An Information-Theoretic Approach to Partially Identified Problems[†]

Amos Golan* and Jeffrey M. Perloff**

December 2025

Abstract

An information-theoretic maximum entropy (ME) model provides an alternative approach to finding solutions to partially identified models. In these models, we can identify only a solution set rather than point-identifying the parameters of interest, given our limited information. Manski (2021) and others propose using statistical decision functions in general, and the minimax-regret (MMR) criterion in particular, to select a unique solution. Using Manski's simulations for a missing data and a treatment problem, including an empirical example, we show that ME performs as well as or better than MMR. In additional simulations, ME dominates various other statistical decision functions. ME has an axiomatic underpinning and is computationally efficient.

Key Words: *Information Theory, Maximum Entropy, Minimax Regret, Partial Identification, Statistical Decision Functions*

[†] Part of this work was done while Golan was visiting the Department of Economics at the University of Chicago. He thanks the department for its hospitality. He benefited from many enchanting discussions with Stephane Bonhomme. He thanks the participants in the University of Chicago's econometrics and quantitative methods workshops, as well as the econometrics workshops at Harvard-MIT, Northwestern, and UC Riverside. He is indebted to Elie Tamer for generously sharing his time and providing many insights, and to Tae-Hwy Lee for his comments. We are grateful to Chuck Manski for his suggestions regarding simulations and to him and Valentyn Litvin for sharing their STATA code. Jeff Wheble helpfully converted the code to allow us to conduct the new simulations presented here.

* Golan: Department of Economics, American University, and the Santa Fe Institute, agolan@american.edu.

** Perloff: Department of Agricultural and Resource Economics, University of California, Berkeley, jperloff@berkeley.edu.

Corresponding Author: agolan@american.edu

Any inference or modeling problem with insufficient information or missing data has multiple possible solutions, each of which is consistent with the available information. The partial identification approach yields estimates or predictions based on various sets of assumptions, rather than exact or probabilistic estimates (e.g., Haavelmo, 1940; Wald, 1945; Leamer, 1985; Tamer, 2010; Manski, 2021, 2025; and Manski, Sanstad, and DeCanio, 2021). Based on this incomplete information, how should one make a prediction, choose a treatment, or make other decisions to maximize welfare? The choice depends on whether the criterion is Bayesian, maximizing minimum welfare (maximin), minimizing maximum risk (Wald, 1945), information-theoretical, or another criterion. Manski (2021) discusses the first three of these approaches, focusing on a statistical decision theory approach that minimizes the maximum risk or regret (minimax regret, or MMR). We propose using an information-theoretical approach, such as maximum entropy (ME), as an alternative and compare it to other commonly used approaches.

The partial identification method begins by asking what inferences a researcher can make using only empirical evidence, which typically yields a range of possibilities due to partial information. The researcher then examines how various assumptions narrow this range (the "identified set") by providing more information (Tamer, 2010). The philosophical underpinning of this process is that conclusions and actions based on empirical models with fewer questionable assumptions are more believable.

Much of the partial identification literature addresses identification problems generated by imperfect data quality, including measurement error and missing data. A planner specifies a state space listing all states of nature deemed feasible. Wald (1945) and others have recommended evaluating the performance of a statistical decision function (SDF) by the state-dependent vector of expected welfare that it yields.

We illustrate an information-theoretic approach and compare it to various alternative decision functions using three sets of simulations: two from Manski (2021) and a six-sided die

problem.[1] Manski's first application is based on a prediction problem from Haavelmo (1944). Manski examined predictions of bounded outcomes under square loss when some outcome data are missing.

His second application concerns the treatment response in randomized trials and observational settings to inform treatment choice with two treatments. Realized outcomes are observable, but counterfactual ones are not. The problem is to choose treatments in a population with the same distribution of treatment response as the study population.

In our third simulation, a six-sided die problem (an unconditional multinomial problem), we compare the MMR of ME to several alternative SDFs: least squares, empirical likelihood, and Rényi (or the equivalent Cressie-Read).

The information-theoretic method has four attractive features. First, it has an axiomatic underpinning (see the Appendix for a summary of the axioms). Second, it is simple to model, compute, and estimate. Third, it can be used with any problem, not just those with a discrete number of choices. Fourth, restrictions can be easily imposed as additional constraints, facilitating bottom-up partial identification comparisons in which one adds assumptions one at a time.

We start by briefly summarizing the MMR and ME approaches. The next two sections replicate Manski's (2021) missing data and treatment problems using ME. The fourth section provides a graphical interpretation of information-theoretic approaches. The fifth section uses a six-sided die problem to compare various SDFs. The final section draws conclusions.

Two Approaches to Predicting or Selecting an Action

Consider the problem of a policymaker who selects an action. That action's effect on welfare depends on an unknown state of nature. The first step is to specify a state space that contains all the states the policymaker believes are possible, which partially identifies the action. The policymaker may observe sample data that might contain information about the true state. However, the policymaker chooses an action without knowing the true state. Two possible

[1] We are grateful to Chuck Manski for his gracious help regarding simulations, and to him and Valentyn Litvin for sharing their STATA code.

approaches are to use a statistical decision function (SDF) that minimizes the maximum risk or maximizes entropy.

Minimize Maximum Regret

To choose an action or make a prediction, Wald (1945) advocated using an SDF that minimizes the maximum risk. He demonstrated that his approach is analogous to game theory, where the statistician plays a game against nature, assuming nature seeks to maximize risk.

In several papers, Manski (e.g., 2005, 2021) reformulated Wald's criterion to partial identification problems into the minimize maximum regret criterion. Several authors have employed this approach, including Stoye (2009), Hirano and Porter (2009), Tetenov (2012), Kitagawa and Tetenov (2018), and Montiel Olea et al. (2025).

The objective of the policymaker or researcher is to employ statistical decision theory to optimize welfare across the entire set of feasible actions. As Manski (2021) explained, the policymaker has a predetermined choice set C and believes the true state of nature s^* lies in state space S. The welfare function, $w(\cdot, \cdot)$: $C \times S \rightarrow R^1$ maps actions and states into welfare. The policymaker cannot maximize $w(\cdot, s^*)$ because s^* is unknown, but can use observed data to shrink the state space.

The MMR criterion solves

$$\min_{c \in C} \max_{s \in S} \left[\max_{d \in C} w(d,s) - w(c,s) \right],$$

where $\max_{d \in C} w(d, s) - w(c, s)$ is the regret for action c in state s. Because we do not know the true state, we evaluate c by its maximum regret over all possible states and select the action that minimizes maximum regret. The maximum regret of an action is a measure of the maximum distance from optimality across states.

This approach is an elegant solution to the partial identification problem. However, it is relatively difficult to use. Manski (2021, pp. 2833 and 2848) observed that, "The primary challenge to use of statistical decision theory is computational… Whereas computation of regret in one state is tractable, finding maximum regret across all states may be burdensome. The state space commonly is uncountable in applications. A pragmatic process is to discretize S, computing regret on a finite subset of states that reasonably approximate the full state space." This approach can be estimated for treatment and missing data problems (with and without instrumental variables) using STATA (Litvin and Manski, 2021).

Maximum Entropy

The modern information-theoretic approach (Jaynes, 1957) is based on Shannon's (1948) communication (information) theory. It has been applied in many disciplines (such as biology, chemistry, computer science, ecology, economics, econometrics, finance, medical sciences, physics, political science, statistics, and visualizations). See, for example, Levine and Tribus (1979), Golan (2018), and Golan and Harte (2022).[2]

We concentrate on the simplest information-theoretic model. It is a constrained optimization approach with an information-theoretic decision function. That decision function is Shannon's (1948) entropy. The Shannon entropy of a random variable is the average level of uncertainty or information associated with the variable's potential states or possible outcomes, s. Let the probability of each state be $p(s) \in [0, 1]$. Shannon used axioms to derive his information measure, which is called (Shannon) entropy:[3]

$$H(p) = -\sum_{s \in S} p(s) \ln p(s),$$

where "ln" stands for the natural logarithm.

