

How to use economic theory to improve estimators

Maximilian Kasy

June 27, 2018

Introduction

- Most regularization methods shrink toward 0, or some other arbitrary point.
- What if we instead shrink toward parameter values consistent with the predictions of economic theory?
- Most economic theories are only approximately correct. Therefore:
 - Testing them always rejects for large samples.
 - Imposing them leads to inconsistent estimators.
 - But shrinking toward them leads to uniformly better estimates.
- Shrinking to theory is an alternative to the standard paradigm of testing theories, and maintaining them while they are not rejected.
- Yields uniform improvements of risk, largest when theory is approximately correct.

- General construction of estimators shrinking to theory:
 - Parametric empirical Bayes approach.
 - Assume true parameters are theory-consistent parameters plus some random effects.
 - Variance of random effects can be estimated, and determines the degree of shrinkage toward theory.
- We apply this to:
 1. Consumer demand
shrunk toward negative semi-definite compensated demand elasticities.
 2. Effect of labor supply on wage inequality
shrunk toward CES production function model.
 3. Decision probabilities
shrunk toward Stochastic Axiom of Revealed Preference.
 4. Expected asset returns
shrunk toward Capital Asset Pricing Model.

Two complementary characterizations of risk (MSE)

1. Approximate, for the high-dimensional case.
 - Variability of hyper-parameters negligible.
 - Simple characterization.
 - Marginal likelihood maximization vs. risk minimization.
2. Exact, using Stein's unbiased risk estimate.
 - In analogy to proof of uniform dominance of James-Stein.
 - Key novelty: Extension to the case of inequality restrictions.

A simple construction of shrinkage-estimators

- Goal: constructing estimators shrinking to theory.
- Preliminary unrestricted estimator:

$$\hat{\beta}|\beta \sim N(\beta, V)$$

- Restrictions implied by theoretical model:

$$\beta^0 \in B^0 = \{b : R_1 \cdot b = 0, R_2 \cdot b \leq 0\}.$$

- Empirical Bayes (random coefficient) construction:

$$\begin{aligned}\beta &= \beta^0 + \zeta, \\ \zeta &\sim N(0, \tau^2 \cdot I), \\ \beta^0 &\in B^0.\end{aligned}$$

Solving for the empirical Bayes estimator

- Marginal distribution of $\hat{\beta}$ given β_0, τ^2 :

$$\hat{\beta} | \beta_0, \tau^2 \sim N(\beta^0, \tau^2 \cdot I + V)$$

- Maximum likelihood estimation of β_0, τ^2 (tuning):

$$\begin{aligned} (\hat{\beta}^0, \hat{\tau}^2) = \underset{b^0 \in B^0, t^2 \geq 0}{\operatorname{argmin}} \quad & \log \left(\det \left(\tau^2 \cdot I + \hat{V} \right) \right) \\ & + (\hat{\beta} - b^0)' \cdot \left(\tau^2 \cdot I + \hat{V} \right)^{-1} \cdot (\hat{\beta} - b^0). \end{aligned}$$

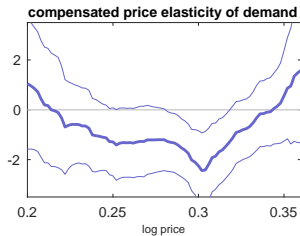
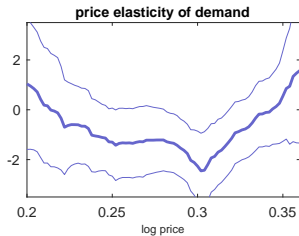
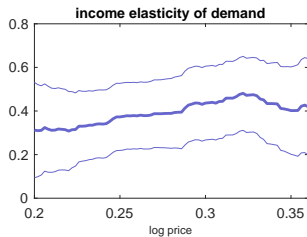
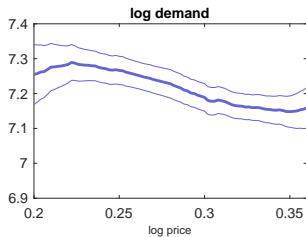
- “Bayes” estimation of β (shrinkage):

$$\hat{\beta}^{EB} = \hat{\beta}^0 + \left(I + \frac{1}{\hat{\tau}^2} \hat{V} \right)^{-1} \cdot (\hat{\beta} - \hat{\beta}^0).$$

