can locally shift $\hat{\theta}_{ML}$ to match the shifted MLE, $T_{ML} + v^*_{\theta_0} + s^*_{\theta_0}$, in the limit experiment. Note that $v^*_{\theta_0}$ and $s^*_{\theta_0}$ depend on f_{θ_0} which is generally unknown. Under further conditions, we can replace these terms by the estimates $v^*_{\hat{\theta}_{ML}}$ and $s^*_{\hat{\theta}_{ML}}$, yielding the following estimator:

$$\hat{\theta}_{BR} = \hat{\theta}_{ML} + \phi_n(v^*_{\hat{\theta}_{ML}} + s^*_{\hat{\theta}_{ML}}).$$

This estimator is best regular, in the sense that its limiting risk is minimal among regular estimators, and it is locally asymptotically minmax in the sense that its worst case limiting risk (suitably defined, see for example van der Vaart (1998)) is minimal among all estimators.

It is important to note that optimality in the local asymptotic minmax sense is conceptually different from finite sample minmaxity, and neither condition implies the other. In the asymptotic analysis, we are considering the normalized risk of the estimator for values of θ in a shrinking neighborhood of θ_0, rather than over the original parameter space Θ.

Example 4 (LAN example, continued) *In the LAN case these considerations lead to familiar asymptotic optimality results. Specifically, the limit experiment consists of observing $X \sim N(h, J_0^{-1})$, and the limiting distribution of the MLE is the same as the distribution of X. Since X is best equivariant and minmax under squared error loss (and other symmetric loss functions), the MLE is asymptotically best equivariant and minmax.*

Example 5 (Auction example, continued) *As noted above, the limit experiment is a shifted exponential distribution where f_{θ_0} is the density of an exponential distribution, as given in (4). The exponential density is maximized at zero for all θ_0, so $v^*_{\theta_0} = 0$. It follows that the MLE has*

$$\phi_n^{-1}\left(\hat{\theta}_{ML} - \theta_0\right) \overset{h}{\rightsquigarrow} X,$$

where X has density $f_{\theta_0}(x-h)$. As discussed above, the limiting distribution is the same as the distribution of the MLE in the shifted exponential experiment. However, in that experiment, the optimal estimator under a location-equivariant loss L is given by $X + s^*_{\theta_0}$, where $s^*_{\theta_0}$ is defined above and depends on both L and f_{θ_0}. For example, under squared error loss the optimal shift can be calculated to be

$$s^*_{\theta_0} = -\frac{\theta_0(m-1)}{m},$$

while under absolute error loss the optimal shift is

$$s^*_{\theta_0} = -\frac{\theta_0(m-1)}{m}\log 2.$$

Putting together these pieces, $\phi_n\left(v^*_{\theta_0} + s^*_{\theta_0}\right)$ forms a "bias-correction" for the MLE (where bias is defined relative to the loss function L). In practice, we can replace $s^*_{\theta_0}$ by a suitable estimator by replacing θ_0 with $\hat{\theta}_{ML}$ in the desired expression above. Noting that $v^*_{\theta_0} = 0$, our locally asymptotically optimal estimator is

$$\tilde{\theta}_n = \hat{\theta}_{ML} + \phi_n s^*_{\hat{\theta}_{ML}}.$$

This estimator will have limiting risk function $R_\infty(h, X + s^*_{\theta_0})$ equal to the limiting risk of the best equivariant estimator in the shifted exponential limit experiment.

3.3.2 Nonhomogeneous Loss

The theory outlined in the previous subsection is the standard approach to showing local asymptotic optimality of point estimators in settings with limiting shift form, including the LAN case. The homogeneity (and location-equivariant form) of the loss function allows us to work with the same loss function in the renormalized limiting version of the estimation problem.

Next, we consider the problem of choosing an estimator when the loss function is not homogeneous. An example of a nonhomogenous loss function is linex loss $L(a-\theta) = \exp(c(a-\theta)) - c(a-\theta) - 1$, as defined in (2). Depending on the value of c, linex loss can be strongly asymmetric, but it is smooth and approximately quadratic at the origin. As a result, the $j = 2$ norming is appropriate, because, for $(\hat{\theta}_n - \theta)$ small,

$$\phi_n^{-2} R(\theta, \hat{\theta}_n) \approx \frac{c^2}{2} E\left[\left(\phi_n^{-1}(\hat{\theta}_n - \theta)\right)^2\right]. \tag{5}$$

Linex loss is approximately quadratic near zero, so if one applies the appropriate correction under squared error loss, the estimator will have asymptotically optimal limiting risk. For example in LAN models, the optimal correction is zero, so that no correction is needed for the MLE to be optimal with respect to local asymptotic risk. This is somewhat unsatisfying, because the asymmetry (at the original global scale) of the linex loss function no longer plays a role in the analysis.

We may wish to apply a small sample correction that reflects the global properties of the nonhomogeneous, possibly asymmetric loss function, but also achieves the usual first-order, local asymptotic optimality for the resulting estimator. We outline one possible approach. For concreteness, we base the corrected estimator on the MLE, but the approach could be applied to other estimators that are asymptoticaly equivalent to the MLE up to a local shift, such as the Bayesian posterior mean.

Suppose we have convergence to a shift experiment as in Theorem 1. Further, assume that the maximum likelihood estimator satisfies $\phi_n^{-1}(\hat{\theta}_{ML} - \theta_0) \overset{\theta_0+\phi_n h}{\rightsquigarrow} T_{ML}(X)$, where $T_{ML}(X) = X - v^*_{\theta_0}$ as defined in Section 3.3.1. Since $X \overset{h}{\sim} V + h$ as in Section 3.2, we use the approximation $\hat{\theta}_{ML} - \theta_0 \underset{h}{\approx} \phi_n V + \phi_n(h - v^*_{\theta_0})$.

Now consider a shifted version of MLE: $\hat{\theta}_{ML} + \phi_n s^*$. We want to use the approximation to the distribution of the MLE to obtain an approximately optimal choice of s^*. The risk of $\hat{\theta}_{ML} + \phi_n s$ is:

$$\begin{aligned}
E_{\theta_0+\phi_n h}\left[L\left((\hat{\theta}_{ML} + \phi_n s) - (\theta_0 + \phi_n h)\right)\right] &= E_{\theta_0+\phi_n h}\left[L\left((\hat{\theta}_{ML} - \theta_0) - \phi_n h + \phi_n s\right)\right] \\
&\approx E_V\left[L\left(\phi_n V + \phi_n(h - v^*_{\theta_0}) - \phi_n h + \phi_n s\right)\right] \\
&= E_V\left[L\left(\phi_n V - \phi_n v^*_{\theta_0} + \phi_n s\right)\right] \\
&= \int \left[L\left(\phi_n v - \phi_n v^*_{\theta_0} + \phi_n s\right)\right] f_{\theta_0}(v)\, dv.
\end{aligned}$$

Choose $s^*_{\theta_0}$ to minimize the above expression, which approximates the risk of locally shifted MLE. Given the shift structure of the limit experiment and the regularity of the MLE, it follows that the approximate risk expression does not depend on the local parameter h. Hence, the approximately optimal local shift $s^*_{\theta_0}$ also does not depend on h. As above, one could implement this correction as: $\hat{\theta}_{ML} + \phi_n\, s^*_{\hat{\theta}_{ML}}$. This estimator would be locally asymptotically optimal under loss L but would include a correction to capture the global curvature of the loss function.

Example 4 (LAN example, continued) *In the LAN case,*

$$\sqrt{n}\left(\hat{\theta}_{ML} - \left(\theta_0 + \frac{h}{\sqrt{n}}\right)\right) \overset{\theta_0+h/\sqrt{n}}{\rightsquigarrow} V,$$

where $V \sim N(0, J_0^{-1})$. Consider linex loss L as introduced in (2). We seek the finite sample correction factor

described above. For simplicity, take θ_0 to be scalar. To obtain $s^*_{\theta_0}$, solve

$$\min_s E\left[L\left(\frac{1}{\sqrt{n}}(V+s)\right)\right]$$

which yields

$$s^*_{\theta_0} = -\frac{1}{2\sqrt{n}}cJ_0^{-1}.$$

Hence, $\tilde{\theta} = \hat{\theta}_{ML} - \frac{1}{2n}c\hat{J}^{-1}$ is an asymptotically optimal estimator under linex loss that reflects the asymmetry in linex loss.

Example 5 (Auction example, continued) *The maximum likelihood estimator in the first price auction is regular, with*

$$n\left(\hat{\theta}_{ML} - \left(\theta_0 + \frac{h}{n}\right)\right) \overset{\theta_0 + h/n}{\rightsquigarrow} V,$$

where $V \sim Exp\left(\frac{m-1}{m}\theta_0\right)$. Again we consider linex loss L as introduced in (2) and want to find the corresponding finite sample correction. To obtain $s^*_{\theta_0}$, solve

$$\min_s E\left[L\left(\frac{1}{n}(V+s)\right)\right]$$

which yields

$$s^*_{\theta_0} = \frac{n}{c}\ln\left(1 - \frac{c(m-1)\theta_0}{nm}\right).$$

Note that the approximate linex risk exists if $\frac{m}{(m-1)\theta_0} > \frac{c}{n}$, which holds for large enough n. The resulting shifted MLE is $\tilde{\theta} = \hat{\theta}_{ML} + \frac{1}{c}\ln\left(1 - \frac{c(m-1)\hat{\theta}_{ML}}{nm}\right)$. For finite n, this differs from the bias correction derived above under squared error loss, but since

$$\frac{n}{c}\ln\left(1 - \frac{c(m-1)\theta_0}{nm}\right) \approx -\frac{(m-1)\theta_0}{m}$$

for large n, the small sample correction for linex loss is asymptotically equivalent to the bias corrected estimator under squared error loss.

3.4 Point Decisions in General Action Spaces

So far in this section we have focused on estimating the entire parameter vector θ and drawn upon classical results. Next we briefly consider other decision problems that also have finite-dimensional action spaces. This applies to point estimation of parameter subvectors, but also to other economically interesting decision problems such as the choice of reserve auction price in Example 2, and the choice of portfolio weights discussed below in Example 3.

Suppose the action space is $\mathscr{A} \subset \mathbb{R}^{k_A}$, and the loss function is $L(\theta, a)$ for $a \in \mathscr{A}$. Here we do not restrict the loss function to have the shift form $L(a-\theta)$. It will be useful to compare action and decisions to the ideal (but typically infeasible) "action rule" $a^*(\theta) = \arg\min_a L(\theta, a)$. Assume that $a^*(\theta)$ is unique for all θ and is smooth in θ. Also, assume that loss is normalized so that $L(\theta, a^*(\theta)) = 0$, as would be true for regret loss. And finally assume that $0 = \frac{\partial}{\partial a} L(\theta, a^*(\theta))$, as would be the case in an interior solution to the minimization problem defining a^*.

Consider a rule $\hat{\delta}$ such that

$$\tilde{\phi}_n^{-1}(\hat{\delta} - a^*(\theta_0 + \phi_n h)) \overset{\theta_0+\phi_n h}{\rightsquigarrow} \mathscr{L}_{\theta_0,h}. \tag{6}$$

for a rate $\tilde{\phi}_n \to 0$. It is straightforward to extend the notion of a regular estimator to rules in this setting. A rule $\hat{\delta}$ satisfying (6) is *regular* if $\mathscr{L}_{\theta_0,h}$ does not depend on h. Let this limiting distribution be denoted $D \sim \mathscr{L}_{\theta_0}$.

Let us expand the loss function around the infeasible rule a^*. For $\gamma = (\gamma_1, \ldots, \gamma_{k_A})$ a multi-index, under appropriate smoothness we can write

$$\begin{aligned} L(\theta, a) &\approx \underbrace{L(\theta, a^*)}_{=0} + \sum_{|\gamma|=1} \frac{1}{\gamma!} \underbrace{\frac{\partial^{|\gamma|}}{\partial a^\gamma} L(\theta, a^*)}_{=0} (a-a^*)^\gamma + \sum_{|\gamma|=2} \frac{1}{\gamma!} \frac{\partial^{|\gamma|}}{\partial a^\gamma} L(\theta, a^*)(a-a^*)^\gamma \\ &= \sum_{|\gamma|=2} \frac{1}{\gamma!} \frac{\partial^{|\gamma|}}{\partial a^\gamma} L(\theta, a^*)(a-a^*)^\gamma. \end{aligned}$$

It follows that the risk of a locally shifted version of the proposed regular rule, $\hat{\delta} + \tilde{\phi}_n b$, can be approximated as follows,

$$\tilde{\phi}_n^{-2} R(\theta_0 + \phi_n h, \hat{\delta} + \tilde{\phi}_n b) \approx \sum_{|\gamma|=2} \frac{1}{2} \frac{\partial^2}{\partial a^\gamma} L(\theta_0, a^*(\theta_0)) E_{\theta_0+\phi_n h}\left[((D+b))^\gamma\right]. \tag{7}$$

This quadratic expansion parallels the familiar expansion of risk for point estimation in (5). We can then minimize this approximate risk function to obtain the asymptotically optimal shift in the rule $\hat{\delta}$.