The simplest information-theoretic approach is Jaynes's (1957) classic maximum entropy (ME) model, which uses only the observed data (if any) and known theoretical information.[4] The ME approach seeks the probability distribution that maximizes Shannon entropy, subject to constraints that capture all available information, thereby determining the identified set. We illustrate how to specify constraints and then solve the optimization problem using Lagrange multipliers in the following sections.

The ME and the more general IT framework—also known as the Generalized Maximum Entropy (GME)[5] approach—are easy to implement. These inferential and modeling methods are

---

[2] For an application of an information-theoretic approach for complete information games under partial identification, see Jun and Pinske (2020).

[3] Independently, Wiener (1948) introduced other logical arguments to derive H. Cover and Thomas (2006) provide a clear discussion and simplification of these axioms.

[4] Shore and Johnson (1980), Skilling (1988), and Csiszar (1991) provide a complementary set of axioms justifying the maximum entropy (or minimum relative entropy) as an inferential method. For details and extensions, see Golan (2018). See the brief summary of the axioms in the Appendix.

[5] GME generalizes the ME approach for cases with greater uncertainty and model ambiguity (Golan, 2018). We do not examine it here.

included in many statistical packages and programming languages (such as GAMS, MATLAB, Mathematica, NLOGIT, R, Python, SAS, SHAZAM, and STATA).[6]

Prediction with Missing Data and an Unknown Observability Rate

Both the MMR and ME approaches can make predictions in situations with missing data and an unknown observability rate. When welfare is measured by square loss, and the distribution is known, the best predictor is the population mean. However, if the distribution is not known, we need an alternative approach.

Minimize Maximum Regret

Manski (2021) addressed predictions under square loss when some outcome data are missing. Consistent with the previous literature, the risk of his predictor based on sample data is the sum of the population variance of the outcome and the mean squared error (MSE) of the predictor as an estimate of the mean outcome. The regret of this predictor is its MSE as an estimate of the mean. Consequently, the MMR predictor minimizes maximum MSE. The MMR prediction of the outcome is equivalent to the minimax estimation of the mean.

Because data are missing, he proposed employing an as-if MMR prediction: Use a point estimate of the model's parameters to make a decision that would be optimal if the estimate were accurate. Suppose we know the population rate of observing outcomes but lack knowledge about the distributions of observed and missing outcomes. We treat the empirical distribution of the observed data as if it were the population distribution of observable outcomes.

We have a fixed number of observed outcomes. Given that we lack knowledge of the distribution of missing outcomes, the population mean is partially identified when the outcome is bounded. In the following simulations, the outcome y is normalized to lie in [0, 1]. The fraction of the population whose outcome is observed is $p(\delta = 1)$, and $p(\delta = 0)$ is the fraction without an observed outcome. Manski showed that the identification region for $E(y)$ is the interval $[E(y|\delta = 1)p(\delta = 1), E(y|\delta = 1)p(\delta = 1) + p(\delta = 0)]$.

The midpoint of this interval would be the MMR predictor if we knew the interval. He estimates the mid-point predictor when $p(\delta)$ is unknown and is estimated by its sample analog.

---

[6] See https://info-metrics.org/code.html for a list of programs and languages and links to codes.

Although he lacks an analytical expression for the maximum regret, Manski and Tabord-Meehan (2017) and Litvin and Manski (2021) provide a STATA algorithm for numerical computation.[7]

Maximum Entropy

An alternative approach is to use the Maximum Entropy (ME) method. Using the same notations as above, let the outcome y and δ be binary variables, each taking the values zero or one. We observe y if $\delta = 1$. Let the joint probability distribution of y and δ be p_{ij} for $i, j = 0,1$. We observe y_{01}, y_{11}, and $p(\delta = 0)$, so $p_{01} + p_{11} + p(\delta = 0) = 1$. Because p_{00} and p_{10} are unobserved, E[y] is partially identified if $p(y \mid \delta = 0) > 0$, and

$$E[y] \in \left[E(y \mid \delta = 1)\, p(\delta = 1),\ E(y \mid \delta = 1)\, p(\delta = 1) + p(\delta = 0) \right].$$

We now show that for the missing data problem (with known or unknown observability rate), the ME solution is $p_{00} = p_{10}$, so $y_0 = \frac{1}{2} p(\delta = 0) + y_{01}$ and $y_1 = \frac{1}{2} p(\delta = 0) + y_{11}$.

To solve this partially identified problem, we convert it to the classic ME formalism (Jaynes, 1957; Levine, 1980, Golan, 2018, Chapter 4). The problem is formulated as a constrained optimization with all observed and known information specified as constraints that determine the identified set, and the Shannon entropy (Shannon, 1948) is used to choose a single solution.

The only two sets of information we have are $p_{i1} = y_{i1}$, where y_{i1} are the observed values of p_{01} and p_{11} for i = 0, 1, and the normalization $\sum_{ij} p_{ij} = 1$. To choose a unique solution from the partially identified set, we maximize the Shannon entropy subject to the two sets of information in the constraints:

$$\begin{aligned} &\underset{\{P\}}{\text{Max}} \left(-\sum_{ij} p_{ij} \ln p_{ij} \right) \\ &\text{subject to} \\ &\quad y_{i1} = p_{i1}, \quad i = 0,1 \\ &\quad \sum_{ij} p_{ij} = 1. \end{aligned} \tag{1}$$

[7] Their algorithms work if y is binary. If y is distributed continuously, the sample frequencies are approximated by Beta distributions.

The corresponding Lagrangian is

$$L = -\sum_{ij} p_{ij} \ln p_{ij} + \sum_{i} \lambda_{i1} \left(y_{i1} - p_{i1} \right) + \mu \left(1 - \sum_{ij} p_{ij} \right). \tag{2}$$

The solution is

$$\hat{p}_{ij} = \frac{\exp\left(-\hat{\lambda}_{ij} \right)}{\sum_{ij} \exp\left(-\hat{\lambda}_{ij} \right)} \equiv \frac{\exp\left(-\hat{\lambda}_{ij} \right)}{\Omega\left(\hat{\lambda}_{ij} \right)}, \tag{3}$$

where λ_{ij} are the Lagrange multipliers associated with the first set of constraints (the real information) $p_{i1} = y_{i1}$, $\Omega(\cdot)$ is a normalization function (known also as the partition function), and the multipliers associated with $\delta = 0$ are zero, $\lambda_{i0} = 0$. Consequently, in the absence of additional information, the estimated probabilities p_{00} and p_{10} must be the same: $p_{00} = p_{10}$. Therefore, the estimated solution (E[y]) is the midpoint. For the missing data problem (with known or unknown observability rate), the ME solution is $p_{00} = p_{10}$, so $y_0 = \frac{1}{2} p\left(\delta = 0 \right) + y_{01}$ and $y_1 = \frac{1}{2} p\left(\delta = 0 \right) + y_{11}$.

Using the Shannon entropy as the decision function means that out of all possible inferences that are consistent with the information in Equation 1, the chosen inference is the least informed (Jaynes, 1957). It is the inference that is the closest to a uniform distribution—a state of perfect uncertainty. One can view it as the inference that it is not affected (biased) by any implicit or explicit information beyond the constraints.

