How to use economic theory to improve estimators: shrinking toward theoretical restrictions

Fessler, Pirmin* Kasy, Maximilian†

February 23, 2018

Abstract

We propose to use economic theories to construct estimators in empirical microeconomics that perform well when the theories' empirical implications are approximately correct, but perform no worse than unrestricted estimators if the theories' implications do not hold. We describe a construction of such estimators using the empirical Bayes paradigm. We implement this construction in various settings, including labor demand and wage inequality, and estimation of consumer demand. We provide theoretical characterizations of the behavior of the proposed estimators, and evaluate them using Monte Carlo simulations. Our approach is an alternative to the use of theory as something to be tested or to be imposed on estimates. Our approach complements uses of theory for identification and extrapolation.

Keywords: microeconometrics, empirical Bayes estimation, shrinkage, labor demand, consumer choice
JEL codes: C11, C52, J23, J31, D01

*Economic Analysis Division, Oesterreichische Nationalbank; Address: Postbox 61, 1011 Vienna, Austria; e-mail: pirmin.fessler@oenb.at. Opinions expressed by the authors of studies do not necessarily reflect the official viewpoint of the Oesterreichische Nationalbank or of the Eurosystem.

†Associate Professor, Department of Economics, Harvard University; Address: 1805 Cambridge Street, Cambridge, MA 02138; e-mail: maximiliankasy@fas.harvard.edu.

1 Introduction

There are various ways economic theory might be put to use in empirical microeconomics.[1] A common role of theory is to provide predictions with empirical content. These predictions might be tested, using statistical tests controlling size at conventional levels such as 5%. A theory that has not been rejected is maintained. The predictions of a theory (which has not been rejected) might then be imposed on estimated parameters. Theory might further provide the assumptions necessary to identify objects such as causal effects or economic primitives, which would not be identified based on observation alone. It might also be used to extrapolate to counterfactual settings. Finally, theory may provide guidance for researchers in terms of what questions to take to the data, in a way that is harder to formalize.

We propose a further, alternative use of economic theory in empirical research. For the purposes of this paper, we consider as "theory" any argument that leads to prior restrictions on a parameter vector β, where β would be identified even in the absence of these restrictions. We suggest a framework for the construction of estimators which perform particularly well when the empirical implications of a theory under consideration are approximately correct. By "approximately correct" we mean that deviations from the theory's predictions are of the same order of magnitude as the standard errors of unrestricted estimates. Estimators constructed in the proposed way tend to outperform estimators ignoring the theory, regardless of what the true data generating process is and whether the theory is correct or not. Our approach provides an alternative to the testing and imposition of theories. We will argue that it is well suited for theories that are only approximately correct, as might be the case for many theories in economics. If the restrictions implied by theory do not hold, their rejection by tests is only a matter of sample size, and their imposition might cause estimators to be biased and inconsistent. Our approach is complementary to the roles of theory in identification and in guiding the choice of research questions.

Estimator construction Our approach is based on estimators shrinking toward the theory in a data dependent way. Such estimators can be constructed as follows. Our construction uses the empirical Bayes paradigm, which requires a family of priors. We

[1]We thank Alberto Abadie, Isaiah Andrews, Gary Chamberlain, Ellora Derenoncourt, Liran Einav, Yuriy Gorodnichenko, Kei Hirano, Michael Kummer, José Montiel Olea, Ashesh Rambachan, Neil Shephard, several anonymous referees, as well as seminar participants at Duke, Harvard, UBC, SFU, Hebrew University, Tel Aviv University, University of Graz, JKU Linz, University of Rochester, and Bank of Canada for helpful discussions and comments.

consider families of priors for the parameters of interest, where the priors are centered on the set of parameters consistent with the predictions of the theory. These priors are further governed by a parameter of dispersion, providing a measure for how well the theory appears to describe the data. A prior with a dispersion of zero would correspond to imposing the theory; an infinite dispersion to an uninformative prior, ignoring the theory, and estimating an unrestricted model.

Estimation proceeds in three steps. In a first step, the parameters of interest are estimated in an unrestricted way, ignoring the predictions of economic theory. This yields noisy but consistent preliminary estimates. This first step requires that the parameters of interest are identified even when ignoring the theory. In a second step, the hyper-parameters governing the family of priors are estimated. The hyper-parameters include both the parameters of the restricted model and the measure of dispersion, where the latter provides a measure of model fit. The hyper-parameters can be estimated by maximizing the marginal likelihood for the preliminary estimates, or alternatively by using a method-of-moments estimator, or by minimizing Stein's unbiased risk estimate. In a third step, "posterior means" for the parameters of interest are calculated, conditioning on the preliminary estimates and on the estimated values for the hyper-parameters. These posterior means are "shrinking" the preliminary estimates toward the restricted model.

