Robust Forecasting under Partial Identification and Misspecification

Timothy Christensen
New York University

Hyungsik Roger Moon
Univ. of Southern California and Yonsei University

Frank Schorfheide*
University of Pennsylvania CEPR, NBER, and PIER

This Version: May 30, 2019

Preliminary and incomplete.

Please do not circulate or cite without permission.

Abstract

We study the problem of forecasting individual-level outcomes in dynamic discrete choice models when model parameters are only set-identified. In this environment, different parameters in the identified set lead to different forecasts, some more accurate than others. We seek to construct forecasts that are "robust" to set identification of parameters of the forecasting model. We characterize forecasts that minimize maximum risk or maximum regret as model parameters vary over the identified set. The optimal forecasts under either robustness criterion depends in a natural way on two extremum problems which can be solved, at least in large part, by duality arguments, making computation of the robust forecasts computationally light. Extensions to model misspecification and structural breaks are also discussed.

JEL CLASSIFICATION: C11, C14, C23, C53

KEY WORDS: Decision Theory, Dynamic Discrete Choice, Forecasting, Identification, Minimax Loss, Minimax Regret, Panel Data Models, Robustness, Structural Breaks.

*Correspondence: T. Christensen: Department of Economics, 19 W. 4th Street, 6FL, New York University, New York, NY 10012. E-mail: timothy.christensen@nyu.edu. H.R. Moon: Department of Economics, University of Southern California, KAP 300, Los Angeles, CA 90089. E-mail: moonr@usc.edu. F. Schorfheide: Department of Economics, 3718 Locust Walk, University of Pennsylvania, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu. Moon and Schorfheide gratefully acknowledge financial support from the National Science Foundation under Grants SES 1625586 and SES 1424843, respectively.

1 Introduction

Suppose that a subset of parameters of a forecasting model are only set-identified. Should the lack of point identification be a concern for the forecaster? At first glance the answer appears to be "no." If the parameters in the identified set generate different forecasts and some of these forecasts are less accurate than others, then we should be able to discriminate among the parameters based on the observed data. To the extent that we are unable to do, the parameterizations should be observationally equivalent and therefore generate the same forecasts. This intuition is confirmed in the context of vector autoregressions (VARs): while the structural form of the VAR may only be set-identified, forecasts only utilize the reduced form of the VAR which is directly identifiable from the observed time series. This intuition is also confirmed in the context of dynamic linear factor models. The parameters are only identified up to a particular normalization of the latent factors, but each normalization leads to identical forecasts.

In this paper, we consider the problem of forecasting with set-identified panel dynamic discrete choice models and show that the VAR intuition does not apply. As is well known (Honoré and Tamer, 2006; Chamberlain, 2010; Chernozhukov, Fernández-Val, Hahn, and Newey, 2013), the homogeneous parameters and the correlated random effects distribution are set-identified when no parametric assumptions are made about the random effects distribution. Our paper makes several contributions. First, we demonstrate that in the panel dynamic discrete choice setting different parameters in the identified set lead to different forecasts, some more accurate than others.

Second, we construct forecasts that are "robust" to set identification of parameters of the forecasting model. By robust, we mean forecasts that minimize either maximum risk or maximum regret (i.e. risk relative to the infeasible Bayes decision if the true model parameter was known) as model parameters vary over the identified set. We show that for binary loss, quadratic loss, and a log-scoring rule, the optimal forecasts under either robustness criterion depends on two extremum problems which characterize the smallest and largest conditional probabilities for the discrete outcomes being forecasted as the model parameters vary over the identified set. These extremum problems can be solved, at least in large part, by duality arguments, making computation of the robust forecasts computationally light.

Third, we show how our approach can be extended to deal with potential structural breaks between the in-sample and forecasting period and misspecification of various aspects of the forecasting model, including the distribution of error terms. This problem is not restricted

to models in short panels. Structural breaks and model misspecifications are concerns for any forecast model.[1]

Our paper is related to several strands of the literature. For forecasting short time-series using panel data see, e.g., Baltagi (2008), Gu and Koenker (2016), Liu, Moon, and Schorfheide (2018b), Liu (2019), and Liu, Moon, and Schorfheide (2018a). Applications of partial identification in nonlinear panel data analysis include Honoré and Tamer (2006) who show that homogeneous parameters in dynamic discrete choice models are set-identified from short panels when the random coefficient distribution isn't specified parametrically. They also propose simple computational methods based on linear programming. Chernozhukov et al. (2013) characterize the identified set of the average treatment effects in semiparametric nonlinear panel models and study inference.

There is an extensive literature on statistical decision theory following Wald (1950). Most closely related to our approach are the notions of Γ-minimax (or Γ-minimax regret) decisions in robust Bayes analysis, in which "robust" decision rules minimize the maximum posterior risk (or regret) over a set of priors (Robbins, 1951; Berger, 1985). In economics, this approach is also related to the multiple priors framework of Gilboa and Schmeidler (1989) and the robustness literature following Hansen and Sargent (2001).

In short panels, a posterior distribution over model parameters will be supported asymptotically on the identified set provided the identified set is contained in the support of the prior. The data does not revise the prior over the identified set. Thus, the posterior-predictive distribution for the discrete variable being forecast conditional on the sample may be affected by the forecaster's prior, even asymptotically. In this respect, our robust forecasts may be viewed as asymptotically Γ-minimax (or Γ-minimax regret) forecasts when the set of priors over which the econometrician is seeking robustness is sufficiently rich that it accommodates any posterior supported on the identified set.

Other related econometric applications of minimax decision rules in econometrics include Chamberlain (2000) who applied the idea of minimax decision making to forecasting with a linear dynamic panel data model and Chamberlain (2001) for forecasting with a time-homogeneous AR(1) model. For related robust Bayesian approaches to inference under partial identification, see Kitagawa (2012) and Giacomini and Kitagawa (2018).

The remainder of this paper is organized as follows. Section 2 introduces the robust forecasting problem in a panel dynamic discrete choice model under the assumption of no

[1]See Clements and Hendry (1998) for a taxonomy of forecast errors.

sampling uncertainty. Section 3 provides a numerical illustration of robustifying forecasts against partial identification. Section 4 considers extensions to model misspecification and structural breaks. In Section 5 we account for sampling uncertainty. Finally, Section 6 concludes.

2 A Panel Dynamic Discrete Choice Model

Let $\mathbb{I}\{y \geq a\}$ be the indicator function that is one if $y \geq a$ and zero otherwise. Throughout this section we shall consider the following basic dynamic discrete choice model:

$$Y_{it+1} = \mathbb{I}\big\{\lambda_i + \beta Y_{it} \geq U_{it+1}\big\}, \quad \mathbb{P}\{U_{it+1} \leq u | Y_i^t = y^t, \lambda_i = \lambda\} = \Phi_{t+1}(u), \tag{1}$$

where $Y_i^t = (Y_{i1}, ..., Y_{it})'$ and y^t is a realized value of Y_i^t. The econometrician observes $Y_i^T = (Y_{i1}, ..., Y_{iT})'$ for $i = 1, \ldots, N$ and wishes to forecast Y_{iT+1}. We treat the initial condition Y_{i0} as unobserved and specify a joint distribution between Y_{i0} and the heterogeneous intercept λ_i. We denote the joint and marginal distributions of (λ_i, Y_{i0}) by $\Pi_{\lambda,y}$, and Π_λ, Π_y, respectively. The corresponding densities are denoted by $\pi_\bullet$. The cumulative density function (cdf) of U_{it} is potentially time-varying. We collect the sequence of cdfs in the set $\Phi_T^\otimes = \big[\Phi_1, \ldots, \Phi_T\big]$. To forecast Y_{iT+1} we also require a cdf Φ_{T+1}. Models are indexed by $\theta = \big(\Phi_{T+1}, \Phi_T^\otimes, \Pi_{\lambda,y}, \beta\big)$. Note that unless one assumes that $\Phi_{T+1} = \Phi_t$ for some $t < T+1$ the sample does not contain any identifying information about Φ_{T+1}. The model (1) can be used to generate point and density forecasts of the form $\widehat{Y}_{iT+1} \in \{0, 1\}$ and $\widehat{p}_{iT+1} \in [0, 1]$, respectively.

In what follows, we shall abstract from sampling uncertainty and focus on the case with $N \longrightarrow \infty$ to illustrate the role of partial identification. Forecasting rules will be rules that condition on the observed vector Y_i^T, rather than the full sample, as the forecaster knows the identified set Θ_0.

2.1 Loss Functions and Bayes Forecasts

We adopt a decision-theoretic approach to analyzing our forecasting problem. We denote the forecast for Y_{iT+1} by $d(Y_i^T)$, which takes values in a decision space $\mathcal{D}$ and is based on an observed vector Y_i^T. Let $\theta = (\Phi_{T+1}, \Phi_T^\otimes, \beta, \Pi_{\lambda|y})$. In this section we derive the forecast that minimizes the posterior risk (expected loss), treating θ as known and λ_i as unknown. Having observed Y_i^T, the forecaster forms a posterior predictive distribution over Y_{iT+1}, with $\Pi_{\lambda|y}$

treated as a conditional prior for λ_i. The forecast $d(Y_i^T)$ is then obtained by minimizing the posterior risk. We relax the assumption of known θ in Section 2.2 and allow θ to be set-identified. Minimax risk and regret arguments will be used to cope with the non-uniqueness of θ. Throughout this paper, we consider binary and quadratic loss functions and the log predictive probability score to evaluate the forecast accuracy.