Simulations

Manski (2021) conducted two simulations for a binary outcome with missing data. The first computes the maximum regret of the midpoint predictor (Table I). The second (Table II) provides the maximum regret for prediction by the sample average of the observed outcome.

For each simulation, he considered two possibilities. Either all distributions are feasible within the identified set, or we have additional bounds on the difference between the observed and missing outcomes distributions, shrinking the identified set. He imposed the bounds of $-½ \leq p(y = 1|\delta = 1) - p(y = 1|\delta = 0) \leq ½$. The identification problem is more challenging in the first case because the bounds provide additional information.

His simulations had sample sizes of 25, 50, 75, and 100, and observability rates in increments of 0.1 from 0.1 to 1. As he observed, the maximum MSE of a predictor depends on statistical imprecision and the identification problem posed by missing data. The maximum variance decreases with sample size, and the maximum squared bias falls with the observability rate $p(\delta = 1)$.

We use ME to replicate his simulation of the maximum regret of the midpoint predictor and calculate the MMR (MSE). We do not report our results because the ME and MMR results are identical (minor rounding aside) to Manski's Table I. We expect this result for the missing data problem because the midpoint is the solution for both ME and MMR.

His second simulation of the sample average predictor imposes an additional assumption: independence between y and δ. For the ME, we impose that information as an additional constraint $p_{ij} = p_i(y) \times p_j(\delta)$ or equivalently $(p_{11}/p_{01}) = (p_{10}/p_{00})$. With this additional information, our ME simulations are equivalent to his Table II, as the problem is no longer underdetermined.

Treatment Choice

We now turn to a treatment choice problem. Suppose that a policymaker wants to assign people to different potential treatments. The problem depends on the number and types of treatments as well as how the experiment is conducted, which affects the observable information.

Here, we use the problem from Manski (2021, Section 5.2). The policymaker wants to select a single treatment for the entire population using knowledge of the distribution of realized outcomes from a sample.

Unlike in the previous missing data problem simulation, the ME and Manski's MMR results are not quantitatively identical. However, both approaches make the same treatment recommendation for this specific empirical problem, as our simulation results show. That is, they are qualitatively the same.

In Manski's problem, individuals are assigned to different treatments a or b based on knowledge of the distribution of observed outcomes in a sample. Each person in the study population has the potential outcome or welfare $y(a)$ or $y(b)$. A binary indicator $[\delta(a), \delta(b)]$ indicates whether these outcomes are observed.

Because we do not observe the counterfactual outcomes, the possible indicator values are $[\delta(a) = 1, \delta(b) = 0]$ and $[\delta(a) = 0, \delta(b) = 1]$. We observe only realized outcomes, so the

probabilities add to one: $p[\delta(a) = 1] + p[\delta(b) = 1] = 1$. The state s indicates a possible distribution $p_s[y(a), y(b), \delta(a), \delta(b)]$ of outcomes and observability. The decision maker chooses treatments in a population with the same distribution of treatment response as in the study population to maximize welfare.

To illustrate this problem, Manski used data from Manski and Nagin's (1998) analysis of two sentencing options for juvenile offenders in Utah. At a judge's discretion, some offenders are sentenced to residential confinement. a, while others are not confined, b. They observed recidivism, y, within two years of sentencing, where $y = 1$ if a youth did not commit a new offense, and $y = 0$ otherwise.

Within the sample, 11% of offenders were confined, a. The sample probability (frequency) was $p[(y(a)|\ (\delta(a) = 1) = 1] = 0.23$ for those who were confined and did not offend again. Of the remaining 89% who were not confined, $p[(y(b)|\ (\delta(b) = 1) = 1] = 0.41$.

Two possible counterfactual policies could replace judicial discretion with a mandate that either all offenders or none are confined. Manski (2021) used what he called the asymptotic minimax-regret (AMMR) rule. The AMMR chooses a treatment if its probability exceeds one-half.

We now show how to determine a rule using ME. The partial information consists of the individuals' choices given the treatments. We do not include any additional information or assumptions. We can approach this problem in two ways. We can estimate the model for both treatments simultaneously or examine each treatment separately. We do the latter because the presentation is simpler.

We use the treatment information to construct the distributions of a and b over zero and one. Then, we obtain the joint distribution of a and b given independence. For treatment a, let $a_0 = a^*_{01} + a_{02}$ and $a_1 = a^*_{11} + a_{12}$ where the star stands for observed information, and the second element in each equation is the unobserved counterfactual. For example, a_{02} is the probability that a person who received treatment a would have chosen action zero if they received treatment b. Similarly, a_{12} is the unknown counterfactual for an individual who received treatment b and chose 1 if they received treatment a. We do not know the counterfactuals a_{02} and a_{12}. Let $p[y(a) = 0] = p[(y(\text{a})|\delta(\text{a}) = 1) = 0] + p[y(b)|\delta(b) = 1]$, $p[y(a) = 1] = p[(y(a)|\delta(a) = 1) = 0] + p[y(b)|\delta(b) =$

1] and $p[y(a) = 0] + p[y(a) = 1] = 1$. We use analogous notation for $y(b)$.The p's are unknown and the y's are observed.

We now connect the observables and unobservable (counterfactuals) $y_0(a) = y_{01}(a) + y_{02}(a)$ and $y_1(a) = y_{11}(a) + y_{12}(a)$, where the first term to the right of each equality is observed, and the other terms are not (counterfactuals). For $i = 0, 1$ and $j = 1, 2$ (and suppressing the "a" for now), we specify the information we have in the two constraints: $y_i = y_{i1} - y_{i2}$ and $\sum_i y_i = 1$. Reorganizing, $y_{i1} = p_i - p_{i2} = \sum_j p_{ij} - p_{i2}$ because $p_i = \sum_j p_{ij}$ and $\sum_{ij} p_{ij} = 1$.

The ME model is

$$\begin{aligned} &\underset{\{p_{ij}\}}{\text{Max}} - \sum_{ij} p_{ij} \log\left(p_{ij}\right) \\ &\text{subject to} \\ &\quad y_{i1} = \sum_j p_{ij} - p_{i2},\ i = 0,1 \\ &\quad 1 = \sum_{ij} p_{ij}. \end{aligned} \tag{4}$$

The corresponding Lagrangian is

$$L = -\sum_{ij} p_{ij} \log\left(p_{ij}\right) + \sum_i \lambda_i \left(y_{i1} - \sum_j p_{ij} + p_{i2}\right) + \mu\left(1 - \sum_{ij} p_{ij}\right).$$

The estimated probabilities are

$$p_{ij} = \begin{cases} \exp\left(-\lambda_i\right)/\Omega \ \text{ for } j = 1 \\ \exp\left(0\right)/\Omega \ \text{ for } j = 2, \end{cases}$$

where $\Omega = \sum_{ij} \exp\left(\cdot\right)$ is the normalization factor.

In this empirical example, $y_{01} = p(0 \mid a) \times p(a) = 0.77 \times 0.11 = 0.0847$ in Equation (4). The observed y_{11} is calculated from the data in a similar way.

Substituting b for a gives a similar model for treatment b. Using the data, the solutions are:

$\hat{p}_{a_0} = 0.5297 = 0.0847 + 0.4450$ and $\hat{p}_{a_1} = 0.4703 = 0.0253 + 0.4450$

$\hat{p}_{b_0} = 0.5801 = 0.5251 + 0.0550$ and $\hat{p}_{b_1} = 0.4199 = 0.3649 + 0.0550$.

By construction, the estimated values of the unknown entities for a and b (say, p_{02}(a) and p_{12}(a), or p_{02}(b) and p_{12}(b)) are distributed uniformly. Because we lack information in the model about that missing information, it must be distributed equally.

