Robust Forecasting

Timothy Christensen [1] Hyungsik Roger Moon [2] Frank Schorfheide [3]

[1]New York University
[2]University of Southern California
[3]University of Pennsylvania, CEPR, NBER, and PIER

June 2021

Robustness

- Ideally, we would like forecasts to be robust against:
  - model misspecification
  - structural breaks
  - outliers
  - $\cdots$
- Robustness can be achieved by minimax considerations: try to guarantee good performance under worst-case scenarios.
- Perennial problem: paranoia can lead to weak performance in regular periods.
- We will focus on a problem in which set identification generates bounds on the worst-case scenario.

Set identification and forecasting

- VAR and factor model intuition: only reduced-form matters for forecasting.
- In this paper, we consider a panel setting (large N, small T) in which
  - the size of the reduced-form parameter space grows over time,
  - the identified set shrinks over time,
  - ex post some parameters in the identified set lead to better forecasts than others.

This paper: decision-theoretic approach to robust forecasting

- Forecaster wishes to forecast a discrete outcome Y with a model $\mathbb{P}_\theta$
- Forecaster is unable to discriminate among a set of plausible parameterizations Θ_0
- Confront
  1. model uncertainty: $\theta \in \Theta_0$,
  2. sampling uncertainty: estimate Θ_0.
- This paper:
  - Characterize robust forecasts which deal with model uncertainty
  - Characterize efficient robust forecasts which deal with model uncertainty and sampling uncertainty
  - Develop a suitable asymptotic efficiency theory
  - Provide computationally efficient implementation based on linear/convex programming

General setup

- Forecaster wishes to forecast a discrete outcome Y with a model $\mathbb{P}_\theta$
- Prior to forecast, observe data $X_n \sim F_{n,P}$ where $P \in \mathcal{P} \subseteq \mathbb{R}^k$ is point-identified, regularly estimable
- A model specifies the following.
  - X_n and Y are linked via

$$\mathbb{P}_\theta(Y = y|X_n, P) = \mathbb{P}_\theta(Y = y|X_n), \quad X_n|\theta, P \sim F_{n,P}.$$

  - θ and P are linked via set-valued function $P \mapsto \Theta_0(P)$.
- For notational simplicity, we write

$$\mathbb{P}_\theta(Y = y) := \mathbb{P}_\theta(Y = y|X_n).$$

Running example: panel model for dynamic binary choice

$$Y_{it+1} = \mathbb{I}\left[\lambda_i + \beta Y_{it} \geq U_{it+1}\right], \quad \mathbb{P}\left(U_{it+1} \leq u | Y_i^t = y^t, \lambda_i = \lambda\right) = \Phi(u)$$

- Observe short panel: $(Y_{it})_{t=1}^{T}$, $i = 1, \ldots, n$ with T fixed, $n \to \infty$
- Y_{it} could be employment status, health status, ...
- Objective: forecast outcome Y_{iT+1} conditional upon a history $Y_i^T = y^T$
- Parameters: $\theta = \left(\beta, \Pi_{\lambda,y}\right)$ where $\Pi_{\lambda,y}$ is the joint distribution for (λ_i, Y_{i0}) (cf. Honoré & Tamer, 2006)

Running example: panel model for dynamic binary choice

- $\mathbb{P}_\theta$ denotes the conditional probability over $Y \equiv Y_{iT+1}$ given $Y_i^T = y^T$:

$$\mathbb{P}_\theta(Y=1) = \frac{\int \Phi(\beta y_{iT} + \lambda) p(y^T | y_0, \lambda; \beta) \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)}{\int p(y^T | y_0, \lambda; \beta) \mathrm{d}\Pi_{\lambda,y}(\lambda, y_0)} .$$

- Identified set is

$$\Theta_0(P) = \left\{ \theta = (\beta, \Pi_{\lambda,y}) \in \Theta : \underbrace{p(y^T | \beta, \Pi_{\lambda,y})}_{\text{model}} = \underbrace{\Pr(Y_i^T = y^T)}_{\text{data}} \ \ \forall y^T \in \{0,1\}^T \right\}$$

- Reduced-form parameter: $P = (\Pr(Y_i^T = y^T))_{y^T \in \{0,1\}^T}$, consistently estimable as $n \to \infty$

Why does partial identification matter for forecasting?