Binary Loss. The binary loss function for the decision space $\mathcal{D} = \{0, 1\}$ takes the form

$$\ell_b\big(Y_{iT+1}, d(Y_i^T)\big) = a_{10}\mathbb{I}\{Y_{iT+1} = 1,\, d(Y_i^T) = 0\} + a_{01}\mathbb{I}\{Y_{iT+1} = 0,\, d(Y_i^T) = 1\}\,. \tag{2}$$

It is straightforward to verify that the optimal point forecast in this environment is

$$d^*_{b,\theta}(Y_i^T) = \mathbb{I}\left\{\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} \geq \frac{a_{01}}{a_{01} + a_{10}}\right\}. \tag{3}$$

The optimal binary forecast is not unique for values of Y_i^T such that $\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} = \frac{a_{01}}{a_{01}+a_{10}}$. All optimal binary forecasts differ only in their handling of ties. Each optimal binary forecast has risk

$$a_{10} \cdot \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} \,\wedge\, a_{01} \cdot (1 - \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\})\,, \tag{4}$$

where $a \wedge b$ denotes the minimum of a and b.

Quadratic Loss. Now consider the quadratic loss function

$$\ell_q\big(Y_{iT+1}, d(Y_i^T)\big) = \big(Y_{iT+1} - d(Y_i^T)\big)^2 \leq 1. \tag{5}$$

The point forecast with decision space $\mathcal{D} = [0, 1]$ that minimizes posterior risk and integrated risk is the posterior mean

$$d^*_{q,\theta}(Y_i^T) = \mathbb{E}_\theta[Y_{iT+1}|Y_i^T] = \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}\,. \tag{6}$$

Log Predictive Probability Score. The loss function for the decision space $\mathcal{D} = [0, 1]$ takes the form

$$\begin{aligned}&\ell_p\big(Y_{iT+1}, d(Y_i^T)\big)\\ &= \mathbb{I}\{Y_{iT+1} = 1\}\log\left(\frac{\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}}{d(Y_i^T)}\right) + \mathbb{I}\{Y_{iT+1} = 0\}\log\left(\frac{\mathbb{P}_\theta\{Y_{iT+1} = 0|Y_i^T\}}{1 - d(Y_i^T)}\right),\end{aligned} \tag{7}$$

with the understanding that $\ell_p\big(Y_{iT+1}, d(Y_i^T)\big) = +\infty$ if $\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} > 0$ and $d(Y_i^T) = 0$, or if $\mathbb{P}_\theta\{Y_{iT+1} = 0|Y_i^T\} > 0$ and $d(Y_i^T) = 1$.

Under the log-scoring rule, the conditional risk $\mathbb{E}_\theta\left[\ell_p\big(Y_{iT+1}, d(Y_i^T)\big)|Y_i^T\right]$ is the Kullback–Leibler divergence between the conditional distribution of Y_{iT+1} given Y_i^T and a Bernoulli distribution with success probability $d(Y_i^T)$:

$$\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} \log\left(\frac{\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}}{d(Y_i^T)}\right) + \mathbb{P}_\theta\{Y_{iT+1} = 0|Y_i^T\} \log\left(\frac{\mathbb{P}_\theta\{Y_{iT+1} = 0|Y_i^T\}}{1 - d(Y_i^T)}\right).$$

It follows that the optimal point forecast in this environment is

$$d^*_{p,\theta}(Y_i^T) = \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}. \tag{8}$$

The optimal point forecasts under quadratic loss and the log-scoring rule are the same but their risks are different. The risk under the log-scoring rule is zero, whereas the risk under quadratic loss is $\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}(1 - \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\})$.

2.2 Robust Forecasts

We now consider forecasts that are robust with respect to the parameterization of the forecast model. Suppose that the goal is to robustify the forecast over the subset $\Theta_0 \subset \Theta$. We take Θ_0 to be the set of parameters that can be identified at time T based on the observable choice probabilities. We consider two notions of forecast robustness, namely minimizing maximum risk (minimax) and minimizing maximum regret (minimax regret), where the regret of a forecast is its risk in excess of the (infeasible) Bayes forecast if the true θ were known.

As we shall see, both the minimax forecasts and minimax regret forecasts will depend on the lower and upper values of the forecast probabilities $\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}$ as θ varies over the identified set Θ_0:

$$p_L(Y_i^T) := \min_{\theta\in\Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}, \text{ and} \tag{9}$$

$$p_U(Y_i^T) := \max_{\theta\in\Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}. \tag{10}$$

The challenge in implementing the robust forecasts is to solve these extremum problems. Section 3.2 below shows how computation of these terms may be simplified using duality techniques, making implementation computationally light.

2.2.1 Minimax Forecasts

Binary Loss. We first derive the binary forecast that achieves the conditional minimax risk:

$$\min_{d(Y_i^T)\in\{0,1\}} \max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_b(Y_{iT+1}, d(Y_i^T))\big|Y_i^T\big]\,, \tag{11}$$

where ℓ_b denotes the binary loss function (2) and the decision space is $\mathcal{D} = \{0,1\}$. If the forecaster chooses $d(Y_i^T) = 1$, then her adversary solves the following problem:

$$\max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_b(Y_{iT+1}, 1)\big|Y_i^T\big] = a_{01} - a_{01}\cdot\min_{\theta\in\Theta_0} \mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\} = a_{01}(1-p_L(Y_i^T)).$$

If the forecaster chooses $d(Y_i^T) = 0$, then

$$\max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_b(Y_{iT+1}, 0)\big|Y_i^T\big] = a_{10}\cdot\max_{\theta\in\Theta_0} \mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\} = a_{10}p_U(Y_i^T).$$

Thus, we can deduce that a minimax-optimal forecast is given by

$$d_{b,mm}(Y_i^T) = \mathbb{I}\left\{a_{01} \le a_{01}p_L(Y_i^T) + a_{10}p_U(Y_i^T)\right\} \tag{12}$$

and the minimax conditional risk is

$$\mathcal{R}^*_{b,mm}(Y_i^T) = \left(a_{01} - a_{01}p_L(Y_i^T)\}\right) \wedge \left(a_{10}p_U(Y_i^T)\right)\,. \tag{13}$$

Like the point-identified case, the minimax-optimal binary forecast is not necessarily unique. Non-uniqueness arises for values of Y_i^T such that

$$a_{01} = a_{01}p_L(Y_i^T) + a_{10}p_U(Y_i^T)\,.$$

If so, each minimax-optimal forecast differs only in its handling of ties. Each minimax-optimal forecast has the same maximum conditional risk.

Quadratic Loss. We now derive the forecast that achieves the conditional minimax risk:

$$\min_{d(Y_i^T)\in[0,1]} \max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_q(Y_{iT+1}, d(Y_i^T))\big|Y_i^T\big]\,, \tag{14}$$

where ℓ_q denotes the quadratic loss function (5) and the decision space is $\mathcal{D} = [0,1]$. Note

$$\mathbb{E}_\theta\big[\ell_q(Y_{iT+1}, d)\big|Y_i^T\big] = \mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\}(1-2d) + d^2\,.$$

Therefore, if the forecaster chooses $d \in [0,1]$, her maximum risk is

$$\max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_q(Y_{iT+1}, d)\big|Y_i^T\big] = \begin{cases} p_U(Y_i^T)(1-2d) + d^2 & \text{if } d < \frac{1}{2}\,, \\ p_L(Y_i^T)(1-2d) + d^2 & \text{if } d > \frac{1}{2}\,, \\ \frac{1}{4} & \text{if } d = \frac{1}{2}\,. \end{cases}$$

It follows that the minimax-optimal forecast under quadratic loss is given by

$$d_{q,mm}(Y_i^T) = \begin{cases} p_U(Y_i^T) & \text{if } p_U(Y_i^T) \le \frac{1}{2}\,, \\ p_L(Y_i^T) & \text{if } p_L(Y_i^T) \ge \frac{1}{2}\,, \\ \frac{1}{2} & \text{otherwise} \end{cases} \tag{15}$$

and the minimax conditional risk is

$$\mathcal{R}^*_{q,mm}(Y_i^T) = \begin{cases} p_U(Y_i^T)(1-p_U(Y_i^T)) & \text{if } p_U(Y_i^T) \le \frac{1}{2}\,, \\ p_L(Y_i^T)(1-p_L(Y_i^T)) & \text{if } p_L(Y_i^T) \ge \frac{1}{2}\,, \\ \frac{1}{4} & \text{otherwise}\,. \end{cases} \tag{16}$$

Log Predictive Probability Score. We now derive the forecast that achieves the conditional minimax risk:

$$\min_{d(Y_i^T)\in[0,1]} \max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_p(Y_{iT+1}, d(Y_i^T))\big|Y_i^T\big]\,, \tag{17}$$

where ℓ_p denotes the loss function (7) corresponding to the log-scoring rule and the decision space is $\mathcal{D} = [0,1]$.

We first rule out a few pathological cases. First, suppose the forecaster chooses $d = 0$. If $p_U(Y_i^T) > 0$ then the adversary can set the forecaster's maximum risk to $+\infty$ by choosing any θ with $\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} > 0$. Thus, it is only optimal for the forecaster to choose $d = 0$ if $p_L(Y_i^T) = p_U(Y_i^T) = 0$, in which case the model is point identified and $d = 0$ is also the Bayes decision. Similarly, it is only optimal for the forecaster to choose $d = 1$ if $p_L(Y_i^T) = p_U(Y_i^T) = 1$.