The joint distribution is

	0	1
a	0.2806	0.2212
b	0.3365	0.1763

Table 1 shows the conditional probabilities under two scenarios: sentencing as in the sample or random sentencing. The second column shows the results based on the sample's assignment frequencies (11% and 89%), where judges have discretion in sentencing. To simplify notations, let $p(1|a)$ be the conditional probability of $y = 1$ given that the individual was confined, treatment a. Similarly, let $p(1|b)$ be the conditional probability of $y = 1$ given treatment b. Because $p(1|a) > p(1|b)$, treatment a (confinement) is preferred to treatment b (no confinement) if the objective is to minimize recidivism (bold numbers, column 2). The MMR analysis draws the same conclusion.

Table 1

The Conditional Probabilities Under Two Scenarios:

Sample's Frequencies and Random Assignments

Conditional Probability	Sample Sentencing	Random Sentencing
$P(0|a)$	0.5592	0.7517
$P(1|a)$	**0.4408**	**0.2483**
$P(0|b)$	0.6562	0.5893
$P(1|b)$	**0.3438**	**0.4107**

In the third column, judges sentence offenders randomly to treatments a and b. Here, $p(a) = p(b) = 0.5$. The ME results (the conditional probabilities in column 3) change so that $p(1|b) > p(1|a)$. Again, we draw the same recommendation as the MMR analysis of the same problem.

In Table 2, we apply our ME approach to the simulations in Manski's Table III, using his random-treatment-assignment assumption. In panel A, all distributions are feasible. Panel B

imposes Manski's additional restriction that $-\frac{1}{2} \leq p[y = 1 \mid \delta = 1] - p[y = 1 \mid \delta = 0] \leq \frac{1}{2}$, which shrinks the identification set. The sample size (*N*) in both panels ranges from 25 to 100 in increments of 25. Each column has a value of *p* ranging from 0.5 to 0.9 in increments of 0.1.[8]

The top number in each cell is the MMR from the ME simulation. The number in parentheses below it is the difference between the mean squared error calculated using AMMR (Manski, Table III) and the MSE using ME. Thus, a positive number indicates that the ME approach has a lower MSE than the AMMR approach. Thus, the ME approach has a lower MSE (AMMR) for all cases except the first two columns of the *N* = 25 row of panel B (with the extra restriction).

The ME and AMMR tables share certain properties. Maximum regret does not vary substantially with *p*. For a given *p*, maximum regret rises with *N* in Manski's Table III (see Manski for an explanation), while decreasing slightly in most ME columns. The MSE is smaller in panel B than in panel A because the restriction in panel B provides more information, which reduces the identification set. For the ME, the restriction reduces the MMR only moderately for *p* = 0.5, 0.6, and 0.7 and *N* = 25, but has no effect for all other *p*'s and *N*'s. In contrast, the additional information has a greater effect on the MMR approach.

Table 2

Maximum Regret (MSE) Using Maximum Entropy

Panel A

	p				
Sample Size	0.5	0.6	0.7	0.8	0.9
25	0.2980	0.2974	0.2971	0.2888	0.2766
	(0.0436)	(0.0467)	(0.0450)	(0.0540)	(0.0703)
50	0.2934	0.2937	0.2956	0.2868	0.2749
	(0.0813)	(0.0845)	(0.0817)	(0.0924)	(0.1028)
75	0.2886	0.2915	0.2949	0.2886	0.2724
	(0.1019)	(0.0972)	(0.0950)	(0.1066)	(0.1175)
100	0.2869	0.2947	0.2972	0.2918	0.2766
	(0.1154)	(0.1074)	(0.1039)	(0.1108)	(0.1256)

Panel B

Imposing $-\frac{1}{2} \leq p[y = 1 \mid \delta = 1] - p[y = 1 \mid \delta = 0] \leq \frac{1}{2}$

[8] As Manski notes, we do not need to consider *p* values below 0.5 because the state space in each panel views the treatments symmetrically, so the maximum regret is the same for *p* and 1 – *p*.

	p				
Sample Size	0.5	0.6	0.7	0.8	0.9
25	0.2980	0.2958	0.2954	0.2876	0.2766
	(–0.0208)	(–0.0053)	(0.0087)	(0.0202)	(0.0547)
50	0.2934	0.2937	0.2956	0.2868	0.2749
	(0.0192)	(0.0353)	(0.0465)	(0.0666)	(0.0897)
75	0.2886	0.2915	0.2948	0.2886	0.2724
	(0.0391)	(0.0482)	(0.0573)	(0.0762)	(0.1050)
100	0.2869	0.2947	0.2972	0.2918	0.2766
	(0.0533)	(0.0630)	(0.0759)	(0.0934)	(0.1161)

A Graphical Interpretation

We now use figures to illustrate how the information-theoretic model works in an unconditional multinomial problem. Let x_k, $k = 1, 2,$ or 3, be realizations of a discrete random variable X. We want to infer the probability distribution p.

Although we have no uncertainty, the following problem is partially identified because we have insufficient information. Our only information is that the (nonnegative) probabilities sum to one, and the arithmetic mean value, after N trials, is $y = \sum_k p_k x_k = \sum_k p_k k$.

Like any partially identified (or underdetermined) problem, this one can be transformed into a well-posed, decision-theoretic constrained optimization problem. Assuming the decision function is well-behaved (e.g., it is concave), the estimated solution will be unique. Using the Shannon entropy as our decision function, the ME is

$$\begin{aligned} &\underset{\{P\}}{\text{Maximize}} \left\{-\sum_{k=1}^{3} p_k \ln\left(p_k\right)\right\} \\ &\text{subject to} \\ &\quad y = \sum_k p_k k \\ &\quad \sum_{k=1}^{K} p_k = 1 \\ &\quad p_k \geq 0, \text{ for } k = 1, 2, 3. \end{aligned} \tag{5}$$

The equality conditions, also known as *conservation rules* or *conservation laws*, capture all the information used in the inference. These are the rules that govern the behavior of the underlying system or distribution. If we know the data generating process and specify the constraints accordingly, then these constraints are sufficient statistics.

Constructing the Lagrangian and solving yields the optimal solution,

$$\hat{p}_k = \frac{\exp\left(-\hat{\lambda}k\right)}{\sum_k \exp\left(-\hat{\lambda}k\right)} \equiv \frac{\exp\left(-\hat{\lambda}k\right)}{\Omega\left(\hat{\lambda}\right)}, \tag{6}$$

where $\hat{\lambda}$ is the estimated Lagrange multiplier associated with the mean constraint, $\equiv$ indicates a definition, and $\Omega(\lambda)$ is the normalization. The Lagrange multiplier associated with the normalization constraint, λ_0, is a function of the other multiplier.

Figure 1 shows the three-dimensional discrete choice problem and solution over the complete simplex. The vertices V_1, V_2, and V_3 are the extreme distributions (1, 0, 0), (0, 1, 0), and (0, 0, 1) for $k = 1, 2, 3$, respectively. Every point in the simplex (including on the boundaries) is a normalized probability distribution, $\sum_k p_k = 1$, where the values p_k, $k = 1, 2, 3$, correspond to the distances from the sides V_2V_3, V_1V_3, and V_1V_2, respectively.

