Notes 1

1 Synopsis

Most estimators used in economics (and causal inference more broadly) presume *point identification*: if we knew the distribution from which the data is drawn (i.e. abstracting from statistical error), then the estimator would return a single number. This can be constraining in the sense that ensuring point identification often requires some combination of (i) making assumptions that are stronger than desired, (ii) estimating a parameter that is not of primary interest ("moving the goalposts"), (iii) restricting application to settings in which there is a lot of variation in the data, and (iv) hand-waving (aka making shit up). Much of my recent work considers *partial identification* settings, which improve on some combination of (i)–(iv), at the cost of potentially having more than a single number (a set of numbers) to deal with.[1]

Traditional partial identification approaches involve deriving bounds. A classic contribution, which I recommend you read, is **?**. You will notice that the model considered in that paper is extremely simple. There's a reason for this: deriving good bounds ("sharp" bounds) is difficult, and becomes increasingly difficult in more complicated models. Much of my recent work tries to circumvent this difficulty by casting the problem of deriving bounds as an abstract computational problem:

$$\begin{aligned} &\min_{\theta}/\max \left(\text{quantity we want to know}\right)(\theta) \\ &\quad \text{s.t.} \quad \theta \text{ is consistent with the data} \\ &\quad \text{and} \quad \theta \text{ satisfies the assumptions of the model} \end{aligned} \tag{1}$$

where θ are the parameters of the model. This replaces the laborious case-by-case analysis required in deriving bounds into a consideration of situations in which solving a problem like (1) is practical. The latter requires thinking carefully about the formulation of the model,

[1] Some papers: **?**, **?**, **??**, **?**, and **?**.

the formulation of the three abstract notions of "quantity we want to know," "is consistent with the data," and "satisfies the assumptions," and the algorithmic tools available in the giant fields of operations research and computer science.

The benefit of this computational approach relative to analytically deriving bounds is that we can dramatically expand the range of partially identified models that we can estimate. The drawback—which is sort of inextricably linked—is that we know a lot less about the properties of the solution of (1) than we would about a derived bound, which would have a form that we can analyze directly. This drawback becomes more important when we stop abstracting from statistical error and start considering how we would use a characterization like (1) to build an estimator or conduct statistical inference.

The goal of this first project is to think about the estimation part of this problem.

2 Problem (1) in Two Example Models

2.1 Missing Data

This model is a simple extension of the one considered in **?** (and many others). The bounds can be constructed analytically in this model, making it too simple to be a substantive motivation for (1). Nevertheless, we can still set it up in form (1), which is going to be helpful since it gives us an example where we have two ways of doing something.

Suppose that $Y \in \{y_0, y_1, \ldots, y_J\}$ for some $J \geq 1$ with $y_j < y_{j+1}$ for all j. We observe $D \in \{0, 1\}$, and we observe Y if and only if $D = 1$. We want to know $\mathbb{E}[Y]$. By the law of iterated expectations:

$$\mathbb{E}[Y] = \mathbb{E}[Y|D=1]\,\mathbb{P}[D=1] + \underbrace{\mathbb{E}[Y|D=0]}_{\text{not observed}}\,\mathbb{P}[D=0]$$

Taking $\mathbb{E}[Y|D=0]$ to be both its smallest and largest logical values (y_0 and y_J), the sharp bounds for $\mathbb{E}[Y]$ can be determined analytically to be

$$\left[y_0\,\mathbb{P}[D=0] + \sum_{j=0}^{J} y_j\,\mathbb{P}[Y=y_j, D=1], \quad y_J\,\mathbb{P}[D=0] + \sum_{j=0}^{J} y_j\,\mathbb{P}[Y=y_j, D=1] \right]. \tag{2}$$

Notice that everything in the bounds is either known (the y_j) or is a feature of the distribution of observables, so is something we can estimate.

Here's how one would do this in (1). For each (j, d), let $\theta_{j,d}$ represent $\mathbb{P}[Y = y_j, D = d]$. Collect all $\theta_{j,d}$ into a vector θ. This is the parameter in (1). The quantity we want to

know is $\mathbb{E}[Y]$, which can be expressed in terms of θ as $\sum_{j=0}^{J} y_j(\theta_{j,0} + \theta_{j,1})$. This is the objective function in (1). In order to be consistent with the observed data, a θ should satisfy $\theta_{j,1} = \mathbb{P}[Y = y_j, D = 1]$ for all j. There are no other assumptions in the model, but θ are probabilities, so they should also live in the set $\Theta \equiv \{\theta \in [0,1]^{2(J+1)} : \sum_{j=0}^{J} \theta_{j,0} + \theta_{j,1} = 1\}$. Then the specific instance of (1) for this model is

$$\min_{\theta \in \Theta} / \max_{\theta \in \Theta} \sum_{j=0}^{J} y_j(\theta_{j,0} + \theta_{j,1}) \quad \text{s.t.} \quad \theta_{j,1} = \mathbb{P}[Y = y_j, D = 1] \text{ for all } j = 0, 1, \ldots, J. \tag{3}$$