Now suppose that $p_L(Y_i^T) < p_U(Y_i^T)$. Here the forecaster will choose some $d \in (0,1)$. By convexity of Kullback–Leibler divergence, the maximum risk of the forecast must be obtained

at either $p_L(Y_i^T)$ or $p_U(Y_i^T)$:

$$\max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_p(Y_{iT+1},d)\big|Y_i^T\big] = \left(p_L(Y_i^T)\log\left(\frac{p_L(Y_i^T)}{d}\right) + (1-p_L(Y_i^T))\log\left(\frac{1-p_L(Y_i^T)}{1-d}\right)\right) \vee \left(p_U(Y_i^T)\log\left(\frac{p_U(Y_i^T)}{d}\right) + (1-p_U(Y_i^T))\log\left(\frac{1-p_U(Y_i^T)}{1-d}\right)\right).$$

By convexity, the maximum risk is minimized by choosing d to equate the two terms. The minimax-optimal forecast under the log-scoring rule is therefore

$$d_{p,mm}(Y_i^T) = \frac{\exp\big(f(p_L(Y_i^T),p_U(Y_i^T))\big)}{1+\exp\big(f(p_L(Y_i^T),p_U(Y_i^T))\big)},$$

where

$$f(p_L(Y_i^T),p_U(Y_i^T)) = \frac{h(p_L(Y_i^T)) - h(p_U(Y_i^T))}{p_L(Y_i^T) - p_U(Y_i^T)},$$

and $h(p) = p\log p + (1-p)\log(1-p)$. The probability $d_p(Y_i^T)$ minimizes the maximum Kullback–Leibler divergence between the conditional distribution of Y_{iT+1} given Y_i^T and the forecast distribution as θ varies over Θ_0.

2.2.2 Minimax Regret Forecasts

The minimax regret criterion measures maximum risk relative to the infeasible first-best Bayes forecast which the forecaster would make if she knew the true value of θ.

Binary Loss. We first derive the binary forecast that minimizes the conditional maximum regret under binary loss:

$$\min_{d(Y_i^T)\in\{0,1\}} \max_{\theta\in\Theta_0} \left(\mathbb{E}_\theta\big[\ell_b(Y_{iT+1},d(Y_i^T))\big|Y_i^T\big] - \mathbb{E}_\theta\big[\ell_b(Y_{iT+1},d^*_{b,\theta}(Y_i^T))\big|Y_i^T\big]\right), \tag{18}$$

where $d^*_{b,\theta}$ is the Bayes forecast from (3), the loss function ℓ_b is the binary loss function (2), and the decision space is $\mathcal{D} = \{0,1\}$.

If the forecaster chooses $d(Y_i^T) = 1$, then her adversary solves the problem (cf. (4)):

$$\max_{\theta\in\Theta_0}\left(\mathbb{E}_\theta\big[\ell_b(Y_{iT+1},1)\big|Y_i^T\big] - a_{10}\cdot\mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\}\wedge a_{01}\cdot(1-\mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\})\right)$$
$$= \big(a_{01} - (a_{01}+a_{10})p_L(Y_i^T)\big)\vee 0,$$

where $a \vee b$ denotes the maximum of a and b. If the forecaster chooses $d(Y_i^T) = 0$, then

$$\begin{aligned}&\max_{\theta\in\Theta_0}\left(\mathbb{E}_\theta\big[\ell_b(Y_{iT+1},0)\big|Y_i^T\big] - a_{10}\cdot\mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\}\wedge a_{01}\cdot(1-\mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\})\right)\\&= \big((a_{01}+a_{10})p_U(Y_i^T) - a_{01}\big)\vee 0\,.\end{aligned}$$

Thus, we can deduce that a minimax regret-optimal binary forecast is given by

$$d_{b,mmr}(Y_i^T) = \mathbb{I}\left\{\left(\left(\frac{a_{01}}{a_{01}+a_{10}} - p_L(Y_i^T)\right)\vee 0\right) \leq \left(\left(p_U(Y_i^T) - \frac{a_{01}}{a_{01}+a_{10}}\right)\vee 0\right)\right\} \tag{19}$$

and the minimax regret is

$$\mathcal{R}^*_{b,mmr}(Y_i^T) = \big(\big(a_{01} - (a_{01}+a_{10})p_L(Y_i^T)\big)\vee 0\big)\wedge\big(\big((a_{01}+a_{10})p_U(Y_i^T) - a_{01}\big)\vee 0\big)\,.$$

As with the other binary forecasts, the minimax regret-optimal forecast is not necessarily unique. Non-uniqueness arises for values of Y_i^T such that

$$\left(\frac{a_{01}}{a_{01}+a_{10}} - p_L(Y_i^T)\right)\vee 0 = \left(p_U(Y_i^T) - \frac{a_{01}}{a_{01}+a_{10}}\right)\vee 0\,.$$

If so, each minimax regret-optimal forecast differs only in its handling of ties. Each minimax regret-optimal forecast has the same maximum regret.

Quadratic Loss. We now derive the forecast that minimizes the conditional maximum regret under quadratic loss:

$$\min_{d(Y_i^T)\in[0,1]}\ \max_{\theta\in\Theta_0}\left(\mathbb{E}_\theta\big[\ell_q(Y_{iT+1},d(Y_i^T))\big|Y_i^T\big] - \mathbb{E}_\theta\big[\ell_q(Y_{iT+1},d^*_{q,\theta}(Y_i^T))\big|Y_i^T\big]\right), \tag{20}$$

where $d^*_{q,\theta}$ is the Bayes forecast from (6), the loss function ℓ_q is the binary loss function (5), and the decision space is $\mathcal{D} = [0,1]$. In fact, it is without loss of generality to restrict the decision space to the interval $[p_L(Y_i^T), p_U(Y_i^T)]$. Note

$$\begin{aligned}&\mathbb{E}_\theta\big[\ell_q(Y_{iT+1},d)\big|Y_i^T\big] - \mathbb{E}_\theta\big[\ell_q(Y_{iT+1},d^*_{q,\theta}(Y_i^T))\big|Y_i^T\big]\\&= \mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\}\left(\mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\} - 2d\right) + d^2\\&= \big(\mathbb{P}_\theta\{Y_{iT+1}=1|Y_i^T\} - d\big)^2\,.\end{aligned}$$

Therefore, if the forecaster chooses $d \in [p_L(Y_i^T), p_U(Y_i^T)]$, her maximum regret is[2]

$$\max_{\theta\in\Theta_0} \mathbb{E}_\theta\big[\ell_q(Y_{iT+1}, d)\big|Y_i^T\big] = \begin{cases} (p_U(Y_i^T) - d)^2 & \text{if } p_U(Y_i^T) - d \geq d - p_L(Y_i^T)\,, \\ (p_L(Y_i^T) - d)^2 & \text{if } p_U(Y_i^T) - d \leq d - p_L(Y_i^T)\,. \end{cases}$$

The minimax regret-optimal forecast is therefore

$$d_{q,mmr}(Y_i^T) = \frac{p_L(Y_i^T) + p_U(Y_i^T)}{2}$$

and the minimax regret is

$$\mathcal{R}^*_{q,mmr}(Y_i^T) = \left(\frac{p_U(Y_i^T) - p_L(Y_i^T)}{2}\right)^2 .$$

Log Predictive Probability Score. The risk of the Bayes forecast $d^*_{p,\theta}$ from (8) is zero. Therefore, under this loss the risk of any forecast is equal to its regret. The minimax-optimal and minimax regret-optimal forecasts are therefore identical: $d_{p,mm}(Y_i^T) = d_{p,mmr}(Y_i^T)$.

2.2.3 Equivalence of Robust Binary Forecasts under Symmetric Loss

The minimax regret-optimal and minimax-optimal binary forecasts are identical under symmetric loss (i.e. $a_{01} = a_{10}$). To see this, first suppose $p_L(Y_i^T) > \frac{1}{2}$. In this case, the Bayes decision is $d^*_{b,\theta}(Y_i^T) = 1$ for all $\theta \in \Theta_0$ and so $d_{b,mmr}(Y_i^T) = d_{b,mm}(Y_i^T) = 1$. Similarly, when $P_U(Y_i^T) < \frac{1}{2}$ the Bayes decision is $d^*_{b,\theta}(Y_i^T)$ for all $\theta \in \Theta_0$ and so $d_{b,mmr}(Y_i^T) = d_{b,mm}(Y_i^T) = 0$. It remains to consider the case in which the inequalities

$$p_L(Y_i^T) \leq \frac{1}{2}\,, \quad p_U(Y_i^T) \geq \frac{1}{2}$$

[2] Without loss of generality, suppose that $d \in [0, p_L(Y_i^T))$. Then,

$$\begin{aligned} &\max_{\theta\in\Theta_0} \left(\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} - d\right)^2 = \left(p_U(Y_i^T) - d\right)^2 \\ &\quad \geq \quad \left(p_U(Y_i^T) - p_L(Y_i^T)\right)^2 \geq \min_{d\in[p_L(Y_i^T), p_U(Y_i^T)]} \max_{\theta\in\Theta_0} \left(\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\} - d\right)^2 . \end{aligned}$$

both hold. This is the case in which the Bayes decision will be different for different $\theta \in \Theta_0$. Here it is straightforward to deduce that

$$d_{b,mmr}(Y_i^T) = \mathbb{I}\left\{\tfrac{1}{2} - p_L(Y_i^T) \leq p_U(Y_i^T) - \tfrac{1}{2}\right\} = \mathbb{I}\left\{1 \leq p_L(Y_i^T) + p_U(Y_i^T)\right\} = d_{b,mm}(Y_i^T)\,.$$

The minimax regret-optimal and minimax-optimal binary forecasts may be different under asymmetric loss (i.e. $a_{01} \neq a_{10}$), however.