When the decision rule is a plug-in rule based on an estimator $\hat{\theta}$,

$$\hat{\delta}(Z) = \arg\min_a L(\hat{\theta}, a) = a^*(\hat{\theta}).$$

If $a^*(\cdot)$ is continuously differentiable, and the estimator $\hat{\theta}$ satisfies

$$\phi_n^{-1}(\hat{\theta} - (\theta_0 + \phi_n h)) \overset{\theta_0+\phi_n h}{\rightsquigarrow} V,$$

then the Delta Method yields

$$\phi_n^{-1}(\hat{\delta} - a^*(\theta_0 + \phi_n h)) \overset{\theta_0+\phi_n h}{\rightsquigarrow} \left[\frac{\partial}{\partial \theta} a^*(\theta_0)\right] V,$$

So, regularity of $\hat{\theta}$ implies regularity of the plug-in rule $\hat{\delta}$.

When the partial derivative matrix $\left[\frac{\partial}{\partial\theta}a^*(\theta_0)\right]$ has full row rank, then we can connect the search for an optimally shifted estimator to the optimal shift for the general decision rule. In particular, we can show that finding the optimal shift b to $\hat{\delta}$ through the quadratic risk approximation in (7) is equivalent to a two-step approach. First, define the following loss function for parameter estimation:

$$\bar{L}(\theta,\hat{\theta}) = (\hat{\theta}-\theta)'\bar{A}(\hat{\theta}-\theta),$$

where

$$\bar{A} = \left[\frac{\partial}{\partial\theta}a^*(\theta_0)\right]\left[\frac{1}{2}\frac{\partial^2}{\partial a a'}L(\theta_0,a^*(\theta_0))\right]\left[\frac{\partial}{\partial\theta}a^*(\theta_0)\right]'.$$

Use this quadratic form estimator loss function to obtain the optimally shifted MLE,

$$\min_{s^*_{\theta_0}} E\left[(V^{ML}+\phi_n s^*_{\theta_0})'\bar{A}(V^{ML}+\phi_n s^*_{\theta_0})\right]$$

where $\phi_n^{-1}\left(\hat{\theta}_{ML}-(\theta_0+\phi_n h)\right) \overset{\theta_0+\phi_n h}{\rightsquigarrow} V^{ML}$.

Second, the optimally shifted decision rule based on the loss approximation in (7) is

$$\tilde{\delta} = \arg\min_a L(\hat{\theta}_{ML}+\phi_n s^*_{\theta_0},a) = a^*(\hat{\theta}_{ML}+\phi_n s^*_{\theta_0}) \quad \left(\approx a^*(\hat{\theta}_{ML})+\phi_n\left[\frac{\partial}{\partial\theta}a^*(\theta_0)\right]' s^*_{\theta_0}\right).$$

This estimator could be implemented by replacing $s^*_{\theta_0}$ with $s^*_{\hat{\theta}_{ML}}$, as above. In the special case where the action and parameter space are both scalar, the optimally shifted rule can be equivalently obtained as the plug-in rule based on the quadratic loss optimally shifted MLE; see the continuation of Example 2 below.

One could consider a further expansion of risk beyond (7) to obtain third-order derivative terms that would capture asymmetries in loss, as shown in previous sections. However, in this more general setting, the loss function cannot be expressed in terms of a difference between parameter θ and an action a. As a result, the higher order terms in the risk expansion will typically include terms that depend on the local parameter h. Then the idea of obtaining an optimal shift based on an approximate risk that does not depend on the local parameter is no longer available. One could consider obtaining an optimal minimax shift based on this risk approximation, but we do not pursue the properties of such a correction factor here.

Example 5 (Auction example, continued) *In this auction example, we can also consider other decision problems. For example, we may wish to use data from past auctions to design future auctions. A simple auction design problem is to select the reserve price (in the case of procurement auctions, the maximum acceptable bid). Our discussion is based on Kim (2010).*

Let $a \in \mathscr{A} = [0,\infty)$ be the reserve price. Suppose the auctioneer faces a cost of c_0 and wants to maximize profits from the auction. With m bidders, the expected profit from the auction with a reserve price of a is

$$W(\theta, a) = c_0[1 - F_\theta(a)]^m + m\int_0^a \left[c f_\theta(c) + F_\theta(c)\right](1 - F_\theta(c))^{m-1} dc. \tag{8}$$

Kim (2010, 2013, 2015) develops Bayesian methods for auction design, which are Bayes-welfare optimal by construction. Aryal and Kim (2013) considers Γ-maxmin expected utility (set of priors) auction design in a partially identified setting. Morgenstern and Roughgarden (2015) construct nearly optimal auction design by bounding revenue loss. See also Chawla, Hartline, and Nekipelov (2017) who consider adaptive auction design for revenue maximization.

Given $\hat{\theta}_{ML}$, a plug-in reserve price rule is

$$\hat{a}^* = \arg\max_a W(\hat{\theta}_{ML}, a).$$

However, we could improve this reserve price rule through a second-order approximation as in (7). With exponentially distributed costs, the parameter and reserve price action space are both scalar. As discussed above, an equivalent way of improving the ML plug-in rule, is to use a plug-in rule based on the MLE shifted optimally for quadratic loss. As noted previously, the quadratic loss optimally shifted MLE is

$$\tilde{\theta} = \hat{\theta}_{ML}\left(1 - \frac{m-1}{nm}\ln 2\right).$$

Then, the optimally shifted reserve price is equivalent to

$$\tilde{a}^* = \arg\max_a W(\tilde{\theta}, a).$$

Example 6 *Another point decision problem in econometrics is portfolio allocation. Suppose X is an m-vector of risky assets, $X \sim N(\mu, \Sigma)$. For allocation $a \in \mathbb{R}^m$ s.t. $\sum a_j = 1$, a mean-variance objective function is given by*

$$W(\theta, a) = a'\mu - c a'\Sigma a,$$

for some $c \geq 0$. The parameters of the model are $\theta = (\mu, \Sigma)$.

A common approach is to calculate a plug-in rule based on historical data Z:

$$\delta(Z) = \arg\max_a W(\hat{\theta}(Z), a),$$

where $\hat{\theta}$ may be the MLE or a shrinkage type estimator. Plug-in rules are widely used, but it is well known that they may work poorly due to estimation error (estimation risk) in $\hat{\theta}$.

There is a large literature on Bayesian portfolio choice, including Zellner and Chetty (1965), Bawa, Brown, and Klein (1979), Kandel and Stambaugh (1996), Barberis (2000), and Avramov and Zhou (2010). Mori

(2001, 2004) and Kan and Zhou (2007) work with regret loss

$$W(\theta, a^*(\theta)) - W(\theta, \delta(Z))$$

and analyze the ex ante expected welfare properties of different rules. Mori (2004) shows that MLE and shrinkage plug-in rules are inadmissible.

While the form of the welfare function is simple quadratic, an additional complication comes from the constrained action space $\sum a_j = 1$. *Suppose the optimal rule* a^* *and MLE rule* $\hat{\delta}_{ML}$ *are in the constrained space, so that* $\mathbf{1}'a = 1$ *and* $\mathbf{1}'\hat{\delta}_{ML} = 1$, *where* $\mathbf{1}$ *is a vector of ones. To stay in the constrained action space, we require that the shifts* s *of the MLE satisfy* $\mathbf{1}'s = 0$. *In this case, the risk expansion only occurs along directions that stay in the constrained action space. For these directions the first derivative of loss regret is zero (though not other directions in general), and the approximation based on second order derivatives (within the constrained action space directions) simplifies nicely. In particular, the asymptotic distribution of* $\hat{\delta}_{ML}$ *is mean zero Gaussian, so that the quadratic loss approximation yields zero shift adjustment to the MLE rule. Interestingly, Mori (2004) contains small sample approximation to the MLE rule, which could potentially be used to derive approximately optimal adjustments which could improve its performance.*

3.5 Variable Selection and Shrinkage

So far we have mainly focused on point decision rules that are translation equivariant or regular (i.e., locally asymptotically translation equivariant). Such procedures can have minmax or average risk optimality properties. However, there are often good reasons to consider estimators and point decision rules that are not equivariant. In this section we review some results for shrinkage estimators, which are used in a variety of guises in empirical work in economics, including recent applications of machine learning regression techniques.

As in Section 3.1, we first consider a simple finite-sample settting. Suppose that Z is a k-dimensional normal vector with

$$Z \sim N(\mu, \sigma^2 I_k),$$

where σ^2 is known. Equivalently, we observe independent random variables Z_j for $j = 1, \ldots, k$ with $Z_j \sim N(\mu_j, \sigma^2)$. A natural point estimator for μ is $\hat{\mu} = Z$, with components $\hat{\mu}_j = Z_j$. When the parameter space for μ is unrestricted, this estimator is the maximum likelihood estimator and the Bayes posterior mean under a flat prior; it is also best equivariant under squared error loss, minmax, and average risk optimal under a flat prior.

Consider estimators of the form

$$\tilde{\mu}_j = Z_j \cdot \Gamma_j(Z),$$

where $0 \leq \Gamma_j(z) \leq 1$. This estimator shrinks each component of $\hat{\mu} = Z$ towards the origin, by a factor that

could depend on the data. (We could also consider estimators that shrink towards some other point in $\mathbb{R}^k$ with minor notational modifications.) For example, if

$$\tilde{\mu}_j = Z_j \mathbf{1}(|Z_j| > c),$$

for some c, the estimator $\tilde{\mu}_j = 0$ when the initial estimate Z_j is sufficiently close to 0. Pre-test estimators (also called hard thresholding estimators) that zero out components for which a standard test does not reject the hypothesis that the component is zero, are of this form. Alternatively, the shrinkage towards zero could be more smooth. For example, the positive-part James-Stein estimator for $k \geq 3$ has

$$\tilde{\mu}_j = Z_j \cdot \left(1 - \frac{k-2}{\|Z\|^2}\right)^+,$$

and shrinks $\hat{\mu}_j$ towards zero by a factor that depends on the sum of squares of the elements of Z. (See James and Stein (1961) and Baranchik (1964).) Other point estimators, such as the Bayesian posterior mean under some choices of the prior, ridge regression (Hoerl and Kennard, 1970), and parametric empirical Bayes methods (Morris, 1983), are also of the smooth shrinkage form. Lasso-type estimators (Tibshirani, 1996) combine variable selection with smooth shrinkage. Bagging (Breiman, 1996) converts hard thresholding estimators into a smooth estimator, by averaging the thresholding estimators over bootstrap replications.

For point estimators with action space $\mathscr{A} = \mathbb{R}^k$, consider the sum-of-squares loss function

$$L(\mu, a) = \sum_{j=1}^{k} (a_j - \mu_j)^2.$$

The estimator $\hat{\mu} = Z$ has constant risk $R(\mu, \hat{\mu}) = k\sigma^2$, and is minmax. In contrast, the risk functions of shrinkage estimators typically vary with μ. The risk of the positive-part James-Stein estimator, and a number of other shrinkage estimators, can be shown to satisfy

$$R(\mu, \tilde{\mu}) \leq k\sigma^2 \quad \forall \mu,$$

with strict inequality for some values of μ, when $k \geq 3$.[5] In other words, $\tilde{\mu}$ dominates $\hat{\mu}$, and $\hat{\mu}$ is inadmissible.

Thus it is possible, in principle, to improve upon the conventional maximum likelihood estimator in this setting by using some type of shrinkage estimator. The potential gains are large when μ is high-dimensional and there are natural restrictions on the parameter space. For this reason, shrinkage methods may be useful in nonparametric and high-dimensional regression methods, forecasting methods using high dimensional vector autoregressions as in Doan, Litterman, and Sims (1984), and problems involving many units with limited data on each unit as in Graham and Hirano (2011). The shrinkage

[5] For results on the risk of James-Stein and related estimators, see James and Stein (1961), Baranchik (1964), Egerton and Laycock (1982), Robert (1988), among others.

may be towards specific points in the parameter space suggested by economic theory, as in Ingram and Whiteman (1994), Del Negro and Schorfheide (2004), and Fessler and Kasy (2018), among others.

On the other hand, it is more difficult to obtain sharp optimality results that hold in wide generality, particularly if the assumption of homoskedasticity is dropped.[6] In general different shrinkage estimators may have risk functions that cross, because they trade off performance across the parameter space differently. It can also be difficult to calculate the exact risk function for some procedures based on variable selection or shrinkage, at least under a realistically wide range of data generating processes.

In the remainder of this section we will focus on obtaining useful large-sample approximations to the risk functions of shrinkage-type procedures. We seek approximations that are sufficiently tractable to allow comparisons between procedures, and in some cases suggest ways to improve upon them. Local asymptotic approximations of the type we consider here have been used by a number of authors to explore model selection and shrinkage estimators; see for example Knight and Fu (2000), Bühlmann and Yu (2002), Inoue and Kilian (2006), Claeskens and Hjort (2008), and Hansen (2016). Some of the following results are based on Hirano and Wright (2017).

To fix ideas, consider the classical regression model, where we have i.i.d. draws from a distribution for (y_i, x_i), with

$$y_i = \beta' x_i + u_i, \quad E[u_i|x_i] = 0, V[u_i|x_i] = \sigma^2,$$

where x_i is $k \times 1$. Assume $E[x_i x_i'] = I_k$ for simplicity. (Many of these restrictions can be relaxed.) Let $\hat{\beta}$ be the usual LS estimator. In this setting, the positive-part James-Stein estimator is

$$\tilde{\beta} = \hat{\beta} \times \max\left\{1 - \frac{k-2}{n\hat{\beta}'\hat{V}^{-1}\hat{\beta}}, 0\right\},$$

where $\hat{V}$ is an estimate of $V(\hat{\beta})$.