Notice that (3) is a *linear program* because the objective function is a linear function of θ, and all of the constraints (including those embedded in Θ) are linear equalities or inequalities. Linear programs will show up a lot, essentially because expectation/probability is linear. In this case, the program is so simple that you can actually reason through the solution starting at (3): substitute the $\theta_{j,1}$ equalities into the objective function, then impose the adding-up-to-1 constraint and reason through how you would choose $\theta_{j,0}$ to make the objective as large or small as you can. You will arrive at the bounds in (2).

To put (3) in the computer, we need a data generating process, so that we get values of $\mathbb{P}[Y = y_j, D = 1]$. Here's a simple one with a nice property. Suppose that $\mathbb{P}[D = 0] = \mathbb{P}[D = 1] = 1/2$ and that Y and D are independent. Also, assume that $\mathbb{P}[Y = y_j] = 1/(J+1)$ for each j and take $y_j = j/J$ for $j = 0, 1, \ldots, J$. The nice property is that then the lower bound on $\mathbb{E}[Y]$ is

$$y_0 \, \mathbb{P}[D = 0] + \sum_{j=0}^{J} y_j \, \mathbb{P}[Y = y_j, D = 1] = \sum_{j=0}^{J} \frac{y_j}{2(J+1)} = \frac{1}{2J(J+1)} \sum_{j=1}^{J} j = \frac{1}{4}$$

for any choice of J. Similarly, the upper bound is

$$y_J \, \mathbb{P}[D = 0] + \sum_{j=0}^{J} y_j \, \mathbb{P}[Y = y_j, D = 1] = \frac{1}{2} + \frac{1}{4} = \frac{3}{4},$$

also independently of J. The reason this is nice is because we might expect (3) to get harder and harder to solve/estimate algorithmically as J increases, even while conceptually (with human inspection) it does not become any more difficult. The fact that the bounds don't vary with J means we can compare apples to apples as we vary J.

2.2 A Mixed Logit Model

Here's a more interesting model, one in which it is not feasible to compute bounds directly with an approach like (2).

The observable variables are $Y \in \{0, 1\}$ and W. For concreteness, think of $X > 0$ as price and Y as the decision whether to buy a good. Choices are determined according to

$$Y = \mathbb{1}[B_0 + B_1 X \geq U],$$

where $B \equiv (B_0, B_1)$ and U are unobserved random variables. Assume that X is independent of (B_0, B_1, U). Assume further that U follows a standard logistic distribution and is independent of all other random variables.

If B_0 and B_1 were deterministic coefficients, instead of random variables, then this would be a simple logit model. When they are proper random variables, it is called a *mixed logit* model. A common approach to implementing mixed logit models is to assume that B follows a parametric distribution, typically a bivariate normal distribution. This is a bit adhoc, because it imposes a strong assumption about how unobserved heterogeneity looks *across* individuals, which is not something theory (economic or otherwise) has much to say about. **?** advocate a nonparametric approach in which the support of B is discrete, say $\{b_r\}_{r=1}^R$, with unknown probabilities $\{\pi_r\}_{r=1}^R$. It is possible to provide some conditions under which the resulting model is point identified (**?**), but these are high-level, and rather hard to interpret (and thus hard to justify in empirical work). The conditions also require continuous variation in X, which is often something we don't have in empirical work.

Instead of assuming point identification, we will use (1) to compute bounds that allow for partial identification. Here's how it maps to (1).

- The unknown parameters of the model (θ in the notation of (1)) are the probabilities $\pi \equiv \{\pi_r\}_{r=1}^R$, which characterize the distribution of B.

- The maintained independence conditions imply that

$$\mathbb{P}[Y = 1 | X = x, B = b] = \frac{1}{1 + \exp(-b_0 - b_1 x)} \equiv \lambda(b, x),$$

  where $b \equiv (b_0, b_1)$. If we average over the distribution of $B|X = x$, which is equal to

the distribution of B by assumption, we get

$$\mathbb{P}[Y = 1|X = x] = \sum_{r=1}^{R} \pi_r \lambda(b_r, x).$$

A π is "consistent with the data" if and only if it satisfies this constraint.