3 Robustness to Partial Identification

We shall illustrate the relation between the different forecasts (minimax, minimax regret, and Bayes) using a numerical example from Honoré and Tamer (2006). The model is a panel probit model: Φ_t is the standard normal cdf for all t. The distribution $\Pi_{\lambda,y}$ is unspecified, but λ is assumed to be supported on the discrete evenly-spaced grid $\{-3, -2.8, \ldots, 2.8, 3\}$. Under the true data-generating process, λ and y_0 are independent with $\Pi_{y_0}(Y_{0i} = 1) = \frac{1}{2}$ and the probability mass for λ under Π_λ is assigned by interpolating a $N(0,1)$ distribution on the support points.

3.1 Forecast Comparisons in a Numerical Illustration

We are interested in comparing the minimax-optimal binary forecast, the minimax regret-optimal binary forecast, and the (infeasible but first-best) binary Bayes forecast. To implement the robust forecasts, the extremum problems are solved using the linear programming techniques described below in Section 3.2. We also consider a naive forecast assuming Markovianity. The naive probability is computed by calculating $P(Y_{iT} = 1|Y_{iT-1})$ from the observed data, and then iterating forward one period assuming it is the transition distribution: $P_N(Y_{i,T+1} = 1|Y_{iT} = y_{iT}) = P(Y_{iT} = 1|Y_{iT-1} = y_{iT})$. The naive forecast is the Bayes decision under the naive probability:

$$d_N(Y_i^T) = \mathbb{I}\left\{P_N(Y_{i,T+1} = 1|Y_{iT}) \geq \frac{a_{01}}{a_{01} + a_{10}}\right\}.$$

Figure 1 illustrates how in dynamic discrete choice forecasting problems when model parameters are only set-identified, different parameters in the identified set may lead to different forecasts, some of which are more accurate than others. Each plot is based on a

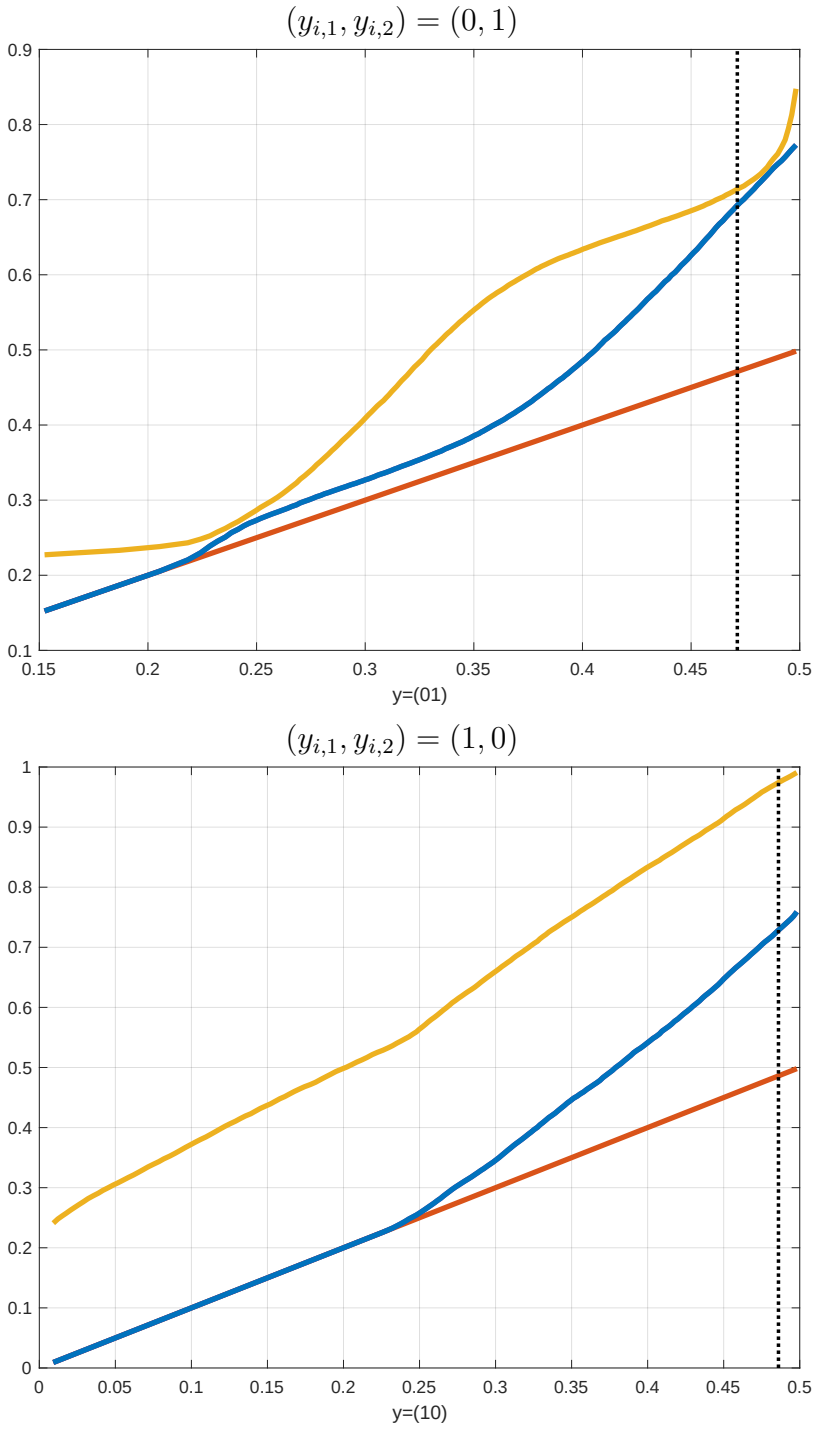

Figure 1: QQ plot of the risk of the naive forecast d_N (yellow line) and minimax forecast d_{mm} (blue line) for Y_{i3} having observed $(y_{i,1}, y_{i,2}) = (0, 1)$ against the infeasible Bayes forecast $d^*_{b,\theta}$ for draws from the identified set Θ_0. The orange line is the 45 degree line and represents the risk associated with $d^*_{b,\theta}$.

large number of draws from the identified set for $(\beta, \Pi_{\lambda,y})$ when $T = 2$ and the true value of β is 0.2. The draws for $(\beta, \Pi_{\lambda,y})$ are obtained by first drawing uniformly from the identified set for β, then drawing from the identified set for $\Pi_{\lambda,y}$ given β. To do so, we first draw a vector of probabilities, say $\tilde{p}$, uniformly from the simplex in $\mathbb{R}^{62}$ (the number of points of support of the discretized prior). We then use exponential tilting to compute the probability distribution closest to $\tilde{p}$ that solves the moment conditions at the draw $\tilde{\beta}$. The resulting distribution $\tilde{\Pi}_{\lambda,y}$ and $\tilde{\beta}$ are in the identified set for $(\beta, \Pi_{\lambda,y})$. Thus, each draw explains the data up to date T.

Consider the problem of forecasting Y_{i3} under symmetric loss, having observed $(y_{i1}, y_{i2}) = (0, 1)$ or $(1, 0)$ (the Bayes, minimax, and naive forecasts are identical when we observe $(y_{i1}, y_{i2}) = (0, 0)$ or $(1, 1)$). For each draw, say θ, from the identified set Θ_0 we compute the Bayes forecast $d^*_{\theta,b}$. This is an infeasible first-best, as it requires knowledge of θ. We calculate the risk of this forecast under $\mathbb{P}_\theta$ and compare it with the risk of two feasible forecasts, namely the minimax forecast (which is also the minimax-regret forecast as the loss function is symmetric) and the naive forecast assuming Markovianity. Figure 1 depicts QQ plots of the risk of the two feasible forecasts relative to the Bayes forecast. In both panels, the 45 degree line (in orange) corresponds to the infeasible Bayes decision, the blue line is the minimax forecast and the yellow line is the naive forecast. The black vertical line denotes the Bayes risk at the true parameter.

We see that the risk of the forecasts are different for different draws $\theta \in \Theta_0$ because distribution of future observations under $\mathbb{P}_\theta$ is different. The minimax and Bayes forecasts are the same for some of the draws from the identified set (the proportion for which the blue line runs along the 45 degree line). The blue line then departs from the orange line for draws where the minimax and Bayes forecasts are different and, consequently, have different posterior risks.

We shall now study how the properties of the forecasts vary as we vary the true β. Note that in varying β we vary the moments observed in the data and, as a consequence, we vary the identified set Θ_0. For each true value of β, we calculate the maximum risk and maximum regret over Θ_0 for each forecast as a function of conditioning variables y^T.

Figures 2 displays the maximum risk and maximum regret for the four forecasts under symmetric loss for $T = 2$. As predicted, the minimax-optimal and minimax-regret optimal forecasts are identical and minimize both maximum risk and maximum regret. The next best-performing forecast under both criteria is the infeasible Bayes forecast, followed by the

naive forecast. The results are quite different depending on the conditioning information and true value of β, however.