- Consider binary (classification) loss $\ell : \{0,1\} \times \{0,1\} \to \mathbb{R}$

$$\ell(y, d) = \mathbb{I}[y \neq d]$$

- The risk of a forecast d under any $\theta \in \Theta_0$ is

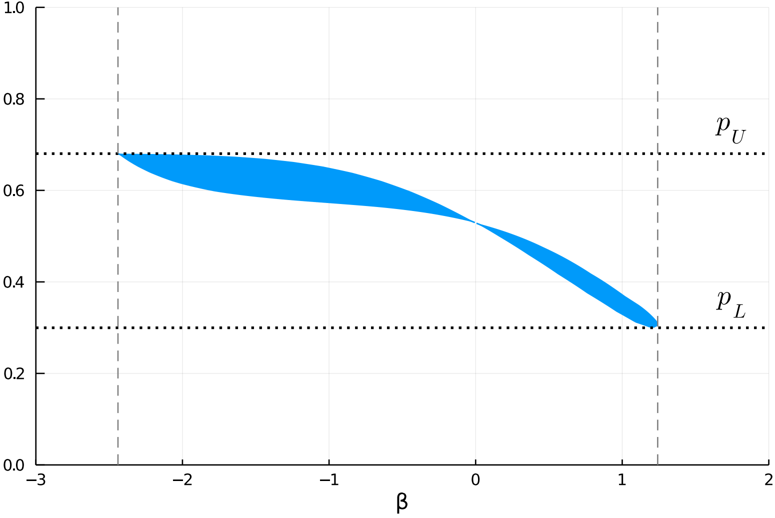
$$\mathbb{E}_\theta[\ell(Y, d)] = d(1 - \mathbb{P}_\theta(Y = 1)) + (1 - d)\mathbb{P}_\theta(Y = 0)$$

- If θ were known, the optimal forecast would minimize risk:

$$d^*_\theta = \operatorname*{argmin}_d \mathbb{E}_\theta[\ell(Y, d)] = \mathbb{I}\left[\mathbb{P}_\theta(Y = 1) \geq \frac{1}{2}\right]$$

Why does partial identification matter for forecasting?

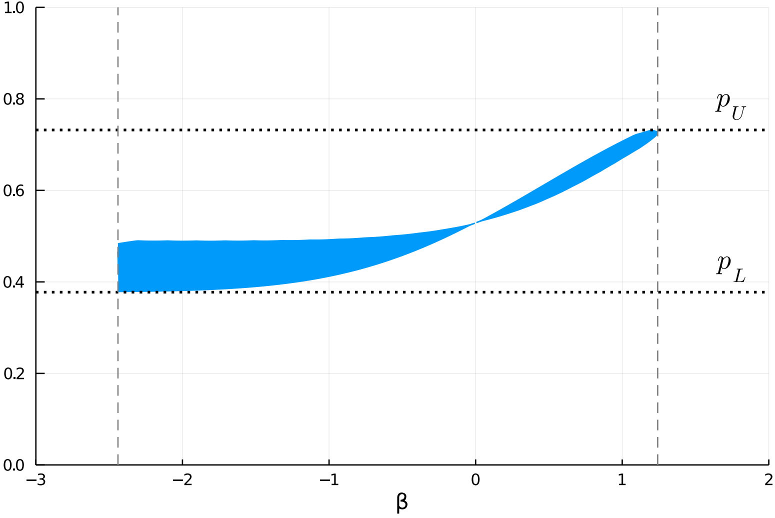
$\mathbb{P}_\theta(Y = 1 | Y_i^T = (0,0))$



- Honoré–Tamer (2006) parameterization
- $T = 2$
- $\theta = (\beta, \Pi_{\lambda,y})$
- $p_U := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)$
- $p_L := \inf_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)$

Why does partial identification matter for forecasting?