- A quantity we might be interested in is the distribution of elasticity. For a given type $B = b$ facing price $\bar{x}$, the elasticity is (check this)

$$\epsilon(b, \bar{x}) \equiv \left(\frac{\partial}{\partial x} \lambda(b, x)|_{x=\bar{x}} \right) \frac{\bar{x}}{\lambda(b, \bar{x})} = b_1 \bar{x}(1 - \lambda(b, \bar{x}))$$

  The cdf of this elasticity across random B is then

$$\mathbb{P}[\epsilon(B, \bar{x}) \leq \bar{e}] = \sum_{r=1}^{R} \pi_r \mathbb{1}[\epsilon(b_r, \bar{x}) \leq \bar{e}]$$

- There are no additional assumptions except the requirement that π constitutes a valid distribution for a discrete random variable: $\pi \in \Pi \equiv \{\pi \in [0, 1]^R : \sum_{r=1}^{R} \pi_r = 1\}$.

Putting the above together gives us the optimization problem:

$$\min_{\pi \in \Pi} / \max_{\pi \in \Pi} \sum_{r=1}^{R} \pi_r \mathbb{1}[\epsilon(b_r, \bar{x}) \leq \bar{e}] \quad \text{s.t.} \quad \sum_{r=1}^{R} \pi_r \lambda(b_r, x) = \mathbb{P}[Y = 1|X = x] \text{ for all } x. \tag{4}$$

This is again a linear program.

Here's a DGP to try this with:

- Suppose that X takes one of four values $0, 1, 2, 3$ with equal probability.
- Suppose that the support of B_0 has ten points $\{.5 \times j/9\}_{j=0}^{9}$ with equal probability.
- Suppose that the support of B_1 has ten points $\{-3 \times j/9\}_{j=0}^{9}$ with equal probability
- Then the support of (B_0, B_1) has 100 points, also with equal probability.

Then for the elasticity, try taking $\bar{x} = 1$ and $\bar{e} = -1$.

Harder: Now suppose that we relax the assumption the assumption that the distribution of B and X are independent. How would the formulation of the bounds problem change?

Compute bounds on average elasticity again: do they get wider or narrower? Suppose we have an instrumental variable $Z \in \{0, 1\}$ that is correlated (dependent) with X, but independent of (B, U). How would the formulation of the bounds problem change? Compute bounds on average elasticity again: do they get wider or narrower, and how does this depend on the strength of the dependence between X and Z [2]?

3 Estimation

Now we're going to consider what happens when the constraint "consistent with the data" in (1) is contaminated with sampling error. The discussion here is going to be applicable to the missing data, mixed logit, and many other models. So let's first adopt a general notation for (1).

3.1 General setup

Let's call the parameters θ and assume that they are elements of $\mathbb{R}^{d_\theta}$ for some integer d_θ.

Then the components of (1) can be formalized as:

- "Quantity we want to know" will be a scalar function of θ called the **target parameter**, denoted by $t : \mathbb{R}^{d_\theta} \to \mathbb{R}$. In the missing data example this was $\mathbb{E}[Y]$ and in the mixed logit model it was $\mathbb{P}[\epsilon(B, \bar{x}) \leq \bar{e}]$.

- "Is consistent with the data" will mean setting a **criterion function** equal to 0. Call the criterion function $Q : \mathbb{R}^{d_\theta} \to \mathbb{R}_+$. Then θ is consistent with the data if and only if $Q(\theta) = 0$. In the missing data example, Q could be any non-negative function that is 0 if and only if $\theta_{j,1} = \mathbb{P}[Y = y_j, D = 1]$ for all j. So, for example, if we stacked this over j and viewed it as a vector equality, then any norm on the difference between the two sides of the equality would be a valid choice of Q. Similarly in the mixed logit example, except the object on the right-hand side of the equality would be $\mathbb{P}[Y = 1|X = x]$ stacked into a vector over all x. More detail is given ahead.

- "Satisfies the assumptions of the model" will mean $\theta \in \Theta$ for some known constraint set Θ. We will always assume that Θ is a convex set, since this is important for computation.

[2] See Notes1a for details.

With this notation we can phrase (1) as

$$\min_{\theta \in \Theta} / \max_{\theta \in \Theta} t(\theta) \quad \text{s.t.} \quad Q(\theta) = 0. \tag{5}$$

Call the minimizer $t^\star_{\text{lb}}$ and the maximizer $t^\star_{\text{ub}}$. These are the **population bounds**. Under some mild regularity conditions, the interval formed by $t^\star_{\text{lb}}$ and $t^\star_{\text{ub}}$ forms the set of all values of $t(\theta)$ that are consistent with both the data and our assumptions. This is called the **identified set** for t.

3.2 The problem

We don't have enough information to solve (3) in an actual empirical setting. The reason is that we don't know Q, which depends on the joint distribution of the observables, so (Y, D) in the missing data example, or (Y, X) in the mixed logit example. We did have enough information in the *simulations*, but only because we were generating the data, and thus knew the joint distribution of the observables. When we apply this to real data, we don't have that luxury.