Figure 3 presents corresponding plots repeating the exercise under asymmetric loss with $a_{01} = 2$, $a_{10} = 1$. Here the minimax-optimal and minimax-regret optimal forecasts can differ quite substantially, and their behavior under one optimality criterion can be quite different under the other. From top panel we see that the minimax regret-optimal forecast is the worst-performing forecast in terms of maximum risk for certain values of y^T and β_0. On the other hand, the bottom panel shows that the minimax-optimal forecast can be the worst-performing forecast in terms of maximum regret for certain values of y^T and β_0. The ordering with respect to the other forecasts is preserved, however: the minimax-optimal forecast dominates the Bayes forecast in terms of maximum risk which, in turn, dominates the naive forecast. Similarly, the the minimax regret-optimal forecast dominates the Bayes forecast in terms of maximum regret which, in turn, dominates the naive forecast.

3.2 Computing Extreme Probabilities

The challenge in implementing the minimax-optimal and minimax regret-optimal forecasts is to solve the extremum problems:

$$p_L(Y_i^T) := \min_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} \quad \text{and} \quad p_U(Y_i^T) := \max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\}\,.$$

In the examples we study, θ may be partitioned as $\theta = (\phi, \nu)$ where ϕ is a low-dimensional parameter and ν is a probability measure. The probability $\mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\}$ and the restrictions defining Θ_0 are linear functionals of ν. Therefore, the problem of minimizing and maximizing $\mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\}$ over Θ_0 can be split into an inner optimization over the high-dimensional parameter ν and an outer optimization over the low-dimensional parameter ϕ. We shall use duality techniques to simplify computation of the inner optimization over the high-dimensional component.

To fix ideas, suppose $\Phi_t = \Phi$, a known cdf, for all t, but we do not wish to specify the distribution $\Pi_{\lambda,y}$. The parameter space reduces to $\Theta = \{(\beta, \Pi_{\lambda,y})\}$. The identified set is

$$\Theta_0 = \{\theta = (\beta, \Pi_{\lambda,y}) \in \Theta : p(y^T | \beta, \Pi_{\lambda,y}) = p(y^T) \;\; \forall\, y^T\}\,,$$

where $p(y^T)$ is the actual probability of $Y_i^T = y^T$ that the econometrician observes in the

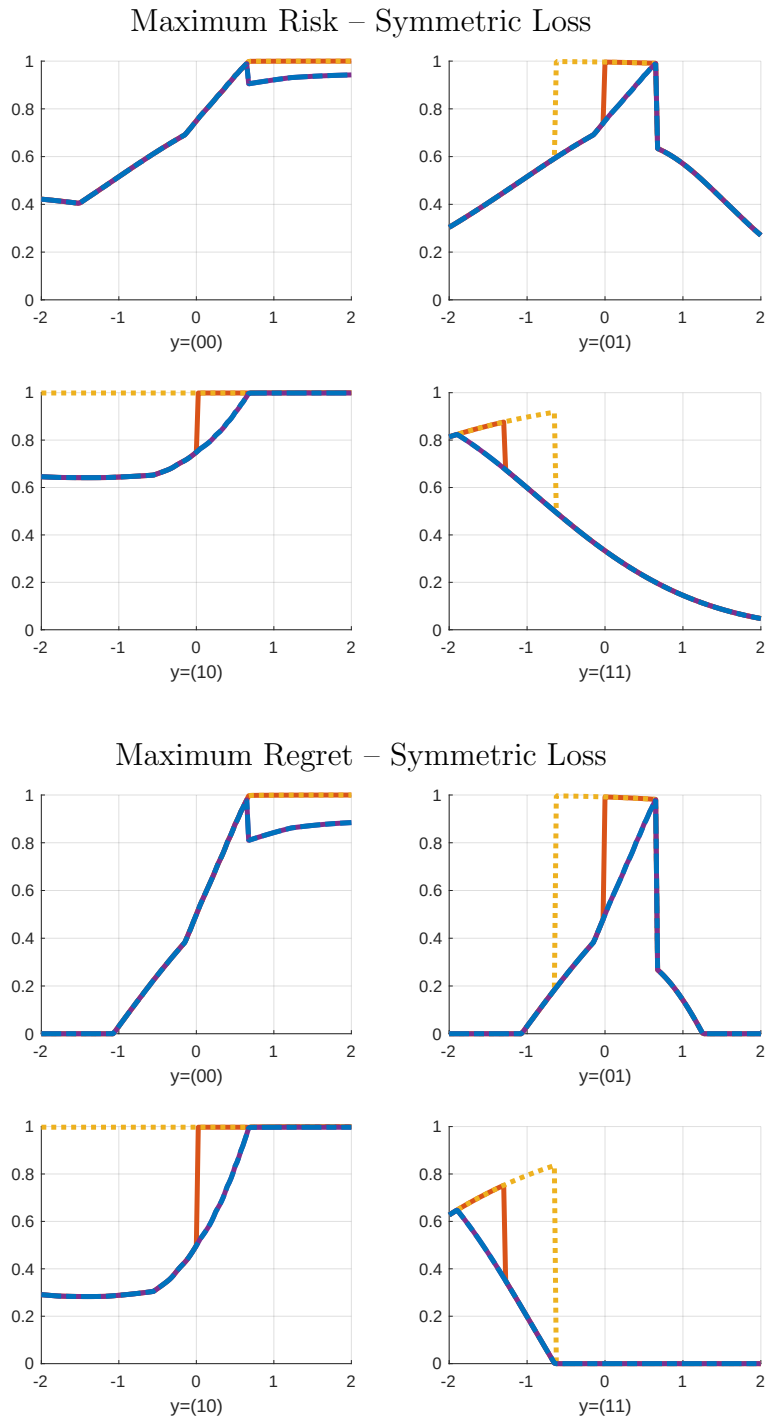

Figure 2: Maximum risk and regret over Θ_0 as a function of β_0 for $T = 2$ under symmetric loss. Results are plotted by $(y_{i,1}, y_{i,2})$. Dot-dashed blue lines are the minimax forecast, solid purple lines are the minimax regret forecast, solid orange lines are the Bayes forecast, and dotted yellow lines are the naive forecast.

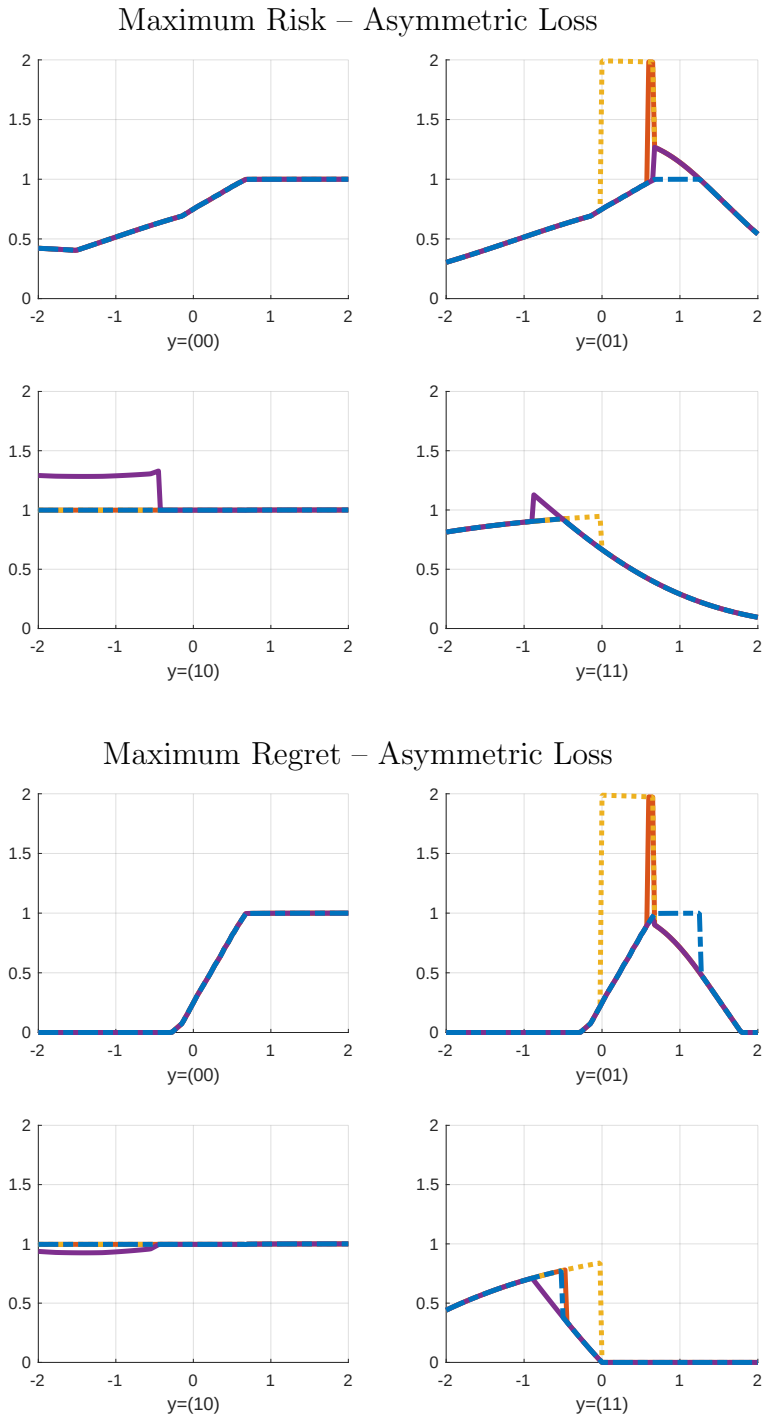

Figure 3: Maximum risk and regret over Θ_0 as a function of β_0 for $T = 2$ under asymmetric loss ($a_{01} = 2, a_{10} = 1$). Results are plotted by $(y_{i,1}, y_{i,2})$. Dot-dashed blue lines are the minimax forecast, solid purple lines are the minimax regret forecast, solid orange lines are the Bayes forecast, and dotted yellow lines are the naive forecast.