Post-model selection estimators can be viewed as shrinkage estimators, which set some elements of β to zero based on the data. One approach popular in forecasting applications is to use a pseudo out-of-sample criterion to select elements of β to set to zero. One version of the out-of-sample (OOS) model selection approach is the following. For each model m which picks a subset of the k regressors, estimate the model recursively starting a fraction $\pi \in (0,1)$ of the way through the sample. At each recursion calculate a one step ahead point forecast, and compare the forecast to the realized value of y. Calculate the empirical mean squared error of the one step ahead forecasts for observations $y_{[\pi n]}, \ldots, y_n$. The model m with the lowest empirical mean squared error is chosen, and re-estimated using the entire sample.

The pseudo out of sample approach is intuitive and popular in practice. But it is somewhat difficult to analyze because of its recursive structure. To analyze it and other procedures that use different types of sample splitting to select the model, such as v-fold cross-validation, it is helpful to use local asymptotics

[6] Recent work on shrinkage estimation in the heteroskedastic case includes Xie, Kou, and Brown (2012).

and a representation result tailored to the problem.

We assume the standard LAN conditions and adopt the following local parametrization:

$$\beta_n = \underline{0} + \frac{h}{\sqrt{n}}.$$

Here we have centered the local sequence at $\beta_0 = 0$ to capture the situation where it is difficult to tell which elements of x_i are nonzero. Then it can be shown that the OLS estimator (using all k regressors) has the following limit under under $\beta_n = h/\sqrt{n}$:

$$\sqrt{n}\hat{\beta} \rightsquigarrow Y, \quad \text{where } Y \sim N(h, \Omega).$$

Thus the OLS estimator acts like an observation in the limiting $N(h,\Omega)$ experiment, and it is regular.

The James-Stein estimator has the following limit:

$$\sqrt{n}\tilde{\beta} \rightsquigarrow Y \times \max\left\{1 - \frac{k-2}{Y'Y}, 0\right\}.$$

So $\tilde{\beta}$ acts like a James-Stein estimator in the limit experiment. It is not regular because its distribution depends on h, but its distributions under h (and associated risks) can be easily computed.

The estimator based on pseudo out-of-sample model selection is less simple to approximate, because it cannot be written as a function solely of the unrestricted OLS estimator $\hat{\beta}$. However, it (and many other post-model selection and shrinkage estimators) can be viewed as functions of the partial sums

$$\sum_{i=1}^{s} x_i x_i', \quad \sum_{i=1}^{s} x_i y_i, \quad s = 1, \ldots, T.$$

If we can obtain useful approximations to the partial sums under the local sequences $\beta_n = h/\sqrt{n}$, we can approximate the distributions of a wide range of estimators. To this end, consider the partial sums of $x_i y_i$. For $r \in [0,1]$, let $[nr]$ be the largest integer less than nr. Then, under standard regularity conditions, a functional central limit theorem gives

$$n^{-1/2} \sum_{i=1}^{[nr]} x_i y_i \rightsquigarrow rb + \sigma B(r),$$

where $B(r)$ is a standard k-dimensional Brownian motion. By a standard construction of the Brownian bridge, we can re-express the limit on the right as

$$rb + \sigma B(r) \sim rY + \sigma U(r),$$

where $Y \sim N(h,\Omega)$ and $U(r)$ is a Brownian bridge, independent of Y and h. This highlights the connection between the limit of the partial sum and the shifted normal limit experiment. Procedures such

as pseudo out-of-sample model selection can be viewed as randomized estimators in the experiment of observing $Y \sim N(h,\Omega)$, depending on both Y and an additional random component $U(\cdot)$. Hirano and Wright (2017) use this result to numerically calculate and compare asymptotic risk functions for a variety of post-model selection and shrinkage estimators.

4 Treatment Assignment Rules

A natural application of statistical decision theory is to choose among different treatments or programs to assign to an individual. Recent work in economics, medical statistics, and other fields has considered treatment assignment, building on methods for causal inference and program evaluation. The classic example is assignment to a medical treatment based on patient characteristics. In economics, examples have included assignment to a job training program (Black, Smith, Berger, and Noel, 2003; Frölich, 2008), enrollment in government welfare programs (Dehejia, 2005), and sentencing by judges (Bushway and Smith, 2007).

4.1 Treatment Assignment as a Decision Problem

Suppose that the individuals to be assigned are drawn from some target population, and have observable background characteristics X on a space $\mathscr{X}$. An individual can be assigned one of two treatments, labeled 0 and 1, based on their value of X. A treatment assignment rule is a mapping from $\mathscr{X}$ to $\{0,1\}$. We could consider all possible such rules, or some constrained set of rules. In the notation of Section 2, the set of possible treatment rules is our set of actions: $\mathscr{A} \subset 2^{\mathscr{X}}$, where $2^{\mathscr{X}}$ is the power set of all binary-valued functions with domain $\mathscr{X}$.[7]

We wish to select a treatment rule based on data $Z \sim P_\theta$. A *statistical treatment assignment rule* maps data into the choice of rule: $\delta : \mathscr{Z} \to \mathscr{A}$. Since the object chosen by δ is itself a function of X, we simplify the notation by writing

$$\delta(x;z),$$

with the interpretation that, given a data set $Z = z$, the rule $\delta(\cdot;z)$ maps covariate X to the treatment space $\{0,1\}$. For notational simplicity we also use $\delta(x)$ to denote particular treatment rules in the action space; so $\delta(x) \in \mathscr{A}$ while $\delta(\cdot;z)$ is a mapping from $\mathscr{Z}$ to $\mathscr{A}$.

This framework is fairly general and encompasses a number of cases considered in the recent literature, though it could be extended to handle other settings such as those with a richer set of treatments. (We discuss some possibilities below.) An important special case occurs when the data Z represent a randomized controlled trial conducted on the same population as the target population for future treat-

[7] We could also allow the range of the rule to be $[0,1]$, with the interpretation that $a(x) \in [0,1]$ gives the probability of assigning an individual with $X = x$ to treatment 1.

ment assignment. In general the data $Z \sim P_\theta$ do not have to come from the same population as the target population for future treatment, nor do they have to come from a randomized experiment, provided that the data are informative about a parameter θ that in turn encompasses all welfare-relevant information about the target population.

To define the welfare and associated criteria for evaluating statistical treatment rules, it is convenient to introduce potential outcomes. In the binary treatment case, let $Y(0)$ and $Y(1)$ be potential outcomes under treatment 0 and 1 respectively. Given treatment $T = 0, 1$, the realized outcome is $Y = TY(1) + (1 - T)Y(0)$. The performance of a given decision rule can then be evaluated based on the distribution of outcomes induced by the treatment assignment. Let the distributions of $Y(0)$ and $Y(1)$ conditional on characteristics be denoted by $F_0(\cdot|x,\theta)$ and $F_1(\cdot|x,\theta)$, where θ is the parameter indexing the distributions of Z. Let $F_X(x)$ denote the marginal distribution of X in the target population.

Suppose the welfare associated with a rule $\delta(x)$ is given by a functional of the potential outcome distributions:

$$W(\theta, \delta(\cdot)) = \int w(F_0(\cdot|x,\theta), F_1(\cdot|x,\theta), \delta(x)) dF_X(x).$$

Then an ideal, but generally infeasible, assignment rule is the following:

$$\delta^*(x) = \begin{cases} 1 & \text{if } w(F_0(\cdot|x,\theta_0), F_1(\cdot|x,\theta_0), 1) \geq w(F_0(\cdot|x,\theta_0), F_1(\cdot|x,\theta_0), 0) \\ 0 & \text{if } w(F_0(\cdot|x,\theta_0), F_1(\cdot|x,\theta_0), 1) < w(F_0(\cdot|x,\theta_0), F_1(\cdot|x,\theta_0), 0). \end{cases} \tag{9}$$

(Here, we assign treatment 1 if the two treatments give equal welfare given X.) A special case is the utilitarian welfare criterion, which defines welfare as the expected utility associated with the outcomes or the expected outcomes themselves. The utilitarian welfare of a rule δ can be expressed as

$$W(\theta, \delta) = \int \int u(y) dF_\delta(y|x,\theta) dF_X(x),$$

where u is outcome utility and $F_\delta(y|x,\theta) = \delta(x)F_1(y|x,\theta) + (1 - \delta(x))F_0(y|x,\theta)$. In this case, if we rescale the outcome into utils so that $u(y) = y$, then the welfare becomes

$$W(\theta, \delta) = \int \delta(x) E[Y(1)|X = x] + (1 - \delta(x)) E[Y(0)|X = x] \, dF_X(x). \tag{10}$$

An infeasible optimal rule chooses, for each x, the treatment that maximizes the expected outcome:

$$\delta^*(x) = \mathbf{1}\{E[Y(1)|X = x] \geq E[Y(0)|X = x]\}. \tag{11}$$

Example 7 *Suppose that the data Z come from a randomized experiment on the same population as the target population for treatment assignment. Thus $Z = \{Y_i, T_i, X_i; i = 1, \ldots, n\}$, where $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$, the potential outcomes $Y_i(0)$ and $Y_i(1)$ are distributed $F_0(\cdot|x)$ and $F_1(\cdot|x)$ conditional on $X_i = x$, and T_i is independent of the potential outcomes (unconditionally, or conditonal on X_i). In this case we identify θ with the entire joint distribution of Z_i.*

If X_i is discrete, and welfare is utilitarian with Y measured in utils, the sample analog to Equation (11) is

$$\hat{\delta}(x;Z) = \mathbf{1}\left\{\frac{\sum_i T_i 1(X_i = x) Y_i}{\sum_i T_i 1(X_i = x)} \geq \frac{\sum_i (1-T_i) 1(X_i = x) Y_i}{\sum_i (1-T_i) 1(X_i = x)}\right\}.$$

Manski (2004) called this the conditional empirical success rule and studied its properties.

Regardless of how we specify the welfare function $W(\theta,\delta)$, we still need to decide on a specific criterion for evaluating statistical treatment rules. As in Section 2.4, we can take the expectation of welfare with respect to the sampling distribution of $\delta(x;\cdot)$,

$$E_\theta\left[W(\theta,\delta)\right]$$

and use either a Bayes or maxmin criterion. In a parametric setting, for example, Dehejia (2005) considers a Bayesian approach to treatment assignment to a labor market program; see also Chamberlain (2011), and for dynamic treatment rules, Arjas and Saarela (2010) and Zajonc (2012). However, when the underlying parameter space and the set of possible rules is large, Bayes and maxmin approaches may be difficult to implement or sensitive to certain aspects of the problem setup. The literature has therefore considered other criteria, especially minmax regret, for which bounds and approximations are relatively tractable. For minmax regret, define the welfare regret loss as

$$L^R(\theta,\delta) = W^*(\theta) - W(\theta,\delta),$$

where $W^*(\theta) = \arg\max_\delta W(\theta,\delta)$ is the welfare obtained by the infeasible optimal rule. The minmax regret criterion is

$$\sup_{\theta\in\Theta} E_\theta\left[L^R(\theta,\delta)\right].$$

This evaluates δ by its worst-case performance relative to the unknown optimal rule. In economics, the minmax regret approach to treatment assignment was pioneered by Manski (2004). Other choices for a loss function based on $W(\theta,\delta)$ are also possible. For example we could attach a fixed, possibly asymmetric penalty to each type of incorrect assignment, by using a loss function of the form

$$L(\theta,\delta) = \begin{cases} (1-\delta) & \text{if } W(\theta,1) > W(\theta,0) \\ \delta K & \text{otherwise,} \end{cases}$$

where $K > 0$. Then we could evaluate rules by their minmax expected loss. This type of criterion, which can lead to rules similar to classical hypothesis tests, was proposed and studied by Tetenov (2012), and also studied by Hirano and Porter (2009).

The following simple example illustrates minmax regret analysis in a simple parametric case.

Example 8 *Suppose the target population has no observable characteristics so that we can suppress the notational dependence on x. The potential outcome distributions are given by $Y(0) \sim N(\mu_0, \sigma_0^2)$ and $Y(1) \sim$*

$N(\mu_1, \sigma_1^2)$. *Suppose* σ_0^2 *and* σ_1^2 *are known so that unknown parameters are* $\theta = (\mu_1, \mu_2)$. *Suppose welfare is utilitarian with* Y *measured in utils:* $W(\theta, \delta) = \delta\mu_1 + (1-\delta)\mu_0$. *Then regret loss is*

$$L^R(\theta, \delta) = \begin{cases} (1-\delta)(\mu_1 - \mu_0) & \text{if } \mu_1 > \mu_0 \\ \delta(\mu_0 - \mu_1) & \text{otherwise} \end{cases}$$

Asymmetric "hypothesis testing" loss is

$$L^H(\theta, \delta) = \begin{cases} (1-\delta) & \text{if } \mu_1 > \mu_0 \\ \delta K & \text{otherwise} \end{cases}$$

where $K > 0$. *Let* $\Delta = \mu_1 - \mu_0$ *be the average treatment effect. We can write the corresponding risks of a rule* $\delta(Z)$ *as:*

$$E_\theta[L^R(\theta, \delta)] = E_\theta[\delta(Z)]\Delta\mathbf{1}\{\Delta \leq 0\} + (1 - E_\theta[\delta(Z)])\Delta\mathbf{1}\{\Delta > 0\}$$

and

$$E_\theta[L^H(\theta, \delta)] = E_\theta[\delta(Z)]K\mathbf{1}\{\Delta \leq 0\} + (1 - E_\theta[\delta(Z)])\mathbf{1}\{\Delta > 0\}.$$

Now, suppose we have an estimator $\hat{\Delta} = \hat{\Delta}(Z)$ *of the average treatment effect based on the data* Z*, that satisfies* $\hat{\Delta} \sim N(\Delta, \sigma^2)$. *For simplicity consider threshold rules of the form* $\delta_c = \mathbf{1}(\hat{\Delta} \geq c)$. *It is straightforward to evaluate the maximum expected loss of such rules, and to find the minmax value of* c *(see Hirano and Porter (2009)). In the case of welfare regret loss* L^R*, the optimal value can be shown to be* $c^* = 0$*, leading to the rule* $\delta_R(Z) = \mathbf{1}(\hat{\Delta} \geq 0)$. *For* L^H*, the optimal value can be shown to be*

$$c^* = \sigma\Phi^{-1}\left(\frac{K}{K+1}\right),$$

where $\Phi(\cdot)$ *is the standard normal CDF.*

4.2 Welfare and Risk Analysis of Treatment Assignment Rules

Treatment assignment problems are naturally related to estimation of conditional average treatment effects. In practice, researchers often wish to avoid restrictive parametric assumptions when evaluating treatments. In our setup, this would lead to θ being high-dimensional. In such cases it can be challenging to evaluate decision rules and find optimal rules. Here we discuss some methods for analyzing decision rules. We do not consider the most general possible settings but aim to illustrate some of the techniques that can be used under relatively mild distributional assumptions.