Instead, we observe a realization of a sample of n observations, so $\{Y_i, D_i\}_{i=1}^n$ for missing data, or $\{Y_i, X_i\}_{i=1}^n$ for mixed logit. We can use this sample to construct an estimate $\hat{Q}$ of Q. We expect that $\hat{Q}$ will converge to Q in probability as $n \to \infty$, so that $\hat{Q}$ is a consistent estimator of Q.

The immediate temptation is to replace Q by $\hat{Q}$ in (5):

$$\min_{\theta \in \Theta} / \max_{\theta \in \Theta} t(\theta) \quad \text{s.t.} \quad \hat{Q}(\theta) = 0. \tag{6}$$

The problem is that this might be infeasible: how do we know that there is any $\theta \in \Theta$ that sets the sample criterion to 0? For (5) it's reasonable/required to assume that there is some $\theta \in \Theta$ that sets $Q(\theta) = 0$ or else the model would be **misspecified**. There's not much constructive that one can do starting from the position that the model is misspecified.[3] But even if the model is not misspecified, there could be sufficient difference between the sample distribution in the realized data and the population distribution that it's not possible to set $\hat{Q}(\theta) = 0$ for any θ.[4]

[3] This starts getting into deep philosophical equestions about how we conduct empirical science ...

[4] In the missing data model, it *is* always the case that $\hat{Q}(\theta)$ can be set to 0. In the mixed logit model, it will usually *not* be the case that $\hat{Q}(\theta) = 0$ for any θ. "Usually" here means in terms of different draws of the data. This difference means that the mixed logit model is falsifiable, whereas the missing data model is not.

3.3 Potential solutions

3.3.1 Minimum criterion

We can tell "how infeasible" the problem (6) is by finding the smallest value that $\hat{Q}(\theta)$ can take:

$$\hat{Q}^\star \equiv \min_{\theta \in \Theta} \hat{Q}(\theta), \tag{MinQ}$$

Then (6) is feasible if and only if $\hat{Q}^\star = 0$. If we assume that the model is correctly specified, then we would expect that $\hat{Q}^\star$ is converging to 0, even if it is strictly positive in the sample.

Knowing whether (6) is feasible or not doesn't really help us if the realization of data we have in front of us is such that (6) is infeasible. But we can use $\hat{Q}^\star$ to construct an alternative problem that is always feasible by construction:

$$\min_{\theta \in \Theta} t(\theta) \quad \text{s.t.} \quad \hat{Q}(\theta) = \hat{Q}^\star. \tag{7}$$

I say this is always feasible because we know, by definition of $\hat{Q}^\star$, that there is some θ that yields $Q(\theta) = \hat{Q}^\star$, namely whichever θ achieved the minimum in (MinQ).

There are different ways we can implement (7) that will likely have different statistical properties. The two that we want to explore initially are the hard and soft constraint approaches.

3.3.2 Hard constraint

The hard constraint approach involves replacing (7) by

$$\min_{\theta \in \Theta} t(\theta) \quad \text{s.t.} \quad \hat{Q}(\theta) \leq \hat{Q}^\star + \epsilon, \tag{H}$$

where $\epsilon \geq 0$ is a tuning parameter that slacks the equality constraint in (7) to allow for "near" equality. The optimal value $\hat{T}^{\text{H}}_{\text{lb}}(\epsilon)$ is a seemingly-reasonable estimator of the lower bound of the population lower bound, $t^\star_{\text{lb}}$. The role of ϵ is to potentially mitigate some bias by making $\hat{T}^{\text{H}}_{\text{lb}}(\epsilon)$ smaller than it would be with $\epsilon = 0$. Setting $\epsilon > 0$ will also have some computational stability benefits for certain choices of $\hat{Q}$.

3.3.3 Soft constraint

The soft constraint approach means replacing the constraint in (7) by a penalty in the objective:

$$\min_{\theta\in\Theta} t(\theta) + \kappa\left(\hat{Q}(\theta) - \hat{Q}^\star\right), \tag{S}$$

where $\kappa \geq 0$ is a tuning parameter that controls how severe the penalty is. The estimator of the lower bound is then

$$\hat{T}^{\mathrm{S}}_{\mathrm{lb}}(\kappa) \equiv t(\hat{\theta}^\star) \quad \text{where } \hat{\theta}^\star \text{ is any minimizer of (S).}$$

The reason for centering around $\hat{Q}^\star$ is to handle the case of $\kappa = +\infty$.[5] This lets (S) nest (7), since when $\kappa = +\infty$, any θ that does not satisfy $\hat{Q}(\theta) = \hat{Q}^\star$ cannot be optimal. This also suggests that $\hat{T}^{\mathrm{S}}_{\mathrm{lb}}(\kappa)$ should be increasing in κ. The following argument shows that this is true.