data, and $p(y^T|\beta, \Pi_{\lambda,y})$ is the model-implied probability of observing $Y_i^T = y^T$:

$$p(y^T|\beta, \Pi_{\lambda,y}) = \int p(y^T|y_0, \lambda; \beta)\, \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)\,,$$

with

$$p(y^T|y_0, \lambda; \beta) = \prod_{t=1}^{T} \Phi(\beta y_{t-1} + \lambda)^{y_t}(1 - \Phi(\beta y_{t-1} + \lambda))^{1-y_t}\,. \tag{21}$$

Because $p(y^T|\beta, \Pi_{\lambda,y}) = p(y^T)$ for any $\theta \in \Theta_0$, the forecast probability is

$$\begin{aligned} &\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T = y^T\} \\ &= \frac{\int \Phi(\beta y_T + \lambda)\left(\prod_{t=1}^{T} \Phi(\beta y_{t-1} + \lambda)^{y_t}(1 - \Phi(\beta y_{t-1} + \lambda))^{1-y_t}\right) \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)}{p(y^T)}\,. \end{aligned} \tag{22}$$

We may write the forecast probability more abstractly with $x = (\lambda, y_0)$ and $\nu = \Pi_{\lambda,y}$ as

$$\int m(x, \beta)\, \mathrm{d}\nu(x)\,,$$

where

$$m((\lambda, y_0), \beta) = \frac{\Phi(\beta y_T + \lambda)\left(\prod_{t=1}^{T} \Phi(\beta y_{t-1} + \lambda)^{y_t}(1 - \Phi(\beta y_{t-1} + \lambda))^{1-y_t}\right)}{p(y^T)}\,.$$

The $K = 2^T$ constraints defining Θ_0 can be stacked in a vector of moment conditions:

$$\int g(x, \beta)\, \mathrm{d}\nu(x) = r\,,$$

where we identify each element $k = 1, \ldots, K$ with a unique sequence $y_k^T \in \{0, 1\}^T$ and let

$$g_k((\lambda, y_0), \beta) = p(y_k^T|y_0, \lambda; \beta)\,.$$

The expression for $p(y_k^T|y_0, \lambda; \beta)$ is given in (21), and $r_k = p(y_k^T)$.

In the remainder we will focus on the maximization of $\mathbb{P}_\theta\{Y_{iT+1} = 1|Y_i^T\}$ and defer the solution to the corresponding minimization problem to the end of this section. For now we will also assume that $\Pi_{\lambda,y}$ has finite support, say $x_1, \ldots, x_L$. Thus, in turn the distribution $\Pi_{\lambda,y}$ can be characterized by a vector $\pi \in \mathbb{R}^L$ and the solution can be computed using linear programming, similar to Honoré and Tamer (2006).

Define the $L \times 1$ vector $m(\beta) = [m(x_1, \beta), \ldots, m(x_L, \beta)]'$ and let $m_l(\beta) = m(x_l, \beta)$. The predictive probability and the vector of moment conditions may be written as

$$m(\beta)'\pi\,, \quad \text{and} \quad G(\beta)\pi = r\,, \tag{23}$$

respectively, where the l^{th} column of $K \times L$ matrix $G(\beta)$ is $G_l(\beta) = g(x_l, \beta)$. Using this notation, we show in Appendix A.1 that one can rewrite

$$\begin{aligned} \max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} &= \max_\beta \left(\max_{\Pi_{\lambda,y}:(\beta,\Pi_{\lambda,y}) \in \Theta_0} \int m(x,\beta)\, \mathrm{d}\Pi_{\lambda,y}(x) \right) \\ &= \max_\beta \left(\inf_\mu \max_l \big(m_l(\beta) + \mu'(r - G_l(\beta)) \big) \right)\,, \end{aligned} \tag{24}$$

where μ is a $K \times 1$ vector of Lagrange multipliers for the vector of moment conditions. The inner program in the second line of the right-hand side involves optimizing piecewise-linear functions and may therefore be computed efficiently by linear programming:

$$\inf_\mu \max_l \left(m_l(\beta) + \mu'(r - G_l(\beta)) \right) = \min_v f'v \quad \text{subject to} \quad A(\beta)v \geq -m(\beta)\,, \tag{25}$$

where

$$v \in \mathbb{R}^{K+1}, \quad f' = [0_{1\times K}\,,\, 1] \quad A(\beta) = [G'(\beta) - (\iota_{L\times 1} \otimes r')\,,\, -\iota_{L\times 1}]\,.$$

Here ι denotes a vector of ones and $\otimes$ is the Kronecker product. The above linear program delivers the value of the inner optimization over $\Pi_{\lambda,y}$ for fixed β. The program returns no solution when β is not feasible, i.e., when there does not exist a probability measure $\Pi_{\lambda,y}$ that satisfies the moment conditions at β. In this case, we set the value of the program to $-\infty$.

The extreme probabilities may therefore be efficiently computed by embedding the linear program in an outer optimization with respect to the scalar parameter β:

$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} = \max_\beta \Big(\min_v f'v \quad \text{subject to} \quad A(\beta)v \geq -m(\beta) \Big)\,, \tag{26}$$

with the understanding that the inner optimization problems take the value $-\infty$, when the linear program returns no solution. The minimization of $\mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\}$ can be implemented by replacing the *inf* and *max* operations in (24) with *sup* and *min* operations, respectively, and the *minimization* and $\leq$ in (25) and (26) by a *maximization* and a $\geq$, respectively. This is how the numerical results in Section 3.1 were generated.

We discuss in Appendix A.2 how the analysis can be extended to non-discrete correlated random effects distributions $\Pi_{\lambda,y}$. The generalized version of the dual representation in (24) for the inner optimzation over $\Pi_{\lambda,y}$ takes the form:

$$\max_{\Pi_{\lambda,y}:(\beta,\Pi_{\lambda,y})\in\Theta_0} \int m(x,\beta)\,\mathrm{d}\Pi_{\lambda,y}(x) = \inf_{\mu} \sup_{x} \left(m(x,\beta) + \mu'(r - g(x,\beta))\right) , \tag{27}$$

where, as in the discrete case, μ is a vector of multipliers of dimension $K = 2^T$. Unfortunately, a convenient representation as a linear programming problem is not available.

4 Robustness to Breaks and Misspecification

In the previous section, we generated the parameter subspace Θ_0 with respect to which we robustified the forecasts through a partial-identification argument. We now will consider two alternatives: structural breaks and misspecification. While both structural breaks and model misspecification are relevant concerns for *any* forecasting model, the subsequent exposition continues to use the dynamic discrete choice model as the running example.

Structural breaks in the distribution of the U_{it} and misspecification of the random effects distribution can both be handled by allowing the distribution under consideration to vary over set of distributions. The set of distributions will give rise to a set of model parameters Θ_0 over which we seek to robustify the forecast. The extremum problems characterizing the lower and upper forecast probabilities as θ varies over Θ_0 may be solved in a computationally tractable manner using convex duality when the set of distributions is constrained via statistical divergence. In the exposition below, we shall follow the robustness literature in economics and focus on Kullback–Leibler divergence neighborhoods. In practical terms, this means that the linear program used to characterize the lower and upper forecast probabilities in Section 3.2 is replaced by a convex program involving an exponential tilt.

4.1 Structural Breaks

In the dynamic discrete choice model (1) three types of breaks can, in principle, occur at the forecast origin T: a break in the distribution of the U_{it}s, a break in the λ_is, and a break in β. To fix ideas, suppose that the researcher allows for $\Phi_t = \Phi$, a known cdf, for dates $t = 1, \ldots, T$, but wishes to allow for the possibility that $\Phi_{T+1} \neq \Phi$. Formally, the researcher might like to allow for $\Phi_{T+1} \in N$, a neighborhood of Φ. Even when the homogeneous

parameters β and random effects distribution Π is known at date T, there are still a set of posterior distributions for Y_{iT+1} corresponding to different $\Phi_{T+1} \in N$.