We return to the setup of Example 7. The observed data $Z = \{Y_i, T_i, X_i; i = 1, \ldots, n\}$ represent a (conditionally) randomized experiment comparing two treatments on the target population, with associated potential outcomes $Y_i(0)$ and $Y_i(1)$. Suppose $X_i \sim F_X$, $Y_i(0)|X_i = x \sim F_0(\cdot|x)$, and $Y_i(1)|X_i = x \sim F_1(\cdot|x)$, and suppose the binary treatment T_i is Bernoulli, conditionally independent of $Y_i(0)$ and $Y_i(1)$ given X_i,

with

$$e(x) := \Pr(T_i = 1 | X_i = x).$$

Due to the conditional independence of the treatment,

$$Y_i | T_i = t, X_i = x \sim F_t(\cdot | x),$$

and we can identify $\theta = (F_X, e, F_0, F_1)$ with the joint distribution P of (Y_i, T_i, X_i). We will therefore treat P as the parameter in the sequel.

Under utilitarian welfare (with Y normalized to utils), the welfare $W(\theta, \delta)$ of a treatment rule δ is given in (10). Under our assumptions we can also write welfare in terms of the observed-data distribution P:

$$\begin{aligned} W(P, \delta) &= E_P\left[\delta(X) E_P[Y | T = 1, X = X] + (1 - \delta(X)) E_P[Y | T = 0, X]\right], \\ &= E_P\left[\delta(X) \frac{YT}{e(X)} + (1 - \delta(X)) \frac{Y(1-T)}{1 - e(X)}\right], \end{aligned} \tag{12}$$

where the P subscripts in (12) highlight that the expectations are taken with respect to the joint distribution of (Y_i, T_i, X_i) in the experimental data set.

The plug-in rule replaces P with an estimate $\hat{P}_n(Z)$, and solves:

$$\hat{\delta}_n(\cdot; Z) = \arg\max_{\delta(\cdot) \in \mathscr{A}} W(\hat{P}_n(Z), \delta(\cdot)).$$

Suppose that the support $\mathscr{X}$ of the covariate X is finite, the action space is unrestricted so that $\mathscr{A} = 2^{\mathscr{X}}$, and we use the empirical distribution estimator $\hat{P}_n = \mathbb{P}_n$. Then

$$W(\mathbb{P}_n, \delta) = \frac{1}{n} \sum_{i=1}^{n} \left[\delta(X_i) \hat{\mu}_1(X_i) + (1 - \delta(X_i)) \hat{\mu}_0(X_i)\right],$$

where $\hat{\mu}_0(x)$ is the conditional average of Y_i given $T_i = 0$ and $X_i = x$ and $\hat{\mu}_1(x)$ is the conditional average of Y_i given $T_i = 1$ and $X_i = x$. By straightforward calculations, if every (x, t) cell is nonempty, then we can also write the plug-in welfare as

$$W(\mathbb{P}_n, \delta) = \frac{1}{n} \sum_{i=1}^{n} \left[\delta(X_i) \frac{Y_i T_i}{\hat{e}(X_i)} + (1 - \delta(X_i)) \frac{Y_i (1 - T_i)}{1 - \hat{e}(X_i)}, \right],$$

where $\hat{e}(x)$ is the conditional empirical probability of $T = 1$ given $X = x$. Then the plug-in approach yields Manski's conditional empirical success rule as in Example 7.

Manski (2004) proposed to evaluate statistical treatment rules by their minmax regret risk. Under the further assumption that the outcome Y is bounded, he developed finite-sample bounds on the worst-case regret risk of the conditional empirical success rule when P is otherwise unrestricted, using Hoeffding's Large Deviations Theorem. With the same setup, Stoye (2009) used game-theoretic techniques to solve

for a minmax regret statistical treatment rule. Stoye's minmax regret rule differs slightly from the conditional empirical success rule. In the case where $\mathscr{X}$ is singleton (so we can omit X from the notation), and T is randomized with $\Pr(T=1)=\frac{1}{2}$, Stoye's rule is

$$\delta^*(Z)=\begin{cases} 1 & \text{if } n_1\left(\hat{\mu}_1-1/2\right)-n_0\left(\hat{\mu}_0-1/2\right)>0 \\ 1/2 & \text{if } n_1\left(\hat{\mu}_1-1/2\right)-n_0\left(\hat{\mu}_0-1/2\right)=0 \\ 0 & \text{if } n_1\left(\hat{\mu}_1-1/2\right)-n_0\left(\hat{\mu}_0-1/2\right)<0 \end{cases}$$

where n_t denotes the number of individuals with $T_i=t$, and $\hat{\mu}_t$ denotes the sample average of Y_i in that group.

The conditional empirical success rule and Stoye's minmax regret rules condition fully on X. When $\mathscr{X}$ is a relatively small set, these rules are natural and may work reasonably well regardless of the underlying parameter P. However, if the support of X is large (as would be the case when some elements of X are continuous, or when X is discrete but of relatively high dimension), the full action space $\mathscr{A}=2^{\mathscr{X}}$ is very large, and it may be difficult to learn all of the conditional average treatment effects from the data Z. Yet, the results in Stoye (2009) indicate that a minmax regret rule fully conditions on X regardless of the size of $\mathscr{X}$. In the extreme, if there are no observations in either treatment arm for a given value of X, this can lead to "no-data" rules that employ no smoothing. Intuitively, rules that attempt to smooth across x can have poor worst-case risk if $F_0(\cdot|x)$ or $F_1(\cdot|x)$ vary strongly with x. This suggests that the minmax regret criterion may lead to very conservative rules that guard against extreme, but perhaps unrealistic, cases.

One way to deal with a complex feature space $\mathscr{X}$ is to restrict the set of possible rules $\mathscr{A}$. Kitagawa and Tetenov (2018) argue that in many applications it may be reasonable to restrict the set of treatment rules under consideration. For example, there may be external constraints that prohibit the use of certain variables in X for treatment assignment, or a restriction to rules based on a linear index $\beta' X$ for ease of implementation. One version of Kitagawa and Tetenov's approach constructs the treatment assignment rule as

$$\hat{\delta}_n(x;Z)=\arg\max_{\delta\in\mathscr{A}}\hat{W}(\delta),$$

where

$$\hat{W}(\delta)=\frac{1}{n}\sum_{i=1}^{n}\left[\delta(X_i)\frac{Y_iT_i}{e(X_i)}+(1-\delta(X_i))\frac{Y_i(1-T_i)}{1-e(X_i)}\right]$$

and $\mathscr{A}$ is the constrained set of rules. In the case of a randomized experiment, $e(x)=\Pr(T_i=1|X_i=x)$ is known, and the Kitagawa-Tetenov approach uses these known randomization probabilities in the expression above to avoid having to estimate the propensity score nonparametrically (unlike Manski's conditional empirical success rule, as described above). We can view $\hat{W}(\delta)$ as an estimate of $W(P,\delta)$.

If the constrained set of rules $\mathscr{A}$ is sufficiently small so that it has a finite Vapnik-Chervonenkis (VC) dimension, and some additional conditions (such as boundedness of Y) hold, then one can obtain non-

trivial finite-sample probabilistic bounds on

$$\sup_P \sup_{\delta \in \mathscr{A}} |\hat{W}(\delta) - W(P,\delta)|$$

using concentration inequalities. Kitagawa and Tetenov use this approach to bound the minmax regret welfare of $\hat{\delta}$ and from this show that the regret welfare of their rule converges at an optimal rate.

A drawback of the Kitagawa-Tetenov rule above is that it is based on an inefficient estimate of $W(P,\delta)$. For any fixed δ, the welfare $W(P,\delta)$ in Equation (12) is a weighted average of conditional means of potential outcomes. The estimator $\hat{W}(\delta)$ is consistent for $W(P,\delta)$ and asymptotically normal:

$$\sqrt{n}\big(\hat{W}(\delta) - W(P,\delta)\big) \xrightarrow{d} N(0,V) \quad \text{under } P.$$

The rate of convergence of this estimator is optimal, but the asymptotic variance V is not. Hahn (1998) and others have derived the semiparametric efficiency bound for the asymptotic variance of estimators of quantities like $W(P,\delta)$, and in general the optimal variance is lower than V. This suggests that there may be scope to improve on the treatment rule based on $\hat{W}(\delta)$.

Athey and Wager (2017) derive alternative risk bounds for the welfare of the statistical treatment rules in this setting. Their bounds involve V^*, the semiparametric efficiency bound for estimation of $W(P,\delta^*)$, where $\delta^* = \arg\max_{\delta \in \mathbb{A}} W(P,\delta)$. To achieve their bounds, they suggest the use of alternative estimators of $W(P,\delta)$.

Consider the following alternative estimator, based on the reasoning in Hirano, Imbens, and Ridder (2003):

$$\tilde{W}(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta(X_i)\frac{Y_i T_i}{\hat{e}(X_i)} + (1-\delta(X_i))\frac{Y_i(1-T_i)}{1-\hat{e}(X_i)}\right],$$

where $\hat{e}(\cdot)$ is a suitable nonparametric estimator of the propensity score. In the case where X_i has finite support and $\hat{e}(x)$ is the conditional empirical frequency of $T_i = 1$ given $X_i = x$, we have

$$\tilde{W}(\delta) = W(\mathbb{P}_n, \delta)$$

and maximizing this with respect to δ leads to Manski's conditional empirical success rules as discussed above.

If X_i has continuous or high dimensional support, then additional care must be taken to construct the estimator of $W(P,\delta)$, in order to ensure that the estimator is semiparametrically efficient *uniformly* in $\delta \in \mathscr{A}$. Doubly robust estimators, such as those developed by Chernuzhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2016) may be used. See Athey and Wager (2017) for further discussion of this approach. Other related work includes Mbakop and Tabord-Meehan (2016) and Kitagawa and Tetenov (2017).

4.3 Local Asymptotics for Treatment Rules

As we noted above, when the action space $\mathscr{A}$ is sufficiently restricted, local asymptotic efficiency considerations in the form of semiparametric efficiency bounds became useful for analyzing the minmax regret risk of procedures. Another approach to dealing with a complex feature space is to restrict the set of possible distributions; for example we could restrict P_θ to lie in some regular parametric or semiparametric class. This would also allow us to utilize large-sample distribution theory to simplify the analysis of decision rules. Here we want to explore the possibility for more refined large-sample analysis of treatment assignment rules under statistical regularity conditions. We draw on Hirano and Porter (2009) for the results below.

First consider the case where θ is a finite-dimensional parameter. Suppose the data are i.i.d. with $Z_i \sim P_\theta$ and $Z^n = (Z_1, \ldots, Z_n)$. The decision rule is allowed to depend on these observations, so we write it as $\delta_n(x; Z^n)$. If Assumption 1 holds, then the statistical model for Z^n is LAN. That is, in large samples, the statistical model for the data can be approximated by a simple Gaussian statistical model. A consequence of the LAN property is that any sequence of rules δ_n can also be approximated by a rule in the limiting Gaussian experiment. In particular, suppose $\theta_0 \in \Theta$ and assume Θ is an open subset of $\mathbb{R}^k$. Consider a sequence of statistical models $P_{\theta_{n,h}}$ where $\theta_{n,h} = \theta_0 + h/\sqrt{n}$. If $E_{\theta_{h,n}}[\delta_n(x; Z^n)]$ has a well defined limit for every h and x, then there exists a rule $\delta : \mathscr{X} \times \mathbb{R}^k \to [0,1]$ such that

$$\lim_{n\to\infty} E_{\theta_{h,n}}\left[\delta_n(x; Z^n)\right] = \int \delta(x;\xi)\, dN(\xi|h, J_0^{-1}) \tag{13}$$

for all $h \in \mathbb{R}^k$ and each $x \in \mathscr{X}$, where $N(\xi|h, J_0^{-1})$ is the Gaussian distribution with mean h and variance J_0^{-1} with J_0 defined in Assumption 1.

This asymptotic representation result states that every converging sequence of rules is matched by some rule in the Gaussian shift statistical model. One can therefore analyze the decision problem in this relatively simple limiting model to characterize the set of attainable expected welfare and risk functions under various criteria. For example, suppose that there is no covariate, so we can drop X_i from the notation, and suppose that the model is parametrized so that the difference in welfare between the two treatments is

$$g(\theta) = W(\theta, 1) - W(\theta, 0).$$

Since we will be taking limits as the sample size increases, we choose θ_0 such that $g(\theta_0) = 0$. Then the local parameter sequences $\theta_0 + h/\sqrt{n}$ for $h \in \mathbb{R}^k$ correspond to cases where the "better" treatment cannot be learned perfectly, even as the sample size increases.