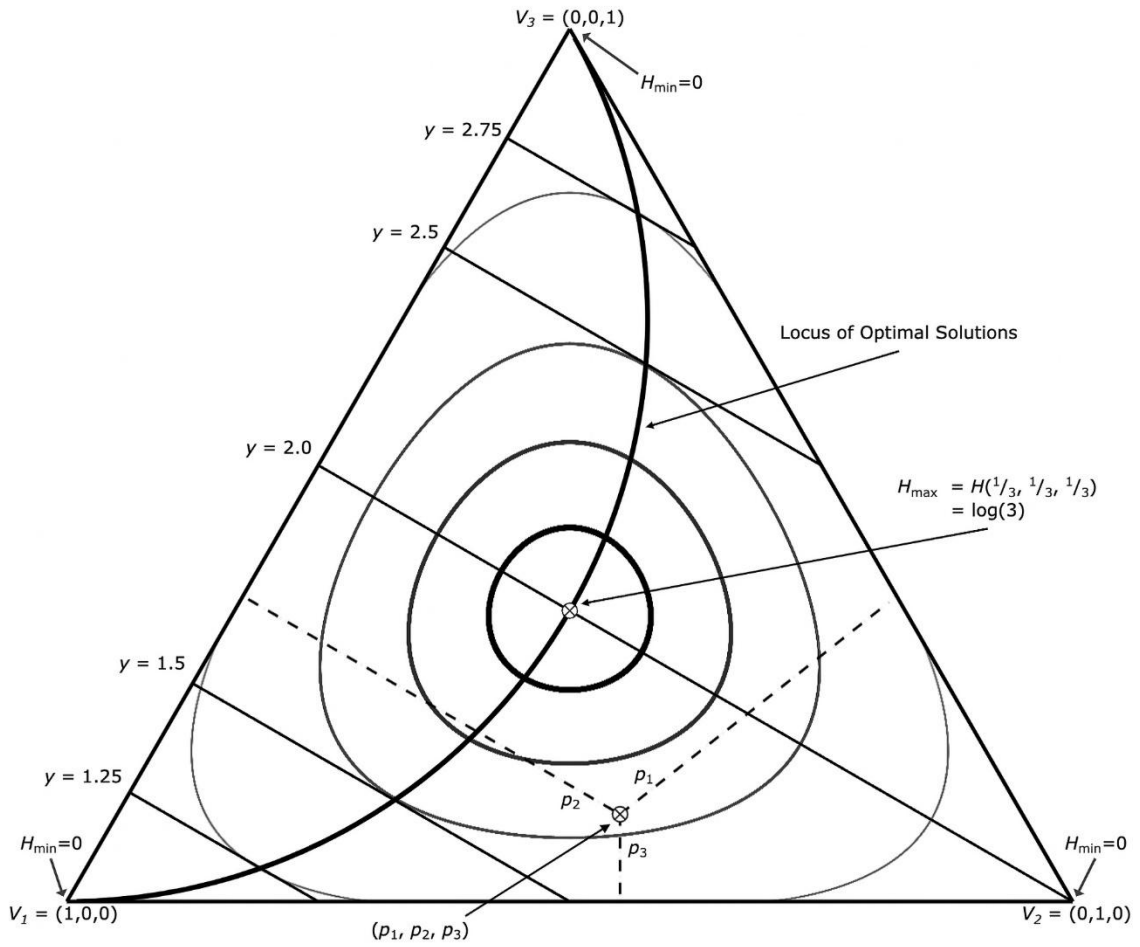
The point (circle) in the lower right of the simplex is one such distribution, $p = (p_1, p_2, p_3)$, where $p_2 > p_1 > p_3$. The midpoint (the center of gravity) of the simplex, $H_{\max}$, represents the uniform distribution ($p_1 = p_2 = p_3$). It has the maximum entropy of $\ln(3) = -\sum_{k=1}^{3} p_k \ln\left(p_k\right) = -3\left(\frac{1}{3}\ln(\frac{1}{3})\right)$. The points at the three vertices have the minimum value of the Shannon entropy, $H_{\min} = 0$, because they have no uncertainty.

The various straight lines across the simplex represent possible linear constraint sets for some values of $y = \sum_k p_k k$. For example, the constraint $y = 2$ is the straight line from the middle of the V_1V_3 side (the left side of the simplex) to the V_2 vertex. It shows all the points with a mean value of 2. The identified set for the $y = 2$ case is the line $2 = \sum_{k=1}^{3} p_k k$.

The contours ("indifference curves") connect distributions with equal entropy. The darker and thicker the contours, the higher the entropy. Contours far away from the center have lower entropy. The dark, heavy curve is the locus of optimal solutions, connecting the distributions with maximum entropy under the constraints for all values of $y \in [1,3]$.

Figure 1

Graphical representation of the three-sided die problem and solution over the complete simplex space



The Decision Function Effect

The ME solution with $H(p)$ as the decision function is the one out of an unlimited number of p's that satisfy the information provided in the constraints (the identified set). It is the least informed solution: The one that is as close as possible to the uniform distribution, which captures maximum uncertainty. It has the flattest possible likelihood that is consistent with the observed expectation values (constraints), as Zellner (1997) argued.

This likelihood is not assumed to be known a priori. Rather, it is a direct consequence of the joint choice of the decision function and the structure of the information imposed. We could use other decision functions to choose the estimated solution. Any other decision function introduces additional information, thus restricting the possible solution space.

In many partial identification problems, we want to compare models and SDFs. We want each model to use the same information except for a single ingredient. Here, that ingredient is the decision function. Using the same problem and information as in Figure 1, Figure 2 adds the loci of optimal solutions for five commonly used approaches, each with a different SDF.

These approaches are all special cases of the Rényi entropy, a generalized information measure. To describe the gain of information, Rényi (1961) developed an entropy measure of order α for incomplete random variables. An incomplete, discrete random variable with K distinct realizations, each with $p_k > 0$ ($k = 1,\ldots, K$), is defined such that $\sum_k p_k \leq 1$, rather than $\sum_k p_k = 1$. We can normalize such an incomplete random variable so it sums to one. Rényi's generalized entropy measure for a normalized probability distribution of order α is

$$H_\alpha^R(p) = \frac{1}{1-\alpha}\log\sum_k p_k^\alpha \,. \tag{7}$$

The Rényi relative entropy (between two distributions p and p^0 for the discrete random variables X and Y) of order α is

$$D_\alpha^R(X \mid Y) = D_\alpha^R\left(p \,||\, p^0\right) = \frac{1}{1-\alpha}\log\sum_k \frac{p_k^\alpha}{p_k^{0,\alpha-1}}. \tag{8}$$

A similar generalized entropy measure is the Cressie-Read (1984) measure:

$$D_\alpha^{CR}\left(p \,|||\, p^0\right) = \frac{1}{\alpha(1+\alpha)}\sum_k p_k\left[\left(\frac{p_k}{p_k^0}\right)^\alpha - 1\right]. \tag{9}$$

From an inferential point of view, the Rényi of order α is equivalent to the Cressie-Read of order α–1: $D_\alpha^R\left(p \,||\, p^0\right) = D_{\alpha-1}^{CR}\left(p \,||\, p^0\right)$.

Figure 2 presents the loci of optimal solutions for the following statistical decision functions (and a uniform p^0), which are all special cases of the Rényi function:

- Shannon entropy, where $\alpha \to 1$,
- Least squares (LS): $D_2^R\left(p \,||\, p^0\right) = D_1^{CR}\left(p \,||\, p^0\right)$,
- Empirical likelihood (EL): $D_0^R\left(p \,||\, p^0\right) = D_{-1}^{CR}\left(p \,||\, p^0\right)$, where $\alpha \to -1$,
- Rényi with α = 3, 4, and 5 (or similarly, Cressie Read with of α = 2, 3, and 4).

Figure 2 illustrates the differences across the complete solution space for all possible values of $y \in [1,3]$. All the loci intersect the center of gravity, the uniform distribution, where the expected value is exactly 2. Elsewhere, the various methods have different solutions. Each locus is symmetric about the center of gravity. This figure illustrates two points.