$\mathbb{P}_\theta(Y = 1 | Y_i^T = (1,1))$



- Honoré–Tamer (2006) parameterization
- $T = 2$
- $\theta = (\beta, \Pi_{\lambda,y})$
- $p_U := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)$
- $p_L := \inf_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)$

Robust forecasts (unknown θ, known $\Theta_0(P_0)$)

- Suppose that true P_0 and hence $\Theta_0 \equiv \Theta_0(P_0)$ is known, but the true $\theta \in \Theta_0$ is unknown
- Given a decision space $\mathcal{D}$, outcome space $\mathcal{Y}$, and loss function $\ell : \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$
- A minimax forecast solves

$$\inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\ell(Y, d)]$$

- A minimax regret forecast solves

$$\inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta_0} \Bigg(\underbrace{\mathbb{E}_\theta[\ell(Y, d)] - \inf_{d' \in \mathcal{D}} \mathbb{E}_\theta[\ell(Y, d')]}_{\text{regret}} \Bigg)$$

Example: binary/classification loss

- Let $\mathcal{D} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$, and

$$\ell(y, d) = \mathbb{I}[y = 1,\ d = 0] + \mathbb{I}[y = 0,\ d = 1]$$

- Define

$$p_L := \inf_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1), \qquad p_U := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(Y = 1)$$

- Minimax forecast

$$d_{mm} = \mathbb{I}\left[1 \leq p_L + p_U\right]$$

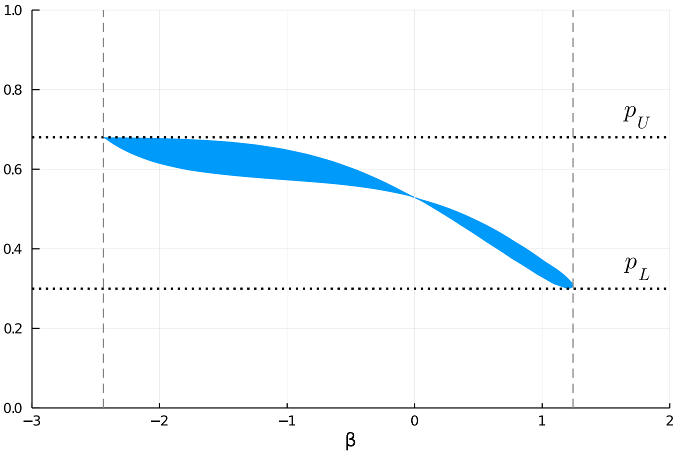
- Minimax regret forecast

$$d_{mmr} = \mathbb{I}\left[\left(\frac{1}{2} - p_L\right)_+ \leq \left(p_U - \frac{1}{2}\right)_+\right]$$

- Minimax (regret) forecasts under other loss functions depend similarly on p_U and p_L (see paper)

Robust forecasts in numerical example

$\mathbb{P}_\theta(Y = 1 | Y_i^T = (0, 0))$



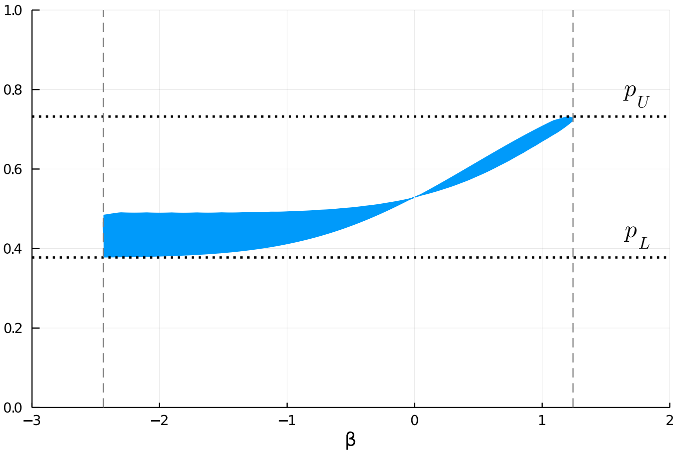
- Wide set of forecast probabilities $\{\mathbb{P}_\theta(Y = 1) : \theta \in \Theta_0\}$: $p_L = 0.2997$ and $p_U = 0.6803$.