Welfare regret loss is

$$L^R(\theta, \delta) = g(\theta)\left[\mathbf{1}(g(\theta) > 0) - \delta\right],$$

Under the local parameter sequence, this will converge to zero, and if the function g is smooth, a scaled

version of the loss has the following limiting form:

$$\sqrt{n}L^R(\theta_0 + h/\sqrt{n}, \delta) \to L^R_\infty(h,\delta) := \dot{g}'h\left[\mathbf{1}(\dot{g}'h - \delta)\right], \tag{14}$$

where $\dot{g}$ is the gradient of g at θ_0.

The asymptotic representation of decision rules in (13) and the limiting representation of the loss function in (14) yield a local asymptotic version of the statistical decision problem. In the limiting problem, one observes a single draw from the shifted Gaussian model:

$$Z \sim N(h, J_0^{-1}).$$

A decision rule δ maps the observation Z into $[0,1]$, with the interpretation that future individuals will be assigned treatment 1 with probability δ. The loss of rule δ is $L^R_\infty(h,\delta)$, which is based on a known linear function $\dot{g}'h$ of the parameter h. This is a mild extension of the decision problem considered in Example 8. It can be shown (see Hirano and Porter (2009)) that the optimal rule is the simple cutoff rule $\delta^*(Z) = \mathbf{1}\left(\dot{g}'Z > 0\right)$. Let

$$R^* = \sup_{h\in\mathbb{R}^k} E_h\left[L^R_\infty(h,\delta^*)\right]$$

be the minmax regret risk of the optimal rule in the limiting problem. In the original problem, any converging sequence of decision rules δ_n can be represented by some rule in the limiting problem. As a result, for any finite set H of possible values of $h \in \mathbb{R}^k$, it follows that

$$\liminf_{n\to\infty} \sup_{h\in H} \sqrt{n} E_{\theta_0 + h/\sqrt{n}}\left[L^R(\theta_0 + h/\sqrt{n}, \delta_n(Z^n))\right] \geq R^*.$$

Furthermore, under suitable conditions, the feasible rule

$$\delta_n(Z^n) = \mathbf{1}\left(g(\hat{\theta}) > 0\right)$$

will achieve the lower bound on expected welfare regret, provided that $\hat{\theta}$ is an asymptotically efficient estimator of θ in the sense that $\sqrt{n}\left(\hat{\theta} - \theta_0 - h/\sqrt{n}\right) \overset{\theta_0 + h/\sqrt{n}}{\rightsquigarrow} N(0, J_0^{-1})$ for every h.

This argument can be extended to semiparametric settings where the distribution $P \in \mathscr{M}$ of the data Z is not constrained to lie in a finite-dimensional parametric family, but the relevant welfare contrast is a smooth functional of P. In analogy with the $g(\theta)$ notation used above, let $g(P)$ denote the difference in welfare between the two treatments. For example, under utilitarian welfare the welfare contrast is the average treatment effect. If the average treatment effect is point-identified from the observed data distribution P, then it can be written as a functional $g(P)$.

We can analyze rules in a similar way to the parametric case. We pick a centering value P_0 of the unknown distribution such that $g(P_0) = 0$. We then need to define local sequences of distributions around

P_0. The *tangent space* contains one-dimensional submodels of $\mathscr{M}$ that pass through P_0.[8] We can view each submodel as a "direction" h away from P_0, and consider sequences of measures $P_{n,h}$ that satisfy a quadratic mean differentiability condition, as in Assumption 1(a), where h is the score function. If the tangent space is a separable linear space, then h can be expressed as an element of ℓ^∞.

Following van der Vaart (1991a), an asymptotic representation result similar to (13) holds in this semiparametric setting and allows us to examine the large sample properties of a sequence of treatment rules through the corresponding problem in the limit experiment. Suppose that a sequence of decision rules δ_n has limits under every local sequence of measures $P_{n,h}$, in the sense that

$$\lim_{n\to\infty} E_{P_{n,h}}\left[\delta_n(x;Z^n)\right] \text{ exists.}$$

Then, there exists a rule δ such that

$$\lim_{n\to\infty} E_{P_{n,h}}\left[\delta_n(x;Z^n)\right] \;=\; E_h\left[\delta(x;\Delta_1,\Delta_2,\ldots)\right],$$

where $(\Delta_1,\Delta_2,\ldots)$ is a sequence of independent random variables with $\Delta_j \overset{h}{\sim} N(h_j,1)$. We can regard the Gaussian sequence model $(\Delta_1,\Delta_2,\ldots)$ as the limit experiment for the analysis of rules δ.

For example, in the case where X_i is singleton and we evaluate rules by welfare regret loss, Hirano and Porter (2009) show that a simple cutoff rule is optimal in the limiting problem. We can then seek a rule in the original problem that matches the welfare regret minimizing rule asymptotically. One such rule based on the data Z^n is

$$\delta_n(Z^n) = \mathbf{1}\left(\hat{g}_n(Z^n) > 0\right)$$

where $\hat{g}_n$ is a semiparametrically efficient estimator of $g(P)$.

4.4 Other Treatment Assignment Problems

We have focused on the case of a binary treatment, but similar analysis is possible in other related assignment or allocation problems. For example, there may be more than two treatment arms available. Then results from the literature on estimating multi-valued treatment effects (e.g., Imbens (2000), Lechner (2001), Cattaneo (2010), Imai and van Dyk (2004), Hirano and Imbens (2004)) could provide a starting point for developing treatment assignment rules mapping some feature space $\mathscr{X}$ into the set of available treatments. The literature on estimating *dynamic treatment regimes*, e.g. Murphy (2003) and Robins (2004), considers rules that choose a sequence of treatments dynamically based on past outcomes of the individual. A related decision problem is to target coupons or other marketing interventions to individual characteristics and past purchase history; see for example Rossi, McCulloch, and Allenby (1996) and Dubé, Fang, Fong, and Luo (2017). In adaptive treatment assignment problems, which can be viewed

[8] See, for example, Bickel, Klaasen, Ritov, and Wellner (1993), van der Vaart (1991a), and van der Vaart (1998) for discussions of tangent spaces in semiparametric statistics.

as combining aspects of treatment assignment in the sense we have described above, and dynamic experimental design, raise additional issues. They have been extensively studied in the form of bandit problems; recent work includes Kock and Thyrsgaard (2017). When there are peer (social interaction) effects, the problem of allocating individuals to different peer groups (for example, assigning students to classrooms), and the problem of assigning treatments to individuals taking into account the potential effect on peers, raise additional conceptual and practical challenges. Recent work on allocation rules under peer effects includes Graham, Imbens, and Ridder (2014), Bhattacharya (2009), and Bhattacharya and Dupas (2012), among others.

5 Other Topics

In this section we consider other applications where viewing procedures as statistical decision rules can provide useful insights.

5.1 Nonstandard Functionals

In some economic applications, the underlying model satisfies conventional regularity conditions, but the nature of the decision problem leads to nonstandard distributional theory. Suppose interest centers on some function of the parameter $\kappa(\theta)$ where $\kappa : \Theta \to \mathbb{R}$. Under conventional smoothness conditions on the sequence of experiments $\mathscr{E}^n = \{P_\theta^n : \theta \in \Theta\}$, the MLE $\hat{\theta}_{ML}$ and other estimators such as the Bayes estimator are asymptotically efficient. However, the limit distributions of derived estimators of $\kappa(\theta)$, and more generally the feasible limit distributions of any estimators of $\kappa(\theta)$, will depend crucially on the smoothness in κ at any θ_0 that is the centering point of possible localized sequences. This may lead to nonstandard limiting distributions for estimators and other decision rules targeting κ.

In some important economic applications, the parameter of interest κ is a directionally differentiable, but not fully differentiable, functional of the distribution of the data. For example, suppose that $\theta = (\theta_1, \theta_2)$ is a two-dimensional parameter and the statistical model is $Z^n = (Z_1, \ldots, Z_n) \sim P_\theta^n$. Suppose the parametrization by θ is such that the model satisfies Assumption 1 so that local asymptotic normality holds, and let $\hat{\theta}$ be a regular estimator of θ as discussed in Section 3. However, suppose we are interested in the following function of θ:

$$\kappa(\theta) := \min\{\theta_1, \theta_2\}.$$

This type of estimand can arise in partially identified econometric models based on moment inequalities, as we discuss further in Section 5.2 below. It can also arise in other applications such as the bounds on valuation distributions, as in Haile and Tamer (2003), or inference on the best treatment as related to treatment assignment in Section 4, see Hirano and Porter (2012) for additional examples. In these examples, the parameter θ is often a reduced form parameter, and κ is a structural or latent quantity that is of primary interest.

The function $\kappa(\cdot)$ is well behaved in many respects, being continuous, homogeneous of degree one, and directionally differentiable, but it is not fully differentiable. Consider the natural analog estimator

$$\hat{\kappa} = \min\{\hat{\theta}_1, \hat{\theta}_2\}.$$

If the true value of (θ_1, θ_2) satisfies $\theta_1 \neq \theta_2$, then by application of the Delta Method, $\hat{\kappa}$ will be asymptotically normal and centered at the true value of κ. However, if θ_1 is close to θ_2, then the estimator $\hat{\kappa}$ will be downward biased and the normal distribution will not provide an accurate approximation to its sampling distribution.

We can capture this situation formally by considering a localization where the two parameters are within $O(1/\sqrt{n})$ of each other:

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \theta_0 + h_1/\sqrt{n} \\ \theta_0 + h_2/\sqrt{n} \end{pmatrix}.$$

Under this localization we will obtain a nonnormal limit distribution for $\hat{\kappa}$, but the implications are deeper. In particular, results in van der Vaart (1991b) imply that there exists *no* regular (locally asympototically translation equivariant) estimator for κ, Hirano and Porter (2012) show that there exist no estimators that are locally asymptotically unbiased or quantile-unbiased. These results can be derived by considering the estimation problem in the limiting normal shift model. Fang and Santos (2014) develop inference methods for directionally differentiable parameters, and Fang (2016) develops locally asymptotically minmax estimators within a class of plug-in estimators.

Another set of econometric models exhibiting nonstandard distributional theory are models exhibiting weak identification. Consider a classical linear instrumental variables model:

$$\begin{aligned} x_i &= \theta_1' w_i + v_{1i}, \\ y_i &= \kappa x_i + \epsilon_i, \end{aligned}$$

where y_i is the outcome of primary interest, x_i is a scalar endogenous explanatory variable, and w_i is a vector of exogenous instruments. We wish to estimate the structural parameter κ. The model can be written in reduced form as

$$\begin{aligned} x_i &= \theta_1' w_i + v_{1i}, \\ y_i &= \theta_2' w_i + v_{2i}, \end{aligned}$$

with $\theta_2 = \kappa\theta_1$. This restricts the possible values of $\theta = (\theta_1, \theta_2)$ in general. Subject to this restriction, we can view κ as a function of the reduced form parameters: $\kappa = \kappa(\theta)$.

If the instrument is weakly correlated with the endogenous regressor, then it is well known that conventional asymptotic distribution theory provides a poor approximation to the distributions of estimators of the structural parameters. Staiger and Stock (1997) proposed to consider a local parameter sequence

in which the coefficient on the instrument in the first stage is local to zero:

$$\theta_{1n} = \frac{h}{\sqrt{n}}.$$

Under this parameter sequence, one obtains non-standard limit distributions for estimators such as 2SLS and LIML that, in some cases, well approximate their finite-sample distributions. A large literature including Stock and Wright (2000), Kleibergen (2002), Moreira (2003), and Andrews, Moreira, and Stock (2006) have used this type of local asymptotic parametrization to study estimation and inference procedures for weakly identified models.

The Staiger-Stock local parametrization chooses a particular centering for θ_{1n} in order to approximate the situation where the correlation of the instrument with the endogenous regressor is close to zero. Subject to this particular choice for the centering, however, the localization is the standard one. Cattaneo, Crump, and Jansson (2012) point out that as a result, conventional LAN theory applies to the reduced form parameter θ, and use the limits of experiments framework to study optimal inference. Nonstandard behavior of the estimator of κ arises because κ depends on θ in a non-smooth fashion when θ_1 is close to zero. Hirano and Porter (2015) show that due to the non-smoothness of $\kappa(\theta)$, no locally asymptotically unbiased estimators for β exist. Andrews and Armstrong (2017) show that unbiased estimation of β is possible if the sign of the first-stage coefficient is known.

For both types of problems considered in this subsection, we can regard the problem as being regular in its reduced form, where we use "reduced form" in the classic econometric sense to mean describing the distibution of the observable data. In our notation, this means that the statistical model $\mathscr{M}$ has a parametrization in terms of some θ such that the model is locally asymptotically normal. The irregularity arises because the "structural" parameter of interest κ is related to θ in a non-smooth way.

While the impossibility results cited above suggest that standard approaches to estimation and inference may not be available in these cases, recent work has proposed creative new approaches. Kaji (2017) develops an alternative criterion for efficiency of weakly identified structural parameters based on minimal sufficiency in the reduced form. And as mentioned in section 2.4.3, Müller and Wang (2017) suggest a numerically attractive approach for proceeding with efficient estimation under constraints in the presence of non-regularity. Chen, Christensen, and Tamer (2017) consider the construction of confidence sets for structural parameters, and use level sets of the quasi-posterior of the quasi-likelihood ratio to obtain correct coverage even when the model exhibits singularities as above.