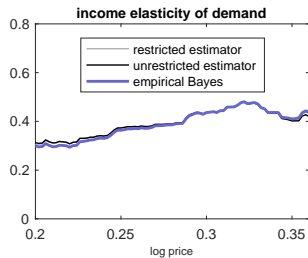
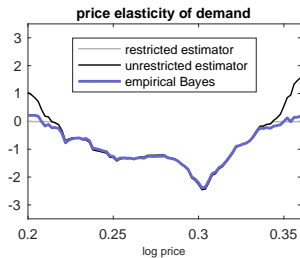
Application 1: Consumer demand

- Consumer choice and the restrictions on compensated demand implied by utility maximization.
- High dimensional parameters if we want to estimate demand elasticities at many different price and income levels.
- Theory we are shrinking to:
 - Negative semi-definiteness of compensated quantile demand elasticities,
 - which holds under arbitrary preference heterogeneity by Dette et al. (2016).
- Application as in Blundell et al. (2017):
 - Price and income elasticity of gasoline demand,
 - 2001 National Household Travel Survey (NHTS).

Unrestricted demand estimation



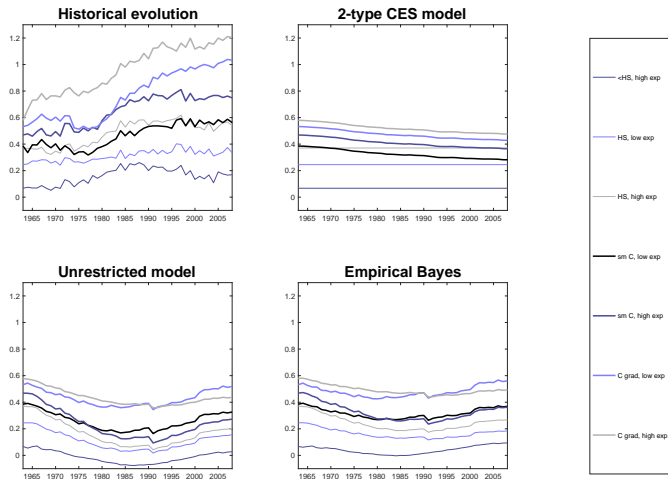
Empirical Bayes demand estimation



Application 2: Wage inequality

- Estimation of labor demand systems, as in literatures on
 - skill-biased technical change, e.g. Autor et al. (2008),
 - impact of immigration, e.g. Card (2009).
- High dimensional parameters if we want to allow for flexible interactions between the supply of many types of workers.
- Theory we are shrinking to:
 - wages equal to marginal productivity,
 - output determined by a CES production function.
- Data: US State-level panel for the years 1960, 1970, 1980, 1990, and 2000 using the Current Population Survey, and 2006 using the American Community Survey.

Counterfactual evolution of US wage inequality



Some theory – canonical coordinates

- By orthogonal change of coordinates, w.l.o.g.

$$\widehat{V} = \text{diag}(v_j).$$

- Then

$$\widehat{\beta}_j^{EB} = \left(\frac{v_j}{\widehat{\tau}^2 + v_j} \right) \cdot \widehat{\beta}_j^0 + \left(\frac{\widehat{\tau}^2}{\widehat{\tau}^2 + v_j} \right) \cdot \widehat{\beta}_j.$$

and

$$(\widehat{\beta}^0, \widehat{\tau}^2) = \underset{b^0 \in B^0, \tau^2}{\text{argmin}} \frac{1}{j} \cdot \sum_j \left(\log(\tau^2 + v_j) + \frac{(\widehat{\beta}_j - b_j^0)^2}{\tau^2 + v_j} \right).$$

Approximate MSE

- Mean squared error for fixed b_0, τ^2 : $MSE(\hat{\beta}^{EB}(b^0, \tau^2), \beta) =$

$$\frac{1}{J} \cdot \sum_{j=1}^J \left[\left(\frac{\tau^2}{\tau^2 + v_j} \right)^2 \cdot v_j + \left(\frac{v_j}{\tau^2 + v_j} \right)^2 \cdot (\beta_j - b_j^0)^2 \right].$$