- For $\theta \in \Theta_0$ such that $\mathbb{P}_\theta(Y = 1) < \frac{1}{2}$ $\Rightarrow d^*_{b,\theta} = 0$.
- For $\theta \in \Theta_0$ such that $\mathbb{P}_\theta(Y = 1) > \frac{1}{2}$ $\Rightarrow d^*_{b,\theta} = 1$.

- As $p_L + p_U < 1$, minimax and minimax regret forecasts are $d_{b,mm} = d_{b,mmr} = 0$.

Robust forecasts in numerical example

$\mathbb{P}_\theta(Y = 1 | Y_i^T = (1,1))$



- Wide set of forecast probabilities $\{\mathbb{P}_\theta(Y = 1) : \theta \in \Theta_0\}$: $p_L = 0.3775$ and $p_U = 0.7320$
- Here $p_L + p_U > 1$ so $d_{b,mm} = d_{b,mmr} = 1$.

Efficient robust forecasts (unknown θ, unknown Θ_0)

- Now dispense with the assumption that P_0, and hence $\Theta_0(P_0)$, is known
- We can learn about P, and therefore $\Theta_0(P)$, from the data X_n
- What's the best way to do this? We will use an asymmetric approach:
    - Use posterior distribution to handle uncertainty about P
    - Use minimax (regret) do handle uncertainty about $\theta \in \Theta_0(P)$.

Efficient robust forecasts (unknown θ, unknown Θ_0)

- Forecast is a function $d_n : \mathscr{X}_n \to \mathcal{D}$
- Forecaster has a prior Π over $\mathcal{P}$
- Evaluate d_n by its integrated maximum risk (or regret):

$$\mathcal{B}^n_{mm}(d_n;\pi) = \int_{\mathcal{P}} \left(\int_{\mathscr{X}_n} \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[\ell(Y, d_n(X_n))] \, \mathrm{d}F_{n,P}(X_n) \right) \mathrm{d}\Pi(P)$$

$$= \int_{\mathscr{X}_n} \underbrace{\left(\int_{\mathcal{P}} \sup_{\theta \in \Theta_0(P)} \mathbb{E}_\theta[\ell(Y, d_n(X_n))] \, \mathrm{d}\Pi_n(P|X_n) \right)}_{\text{posterior maximum risk}} \mathrm{d}F_n(X_n)$$

- Efficient robust forecast minimizes posterior maximum risk (or regret) for each realization X_n

Example: binary/classification loss

- Let $\mathcal{D} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$, and

$$\ell(y, d) = \mathbb{I}[y = 1,\ d = 0] + \mathbb{I}[y = 0,\ d = 1]$$

- Lower and upper probabilities are functions of P:

$$p_L(P) := \inf_{\theta \in \Theta_0(P)} \mathbb{P}_\theta(Y = 1)\,, \qquad p_U(P) := \sup_{\theta \in \Theta_0(P)} \mathbb{P}_\theta(Y = 1)\,,$$

- Recall: minimax forecast with known Θ_0:

$$d_{mm} = \mathbb{I}\left[1 \leq p_L + p_U\right]$$

- Efficient robust forecast (minimax) with unknown Θ_0:

$$d_{mm}(X_n) = \mathbb{I}\left[1 \leq \int p_L(P)\,\mathrm{d}\Pi_n(P|X_n) + \int p_U(P)\,\mathrm{d}\Pi_n(P|X_n)\right]$$

Asymptotic efficiency

- Benchmark: oracle forecast $d^o_{mm}(P)$ (minimax forecast if P were known)
- Excess maximum risk (or regret) of $d_n(X_n)$ is

$$\Delta\mathcal{R}_{mm}(d_n; P, X_n) = \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y, d_n(X_n))] - \sup_{\theta\in\Theta_0(P)} \mathbb{E}_\theta[\ell(Y, d^o_{mm}(P))]$$

- Integrated excess maximum risk (or regret) at P_0

$$\Delta\mathcal{B}^n_{mm}(d_n; P_0, \pi) = \int \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}_{mm}\left(d_n, P_{n,h}; X_n\right)\right] \pi\left(P_{n,h}\right) \mathrm{d}h\,, \quad P_{n,h} = P_0 + n^{-1/2}h$$