5.2 Partial Identification

Recently, a large literature has developed dealing with estimation and inference in partially identified models (Manski (1995, 2003, 2007), Tamer (2012), Molinari (2018)). In these models, we observe $Z \sim P$, but knowledge of P does not pin down the value of the parameters of interest. Formally, let $\theta \in \Theta$ be the

parameter of some underlying latent variable or structural model. The value of θ determines the distribution of the observable data: $Z \sim P = P_\theta$, but different values of the parameter may be observationally equivalent in the sense that

$$P_{\theta'} = P_{\theta''} \text{ for some } \theta', \theta'' \in \Theta \text{ with } \theta' \neq \theta''.$$

Let $\Theta(P) \subset \Theta$ be the set of values for θ consistent with $Z \sim P$:

$$\Theta(P) = \{\theta \in \Theta : P_\theta = P\}.$$

This is the set-valued inverse of the mapping $\theta \mapsto P_\theta$. If interest centers on a subparameter, say $\kappa := \kappa(\theta)$, let $K(P) = \kappa(\Theta(P))$ be the set of values of κ consistent with $Z \sim P$. We say the model is point identified if $\theta' \neq \theta'' \Rightarrow P_{\theta'} \neq P_{\theta''}$, in other words if the mapping $\theta \mapsto P_\theta$ is injective. That is, point identification implies that $\Theta(P_\theta) = \{\theta\}$ for all θ. We say that the model is partially identified if $\Theta(P)$ is not singleton, but is a strict subset of Θ for at least some values of P.

Whether or not the statistical model is point identified, the general statistical decision framework of Section 2 can still be applied. However, lack of point identification raises additional technical and conceptual issues. To illustrate some of these issues in a simple setting, consider the following simple moment inequality problem. Suppose we are interested in a scalar subparameter $\kappa \in \mathbb{R}$, and its identified set can be represented by a set of moment inequalities

$$K(P) = \{\kappa : E_P[m(Z, \kappa)] \geq 0\},$$

where $m(Z, \kappa)$ is a given (vector-valued) moment function.

The standard point estimation problem, where the action space is $\mathscr{A} = \mathbb{R}$, is still well defined. Given a point estimator $\hat{\kappa} : \mathscr{Z} \to \mathbb{R}$, we could apply one of the loss functions considered in Section 3, such as squared error loss $(\kappa - \hat{\kappa})^2$. However, because κ cannot be perfectly learned even in large samples, the risk of the point estimator will not converge to 0 in general, leading to different considerations for the asymptotic analysis. See Aryal and Kim (2013) and Song (2014) for examples of point decisions in partially identified settings.

Alternatively, we can consider "point" estimation of the set $K(P)$ by a set $\hat{K}$. Here the action space $\mathscr{A}$ is a collection of subsets of $\mathbb{R}$. In this case we need to posit a loss function compatible with the action space. For example, we could set $L(P, \hat{K}) = d(K(P), \hat{K})$, where $d(\cdot)$ is some measure of distance between the identified set $K(P)$ and the set estimate $\hat{K}$, such as the Hausdorff metric.

As a concrete example, suppose that (Y_i, V_i, W_i) are i.i.d. for $i = 1, \ldots, n$, and we are interested in $\kappa := E[Y_i]$. Both V_i and W_i are nonnegative random variables, but otherwise we put no restrictions on the joint distribution of (Y_i, V_i, W_i). We can think of the parameter θ as indexing this joint distribution, and

κ as a subparameter of θ. Suppose we only observe $Z_i = (Z_{i1}, Z_{i2})$, where

$$Z_{i1} = Y_i - V_i,$$
$$Z_{i2} = Y_i + W_i.$$

In other words, Z_{i1} and Z_{i2} are downward and upward biased measures of κ, respectively. The observed data is $Z^n = (Z_1, \ldots, Z_n)$ where Z_i are i.i.d. P. Then a natural moment function for κ is

$$m(Z_i, \kappa) = \begin{pmatrix} \kappa - Z_{i1} \\ Z_{i2} - \kappa \end{pmatrix}.$$

Let $\mu_1 = \mu_1(P) = E_P[Z_{i1}]$ and $\mu_2 = \mu_2(P) = E_P[Z_{i2}]$. Then the identified set is

$$K(P) = [\mu_1, \mu_2].$$

Suppose the action space $\mathscr{A}$ consists of closed intervals in $\mathbb{R}$, so that estimators $\hat{K}$ have the form

$$\hat{K}(Z^n) = \left[\delta_1(Z^n), \delta_2(Z^n)\right].$$

In this case a simple loss function that could be used is

$$L(P, \hat{K}) = (\delta_1 - \mu_1)^2 + (\delta_2 - \mu_2)^2.$$

This is the sum of the squared error losses of the estimates of the endpoints of the identified set. In this case, the decision problem is equivalent to a two-dimensional point estimation problem, and an analysis similar to that in Section 3 can be applied.[9]

On the other hand, suppose we have a different moment function:

$$\tilde{m}(Z_i, \kappa) = \begin{pmatrix} Z_{i1} - \kappa \\ Z_{i2} - \kappa \end{pmatrix},$$

corresponding, for example, to having two upward biased measures of κ. Then the identified set is

$$K(P) = (-\infty, \min\{\mu_1, \mu_2\}].$$

Suppose the action space consists of sets of the form $\hat{K} = (-\infty, \delta]$ for $\delta \in \mathbb{R}$. We could use the loss function $L(P, \hat{K}) = (\min\{\mu_1, \mu_2\} - \delta)^2$, which measures the closeness of the estimated upper bound on γ to the true upper bound. Thus the decision problem is equivalent to point estimation of $\min\{\mu_1, \mu_2\}$, but since the function $h(P) = \min\{\mu_1(P), \mu_2(P)\}$ is not differentiable in P, the estimation problem is nonstandard,

[9] However, there is an additional implicit inequality: $\mu_1 \leq \mu_2$. If μ_2 is sufficiently greater than μ_1, then in large samples this constraint will have a negligible impact on the analysis. But if μ_1 is arbitrarily close to μ_2 in the collection of measures P, then the analysis is more subtle.

as we discussed in Section 5.1. For example, no regular estimators of $h(P)$ exist, and locally unbiased estimation of $h(P)$ is not possible.

These simple examples can be generalized to many other decision problems with partially identified parameters. If the subparameter of interest κ is a d-dimensional vector, then its identified set $K(P)$ will be a subset of $\mathbb{R}^d$. If $K = K(P)$ is convex, then it can be characterized by its *support function*

$$h_K(x) = \sup_{v \in K} x'v, \quad x \in \mathbb{R}^d,$$

which encodes its supporting hyperplanes in all directions. The support function approach has been used by Beresteanu and Molinari (2008), Bontemps, Magnac, and Maurin (2012), and Kaido (2016), among others. Then, estimation of $K(P)$ can be viewed as a functional point estimation problem. If the support function of $K(P)$ is smooth (as a function of P), then classical results on efficiency bounds for estimation of functions, such as those in Bickel, Klaasen, Ritov, and Wellner (1993), can be applied. Kaido and Santos (2014) use this approach to obtain asymptotically efficient estimators for models defined by convex moment inequalities, under sufficient smoothness. On the other hand, in many econometric applications $K(P)$ is not differentiable in P. This can arise, for example, in moment inequality problems when the number of binding (or nearly binding) inequalities at a given point on the boundary of the identified set is greater than d. Then $K(P)$ will not be differentiable in P.

5.3 Confidence Intervals and Sets

Confidence intervals and confidence sets can be regarded as rules of the form $\delta : \mathcal{Z} \to \mathcal{A}$, where $\mathcal{A}$ is some collection of subsets of the Euclidean parameter space Θ. For example, if θ is a scalar parameter we could consider rules that take the data Z and produce a closed interval. In the classical Neyman-Pearson approach to constructing confidence intervals, we require that, for any $\theta \in \Theta$, if $Z \sim P_\theta$ then $\delta(Z)$ contains θ with probability equal to or greater to some prespecified value, such as 0.95. This coverage condition amounts to a restriction on the set of decision rules, as discussed in Section 2.4.3. Subject to this restriction, we could seek to find a rule that generates small sets, by using a loss function that depends on the volume of the set or some other criterion.

An alternative decision theoretic approach could entertain a trade-off between coverage and length or volume of the set. Again suppose $\delta : \mathcal{Z} \to \mathcal{A}$ where $\mathcal{A}$ is a collection of subsets of Θ. A generic form of the loss function for confidence sets would be

$$L(\theta, a) = \ell\left(\text{vol}(a), \chi(\theta, a)\right),$$

where $\text{vol}(a)$ denotes the length or volume of the set a, and $\chi(\theta, a)$ measures lack of coverage or precision

of the set. The most common choice for χ is

$$\chi(\theta, a) = \mathbf{1}\{\theta \notin a\}.$$

When loss is additively separable in the arguments $\text{vol}(a)$ and $\chi(\theta, a)$, the part of risk that comes from the $\chi(\theta, a)$ term directly measures coverage. A modification of χ to penalize non-coverage depending on the distance between the parameter θ and the set δ would set

$$\chi(\theta, a) = \inf_{t \in a} \|\theta - t\|$$

where $\|\cdot\|$ is used to denote a Euclidean distance.

There is a long history of work adopting a decision theoretic perspective on the construction of confidence sets. In a setting with a scalar parameter, Winkler (1972) and Cohen and Strawderman (1973) consider specific choices for loss and analyze Bayes rules and admissability. Hwang and Casella (1982) consider restricting the space of confidence set rules to satisfy a minimum coverage requirement and then minimize a loss that depends only on volume. They show that when $\dim(\theta) \geq 4$, confidence sets centered at James-Stein estimators can be constructed that dominate classical confidence intervals. Evans, Hansen, and Stark (2005) consider minimax expected volume (length) subject to a coverage condition for a scalar parameters on a bounded space. Casella, Hwang, and Robert (1993) note that using a loss function that is linear in lack of coverage and volume can lead to unappealing decision rules. Casella, Hwang, and Robert (1994) and Rice, Lumley, and Szpiro (2008) consider various loss functions that avoid the problem pointed out in Casella, Hwang, and Robert (1993) and provide a decision theoretic analysis of confidence set rules for these loss functions.

Given the prevalence of non-regular models and partial identification problems in econometrics, there is potential for further work on decision theoretic analysis of confidence set rules under these conditions. For example, Chamberlain (2007) considers a loss function that depends on volume and coverage and uses invariance properties to develop a Bayesian confidence procedure in an instrumental variables setting with possibly weak instruments. Müller and Norets (2016) consider an alternative bet-proofness criterion to construct confidence sets in nonstandard problems.

5.4 Experimental and Data Collection Design

Experimental design, and the more general problem of choosing a data collection scheme (such as choosing sample weights in a stratified survey) is a classical problem in statistics and there is a vast literature on this topic. Some standard texts on experimental design include Fisher (1935), Cochrane and Cox (1957), and Cox (1958); discussions of survey and sampling design include Kish (1965), Manski and McFadden (1981), and Thompson (2012). There has been renewed interest in data collection design in recent years, for various reasons including: the increasing use of randomized controlled trials in microe-

conomics; the development of new surveys with design input from social scientists; and technological advances that facilitate new data collection, such as online experiments and richer observational data sources. As a result, the design of experiments and surveys remains an active area of research, often with immediate applications in economics and other fields.

The decision theoretic framework we have outlined in this chapter can provide a useful set of organizing concepts for a range of data collection problems. However, there are also some limitations of the framework that we shall discuss below. To apply our framework, we must specify an action space $\mathscr{A}$, a parameter space Θ, a random variable $Z \sim P_\theta$, and an evaluation criterion in the form of a welfare function $W(\theta, a)$ or a loss function $L(\theta, a)$.

The action space embodies the choices available to the designer of the experiment or survey. For example, in a simple randomized control trial, the experimenter could choose the probability of assigning individuals to treatment $p \in (0,1)$. Then we could set $\mathscr{A} = (0,1)$. In a stratified survey, there is a characteristic X, usually discrete taking values in some set $\{\xi_1, \ldots, \xi_K\}$. A stratification scheme consists of a weighting $p = (p_1, \ldots, p_K)$ in the simplex Δ^{k-1}, with the interpretation that individuals with $X = \xi_k$ are included in the sample with probability p_k. Then the action space $\mathscr{A}$ is the simplex Δ^{k-1} or some prespecified subset of the simplex. Other design problems involve different choices for the action space. Economically motivated mechanisms for data design can be considered, as in Philipson (1997), Narita (2018), and others.

The parameter $\theta \in \Theta$ characterizes all relevant features of the population; usually at least some components of θ are unknown. Hence there is a role for considering the different outcomes that may result from a given design under different values of θ.

Recall that $Z \sim P_\theta$ represents the data available *before* the action is chosen. Here, Z would correspond to data available in advance of specification of the survey or experiment. In some cases, no such data may be available. In other cases, there may be data from prior surveys or experiments that are informative about θ. For example, Sukhatme (1935) and Solomon and Zacks (1970) consider the use of prior samples to select a stratified sampling scheme. Hahn, Hirano, and Karlan (2011) consider using data from the initial wave of a two-wave experiment to choose conditional treatment probabilities in the second wave.