- Hyper-parameters maximizing expected LLH:

$$(\beta^0, \tau^{*2}) = \operatorname{argmin}_{b^0 \in B^0, \tau^2} \frac{1}{J} \cdot \sum_{j=1}^J \left[\log(\tau^2 + v_j) + \frac{(\beta_j - b_j^0)^2 + v_j}{\tau^2 + v_j} \right].$$

Theorem

Under [some empirical Bayes assumptions]

$$SE(\hat{\beta}^{EB}, \beta) - MSE(\hat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) \rightarrow^P 0$$

as $J \rightarrow \infty$.

Marginal likelihood vs. MSE

- FOCs for optimal τ^2 in high dimensional limit.
- Minimizer of MSE:

$$\sum_{j=1}^J \left[\frac{v_j^2}{(\tau^{\times 2} + v_j)^3} \cdot (\tau^{\times 2} - (\beta_j - \beta_j^0)^2) \right] = 0.$$

- Maximizer of expected marginal LLH:

$$\sum_{j=1}^J \left[\frac{1}{(\tau^{*2} + v_j)^2} (\tau^{*2} - (\beta_j - \beta_j^0)^2) \right] = 0.$$

- The two differ when β_j and v_j are correlated across j .
- In that case, EB can be inefficient.

Exact characterization of risk: SURE

- Consider canonical coordinates with $V = I$, and restrictions of the form

$$B^0 = \{b : b_1, \dots, b_K = 0, b_{K+1}, \dots, b_L \leq 0\}.$$

- Denote $R = \sum_{j=1}^K \hat{\beta}_j^2 + \sum_{j=K+1}^L \max(\hat{\beta}_j, 0)^2$. Then

$$\hat{\beta}^0 = \begin{cases} 0 & j = 1, \dots, K \\ \max(\hat{\beta}_j, 0) & j = K+1, \dots, L \\ \hat{\beta}_j & j = L+1, \dots, J \end{cases}$$

$$\hat{\tau}^2 = \max\left(\frac{1}{J}R - 1, 0\right)$$

$$\hat{\beta}_j^{EB} = \begin{cases} \frac{\hat{\tau}^2}{\hat{\tau}^2 + 1} \cdot \hat{\beta}_j & j = 1, \dots, K \\ & \text{or } j = K+1, \dots, L \text{ and } \hat{\beta}_j > 0, \\ \hat{\beta}_j & \text{else.} \end{cases}$$

Exact characterization of risk, continued

Theorem

Under these assumptions, $MSE(\hat{\beta}^{EB}, \beta) = 1 + E_{\beta}[\Delta]$, where

$$\Delta = \begin{cases} \frac{1}{R} \cdot [J + 4 - 2J^*] & R > J \\ \frac{1}{J} \cdot [R - 2J^*] & \text{else,} \end{cases} \quad (1)$$

$R = \sum_{j=1}^K \hat{\beta}_j^2 + \sum_{j=K+1}^L \max(\hat{\beta}_j, 0)^2$, and

$J^* = K + \sum_{j=K+1}^L \mathbf{1}(\hat{\beta}_j > 0)$.

Immediate consequence: EB has uniformly lower risk than the unrestricted estimator for all β if

$$J^* > J/2 + 2.$$

Summary

- Proposed estimator construction:
 1. First-stage: estimate neglecting the theoretical predictions.
 2. Assume: True parameter values = parameter values conforming to the theory + noise.
 3. Maximize the marginal likelihood of the data given the hyperparameters. (Variance of noise \approx model fit!)
 4. Bayesian updating | estimated hyperparameters, data \Rightarrow estimates of the parameters of interest.
- Implement for range of applications / theories:
 1. Consumer demand,
 2. Effect of labor supply on wage inequality,
 3. Decision probabilities,
 4. Capital Asset Pricing Model.
- Two characterizations of risk:
 1. High-dimension asymptotics (simple and transparent).
 2. Exact (somewhat more restrictive setting).

Thank you!