- Forecast rule $\{d_n\}_{n\geq 1}$ is asymptotically efficient-robust if it minimizes

$$\lim_{n\to\infty} \Delta\mathcal{B}^n_{mm}(d_n; P_0, \pi) = \pi(P_0) \underbrace{\int \left(\lim_{n\to\infty} \mathbb{E}_{P_{n,h}}\left[\sqrt{n}\Delta\mathcal{R}_{mm}\left(d_n, P_{n,h}; X_n\right)\right]\right) \mathrm{d}h}_{\text{ranking is independent of } \Pi}$$

for each $P_0 \in \mathcal{P}$

Asymptotic efficiency

- Say $\{d_n\}, \{\tilde{d}_n\} \in \mathbb{D}$ are asymptotically equivalent if $d_n(X_n)$ and $\tilde{d}_n(X_n)$ have the same asymptotic distribution under $F_{n,P_{n,h}}$ for all $P_0 \in \mathcal{P}$ and $h \in \mathbb{R}^k$

Theorem

(i) Let $\{\tilde{d}_n\} \in \mathbb{D}$ be ***asymptotically equivalent to the minimax efficient robust forecast (ERF)****.*
Then: for all $P_0 \in \mathcal{P}$,

$$\lim_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(\tilde{d}_n; P_0, \pi) = \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(d_n; P_0, \pi) .$$

(ii) If $p_L(P)$ and $p_U(P)$ are directionally—but not fully—differentiable at P_0, then for any $\{\tilde{d}_n\} \in \mathbb{D}$ that is ***not asymptotically equivalent to the minimax ERF****, we have*

$$\liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(\tilde{d}_n; P_0, \pi) > \inf_{\{d_n\}\in\mathbb{D}} \liminf_{n\to\infty} \Delta\mathcal{B}^n_{b,mm}(d_n; P_0, \pi)$$

for some $P_0 \in \mathcal{P}$.

Implications

- Asymptotic efficient-robustness extends to:
  - ERFs under any positive, smooth prior (not nec. subjective)
  - ERFs under misspecified likelihoods (provided asymptotically correct)
  - Bagged forecasts
- Plug-in rules $d^o_{mm}(\hat{P})$, $d^o_{mmr}(\hat{P})$ are inefficient under directional differentiability of $p_L(P)$, $p_U(P)$
  - $p_L(P)$, $p_U(P)$ typically linear programs or min-max programs
  - Directional differentiability is the rule, rather than the exception (e.g. Milgrom and Segal, 2002)

Simple illustration of plug-in inefficiency

- Suppose $\mathcal{P} = (0,1)$, $p_L(P) = P$, and

$$p_U(P) = \left[\begin{array}{ll} \frac{1}{2} & P < \frac{1}{2}\,, \\ (2P - \frac{1}{2}) \wedge 1 & P \geq \frac{1}{2} \end{array} \right.$$

- Oracle forecast under symmetric binary (classification) loss: $d^o_{mm}(P) = \mathbb{I}[1 \leq p_L(P) + p_U(P)]$

- Suppose that efficient estimator $\hat{P}$ satisfies

$$\hat{P} \overset{P_{n,h}}{\sim} N(P_{n,h}, n^{-1})\,, \qquad P|X_n \sim N(\hat{P}, n^{-1})$$

- ERF

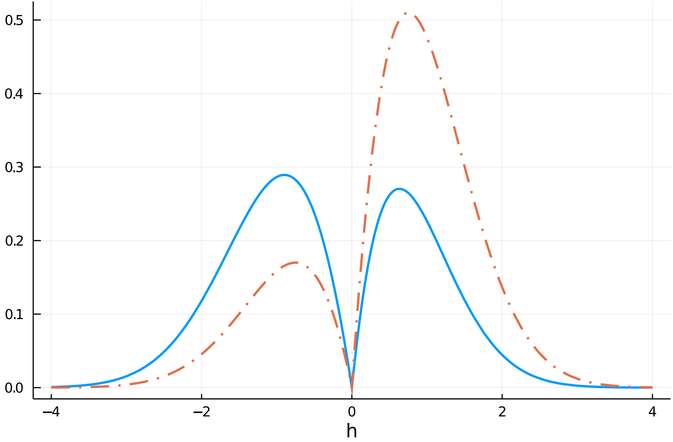
$$d_{mm}(\hat{P}) = \mathbb{I}\left[\sqrt{n}(\hat{P} - \tfrac{1}{2}) \geq -\frac{2\phi(\sqrt{n}(\hat{P} - \frac{1}{2}))}{1 + 2\Phi(\sqrt{n}(\hat{P} - \frac{1}{2}))}\right]$$