To apply the framweork of this chapter we need to specify a welfare function $W(\theta, a)$, or, equivalently, a loss function $L(\theta, a)$. The choice of the evaluation criterion may be motivated by the intended use of the experiment or survey. In practice, there may not be a single obvious choice for W or L. One way to specify the evaluation criterion is to focus on some particular estimator (and estimand) that will be applied to the data collected from the experiment or survey, and evaluate its expected loss under different design choices a.

Example 9 *Suppose the decision maker will run a randomized controlled trial, assigning n_1 individuals to treatment 1 and n_0 individuals to treatment 0. Suppose potential outcomes $Y(0)$ and $Y(1)$ are distributed*

as

$$Y(0) \sim N(\mu_0, \sigma_0^2)$$
$$Y(1) \sim N(\mu_1, \sigma_1^2)$$

The experiment will produce observations on Y_i, T_i. Suppose T_i is Bernoulli with probability $a \in (0,1)$ and Y_i is the observed outcome, satisfying $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. The decision maker plans to estimate the average treatment effect $\mu_1 - \mu_0$ with the simple difference in means estimator

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) Y_i.$$

The estimator is unbiased (conditional on both treatment cells being nonempty), and its large-sample normalized variance is

$$\frac{\sigma_1^2}{na} + \frac{\sigma_0^2}{n(1-a)}.$$

Then we could take $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1)$ and the loss function to be

$$L(\theta, a) = \frac{\sigma_1^2}{na} + \frac{\sigma_0^2}{n(1-a)}. \tag{15}$$

If σ_0 and σ_1 are known (i.e., Θ is degenerate along the relevant dimensions), then it is straightforward to minimize (15) to obtain the optimal randomization probability

$$a^*(\theta) = \frac{\sigma_1}{\sigma_0 + \sigma_1}.$$

In Example 9, there is an optimal action given knowledge of σ_0^2 and σ_1^2. If the variances are not known, then this optimal action is not feasible. Following our framework, we can adopt a Bayesian approach by specifying a prior distribution over the parameters, or a minmax or minmax regret approach. For discussions of Bayesian experimental design see Spiegelhalter, Freedman, and Parmar (1994) and Chaloner and Verdinelli (1995). For minmax-regret approaches to data design, see for example Manski and Tetenov (2016) and Dominitz and Manski (2017).

An important practical issue in experimental design is stratified randomization. If a binary treatment is simply randomized over a sample of individuals, this can lead to *ex post* imbalance of covariates between treated and control groups. As a consequence, estimators of treatment effects can have higher variance than is possible if the treatment is assigned in a way that reduces covariate imbalance. Researchers therefore often use more complicated assignment mechanisms designed to improve covariate balance. One common reaction to covariate imbalance is to re-randomized. Morgan and Rubin (2012) propose to establish rules for re-randomization and account for these rules in inference. In Example 9, for instance, one could use an assignment rule that re-randomizes if either treatment cell is empty. See Athey and Imbens (2017) for a recent survey of approaches to stratified randomization.

When peer effects (or "interference" in the classical experimental design terminology) are of interest, one may wish to design experiments that are informative about the magnitude of such effects. Economic studies that use randomized experiments to measure peer effects include Duflo and Saez (2003), Miguel and Kremer (2003), and Angelucci and De Giorgi (2009). The design of experiments to measure peer effects has been studied in Hirano and Hahn (2010) and Baird, Bohren, McIntosh, and Özler (2017).

The Wald framework we have emphasized in this chapter is a single-agent theory of statistical decision-making. As such, it does not easily handle strategic considerations that could arise from the behavior of the subjects of the experiment or survey, unless they can be subsumed into the loss or welfare function. Some recent work has explicitly modeled design and other statistical decision problems as games between some designer and her human subjects. Somes examples include Chassang, Padró i Miquel, and Snowberg (2012), Tetenov (2016), and Spiess (2018).

6 Conclusion

In this chapter we have discussed the construction and analysis of decision rules in econometrics, focusing on obtaining useful approximations to their risk (or expected welfare) properties. The framework is quite general and provides useful insights not only into conventional statistical activities such as estimation, but also into policy decision-making (through empirical auction design and treatment assignment rules for example), design of experiments and surveys, point forecasting, and other empirical problems. There are many other potential applications of the framework which we have not covered here, such as the construction and comparison of forecast intervals (e.g. Christoffersen (1998) and Askanazi, Diebold, Schorfheide, and Shin (2018)) and forecast densities (e.g. Diebold, Gunther, and Tay (1998), Tay and Wallis (2000), and Hall and Mitchell (2004)).

We focused on local asymptotic methods as a tool for obtaining relatively simple characterizations of the risk/welfare properties of decision rules. Local asymptotics can be used to study not only smooth settings where normality emerges naturally, but also other settings involving parameters that do not change smoothly with the underlying distribution of the data, rare events, and nonstationary time series. Of course, depending on the problem, other approaches such as the use of concentration inequalities to obtain risk bounds, and the use of global nonparametric approximations (as in Brown and Low (1996), Nussbaum (1996), and Armstrong and Kolesár (2018)) may provide more useful characterizations of complex decision rules.

References

ANDREWS, D. W. K., M. J. MOREIRA, AND J. H. STOCK (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74(3), 715–752.

ANDREWS, I., AND T. B. ARMSTRONG (2017): "Unbiased Instrumental Variables Estimation under Known First-Stage Sign," *Quantitative Economics*, 8(2), 479–503.

ANGELUCCI, M., AND G. DE GIORGI (2009): "Indirect Effects of an Aid Program: How do Cash Transfers Affect Ineligibles' Consumption?," *American Economic Review*, 99(1), 486–508.

ANSCOMBE, F. J., AND R. J. AUMANN (1963): "A Definition of Subjective Probability," *Annals of Mathematical Statistics*, 34, 199–205.

ARJAS, E., AND O. SAARELA (2010): "Optimal Dynamic Regimes: Presenting a Case for Predictive Inference," *The International Journal of Biostatistics*, 6(2).

ARMSTRONG, T. B., AND M. KOLESÁR (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86(2), 655–683.

ARYAL, G., AND D.-H. KIM (2013): "A Point Decision for Partially Identified Auction Models," *Journal of Business and Economic Statistics*, 31(4), 384–397.

ASKANAZI, R., F. X. DIEBOLD, F. SCHORFHEIDE, AND M. SHIN (2018): "On the Comparison of Interval Forecasts," working paper.

ATHEY, S., AND G. W. IMBENS (2017): "The Econometrics of Randomized Experiments," in *Handbook of Economic Field Experiments*, ed. by E. Duflo, and A. Banerjee, vol. 1, pp. 73–140. Elsevier.

ATHEY, S., AND S. WAGER (2017): "Efficient Policy Learning," working paper.

AVRAMOV, D., AND G. ZHOU (2010): "Bayesian Portfolio Analysis," *Annual Review of Financial Economics*, 2(1), 25–47.

BAIRD, S., J. A. BOHREN, C. MCINTOSH, AND B. ÖZLER (2017): "Optimal Design of Expeirments in the Presence of Interference," working paper.

BARANCHIK, A. J. (1964): "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Stanford University Department of Statistics Technical Report 51.

BARBERIS, N. (2000): "Investing for the Long Run when Returns Are Predictable," *The Journal of Finance*, 55(1), 225–264.

BERAN, R. (1995): "The Role of Hájek's Convolution Theorem in Statistics," *Kybernetika*, 31, 221–237.

BERESTEANU, A., AND F. MOLINARI (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76(4), 763–814.

BERGER, J. O. (1993): *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 2nd edn.

BHATTACHARYA, D. (2009): "Inferring Optimal Peer Assignment from Experimental Data," *Journal of the American Statistical Association*, 104, 486–500.

BHATTACHARYA, D., AND P. DUPAS (2012): "Inferring Welfare Maximizing Treatment Assignment under Budget Constraints," *Journal of Econometrics*, 167, 168–196.

BICKEL, P. J., C. A. KLAASEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

BLACK, D., J. SMITH, M. BERGER, AND B. NOEL (2003): "Is the Threat of Training More Effective than Training Itself? Experimental Evidence from the UI System," *American Economic Review*, 93(4), 1313–1327.

BONTEMPS, C., T. MAGNAC, AND E. MAURIN (2012): "Set Identified Linear Models," *Econometrica*, 80(3), 1129–1155.

BOUSQUET, O., S. BOUCHERON, AND G. LUGOSI (2004): "Introduction to Statistical Learning Theory," in *Advanced Lectures on Machine Learning*, pp. 169–207. Springer-Verlag.

BREIMAN, L. (1996): "Bagging Predictors," *Machine Learning*, 36, 105–139.

BROWN, L. D., AND M. G. LOW (1996): "Asymptotic Equivalence of Nonparametric Regression and White Noise," *The Annals of Statistics*, 24(6), 2384–2398.

BÜHLMANN, P., AND B. YU (2002): "Analyzing Bagging," *Annals of Statistics*, 30, 927–961.

BUSHWAY, S., AND J. SMITH (2007): "Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us," *Journal of Quantitative Criminology*, 23(4), 377–387.

CASELLA, G., J. T. G. HWANG, AND C. ROBERT (1993): "A Paradox in Decision-Theoretic Interval Estimation," *Statistica Sinica*, 3, 141–155.

——— (1994): "Loss Functions for Set Estimation," in *Statistical Decision Theory and Related Topics V*, ed. by S. S. Gupta, and J. O. Berger. Springer, New York.

CATTANEO, M. D. (2010): "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155, 138–154.

CATTANEO, M. D., R. K. CRUMP, AND M. JANSSON (2012): "Optimal Inference for Instrumental Variables Regression with Non-Gaussian Errors," *Journal of Econometrics*, 167(1), 1–15.

CHALONER, K., AND I. VERDINELLI (1995): "Bayesian Experimental Design: A Review," *Statistical Science*, 10(3), 273–304.

CHAMBERLAIN, G. (2007): "Decision Theory Applied to an Instrumental Variables Model," *Econometrica*, 75(3), 609–652.

——— (2011): "Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, pp. 11–39. Oxford University Press.

CHASSANG, S., G. PADRÓ I MIQUEL, AND E. SNOWBERG (2012): "Selective Trials: A Principal-Agent Approach to Randomized Control Trials," *American Economic Review*, 102(4), 1279–1309.

CHAWLA, S., J. D. HARTLINE, AND D. NEKIPELOV (2017): "Mechanism Redesign," working paper.

CHEN, X., T. M. CHRISTENSEN, AND E. TAMER (2017): "Monte Carlo Confidence Sets for Identified Sets," working paper.

CHERNUZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, AND W. NEWEY (2016): "Double Machine Learning for Treatment and Causal Parameters," Working paper.

CHRISTOFFERSEN, P. F. (1998): "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862.

CHRISTOFFERSEN, P. F., AND F. X. DIEBOLD (1997): "Optimal Prediction Under Asymmetric Loss," *Econometric Theory*, 13(6), 808–817.

CLAESKENS, G., AND N. L. HJORT (2008): *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

COCHRANE, W. G., AND G. M. COX (1957): *Experimental Designs*. Wiley, New York.

COHEN, A., AND W. E. STRAWDERMAN (1973): "Admissible Confidence Interval and Point Estimation for Translation or Scale Parameters," *The Annals of Statistics*, 1(3), 545–550.

COX, D. R. (1958): *Planning of Experiments*. Wiley, New York.

DAVIES, R. B. (1985): "Asymptotic Inference when the Amount of Information is Random," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, ed. by L. M. Le Cam, and R. A. Olshen, vol. II. Wadsworth, Belmont, CA.

DEHEJIA, R. H. (2005): "Program Evaluation as a Decision Problem," *Journal of Econometrics*, 125, 141–173.

DEL NEGRO, M., AND F. SCHORFHEIDE (2004): "Priors from General Equilibrium Models for VARs," *International Economic Review*, 45(2), 643–673.

DIEBOLD, F. X., T. A. GUNTHER, AND A. S. TAY (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management," *International Economic Review*, 39(4), 863–883.

DOAN, T., R. LITTERMAN, AND C. SIMS (1984): "Forecasting and Conditional Projection using Realistic Prior Distributions," *Econometric Reviews*, 3(1), 1–100.

DOMINITZ, J., AND C. F. MANSKI (2017): "More Data or Better Data? A Statistical Decision Problem," *Review of Economic Studies*, 84, 1583–1605.

DUBÉ, J.-P., Z. FANG, N. FONG, AND X. LUO (2017): "Competitive Price Targeting with Smartphone Coupons," *Marketing Science*, 36(6), 944–975.

DUFLO, E., AND E. SAEZ (2003): "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 118, 815–842.

EATON, M. L. (1989): *Group Invariance Applications in Statistics*, Regional Conference Series in Probability and Statistics. IMS-ASA.

EGERTON, M. F., AND P. J. LAYCOCK (1982): "An Explicit Formula for the Risk of James-Stein Estimators," *The Canadian Journal of Statistics*, 10(3), 199–205.

EVANS, S., B. HANSEN, AND P. STARK (2005): "Minimax Expected Measure Confidence Sets for Restricted Location Parameters," *Bernoulli*, 11(4), 571–590.

FANG, Z. (2016): "Optimal Plug-in Estimators of Directionally Differentiable Functionals," working paper.

FANG, Z., AND A. SANTOS (2014): "Inference on Directionally Differentiable Functionals," working paper.

FERGUSON, T. S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.

FESSLER, P., AND M. KASY (2018): "How to Use Economic Theory to Improve Estimators: Shrinking toward Theoretical Restrictions," working paper.

FISHER, R. A. (1935): *The Design of Experiments*. Oliver and Boyd, Edinburgh.