Contributions The main contribution of this paper is to bring together economic theory with the tools of the empirical Bayes paradigm, in order to leverage economic theory for improved estimation in a way that contrasts with the "testing and imposition" approach. Empirical Bayes estimators were originally proposed by Robbins (1956); they are closely related to shrinkage estimators as introduced by James and Stein (1961) and characterized by Stein (1981). Parametric empirical Bayes was introduced by Morris (1983). The empirical Bayes estimators usually considered in the statistics literature shrink toward an arbitrary point in the parameter space, such as 0. We instead modify the construction to shrink toward parameter sets consistent with economic theories such as structural models of labor demand (as in Card 2009), or the theory of consumer choice (as in Blundell et al. (2017)). The supplementary appendix discusses additional applications shrinking toward the predictions of general equilibrium models of asset markets (as in Jensen et al. 1972), to structural discrete choice models of consumer demand (as in Train 2009), or toward the predictions of abstract theories of economic decision making (as in McFadden 2005).

In addition to proposing to use the restrictions implied by economic theory to construct shrinkage estimators, and providing guidelines and examples for implementation, we develop statistical theory results, characterizing the behavior of the proposed estimators. Our estimators are related to but different from the shrinkage estimators discussed in Hansen (2016); our results complement those of Hansen (2016) in a way discussed in greater detail below.

Our approach stands in contrast to other approaches for estimating the parameters of interest, including (i) unrestricted estimation, (ii) estimation imposing the theory, (iii) fully Bayesian estimation, and (iv) pre-testing where the theory is imposed if and only if it is not rejected. There are a number of advantages to our approach relative to these alternatives. (i) The resulting estimates are consistent, i.e., converge to the truth as samples get large, for any parameter values, in contrast to estimation imposing the theory. (ii) The variance and mean squared error of the estimates is smaller than under unrestricted estimation. Simulations, finite sample characterizations, and asymptotic approximations show this is the case uniformly over most of the parameter space.[2] (iii) In contrast to a fully Bayesian approach, no tuning parameters (features of the prior) have to be picked by the researcher. (iv) Our empirical Bayes approach avoids the irregularities (poor mean squared error in intermediate parameter regions) which are associated with testing theories and imposing them if they are not rejected (cf. Leeb and Pötscher, 2005). (v) Counterfactual predictions and forecasts are driven by the data whenever the latter are informative.

After introducing our approach in Section 2, we implement it in two economic contexts in Section 3. These contexts are distinguished in particular by the type of "theory" considered, including parametric structural models of production and labor demand, and the general theory of consumer demand. The applications in the supplement consider general equilibrium models of financial markets, structural models of preferences, and abstract theories of decision making. Let us briefly sketch the settings considered in the main paper.

Labor demand and wage inequality Wage inequality has increased significantly in most industrial countries since the 1980s. There is considerable disagreement over the relative contribution to this increase of alternative factors, such as technical change (Autor et al., 2008), migration (Card, 2009), and institutional factors (Fortin and Lemieux, 1997). Some of these disagreements have methodological roots. The workhorse method

[2] It is however possible to construct counter-examples when hyper-parameters are estimated by maximizing the marginal likelihood, see Section 4.2.

of estimating structural models of labor demand yields results that depend on the specific model chosen and the implied substitutability patterns. Flexible (unrestricted) estimation, on the other hand, results in very noisy estimates. We propose instead estimating flexible systems of labor demand, shrinking toward the predictions of a canonical model such as the 2-type constant elasticity of substitution (CES) model. We apply this method to data from the Current Population Survey (CPS) and the American Community Survey (ACS). We generally find negative but small inverse elasticities of substitution. The explanatory power of changes in labor supply for changes in relative wages appears to be quite small, based on our estimates. The 2-type CES model does not fit our data very well.

Consumer demand and the Slutsky condition The price and income elasticities of consumer demand are key parameters for the design of taxes and other policies. The price and income elasticity of gasoline demand, for instance, matter for the effectiveness and incidence of potential taxes aiming to reduce carbon emissions. A large literature considers the estimation of consumer demand; recent contributions to this literature include Dette et al. (2016) and Blundell et al. (2017).

Assuming exogenous variation of prices and income, the elasticities of demand can be estimated using local linear quantile regression or related nonparametric methods. If demand arises from utility maximization by consumers, then compensated own-price elasticities of demand are non-positive. Dette et al. (2016) show that negative semi-definiteness of compensated demand slopes holds not only for individual demand functions, but also for quantile demand functions. This is a theoretical restriction on price and income elasticities which could be imposed in order to improve estimator precision, as in Blundell et al. (2017). We propose to instead use an estimator that shrinks toward this theoretical restriction. In our application to gasoline demand using the 2001 National Household Travel Survey (NHTS), unrestricted estimates of price elasticities violate this restriction for high and low prices. The empirical Bayes estimator, by contrast, yields elasticities that are close to zero or negative over the entire range of observed prices. Negative compensated elasticities provide a good fit for our data.