Proposition 1. *$\hat{T}^S_{lb}(\kappa)$ is weakly increasing in κ.*

Proof. Suppose that $\kappa_0 < \kappa_1$, and let $\theta_0^\star, \theta_1^\star$ be optimizers of (S) under κ_0 and κ_1. By virtue of being minimizers,

$$t(\theta_0^\star) + \kappa_0\left(\hat{Q}(\theta_0^\star) - \hat{Q}^\star\right) \leq t(\theta_1^\star) + \kappa_0\left(\hat{Q}(\theta_1^\star) - \hat{Q}^\star\right) \tag{8}$$

$$\text{and} \quad t(\theta_1^\star) + \kappa_1\left(\hat{Q}(\theta_1^\star) - \hat{Q}^\star\right) \leq t(\theta_0^\star) + \kappa_1\left(\hat{Q}(\theta_0^\star) - \hat{Q}^\star\right). \tag{9}$$

Putting the two inequalities together implies that

$$t(\theta_0^\star) + \kappa_0\left(\hat{Q}(\theta_0^\star) - \hat{Q}^\star\right) \leq \left(t(\theta_0^\star) + \kappa_1\left(\hat{Q}(\theta_0^\star) - \hat{Q}(\theta_1^\star)\right)\right) + \kappa_0\left(\hat{Q}(\theta_1^\star) - \hat{Q}^\star\right). \tag{10}$$

Rearranging terms,

$$\kappa_0\left(\hat{Q}(\theta_0^\star) - \hat{Q}(\theta_1^\star)\right) \leq \kappa_1\left(\hat{Q}(\theta_0^\star) - \hat{Q}(\theta_1^\star)\right), \tag{11}$$

[5] When $\kappa < \infty$, the centering wouldn't affect the optimal θ, and thus wouldn't affect $\hat{T}^{\mathrm{S}}_{\mathrm{lb}}(\kappa)$.

which in turn implies that $\hat{Q}(\theta_0^\star) \geq \hat{Q}(\theta_1^\star)$, since $\kappa_1 > \kappa_0$.[6] Returning to (8), we have

$$t(\theta_1^\star) - t(\theta_0^\star) \geq \kappa_0 \left(\hat{Q}(\theta_0^\star) - \hat{Q}(\theta_1^\star) \right) \geq 0, \tag{12}$$

which was the claim. *Q.E.D.*

3.4 Specific implementation

Now let's fill in the missing pieces necessary for implementation.

First, let's always assume that $t(\theta) = \tau'\theta$ is a linear function. This makes life easy, since then t is both convex and concave. It also turns out to be not that restrictive, in the sense that many interesting target parameters turn out to be linear functions of θ.

The more involved issue is choosing $\hat{Q}$.

Let's assume that the model is such that $Q(\theta) = 0$ if and only if $\gamma\theta = g$, where g are some population moments of the distribution of observables, and γ is a known matrix. In the missing data example, each element of g would be a $\mathbb{P}[Y = y_j, D = 1]$ for some j, and each row of γ would be zero except for a 1 on the element that corresponds to $\theta_{j,1}$. In the mixed logit example, each element of g would be a $\mathbb{P}[Y = 1|X = x]$ for some x, and each row of γ would consist of $\{\lambda(b_r, x)\}_{r=1}^R$, which are known. We could say that the model has linear (conditional) moment equalities.

We want to construct an estimator $\hat{G}$ of the vector of moments g. For both of these examples, we can just use sample means. So in the missing data example, where $g_j = \mathbb{P}[Y = y_j, D = 1]$, we simply count how many observations i have $Y_i = y_j$ and $D_i = 1$, and let the result be $\hat{G}_j$, with $\hat{G}$ the vector formed across j. For the mixed logit example, we can estimate the component of g corresponding to x with the binning estimator

$$\hat{G}(x) \equiv \frac{\sum_{i=1}^n \mathbb{1}[X_i = x] Y_i}{\sum_{i=1}^n \mathbb{1}[X_i = x]} = \text{sample mean of } Y_i \text{ among } i \text{ with } X_i = x, \tag{13}$$

then let $\hat{G}$ be the vector formed from stacking across x.

We can now specify $\hat{Q}$ from γ and $\hat{G}$ by simply replacing g with $\hat{G}$. However, we want to be careful here, since the way in which we do so is going to affect the computational properties of (H) and (S), as well as the statistical properties of the estimators they produce. We want to make sure that (H) and (S) are still "easy" to solve, meaning that they lie in classes of

[6] Intuitively this makes sense: when κ is larger, we want to put more emphasis on the penalty term, so we want to make $\hat{Q}(\theta)$ smaller.

convex problems for which fast algorithms exist. There are three specific choices of $\hat{Q}$ that are natural for this: the 1–norm, the squared Euclidean norm, and the max–norm.