- Cf. plug-in rule

$$d^o_{mm}(\hat{P}) = \mathbb{I}[\sqrt{n}(\hat{P} - \tfrac{1}{2}) \geq 0]$$

Simple illustration: asymptotic excess maximum risk

Asymptotic excess maximum risk as a function of h at $P_0 = \frac{1}{2}$



Solid lines: Efficient robust forecast. **Dashed lines:** Oracle plug-in rule.

Extensions: structural breaks

Three types of breaks in the running example:

$$Y_{it+1} = \mathbb{I}\left[\lambda_i + \beta Y_{it} \geq U_{it+1}\right], \quad \mathbb{P}\left(U_{it+1} \leq u | Y_i^t = y^t, \lambda_i = \lambda\right) = \Phi(u)$$

1. A break in the distribution of the U_{it+1}:
   suppose $\Phi_t = \Phi$ for dates $t = 1, \ldots, T$, but $\Phi_{T+1} \in \mathcal{N}(\Phi)$. Identified set:

$$\Theta_0 = \left\{\theta = (\beta, \Pi_{\lambda,y}, \Phi_{T+1}) \in \Theta : p(y^T|\beta, \Pi_{\lambda,y}) = p(y^T) \ \forall\, y^T \in \{0,1\}^T \ \text{ and } \ \Phi_{T+1} \in \mathcal{N}(\Phi)\right\},$$

2. A break in the λ_i:
   can be viewed as a location shift of the distribution Φ_t

3. A break in β:
   can be handled by defining

$$\Theta_0 = \left\{\theta = (\beta, \beta_{T+1}, \Pi_{\lambda,y}) \in \Theta : p(y^T|\beta, \Pi_{\lambda,y}) = p(y^T) \ \forall\, y^T \in \{0,1\}^T \ \text{ and } \ |\beta - \beta_{T+1}| \leq \delta\right\},$$

Extensions

- Multinomial forecasts
- Sensitivity analysis:
  generalize certain aspects of the model, e.g., corr. random effects $\Pi_{\lambda,y} = \Pi(\lambda, y_0, \xi)$ for $\xi \in \Xi$.
- Counterfactuals in structural models:
  predict an outcome Y (e.g., firm entry/exit) under an intervention
- Statistical treatment assignment:
  predict optimal treatment Y for individual $n+1$ having observed data on n individuals.

Some related literature

- Binary forecasting: e.g., Elliott and Lieli (2013), Lahiri and Yang (2013), and Elliott and Timmermann (2016)
- Partial identification in nonlinear panels: e.g., Honore and Tamer (2006), Chernozhukov, Fernandez-Val, Hahn, Newey (2013)
- Short panels: Baltagi (2008), Gu and Koenker (2016), Liu (2019), Liu, Moon, Schorfheide (2018,2020)
- Decision theory: Wald (1950), Robbins (1951), Berger (1985), ..., Manski (2007, 2011), Stoye (2011)
- Robustness: Gilboa and Schmeidler (1989), Hansen and Sargent (2001), ..., Chamberlain (2000, 2001)
- Robustness/sensitivity analysis in econometrics: Chamberlain (2000, 2001), Kitagawa (2012), Andrews, Gentzkow, Shapiro (2017), Giacomini and Kitagawa (2018), Armstrong and Kolesar (2018), Bonhomme and Weidner (2019), Christensen and Connault (2019)

Conclusion

- Robust forecasts (minmax risk or minimax regret) to deal with uncertainty about the forecast distribution
- Efficient robust forecasts that deal with estimation of the set of forecast distributions
- Develop a suitable asymptotic efficiency theory
- Provide computationally efficient implementation based on linear/convex programming
- Basic idea is applicable in a wide variety of applications