To map into the earlier setup, we would parameterize the model by $\theta = (\beta, \Pi_{\lambda,y}, \Phi_{T+1})$. The identified set (assuming point-identification of $(\beta, \Pi_{\lambda,y})$) would be $\Theta_0 = (\beta, \Pi_{\lambda,y}) \times N$. Consider the maximum forecast probability as θ varies over Θ_0:

$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} = \max_{\Phi_{T+1} \in N} \int \left(\int \tilde{m}(x, u, \beta) \, \mathrm{d}\Pi_{\lambda,y}(x) \right) \mathrm{d}\Phi_{T+1}(u) \,,$$

where

$$\tilde{m}((\lambda, y_0), u, \beta) = \frac{\mathbb{I}\{\lambda + \beta y_T \geq u\} \left(\prod_{t=1}^T \Phi(\beta y_{t-1} + \lambda)^{y_t} (1 - \Phi(\beta y_{t-1} + \lambda))^{1-y_t} \right)}{p(y^T)} \,.$$

The extremum problem is one of maximizing a linear functional of Φ_{T+1} subject to the constraint that $\Phi_{T+1} \in N$. When

$$N = \big\{ \Phi_{T+1} : K(\Phi_{T+1} \| \Phi) \leq \delta \big\},$$

the maximization problem has a dual representation in terms of optimization over a single scalar parameter η, which is the Lagrange multiplier on the constraint $\Phi \in N$:

$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} = \inf_{\eta > 0} \eta \log \left(\int e^{\eta^{-1} \check{m}(u,\beta)} \, \mathrm{d}\Phi(u) \right) + \eta\delta \,,$$

where $\check{m}(u, \beta) = \int \tilde{m}(x, u, \beta) \, \mathrm{d}\Pi_{\lambda,y}(x)$. Similar duality results underlie the robustness literature in economics (see, e.g, Hansen and Sargent (2007) and references therein), generalized empirical likelihood estimation via exponential tilting (Kitamura and Stutzer, 1997), and latent variable methods in econometrics (Schennach, 2014; Christensen and Connault, 2019). The lower value $\min_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\}$ is computed analogously, replacing the $\inf_{\eta > 0}$ with $\sup_{\eta < 0}$.

Breaks in λ_i can be viewed as a location shift of the distribution Φ_t and are therefore subsumed under breaks in the distribution of U_{it}. Breaks in β do not require the use of the Kullback-Leibler divergence. They can be constrained by defining N as the set

$$N = \big\{ \beta_{T+1} : |\beta_{T+1} - \beta| \leq \delta \big\}.$$

Conditional on Y_{iT} the probability $\mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} = \Phi_{T+1}(\lambda_i + \beta_{T+1} Y_{iT})$ is a monotone

function in β_{T+1}, which means that the extremum is attained either at $\beta - \delta$ or $\beta + \delta$.

4.2 Misspecification

Suppose the forecaster used a parametric correlated random effect model, $\Pi_{\lambda,y} = \Pi(\lambda, y_0; \xi)$ for $\xi \in \Xi$, a set of auxiliary parameters. The forecaster is worried that this parametric specification might be misspecified. Therefore, for each $\xi \in \Xi$ she allows for the possibility that $\Pi_{\lambda,y} \in N(\xi)$, a neighborhood of $\Pi(\lambda, y; \xi)$. TA natural specification of $N(\xi)$ is to view the researcher's parametric correlated random effect model to be approximately correct, in the sense that there is some $\Pi(\lambda, y_0; \xi)$ close to the true distribution $\Pi_{\lambda,y}$. We follow the same approach as in Section 4.1 and constrain $N(\xi)$ to be a Kullback–Leibler neighborhood of $\Pi(\lambda, y; \xi)$. That is, $N(\xi) = \{\Pi : K(\Pi \| \Pi(\,\cdot\,; \xi) \leq \delta\}$ for each $\xi \in \Xi$. We assume the absence of a structural break: $\Phi_t = \Phi$ for all t. The parameter space is therefore $\Theta = \{(\beta, \xi, \Pi_{\lambda,y}) : \Pi_{\lambda,y} \in N(\xi)\}$. The set Θ_0 is given by

$$\Theta_0 = \{\theta = (\beta, \xi, \Pi_{\lambda,y}) \in \Theta : p(y^T | \beta, \Pi_{\lambda,y}) = p(y^T) \ \ \forall\, y^T \ \text{ and } \ \Pi_{\lambda,y} \in N(\xi)\}\,,$$

where the model-implied probabilities $p(y^T | \beta, \Pi_{\lambda,y})$ are as before.

Duality techniques may be used to simplify computation of the lower and upper forecast probabilities as θ varies over Θ_0. Consider the maximum forecast probability

$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} = \max_{\beta,\xi} \left(\max_{\Pi_{\lambda,y} \in N(\xi) : (\beta,\xi,\Pi_{\lambda,y}) \in \Theta_0} \int m(x, \beta)\, \mathrm{d}\Pi_{\lambda,y}(x) \right)\,,$$

where the inner maximum is $-\infty$ if it runs over an empty set. Under a mild constraint qualification condition, the inner extremum problem admits a dual representation in terms of a scalar Lagrange multiplier η on the constraint $\Pi_{\lambda,y} \in N(\xi)$ and a 2^T vector of multipliers μ on the constraint $\int g(x, \beta)\, \mathrm{d}\nu(x) = r$:

$$\max_{\theta \in \Theta_0} \mathbb{P}_\theta\{Y_{iT+1} = 1 | Y_i^T\} = \max_{\beta,\xi} \left(\inf_{\eta > 0, \mu} \eta \log \left(\int e^{\eta^{-1}(m(x,\beta) + \mu'(r - g(x,\beta)))}\, \mathrm{d}\Pi(x; \xi) \right) + \eta\delta \right)\,,$$

see Christensen and Connault (2019) for a formal statement and related regularity conditions. They also study convergence of the above value as $\delta \to \infty$ to the maximum over the identified set without parametric restrictions on $\Pi_{\lambda,y}$.

The lower probability is computed similarly, replacing the $\max_{\beta,\xi}$ with $\min_{\beta,\xi}$ and the $\inf_{\eta>0,\mu}$ with $\sup_{\eta<0,\mu}$.

5 Forecasts Based on Plug-in Estimators

So far we have abstracted from sampling uncertainty, treating the data-moments $p(y^T)$ and extreme probabilities as known to the econometrician. In practice, the data moments $p(y^T)$ may be estimated by their empirical counterparts $\hat{p}(y^T)$ for each $y^T \in \{0,1\}^T$. The estimators $\hat{p}(y^T)$ may be plugged-in in place of $p(y^T)$ in the linear programs from the previous section. In the context of the illustration in Section 2.2, the vector r in (23) needs to be replaced by the vector $\hat{r}$ with elements $\hat{r}_k = \hat{p}(Y_k^T)$ and the optimization problems in (24) and (25) need to be changed accordingly. This will lead to estimates $\hat{p}_L(y^T)$ and $\hat{p}_U(y^T)$ of the lower and upper extreme probabilities $p_L(y^T)$ defined in (9) and $p_U(y^T)$ defined in (10).

For fixed T and $N \longrightarrow \infty$, the estimator $\hat{p}$ of the reduced-form choice probabilities is consistent under a large variety of low-level regularity conditions. Based on this, the consistency of $\hat{p}_U$ and $\hat{p}_L$ is generally straightforward to establish. Suppose the econometrician has symmetric loss, so her minimax-optimal and minimax regret-optimal forecasts will agree. Without loss of generality, normalize $a_{01} = a_{10} = 1$. A natural empirical counterpart to the minimax binary forecast (12) is the plug-in forecast for Y_{iT+1} having observed $Y_i^T = y^T$ is

$$\hat{d}_{mm} = \mathbb{I}\left\{1 \leq \hat{p}_L(y^T) + \hat{p}_U(y^T)\right\} . \tag{28}$$

Conditioning on the data, the maximum risk of $\hat{d}_{mm}(y^T)$ in excess of the maximum risk of $d_{mm}(y^T)$ is zero if $p_L(y^T) + p_U(y^T) - 1$ and $\hat{p}_L(y^T) + \hat{p}_U(y^T) - 1$ both have the same sign. Otherwise, the excess maximum risk is

$$ER(\hat{d}_{mm}(y^T)) = \left[\begin{array}{ll} 1 - p_L(y^T) - p_U(y^T) & \text{if } p_L(y^T) + p_U(y^T) < 1,\ \hat{p}_L(y^T) + \hat{p}_U(y^T) \geq 1\,, \\ p_L(y^T) + p_U(y^T) - 1 & \text{if } p_L(y^T) + p_U(y^T) \geq 1,\ \hat{p}_L(y^T) + \hat{p}_U(y^T) < 1\,. \end{array}\right. \tag{29}$$

This calculation reveals a tradeoff between the excess risk of the plug-in decision and the precision of the plug-in estimators $\hat{p}_U$ and $\hat{p}_L$. When $|1 - p_L(y^T) - p_U(y^T)|$ is large the excess risk of the plug-in forecast is (possibly) larger, but p_U and p_L do not need to be estimated as precisely to control the excess risk. Conversely, when $|1 - p_L(y^T) - p_U(y^T)|$ is small the excess risk of the plug-in forecast is smaller but p_U and p_L need to be estimated more precisely to control the excess risk.

6 Conclusion

Panel data sets generate many challenges for forecasters. In this paper we focused on the set-identification of parameters and correlated random effects distributions in panel dynamic discrete choice models. We showed that different parameterizations that lie in the identified sets can lead to different forecasts that *ex post* are associated with different forecast errors. We proposed use of robust forecasts that are obtained by solving a minimax risk or regret problem. These methods have wide applicability beyond panel discrete choice models, in enviroments in which a forecaster is concerned about structural breaks or model misspecification.

References

Baltagi, B. H. (2008). Forecasting with panel data. *Journal of Forecasting 27*(2), 153–173.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York.

Chamberlain, G. (2000). Econometric applications of maxmin expected utility. *Journal of Applied Econometrics 15*(6), 625–644.

Chamberlain, G. (2001). Minimax estimation and forecasting in a stationary autoregression model. *American Economic Review, Papers & Proceedings 91*(2), 55–59.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica 78*(1), 159–168.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica 81*(2), 535–580.