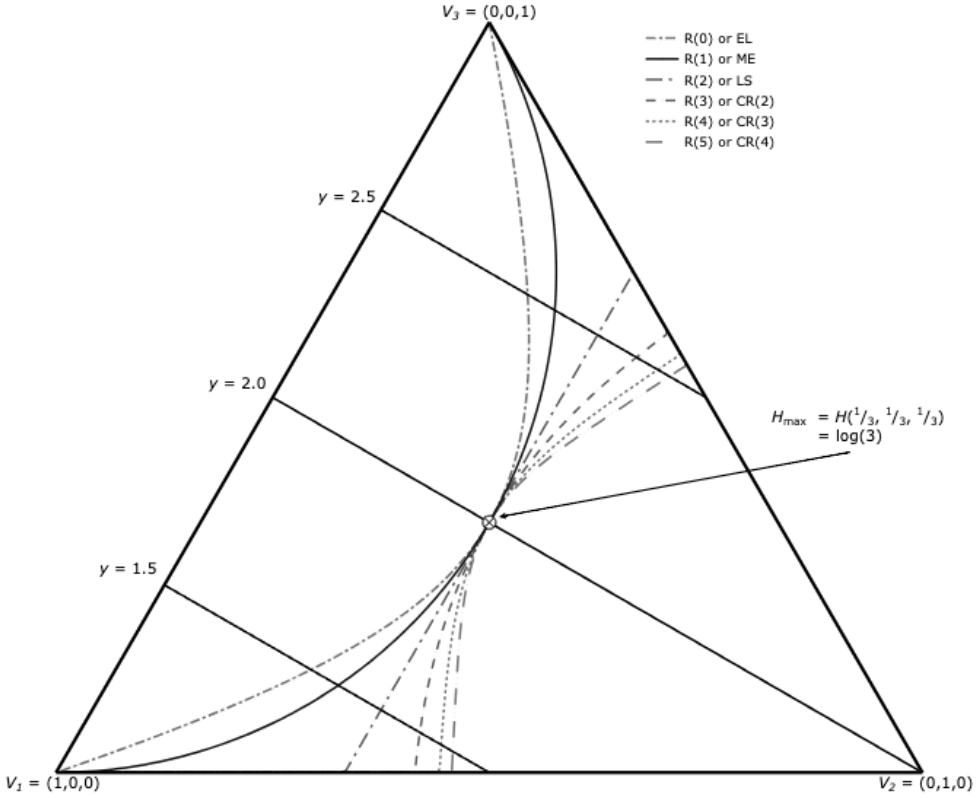
First, the differences among the methods are due only to the differences in the decision functions. The figure provides a visualization of the amount of information introduced by each decision function relative to the ME.

Second, among these six methods, the maximum entropy locus (solid, dark line) is the most uninformed. It is the one that is closest, in information units, to the uniform distribution at each point within the simplex. The empirical likelihood (dashed, with dots, light grey line) has the same basic shape as the ME, but is less uniform due to additional information in its SDF. In fact, empirical likelihood is a power law. All the others have extreme solutions (one of the probabilities is zero) for more extreme values of *y*. In this simple example, the least square solution does not exist for values below 4/3 or above 8/3. These extreme solutions mean that in the inferred distribution, one of the events is ruled out—an unrealistic outcome.

This example demonstrates that information theory provides a means for ranking estimators based on the information embedded in their decision functions or likelihoods, which are composed of the constraints (including assumptions imposed) together with the decision function.

Figure 2

Comparison of the three-sided die problem for six decision functions



Concentrated Model

We can use duality theory to convert the ME problem from a constrained to an unconstrained optimization problem. It is a transformation from the probability space to the Lagrange multiplier space. The concentrated ME model estimates the minimal set of parameters required to fully characterize the system. Because we are dealing with partially identified problems, the number of Lagrange multipliers is much smaller than the dimension of the probabilities. In addition to being computationally simple and efficient, the concentrated model provides a straightforward way to compare the ME with other inferential approaches. These comparisons are straightforward because the concentrated model is formulated similarly to a likelihood.

Six-Sided Die Simulation

To select between multiple solutions in an underdetermined or partially identified problem, a researcher chooses an SDF. The following simulations compare four SDFs: Shannon entropy, empirical likelihood, least squares, and Rényi entropy of order three: $H_3^R = -\frac{1}{2}\ln\left(\sum_{k=1}^{K} p_k^3\right)$.

Suppose we have a discrete random variable, X, with K possible value, where each realization x_k has the probability p_k. If we know or observe only two pieces of information and $K > 2$, the problem is partially identified. For simplicity, and without loss of generality, we assume that $K = 6$, the six-sided die problem: $x_k = k$. The two pieces of information are the normalization of the nonnegative probabilities $\left(\sum_k p_k = 1\right)$ and the first moment or expected value $\left(\sum_k p_k k = y\right)$.

In our simulations, each sample is generated from a uniform or one of three normal distributions. The data generating processes (DGP) for the three normal models are N(3.5, 3), N(3.5, 7), and N(2, 5).

For each of 2,000 samples, we calculate the expected value for the true p_k's. Using a grid within the simplex, we ensure that the data cover the complete parameter space $y \in (1, 6)$ for each sampling experiment. However, the concentration of data points about the mean depends on the DGP's variance.

We use our four SDFs to infer the probability distribution given the two pieces of information. We show the minimum maximum regret (MMR) using the mean squared error (MSE). We also conducted the same experiments using the Kullback-Leibler (KL) criterion with the same qualitative results.[9] We compute the MMR for each SDF by assuming it is correct and comparing the MMR of each of the other SDFs. Then, we select the maximum value for each case and choose the minimum from those, which are bold values in the tables.

We consider two scenarios: misspecified models and correctly specified models. A correctly specified model is one where the functional form of the inferential model is the same as that of the DGP. For example, consider trying to infer the probability distribution based on several known moment conditions. It is a partially identified problem that we formulate as a

[9] All simulation results are available upon request.

constrained optimization problem. To do so, we need to (i) specify the information we have in terms of the constraints and (ii) choose a decision function.

Assume we are trying to infer the probability distribution of data generated from an exponential distribution. Theory tells us that the sufficient statistic is the arithmetic mean. Specifying our constraint as an arithmetic mean and using Shannon entropy as the decision function yields the exponential distribution with a single parameter (the Lagrange multiplier of the mean constraint).

If, instead, we know that the DGP is a power law (say, a Pareto distribution), theory tells us to use the geometric mean as the constraint (in addition to a normalization). Because it is a sufficient statistic, we are using the minimal necessary constraints/information. The ME in that case will yield a single-parameter (the Lagrange multiplier) power law.

For normally distributed data, we must impose more than just the mean constraint. We need both the mean and the variance constraints in addition to the normalization. Maximizing Shannon entropy subject to these three restrictions yields the normal distribution. These examples are correctly specified models. That is, if we know the true underlying distribution and impose the sufficient statistics as constraints, the resulting distribution is correctly specified.

However, with a normal DGP, if we impose only the normalization and mean constraints, but not the variance constraint, the ME is misspecified. Using too few constraints has the same effect on the other models. Since the real DGP is unknown in practice, this experiment on partially identified problems investigates the behavior of both correctly and incorrectly specified models. (We can also think of these two scenarios as completely or incompletely specified models.) Our sampling experiments show that under both scenarios, the ME minimizes the maximum regret.

Table 3 presents the MMR under an MSE loss function for our four SDFs: least squares (LS), $\sum_k p_k^2$; empirical likelihood (EL), $\sum_k \log\left(p_k\right)$; Rényi entropy of order 3; and Shannon entropy. The columns show the true SDF, and the rows show the comparison SDF. The right-hand side column shows the MMR.