3.4.1 1–norm

Define

$$\hat{Q}(\theta) \equiv \|\gamma\theta - \hat{G}\|_1, \tag{14}$$

where for any k–dimensional vector x, $\|x\|_1 \equiv \sum_{j=1}^{k} |x_k|$ is the 1–norm of x. The 1–norm allows each of (MinQ), (H), and (S) to be set up as linear programs. This is not so obvious. The trick is to notice that any real number x can be written as the difference of its positive and negative parts:

$$x = \max\{x, 0\} - \max\{-x, 0\} \equiv x^+ - x^-.$$

For example,

$$\begin{aligned} 5 &= \max\{5, 0\} - \max\{-5, 0\} = 5 - 0 = 5 \\ -6 &= \max\{-6, 0\} - \max\{6, 0\} = 0 - 6 = -6. \end{aligned}$$

At the same time, $|x| = x^+ + x^-$, e.g.

$$\begin{aligned} |5| &= \max\{5, 0\} + \max\{-5, 0\} = 5 + 0 = 5 \\ |-6| &= \max\{-6, 0\} + \max\{6, 0\} = 0 + 6 = 6. \end{aligned}$$

The idea then is to replace each component of $\gamma\theta - \hat{G}$ by a pair of "slack variables" that represent its positive and negative parts.

Consider (MinQ) first. The claim is that $\hat{Q}_1^\star = \hat{Q}^\star$, where

$$\hat{Q}_1^\star \equiv \min_{\theta\in\Theta, s^+\in\mathbb{R}_+^{d_g}, s^-\in\mathbb{R}_+^{d_g}} \sum_{j=1}^{d_g} s_j^+ + s_j^- \quad \text{s.t.} \quad s^+ - s^- = \gamma\theta - \hat{G}, \tag{MinQ$_1$}$$

where d_g is the dimension of $\hat{G}$. The slack variables are s^+ and s^-; note that they are restricted to be non-negative.

To see that this is true, suppose that we solved (MinQ) directly somehow. Let $\bar{\theta}$ denote

a minimizer, so that $\hat{Q}(\bar{\theta}) = \hat{Q}^\star$. Then rewrite the components of $\gamma\bar{\theta} - \hat{G}$ in terms of its positive and negative parts, denoting them $\bar{s}^+$ and $\bar{s}^-$. Then the tuple $\bar{\theta}, \bar{s}^+, \bar{s}^-$ is feasible in (MinQ_1) and produces an objective value of $\hat{Q}^\star$. We don't necessarily know that this tuple is optimal, but we can conclude that $\hat{Q}_1^\star \leq \hat{Q}^\star$. Conversely, suppose that $\bar{\theta}, \bar{s}^+, \bar{s}^-$ are optimal in (MinQ_1). Then $\bar{\theta}$ is feasible in (MinQ), so

$$\hat{Q}^\star \leq \hat{Q}(\bar{\theta}) = \|\gamma\bar{\theta} - \hat{G}\|_1 = \|\bar{s}^+ - \bar{s}^-\|_1 \leq \sum_{j=1}^{d_g} |\bar{s}_j^+| + |\bar{s}_j^-| = \sum_{j=1}^{d_g} \bar{s}_j^+ + \bar{s}_j^- = \hat{Q}_1^\star. \tag{15}$$

Now we have $\hat{Q}^\star \geq \hat{Q}_1^\star$ and $\hat{Q}^\star \leq \hat{Q}_1^\star$, so it must be that $\hat{Q}^\star = \hat{Q}_1^\star$, as claimed.

The reasoning for (H) and (S) is similar. The optimal value of the following linear program will also be $\hat{T}_{\text{lb}}^{\text{H}}(\epsilon)$:

$$\min_{\theta\in\Theta, s^+\in\mathbb{R}_+^{d_g}, s^-\in\mathbb{R}_+^{d_g}} \tau'\theta \quad \text{s.t.} \quad s^+ - s^- = \gamma\theta - \hat{G} \quad \text{and} \quad \sum_{j=1}^{d_g} s_j^+ + s_j^- \leq \hat{Q}^\star + \epsilon \tag{H_1}$$

And the following linear program will produce $\hat{T}_{\text{lb}}^{\text{S}}(\kappa)$ as $\tau'\theta^\star$ for any optimal $\theta^\star$:

$$\min_{\theta\in\Theta, s^+\in\mathbb{R}_+^{d_g}, s^-\in\mathbb{R}_+^{d_g}} \tau'\theta + \kappa\left(\sum_{j=1}^{d_g} s_j^+ + s_j^- - \hat{Q}^\star\right) \quad \text{s.t.} \quad s^+ - s^- = \gamma\theta - \hat{G} \tag{S_1}$$

3.4.2 Squared Euclidean norm

Define

$$\hat{Q}(\theta) = \|\gamma\theta - \hat{G}\|_2^2 = (\gamma\theta - \hat{G})'(\gamma\theta - \hat{G}), \tag{16}$$

where $\|\cdot\|_2$ is the usual Euclidean norm, but we've squared it here to remove the square root. This case is more straightforward since it doesn't require any reformulation.