Christensen, T. and B. Connault (2019). Counterfactual sensitivity and robustness. *Manuscript, New York University*.

Clements, M. P. and D. F. Hendry (1998). *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

Csiszár, I. and F. Matúš (2012). Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika 48*(4), 637–689.

Giacomini, R. and T. Kitagawa (2018). Robust bayesian inference for set-identified models. *Manuscript, University College London*.

Gilboa, I. and D. Schmeidler (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics 18*(2), 141–153.

Gu, J. and R. Koenker (2016). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics (Forthcoming)*.

Hansen, L. P. and T. J. Sargent (2001). Robust control and model uncertainty. *The American Economic Review 91*(2), 60–66.

Hansen, L. P. and T. J. Sargent (2007). *Robustness*. Princeton University Press.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*(5), 611–629.

Kitagawa, T. (2012). Estimation and inference for set-identified parameters using posterior lower probabilities. *Manuscript, University College London*.

Kitamura, Y. and M. Stutzer (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica 65*(4), 861–874.

Liu, L. (2019). Density forecasts in panel data models: A semiparametric bayesian perspective. *Manuscript, Indiana University*.

Liu, L., H. R. Moon, and F. Schorfheide (2018a). Forecasting with a panel tobit model. *Manuscript, University of Pennsylvania*.

Liu, L., H. R. Moon, and F. Schorfheide (2018b). Forecasting with dynamic panel data models. *Manuscript, University of Pennsylvania*.

Robbins, H. (1951). Asymptocially subminimax solutions of compound decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Volume I. University of California Press, Berkeley and Los Angeles.

Schennach, S. M. (2014). Entropic latent variable integration via simulation. *Econometrica 82*(1), 345–385.

Wald, A. (1950). *Statistical Decision Functions*. John Wiley, New York.

Online Appendix: Robust Forecasting under Partial Identification and Misspecification

Timothy Christensen, Hyungsik Roger Moon, and Frank Schorfheide

A Duality Results

A.1 Finite-Dimensional Case

Suppose that $\pi \in \mathbb{R}^L$ is a discrete probability measure on a finite support $\{x_1, \ldots, x_L\}$. That is, $\pi \geq 0$ and $1'\pi = 1$. Also, suppose that all moment conditions are equality conditions. The constrained optimization problem of Section 3.2 becomes

$$\max_{\pi} \; m'\pi \quad \text{subject to} \quad G\pi - r = 0, \; \iota'\pi - 1 = 0, \; \pi \geq 0.$$

The Lagrangian problem is

$$\max_{\pi} \min_{\mu,\zeta} \min_{\kappa \geq 0} \; \mathcal{L}(\pi, \mu, \zeta, \kappa).$$

Here μ, ζ, and κ are the Lagrange multipliers on the three constraints and

$$\begin{aligned} \mathcal{L}(\pi, \mu, \zeta, \kappa) &= m'\pi + \mu' (G\pi - r) + \zeta (\iota'\pi - 1) + \kappa'\pi \\ &= (m + G'\mu + \zeta\iota + \kappa)' \pi - (\mu' r + \zeta) \end{aligned}$$

Then, by duality we have

$$\max_{\pi} \min_{\mu,\zeta} \min_{\kappa \geq 0} \; \mathcal{L}(\pi, \mu, \zeta, \kappa) = \min_{\mu,\zeta} \min_{\kappa \geq 0} \max_{\pi} \; \mathcal{L}(\pi, \mu, \zeta, \kappa).$$

Now consider the problem

$$\min_{\kappa \geq 0} \max_{\pi} \; m^*(\kappa; \mu, \zeta)'\pi \quad \text{where} \quad m^*(\kappa; \mu, \zeta) = m + G'\mu + \zeta\iota + \kappa.$$

Conditional on κ, the maximization with respect to π is solved by assigning probability one to the largest element of the vector $m^*(\kappa; \mu, \zeta)$. Conditional on this optimal choice for π, the optimal choice of κ is $\kappa = 0$. Therefore,

$$\min_{\kappa \geq 0} \max_{\pi} \; m^*(\kappa; \mu, \zeta)'\pi = \max_{l \in \{1, \ldots, L\}} \; m_l + \mu' G_l + \zeta,$$

where m_l is the l^{th} element of m and G_l is the l^{th} column of the $K \times L$ matrix G.

Combining the intermediate results, we obtain (24) in the main text:

$$\begin{aligned}\max_{\pi} \min_{\mu,\zeta} \min_{\kappa \geq 0} \mathcal{L}(\pi, \mu, \zeta, \kappa) &= \min_{\mu,\zeta} \left(\max_{l \in \{1,\ldots,L\}} m_l + \mu' G_l \right) + \zeta - (\mu' r + \zeta) \\ &= \min_{\mu} \left(\max_{l \in \{1,\ldots,L\}} m_l + \mu'(G_l - r) \right).\end{aligned}$$

The dual problem may be implemented as a linear program:

$$\min_{v} f'v \quad \text{subject to} \quad Av \geq -m$$

where $v \in \mathbb{R}^{K+1}$, $f' = [0_{1\times K}, 1]$ and $A = [G' - (\iota_{L\times 1} \otimes r'), -\iota_{L\times 1}]$. To see the equivalence, write the Lagrangian problem for the linear program, using the $L \times 1$ vector of Lagrange multipliers ξ:

$$\min_{v_{1:K}} \left(\min_{\nu_{K+1}} \max_{\phi_l \geq 0} v_{K+1} + \sum_{l=1}^{L} \phi_l \big(m_l + v'_{1:K}(G_l - r) - v_{K+1} \big) \right).$$

Note that whenever $m_l + v'_{1:K}(G_l - r) - v_{K+1} > 0$, the inner maximization is solved by setting $\phi_l = +\infty$. Thus, v_{K+1} has to be chosen such that $m_l + v'_{1:K}(G_l - r) - v_{K+1} \leq 0$. If the inequality holds strictly, then the optimal choice for ϕ_l is $\phi_l^* = 0$. In turn, the optimal choice for v_{K+1} is

$$v_{K+1}^* = \max_{l \in \{1,\ldots,L\}} m_l + v'_{1:K}(G_l - r).$$

Therefore, we can deduce that

$$\begin{aligned}&\min_{v_{1:K}} \left(\min_{\nu_{K+1}} \max_{\phi_l \geq 0} v_{K+1} + \sum_{l=1}^{L} \phi_l \big(m_l + v'_{1:K}(G_l - r) - v_{K+1} \big) \right) \\ &= \min_{v_{1:K}} \left(\max_{l \in \{1,\ldots,L\}} m_l + v'_{1:K}(G_l - r) \right),\end{aligned}$$

which establishes that the linear programs solves the original problem.

A.2 Infinite-Dimensional Case

We now turn to the case in which x has continuous support X. We shall optimize with respect to probability measures that have a density with respect to a common σ-finite dominating

measure on $(X, \mathcal{X})$, where $\mathcal{X}$ denotes the Borel σ-field on X. In the example in Section 3.2, we could let the random effects have continuous support $\Lambda \subseteq \mathbb{R}$, take $X = \Lambda \times \{0, 1\}$, and set ν as the product of Lebesgue measure on Λ and counting measure on $\{0, 1\}$.

Formally, we consider the program

$$\sup_{\pi \in \mathcal{P}} \int m(x)\pi(x)\,\mathrm{d}\nu(x) \quad \text{subject to} \quad \int g(x)\pi(x)\,\mathrm{d}\nu(x) = r$$

where $m : X \to \mathbb{R}$ is a bounded function, $g : X \to \mathbb{R}^K$ are moment conditions and $r \in \mathbb{R}^K$ are the value in the population. The supremum is taken over the set $\mathcal{P}$ of all densities on $(X, \mathcal{X})$ that admit densities π with respect to ν and for which the integral $\int g(x)\pi(x)\,d\nu(x)$ is finite. We apply the results of Csiszár and Matúš (2012) who allow the maximum to be taken over all densities π for which the integral $\int g(x)\pi(x)\,\mathrm{d}\nu(x)$ is finite, thereby accommodating a very large class of constraint functions g. For instance, g does not have to be bounded. Appendix B of Christensen and Connault (2019) extends results from Csiszár and Matúš (2012) to setting with moment inequality constraints.

The solution in the discrete-support case may be expressed as

$$\min_{\mu} \left(\max_{x_1, \ldots, x_L} m(x_l) + \mu'(g(x_l) - r) \right).$$

Intuitively, we might expect we could let the support points become dense in X and replace the inner maximum over $x_1, \ldots, x_L$ with a sup over x. This intuition is correct, subject to some measure-theoretic considerations. Formally, the dual program is

$$\inf_{\mu : \sup_x (m(x) + \mu'(g(x) - r)) < +\infty} \left(\sup_x m(x) + \mu'(g(x) - r) \right),$$

where the inf on the right-hand side is to be understood as the ν-essential infimum (i.e. the "almost-everywhere" version of infimum used in measure-theoretic settings). When g consists entirely of bounded functions the outer infimum may be replaced with $\inf_\mu$. The constraint qualification condition guaranteeing equivalence of the primal and dual is

$$r \in \mathrm{ri} \left(\left\{ \int g(x)\pi(x)\,\mathrm{d}\nu(x) : \left| \int g(x)\pi(x)\,\mathrm{d}\nu(x) \right| < \infty \right\} \right),$$

where ri denotes relative interior.