Table 3
MMR Using MSE for the Six-sided Die Problem for Misspecified and Well-specified Models

(The bold "Max" number in the right-hand side column is the global MMR for each scenario. The values of the MMR are multiplied by 1,000. For the normal DGP experiments, "Mean" indicates that only the mean constraint was imposed, so the model is misspecified. "Var" indicates that both the mean and variance constraints were imposed, so the model is correctly specified under ME. Normalization is always imposed.)

MSE Experiments	**True objective function**				
U/Mean	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	0.6376	1.0972	0.1852	1.0972
LS	0.6376	0.	0.1565	0.1355	0.6376
Renyi	1.0972	0.1565	0.	0.4902	1.0972
Shannon	0.1852	0.1355	0.4902	0.	**0.4902**
Norm/Mean N(3.5,3)	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	0.0281	0.0296	0.0399	0.0399
LS	0.0281	0.	0.0373	0.0283	**0.0373**
Renyi	0.0296	0.0373	0.	0.0238	**0.0373**
Shannon	0.0399	0.0283	0.0238	0.	0.0399
Norm/Mean N(3.5,7)	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	0.7763	1.3048	0.2259	1.3053
LS	0.7763	0.	0.1737	0.1647043	0.7763
Renyi	1.3048	0.1737	0.	0.5643095	1.3048
Shannon	0.2259	0.1647	0.5643	0.	**0.5643**
Norm/Mean N(2,5)	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	1.1377	1.8350	0.3311	1.8350
LS	1.1377	0.	0.2126	0.2414	1.1377
Renyi	1.8350	0.2126	0.	0.7432	1.8350
Shannon	0.3311	0.2414	0.7432	0.	**0.7432**
Norm/Var N(3.5,3)	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	0.5098	0.9383	0.1723	0.9383
LS	0.5098	0.	0.1073228	0.1310	0.5098
Renyi	0.9383	0.1073	0.	0.4373	0.9383
Shannon	0.1723	0.1310	0.4373	0.	**0.4373**
Norm/Var N(3.5,7)	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	1.2380	2.1213	0.4298	2.1213
LS	1.2380	0.	0.2380	0.3016	1.2380
Renyi	2.1213	0.2380	0.	0.9362	2.1213
Shannon	0.4298	0.3016	0.9362	0.	**0.9362**

Norm/Var N(2,5)	EL	LS	Renyi	Shannon	Max
Comparison objective function					
EL	0.	1.8033	3.0660	0.6317	3.0660
LS	1.8033	0.	0.3411	0.4435	1.8033
Renyi	3.0660	0.3411	0.	1.3559	3.0660
Shannon	0.6317	0.4435	1.3559	0.	**1.3559**

The Shannon criterion has the lowest MMR in all the simulations where it is correctly specified (normal with mean and variance constraints) and in all the other simulations except for the DGP N(3.5, 3). In that case, the mean lies at the center of the 6-dimensional simplex, and the variance is relatively small, so data points are concentrated about the unbiased mean. Under MSE, all decision functions differ only slightly in their MMRs. The results also show, as expected, that the correctly specified model (Shannon decision function with mean and variance constraints for a normal DGP) dominates the misspecified models (only mean constraint).

Conclusions

Partial identification problems are ubiquitous. Researchers may use several statistical decision function (SDF) approaches. We add one more: an information-theoretical approach. We illustrate it using the classical information-theoretical method, maximum entropy (ME), based on the Shannon entropy SDF.

We use three sets of simulations to compare the ME approach to other SDFs. In the missing data problem simulation from Manski (2021), the ME produces the same result as the Wald-Manski minimum maximum regret (MMR) approach. In Manski's treatment problem, the ME and MMR approaches produced the same qualitative results, but ME had a lower mean squared error (better MMR). In the six-sided die (unconditional multinomial problem) simulations, the Shannon SDF outperformed least squares, empirical likelihood, and Rényi of order three in terms of both MMR (mean squared error) and Kullback-Leibler criteria in all but one simulation experiment.

Of course, we do not know whether these results would hold more generally. However, these results illustrate that ME is a viable alternative to MMR and other SDFs for several partial identification problems.

The information-theoretic method has four attractive features. First, its SDF and inferential method have an axiomatic underpinning.

Second, it is easy to estimate using any of several statistical packages and programming languages. Estimation is quick because (i) it does not require Monte Carlo integration or other time-consuming computational methods, and (ii) it can be formulated and solved in its dual, unconstrained form.

Third, it can be applied to any problem, not just those with a discrete number of choices. Thus, a researcher does not need to discretize the state space.

Fourth, it is easy to impose additional constraints (often simply by writing them explicitly in a single line of the program). This simplicity facilitates bottom-up comparisons where one adds assumptions one at a time, making this method particularly well-suited for partial identification problems.

References

Cover, Thomas M., and Joy A. Thomas (2006): *Elements of Information Theory*. John Wiley & Sons, 2nd edition.

Cressie, Noel, and Timothy R. C. Read (1984): "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society*, 46, 440–464.

Csiszar, Imre (1991): "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems, *The Annals of Statistics*, 19, 2032–2066.

Golan, Amos (2018): *Foundations of Info-Metrics: Modeling, Inference, and Imperfect Information*, Oxford University Press.

Golan, Amos, and John Harte (2022), Information theory: A foundation for complexity science, *Proc. Natl. Acad. Sci. U.S.A.* 119 (33) e2119089119, https://doi.org/10.1073/pnas.2119089119.

Golan, Amos, George G. Judge, Douglas Miller (1996): *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, John Wiley & Sons, Inc.

Golan, Amos, and Jeffrey M. Perloff (2002): "Comparison of Maximum Entropy and Higher-Order Entropy Estimators," *Journal of Econometrics*, 107, 195–211.

Good, Irving J. (1963): "Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables, *Annals of Mathematical. Statistics*, 34, 911–934.

Haavelmo, Trygve (1944): "The Probability Approach in Econometrics," *Econometrica*, 12 (Supplement), iii–vi and 1–115.

Hirano, Keisuke, and Jack R. Porter (2009): "Asymptotics for Statistical Treatment Rules," *Econometrica*, 77, 1683–1701.

Jaynes, Edwin T. (1957): "Information Theory and Statistical Mechanics," *Physics Review*, 106, 620–630.

Jun, Sung Jae, and Joris Pinske (2020): "Counterfactual prediction in complex information games: Point prediction under partial identification," *Journal of Econometrics*, 216, 394–429.

Kitagawa, Toru, and Aleksey Tetenov (2018): "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 86, 591–616.

Kullback, Solomon, and Richard A. Leibler (1951): "On Information and Sufficiency," *Annals of Mathematical Statistics,* 22, 79–86.

Leamer, Edward E. (1985): “Sensitivity Analyses Would Help,” *American Economic Review*, 75, 308–313.

Levine, Raphael D. (1980): “An Information Theoretical Approach to Inversion Problems,” *Journal of Physics A: Mathematical and General*, 13, 91–108.

Levine, Raphael D., and Myron Tribus (1979): *The Maximum Entropy Formalism*. MIT Press, Cambridge, MA.

Litvin, Valentyn, and Charles F. Manski (2021): “Evaluating the Maximum Regret of Statistical Treatment Rules with Sample Data on Treatment Response,” *STATA Journal*, 21,: 97–122.

Manski, Charles F. (2004): “Statistical Treatment Rules for Heterogeneous Populations,” *Econometrica*, 72, 1221–1246.