The first problem, (MinQ), is a quadratic program (quadratic objective, linear constraints), which is only marginally more difficult to solve than a linear program. This is also true of the soft constraint problem, (S).

The hard constraint problem, (H), is a convex quadratically-constrained quadratic program (QCQP) because it has linear and quadratic constraints and a quadratic (special case here: linear) objective. These are more difficult to solve, but Gurobi can still handle them pretty well because (16) is a positive semi-definite quadratic form, and thus a convex func-

tion. Setting $\epsilon > 0$ (vs. $\epsilon = 0$) is often important for computational stability. It's also a good idea to explicitly tell Gurobi that this is a convex problem by setting `NonConvex` to 0, although usually Gurobi can detect this on its own.

3.4.3 Max–norm

Define

$$\hat{Q}(\theta) \equiv \left\| \gamma\theta - \hat{G} \right\|_\infty \equiv \max_{j=1,\ldots,d_g} \left| \gamma_j'\theta - \hat{G}_j \right| \tag{17}$$

where γ_j is the jth row of γ.[7] So in words, $\hat{Q}(\theta)$ is the largest (in magnitude) element of the vector $\gamma\theta - \hat{G}$.

Like the 1–norm, the max-norm can also be reformulated into something linear using a slack variable. Start with (MinQ) and define

$$\hat{Q}_\infty^\star = \min_{\theta\in\Theta, s\in\mathbb{R}} s \quad \text{s.t.} \quad -s \leq \gamma_j'\theta - \hat{G}_j \leq s \text{ for all } j. \tag{MinQ$_\infty$}$$

To see that $\hat{Q}_\infty^\star = \hat{Q}^\star$, profile the problem so that we optimize over s first:

$$\hat{Q}_\infty^\star = \min_{\theta\in\Theta} \left(\min_{s\in\mathbb{R}} s \quad \text{s.t.} \quad -s \leq \gamma_j'\theta - \hat{G}_j \leq s \text{ for all } j \right). \tag{18}$$

The constraints in the inner problem ensure that $s \geq \|\gamma\theta - \hat{G}\|_\infty$ for any feasible s and θ. Since we are trying to minimize s, at the optimum s will be pushed down to be exactly equal to $\|\gamma\theta - \hat{G}\|_\infty$, for any value of θ. Thus:

$$\hat{Q}_\infty^\star = \min_{\theta\in\Theta} \|\gamma\theta - \hat{G}\|_\infty = \hat{Q}^\star, \tag{19}$$

which was the claim. The reformulation of the soft-constraint problem (S) follows similar reasoning:

$$\min_{\theta\in\Theta, s\in\mathbb{R}} \tau'\theta + \kappa\left(s - \hat{Q}^\star\right) \quad \text{s.t.} \quad -s \leq \gamma_j'\theta - \hat{G}_j \leq s \text{ for all } j. \tag{S$_\infty$}$$

Reformulating the hard constraint problem follows from similar ideas with a couple of

[7] In my world all vectors are column vectors. So the jth row of γ is a column vector, γ_j, so that the jth element of $\gamma\theta$ is $\gamma_j'\theta$. With this notation, $\gamma = \begin{bmatrix}\gamma_1 & \gamma_2 & \cdots & \gamma_{d_g}\end{bmatrix}'$.

steps. The reformulation is

$$\hat{T}^{\mathrm{H}}_{\mathrm{lb},\infty}(\epsilon) = \min_{\theta\in\Theta, s\in\mathbb{R}} \tau'\theta \quad \text{s.t.} \quad s \leq \hat{Q}^\star + \epsilon \quad \text{and} \quad -s \leq \gamma_j'\theta - \hat{G}_j \leq s \text{ for all } j., \qquad (\mathrm{H}_\infty)$$

and the claim is that $\hat{T}^{\mathrm{H}}_{\mathrm{lb},\infty}(\epsilon) = \hat{T}^{\mathrm{H}}_{\mathrm{lb}}(\epsilon)$. To justify, let $\bar{\theta}$ denote an optimizer of (H), so that $\tau'\bar{\theta} = \hat{T}^{\mathrm{H}}_{\mathrm{lb}}(\epsilon)$. Then taking $\bar{s} = \hat{Q}(\bar{\theta}) = \|\gamma\bar{\theta} - \hat{G}\|_\infty$ yields a feasible pair for (H_∞) with objective value $\hat{T}^{\mathrm{H}}_{\mathrm{lb}}(\epsilon)$, implying that $\hat{T}^{\mathrm{H}}_{\mathrm{lb},\infty}(\epsilon) \leq \hat{T}^{\mathrm{H}}_{\mathrm{lb}}(\epsilon)$. Conversely, if $\bar{\theta}, \bar{s}$ are optimal for (H_∞), then