FRÖLICH, M. (2008): "Statistical Treatment Choice: An Application to Active Labor Market Programs," *Journal of the American Statistical Association*, 103, 547–558.

GILBOA, I., AND D. SCHMEIDLER (1989): "Maxmin Expected Utility with Non-unique Prior," *Journal of Mathematical Economics*, 18, 141–153.

GRAHAM, B. S., AND K. HIRANO (2011): "Robustness to Parametric Assumptions in Missing Data Models," *American Economic Review: Papers & Proceedings*, 101(3), 538–543.

GRAHAM, B. S., G. W. IMBENS, AND G. RIDDER (2014): "Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis," *Quantitative Economics*, 5, 29–66.

HAHN, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66(2), 315–331.

HAHN, J., K. HIRANO, AND D. KARLAN (2011): "Adaptive Experimental Design using the Propensity Score," *Journal of Business and Economic Statistics*, 29(1), 96–108.

HAILE, P., AND E. TAMER (2003): "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 111(1), 1–51.

HÁJEK, J. (1970): "A Characterization of Limiting Distributions of Regular Estimates," *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14, 323–330.

HALL, S. G., AND J. MITCHELL (2004): "'Optimal' Combination of Density Forecasts," working paper.

HANSEN, B. E. (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190, 115–132.

HAYASHI, T. (2008): "Regret Aversion and Opportunity Dependence," *Journal of Economic Theory*, 139(1), 242–268.

HIRANO, K., AND J. HAHN (2010): "Design of Randomized Experiments to Measure Social Interaction Effects," *Economics Letters*, 106, 51–53.

HIRANO, K., AND G. W. IMBENS (2004): "The Propensity Score with Continuous Treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*, ed. by A. Gelman, and X.-L. Meng. Wiley, New York.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HIRANO, K., AND J. R. PORTER (2003): "Asymptotic Efficiency in Parametric Structural Models with Parameter?Dependent Support," *Econometrica*, 71(5), 1307–1338.

——— (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77(5), 1683–1701.

——— (2012): "Impossibility Results for Nondifferentiable Functionals," *Econometrica*, 80(4), 1769–1790.

——— (2015): "Location Properties of Point Estimators in Linear Instrumental Variables and Related Models," *Econometric Reviews*, 34(6-10), 720–733.

HIRANO, K., AND J. H. WRIGHT (2017): "Forecasting with Model Uncertainty: Representations and Risk Reduction," *Econometrica*, 85(2), 617–643.

HOERL, A. E., AND R. W. KENNARD (1970): "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12(1), 55–67.

HWANG, J. T. G., AND G. CASELLA (1982): "Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 10(3), 868–881.

IBRAGIMOV, I., AND R. HASMINSKII (1981): *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

IMAI, K., AND D. A. VAN DYK (2004): "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866.

IMBENS, G. W. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.

INGRAM, B. F., AND C. H. WHITEMAN (1994): "Supplanting the 'Minnesota' Prior: Forecasting Macroeconomic Time Series using Real Business Cycle Model Priors," *Journal of Monetary Economics*, 34, 497–510.

INOUE, A., AND L. KILIAN (2006): "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306.

JAMES, W., AND C. STEIN (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379.

JEGANATHAN, P. (1982): "On the Asymptotic Theory of Estimation when the Limit of the Log-Likelihood Ratios is Mixed Normal," *Sankhya: A*, 44, 173–212.

——— (1995): "Some Aspects of Asymptotic Theory with Applications to Time Series Models," *Econometric Theory*, 11, 818–887.

KAIDO, H. (2016): "A Dual Approach to Inference for Partially Identified Econometric Models," *Journal of Econometrics*, 192(1), 269–290.

KAIDO, H., AND A. SANTOS (2014): "Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities," *Econometrica*, 82(1), 387–413.

KAJI, T. (2017): "Theory of Weak Identification in Semiparametric Models," working paper.

KAN, R., AND G. ZHOU (2007): "Optimal Portfolio Choice with Parameter Uncertainty," *The Journal of Financial and Quantitative Analysis*, 42(3), 621–656.

KANDEL, S., AND R. F. STAMBAUGH (1996): "On the Predictability of Stock Returns: An Asset-Allocation Perspective," *The Journal of Finance*, 51(2), 385–424.

KIM, D.-H. (2010): "Bayesian Econometrics for Auction Models," Ph.D. thesis, University of Arizona.

——— (2013): "Optimal Choice of a Reserve Price Under Uncertainty," *International Journal of Industrial Organization*, 31(5), 5887–602.

——— (2015): "Flexible Bayesian Analysis of First Price Auctions Using a Simulated Likelihood," *Quantitative Economics*, 6(2), 429–461.

KISH, L. (1965): *Survey Sampling*. Wiley, New York.

KITAGAWA, T., AND A. TETENOV (2017): "Equality-Minded Treatment Choice," working paper.

——— (2018): "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86(2), 591–616.

KLEIBERGEN, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70(5), 1781–1803.

KNIGHT, K., AND W. FU (2000): "Asymptotics for LASSO-Type Estimators," *Annals of Statistics*, 28, 1356–1378.

KOCK, A. B., AND M. THYRSGAARD (2017): "Optimal Sequential Treatment Allocation," working paper.

LE CAM, L. (1970): "On the Assumptions Used to Prove Asymptotic Normality of Maximum Likelihood Estimates," *The Annals of Mathematical Statistics*, 41(3), 802–828.

LE CAM, L. M. (1972): "Limits of Experiments," in *Proceedings of the Sixth Berkeley Symposium of Mathematical Statistics*, vol. 1, pp. 245–261.

——— (1986): *Asymptotic Methods in Statistical Theory*. Springer-Verlag, New York.

LE CAM, L. M., AND G. L. YANG (2000): *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, 2nd edn.

LECHNER, M. (2001): "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in *Econometric Evaluations of Active Labor Market Policies in Europe*, ed. by M. Lechner, and F. Pfeiffer. Physica, Heidelberg.

LI, T., I. PERRIGNE, AND Q. VUONG (2002): "Structural Estimation of the Affiliated Private Value Model," *The RAND Journal of Economics*, 33(2), 171–193.

LIESE, F., AND K.-J. MIESCKE (2008): *Statistical Decision Theory: Estimation, Testing, and Selection*. Springer, New York.

MANSKI, C. F. (1995): *Identification problems in the social sciences*. Harvard University Press.

——— (2003): *Partial Identification of Probability Distributions*. Springer-Verlag.

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72(4), 1221–1246.

——— (2007): *Identification for Prediction and Decision*. Harvard University Press.

MANSKI, C. F., AND D. L. MCFADDEN (1981): "Alternative Estimators and Sampling Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data and Econometric Applications*, ed. by C. F. Manski, and D. L. McFadden. MIT Press, Cambridge, MA.

MANSKI, C. F., AND A. TETENOV (2016): "Sufficient Trial Size to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, 113(38), 10518–10523.

MBAKOP, E., AND M. TABORD-MEEHAN (2016): "Model Selection for Treatment Choice: Penalized Welfare Maximization," working paper.

MIGUEL, E., AND M. KREMER (2003): "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 72(1), 159–217.

MILNOR, J. (1954): "Decisions Against Nature," in *Decision Processes*, ed. by R. M. Thrall, C. H. Coombs, and R. L. Davis. Wiley, New York.

MOREIRA, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71(4), 1027–1048.

MORGAN, K. L., AND D. RUBIN (2012): "Rerandomization to Improve Covariate Balance in Experiments," *The Annals of Statistics*, 40(2), 1263–1282.

MORGENSTERN, J., AND T. ROUGHGARDEN (2015): "The Pseudo-Dimension of Near-Optimal Auctions," Working paper.

MORI, H. (2001): "Asymptotic Inadmissibility of Maximum Likelihood Estimator in a Quadratic Programming Problem," *Gakushuin Economic Paper*, 38(1), 35–49.

——— (2004): "Finite Sample Properties of Estimators for the Optimal Portfolio Weight," *Journal of the Japan Statistical Society*, 34(1), 27–46.

MORRIS, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78(381), 47–55.

MÜLLER, U. K. (2011): "Efficient Tests under a Weak Convergence Assumption," *Econometrica*, 79(2), 395–435.

MÜLLER, U. K., AND A. NORETS (2016): "Credibility of Confidence Sets in Nonstandard Econometric Problems," *Econometrica*, 84(6), 2183–2213.

MÜLLER, U. K., AND Y. WANG (2017): "Nearly Weighted Risk Minimal Unbiased Estimation," working paper.

MURPHY, S. A. (2003): "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society, Series B*, 65, 331–366.

NARITA, Y. (2018): "Toward an Ethical Experiment," working paper.

NUSSBAUM, M. (1996): "Asymptotic Equivalence of Density Estimation and Gaussian White Noise," *The Annals of Statistics*, 24(6), 2399–2430.

PAARSCH, H. J. (1992): "Deciding between the common and private value paradigms in empirical models of auctions," *Journal of Econometrics*, 51(1), 191–215.

PHILIPSON, T. (1997): "Data Markets and the Production of Surveys," *Review of Economic Studies*, 64, 47–72.

RICE, K. M., T. LUMLEY, AND A. A. SZPIRO (2008): "Trading Bias for Precision: Decision Theory for Intervals and Sets," UW Biostatistics Working Paper.

ROBERT, C. (1988): "An Explicit Formula for the Risk of the Positive-Part James-Stein Estimator," *The Canadian Journal of Statistics*, 16(2), 161–168.

ROBINS, J. M. (2004): "Optimal Structural Nested Models for Optimal Decisions," in *Proceedings of the Second Seattle Symposium on Biostatistics*, ed. by D. Y. Lin, and P. Heagerty, New York. Springer-Verlag.

ROSSI, P. E., R. E. MCCULLOCH, AND G. M. ALLENBY (1996): "The Value of Purchase History in Target Marketing," *Marketing Science*, 15(4), 321–340.

SAVAGE, L. J. (1951): "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 46(253), 55–67.

——— (1972): *The Foundations of Statistics*. Dover, New York, 2nd edn.

SOLOMON, H., AND S. ZACKS (1970): "Optimal Design of Sampling from Finite Populations: A Critical Review and Indication of New Research Areas," *Journal of the American Statistical Association*, 65(330), 653–677.

SONG, K. (2014): "Point Decisions for Interval-Identified Parameters," *Econometric Theory*, 30(2), 334–356.

SPIEGELHALTER, D. J., L. S. FREEDMAN, AND M. K. B. PARMAR (1994): "Bayesian Approaches to Randomized Trials," *Journal of the Royal Statistical Society, Series A*, 157(3), 357–416.

SPIESS, J. (2018): "Optimal Estimation when Researcher and Social Preferences are Misaligned," working paper.

STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

STOCK, J. H., AND J. H. WRIGHT (2000): "GMM with Weak Identification," *Econometrica*, 68(5), 1055–1096.

STOYE, J. (2009): "Minimax Regret Treatment Choice with Finite Samples," *Journal of Econometrics*, 151, 70–81.

——— (2011): "Axioms for Minimax Regret Choice Correspondences," *Journal of Economic Theory*, 146(6), 2226–2251.

STRASSER, H. (1985): *Mathematical Theory of Statistics*. Walter de Gruyter & Co., New York.

SUKHATME, P. V. (1935): "Contribution to the Theory of the Representative Method," *Supplement to the Journal of the Royal Statistical Society*, 2(2), 253–268.

TAMER, E. (2012): "Partial Identification in Econometrics," *Annual Review of Economics*, 2, 167–195.

TAY, A. S., AND K. F. WALLIS (2000): "Density Forecasting: A Survey," *Journal of Forecasting*, 19, 235–254.

TETENOV, A. (2012): "Statistical Treatment Choice Based on Asymmetric Minmax Regret Criteria," *Journal of Econometrics*, 166, 157–165.

——— (2016): "An Economic Theory of Statistical Testing," working paper.

THOMPSON, S. K. (2012): *Sampling*. Wiley, New York, 3rd edn.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288.

VAN DER VAART, A. W. (1991a): "An Asymptotic Representation Theorem," *International Statistical Review*, 59, 99–121.

——— (1991b): "On Differentiable Functionals," *The Annals of Statistics*, 19, 178–204.

——— (1998): *Asymptotic Statistics*. Cambridge University Press, New York.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

VARIAN, H. (1974): "A Bayesian Approach to Real Estate Assessment," in *Studies in Bayesian Econometrics and Statistics in Honor of L. J. Savage*, ed. by S. E. Feiberg, and A. Zellner. North-Holland.

WALD, A. (1950): *Statistical Decision Functions*. Wiley, New York.

WESLER, O. (1959): "Invariance Theory and a Modified Minimax Principle," *The Annals of Mathematical Statistics*, 30(1), 1–20.

WINKLER, R. L. (1972): "A Decision-Theoretic Approach to Interval Estimation," *Journal of the American Statistical Association*, 67(337), 187–191.

XIE, X., S. C. KOU, AND L. D. BROWN (2012): "SURE Estimates for a Heteroscedastic Hierarchical Model," *Journal of the American Statistical Association*, 107(500), 1465–1479.

ZAJONC, T. (2012): "Bayesian Inference for Dynamic Treatment Regimes: Mobility, Equity, and Efficiency in Student Tracking," *Journal of the American Statistical Association*, 107(497), 80–92.

ZELLNER, A. (1986): "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," *Journal of the American Statistical Association*, 81(394), 446–451.

ZELLNER, A., AND V. K. CHETTY (1965): "Prediction and Decision Problems in Regression Models from the Bayesian Point of View," *Journal of the American Statistical Association*, 60(310), 608–616.