Manski, Charles F. (2005): “Partial Identification with Missing Data: Concepts and Findings,” *International Journal of Approximate Reasoning*, 39, 151–165. h

Manski, Charles F. (2021): “Econometrics for Decision Making: Building Foundations Sketched by Haavelmo and Wald,” *Econometrica*, 89, 2827–2853.

Manski, Charles F., and Max Tabord-Meehan (2017): “Evaluating Maximum MSE of Mean Estimators with Missing Data,” *Stata Journal*, 17, 723–735.

Manski, Charles F., and Daniel S. Nagi (1988): “Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism,” *Sociological Methodology*, 28, 99–137.

Manski, Charles F., Alan H. Sanstad, and Stephen J. DeCanio (2021): “Addressing Partial Identification in Climate Modeling and Policy Analysis,” *Proceedings of the National Academy of Sciences*, 118, e2022886118. doi: 10.1073/pnas.2022886118.

Montiel Olea, José Luis, Chen Qiu, and Jörg Stoye, “Decision Theory for Treatment Choice Problems with Partial Identification,” manuscript, 2025.

Rényi, Alfred (1961): “On Measures of Information and Entropy,” *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, I, 547–561.

Shannon, Claude E. (1948):, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, 27, 379–423.

Shore, John E., and Rodney W. Johnson (1980): "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions on Information Theory*, IT-26, 26–37.

Skilling, John (1988): "The Axioms of Maximum Entropy," Erickson, G.J., Smith, C.R. (eds), *Maximum-Entropy and Bayesian Methods in Science and Engineering. Fundamental Theories of Physics*, 31–32, Springer, Dordrecht, 173–187.

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

Tamer, Elie (2010): "Partial Identification in Econometrics," *Annual Review of Economics* 2, 167–195.

Tetenov, Aleksey (2012): "Statistical Treatment Choice Based on Asymmetric Minimax Regret Criteria," *Journal of Econometrics*, 166, 157–165.

Wald, Abraham (1945): "Statistical Decision Functions Which Minimize the Maximum Risk," *Annals of Mathematics*, 46, 265–280.

Wiener, Norbert (1948): *Cybernetics or Control and Communication in the Animal and the Machine,* The Technology Press, John Wiley & Sons, Inc.

Zellner, Arnold (1997): "The Bayesian Method of Moments (BMOM): Theory and Applications," in *Advances in Econometrics: Applying Maximum Entropy to Econometric Problems*, Vol. 12, T. B. Fomby and R. C. Hill (eds.), Greenwich, CT: JAI Press, 85, 106.

Appendix: Axioms

The following is a brief summary of the axioms behind the entropy functional itself (Shannon 1948; Wiener, 1948) and the IT method of inference. As Golan (2018) shows, various sets of axioms underlie the classical IT approach. Some are defined on the decision function itself, others on the inference itself, and one is based on a symmetry condition (Shore and Johnson, 1980; Skilling, 1988; Csiszar, 1991; and Golan, 2018). Golan and Perloff (2002) presented a modification of the axioms for the more general framework (with ambiguity, misspecification, and flexible constraints).

Here, we use a set of axioms based on Shore and Johnson (1980), Skilling (1988), and Csiszar (1991), and treat the inference as an optimization problem. We start by characterizing the five properties (axioms) we want our method of inference to possess, which determine a unique inference approach: the information-theoretic method we use in this paper.

To simplify the exposition, consider the linear problem $\boldsymbol{y} = X\boldsymbol{p}$, where $\boldsymbol{y}$ is an M-dimensional vector of expectation values, X is an $M \times K$ matrix of rank M, and $\boldsymbol{p}$ is a K-dimensional vector whose components are the unknown probabilities p_k we wish to infer. These axioms are general and apply to other functions and constraints, such as those in this paper.

The five axioms represent a minimum set of requirements for a logically consistent method of inference from a finite data set. Following Skilling, we define a distribution $f(\boldsymbol{x})$ as a positive, additive distribution function (PADF). It is positive by construction: $f\left(x_k\right) = p_k \geq 0$ for each realization x_k, $k = 1, 2, \ldots, K$, and strictly positive for at least one x_k. It is additive in the sense that the probability in some well-defined domain is the *sum* of all the probabilities in *any* decomposition of this domain into sub-domains. A PADF lacks the property of normalization: $\sum_k p_k = 1$.[10] The inference question can be thought of as a search for those PADFs that best characterize the finite data set. Note that working with PADFs is not necessary but it allows us to avoid the complexity of normalizations, which simplifies the analysis.

Our objective is to identify an estimate that is the best according to some criterion. We want a transitive ranking of estimates so we can determine which estimate optimizes a certain

[10] One can work with an improper probability distribution, $\boldsymbol{p}^*$, that sums up to a number $s < 1$, by normalizing so that $p_k = p_k^* / \sum p_k^* = p_k^* / s$.

decision function. The following axioms are used to determine the exact form of that decision function, while requiring that function to be independent of the data. Let $f(I; \boldsymbol{q})$ be the estimates provided by maximizing some function H with respect to the available data $I(\boldsymbol{y}; X)$, given some prior model $\boldsymbol{q}$. Below, we refer to $f(I; \boldsymbol{q})$ as the 'posterior' (or 'post-data'). The five axioms are:

> **A1.** ('Best' posterior: completeness, transitivity, and uniqueness). *All posteriors can be ranked, the rankings are transitive, and, for any given prior and data set, the 'best' posterior (the one that maximizes the decision function H) is unique*.
>
> **A2.** (Permutation or coordinate invariance). *Let H be any unknown criterion, and f(I; **q**) be the estimate obtained by optimizing H subject to the information set I (data) and prior model **q**. For Δ, a coordinate transformation,* $\Delta f\ I, q\ = f\ \Delta I, \Delta q\ $. (This axiom states that if we solve a given problem in two different coordinate systems, both sets of estimates are related by the same coordinate transformation.)
>
> **A3.** (Scaling). *If no additional information is available, the posterior (inferred quantity) must equal that of the prior model*.[11]
>
> **A4.** (Subset independence). *Let* I_1 *be a constraint on f(**x**) in the domain* $\boldsymbol{x} \in B_1$ *Let* I_2 *be another constraint in a different domain* $\boldsymbol{x} \in B_2$. *Then, we require that our inferential method yield* $f\ B_1 | I_1\ \bigcup f\ B_2 | I_2\ = f\ B_1 \bigcup B_2 | I_1 \bigcup I_2\ $ *where f(B|I) is the chosen PADF in the domain B, given the information I*. (Our estimation rule produces the same results whether we use the subsets separately or their union. The information contained in one subset of the data, or a specific data set, should not affect the estimates based on another subset if these two subsets are independent.)
>
> **A5.** (System independence). *The same estimate should result from optimizing independent information (data) from independent systems separately using their respective densities or together using their joint density.*

The following theorem holds for the information-theoretical method of inference.

Theorem 1. For the linear model $\boldsymbol{y} = X\boldsymbol{p}$ with a prior, and with a finite information set; the PADF (or set of PADF's) that satisfy (A1–A5) and that result from an optimization procedure (with respect to the observed data) contains only the information-theoretic Maximum Entropy.

---

[11] Following Skilling, this axiom is used for convenience only. It guarantees the units of the posterior are equivalent (rather than proportional to) those of the priors. If we use normalized probability distributions instead of PADF, this axiom is not necessary, but the proof becomes slightly more complicated.

Proof. See Shore and Johnson (1980) or Skilling (1988), and Golan and Perloff (2002) for the generalized version.[12]

[12] Golan and Perloff (2002) showed under what axioms the decision function is the Rényi generalized entropy (or the Cressie-Read) and the axioms under which the decision function is the Tsallis generalized entropy. That paper also extended the axioms for modeling problems with more uncertainty and flexible constraints.