$$\hat{Q}(\bar{\theta}) \equiv \|\gamma\bar{\theta} - \hat{G}\|_\infty \leq \bar{s} \leq \hat{Q}^\star + \epsilon, \qquad (20)$$

so that $\bar{\theta}$ is feasible in (H) with $\tau'\bar{\theta} = \hat{T}^{\mathrm{H}}_{\mathrm{lb},\infty}(\epsilon)$, implying that $\hat{T}^{\mathrm{H}}_{\mathrm{lb}}(\epsilon) \leq \hat{T}^{\mathrm{H}}_{\mathrm{lb},\infty}(\epsilon)$.

3.5 Estimating upper bounds

We've been focusing on estimating lower bounds for concreteness, but the same ideas apply to estimating upper bounds. The only change in the hard constraint problem (H) is to replace min by max. In the soft constraint problem (S), we also need to change the sign of the penalty:

$$\max_{\theta\in\Theta} t(\theta) - \kappa\left(\hat{Q}(\theta) - \hat{Q}^\star\right), \qquad (\mathrm{H}_{\mathrm{ub}})$$

where $\kappa \geq 0$ still. Notice that if $\theta \mapsto \hat{Q}(\theta)$ is convex, then $\theta \mapsto -\kappa\hat{Q}(\theta)$ is concave, as we want for a maximization problem. Since we are maximizing, and $\hat{Q}(\theta) \geq \hat{Q}^\star$, we are still trying to push $\hat{Q}(\theta)$ close to $\hat{Q}^\star$, with the emphasis on this objective controlled by κ.

4 Monte Carlo simulations

The procedures in the previous section all take as input the sample of n observations that was drawn, which produces $\hat{G}$. Each sample will produce a different realization of $\hat{T}^{\mathrm{H}}_{\mathrm{lb}}(\epsilon)$ and $\hat{T}^{\mathrm{S}}_{\mathrm{lb}}(\kappa)$. We want to study the distributions of these estimators, but this is analytically intractable here (as well as in easier problems, e.g. OLS). Thus, we resort to simulation.

A Monte Carlo simulation for the mixed logit model starts by drawing a sample of n i.i.d. observations from the DGP:

- Draw B_i for $i = 1, \ldots, n$ from the distribution implied by the true value of θ (previously called π).

- Draw U_i for $i = 1, \ldots, n$ from a standard logistic distribution.
- Draw X_i for $i = 1, \ldots, n$ from the marginal distribution, e.g. uniform over $\{0, 1, 2, 3\}$ in the proposed DGP.[8]
- Compute Y_i for $i = 1, \ldots, n$ as $Y_i = \mathbb{1}[B_{i,0} + B_{i,1} X_i \geq U_i]$.
- Throw away B_i and U_i (or just don't touch them)—these aren't observed variables.

From this sample $\{Y_i, X_i\}_{i=1}^n$ we can compute $\hat{G}$, which then implies a realized value of $\hat{T}_{\text{lb}}^{\text{H}}(\epsilon)$ and $\hat{T}_{\text{lb}}^{\text{S}}(\kappa)$. Repeating this process m times gives m realizations of $\hat{T}_{\text{lb}}^{\text{H}}(\epsilon)$ and $\hat{T}_{\text{lb}}^{\text{S}}(\kappa)$, which can then be used to approximate features of the distribution of $\hat{T}_{\text{lb}}^{\text{H}}(\epsilon)$ and $\hat{T}_{\text{lb}}^{\text{S}}(\kappa)$. We generally want $m \geq 1000$ to get a good approximation, but bigger is always better.

Here are some useful features of the distribution of $\hat{T}_{\text{lb}}^{\text{H}}(\epsilon)$ and $\hat{T}_{\text{lb}}^{\text{S}}(\kappa)$ to examine:

- The mean and the bias (for $t_{\text{lb}}^{\star}$).
- The standard deviation.
- The mean-squared error (MSE).
- A histogram (or smoothed density estimate) constructed from the m realizations. This is useful for examining the shape of the distribution, which in this case we expect to *not* be normal.

The overall goal is to compare $\hat{T}_{\text{lb}}^{\text{H}}(\epsilon)$ across the three criterion choices (1-norm, Euclidean norm, max-norm) and different values of ϵ, and $\hat{T}_{\text{lb}}^{\text{S}}(\kappa)$ across the three criterion choices and different values of κ.

[8] If B_i and X_i were dependent in the simulation, then one would want to draw B_i and X_i jointly, or draw X_i first, then B_i conditional on X_i, or vice versa.