Characterizations of estimator properties In Section 4 we provide characterizations of the behavior of the proposed estimators. We first show consistency and characterize the mapping from unrestricted estimates to empirical Bayes estimates. Our key results in this section are Theorem 1 and Theorem 2, which provide complementary characterizations of the mean squared error of the proposed estimators. Theorem 1 uses an asymptotic approximation that is valid whenever the hyper-parameters are estimated with small variance relative to the parameters of interest. This approximation works well when the dimension of hyper-parameters is small, as in our application to labor demand. Theorem 2 provides a characterization of the mean squared error which does not rely on asymptotic approximations, and instead uses Stein's unbiased risk estimate. This characterization assumes normality of unrestricted estimates, but covers the case of theoretical restrictions which allow for high-dimensional hyper-parameters, as in our application to consumer demand where only inequality restrictions are imposed by theory. This characterization allows us to prove uniform dominance of our estimator relative to the unrestricted estimator, under certain conditions. Monte Carlo simulations (in the supplementary online appendix) confirm the validity of our characterizations of risk for realistic specifications.

Roadmap The rest of this paper is structured as follows: Section 2 discusses the empirical Bayes paradigm and introduces our proposed construction of estimators. This section also reviews some related literature. Section 3 implements and adapts this construction to the economic settings described above. Section 4 develops statistical theory for the estimators we consider, including consistency and theoretical characterizations of their risk properties. Section 5 concludes. Appendix A describes a way to construct empirical Bayes confidence sets, based on the heuristic arguments of Laird and Louis (1987). Appendix B contains all proofs. The supplementary online appendix contains additional applications, Monte Carlo simulations, a geometric analysis of the proposed estimators, some additional discussion of labor demand systems, and a discussion of numerical methods for maximizing the marginal likelihood for our application to decision theory.

2 Estimator construction

Throughout this paper, we assume that there is a preliminary estimator $\widehat{\beta}$ of the parameter vector of interest, β, where the preliminary estimator does *not* make use of restrictions implied by economic theory. Economic theory is assumed to provide overidentifying restrictions on β; for simplicity of exposition, we focus on linear equality- and inequality-restrictions in this section. We briefly discuss how smooth restrictions are asymptotically equivalent to linear restrictions. We use these restrictions to construct an estimator $\widehat{\beta}^{EB}$ designed to outperform $\widehat{\beta}$ if the restrictions are approximately correct, and to perform no worse than $\widehat{\beta}$ if they are not. In Section 3 we will then adapt this setting to our applications, detailing in each case where the preliminary estimator and the theoretical restrictions are coming from.

This section is structured as follows. First, we introduce the setup in Section 2.1 and review the general empirical Bayes approach in Section 2.2. Section 2.3 presents our proposed empirical Bayes estimator. Section 2.4 reviews some of the relevant literature on empirical Bayes estimation and shrinkage.

2.1 Setup

Throughout this section, we consider as our object of interest a J-vector β. We assume the availability of a preliminary, unrestricted estimator

$$\widehat{\beta} \sim N(\beta, V), \tag{1}$$

of β, with consistently estimable variance V. This assumption implies that β is identified. The assumption of normality is best thought of as an asymptotic approximation. We will use the assumption of normality in order to construct estimators within the empirical Bayes paradigm. Most of our discussion of the theoretical properties of these estimators will not use normality. Asymptotically normal estimators $\widehat{\beta}$ might for instance be obtained using linear regressions, $Y = X \cdot \beta + \epsilon$, which might be estimated using ordinary least squares, instrumental variables, panel variation, etc.

The second key ingredient to our setting is the availability of overidentifying restrictions implied by economic theory. In this section, we will focus on the case where a theoretical model implies that

$$\beta^0 \in B^0 = \{b : \ R_1 \cdot b = 0, \ R_2 \cdot b \leq 0\}. \tag{2}$$

Here B^0 is the set of parameter vectors β^0 satisfying the restrictions implied by theory, including a set of linear equality restrictions $R_1 \cdot b = 0$, and a set of linear inequality restrictions $R_2 \cdot b \leq 0$. The inequality is to be understood componentwise, and the matrices R_1 and R_2 are known. More generally, shrinkage toward non-linear smooth restrictions (imposing that β^0 lies in some smooth manifold) could be considered. Theorem 1 in Hansen (2016) provides the type of result needed for such a generalization, using a local asymptotic framework.

The results reviewed in chapter 7 of van der Vaart (2000) imply that, under sufficient regularity conditions and i.i.d. asymptotics, the likelihood ratio process of any parametric model converges to the likelihood ratio process of the normal means problem as considered here. Additionally, smooth restrictions on the mean vector asymptotically become linear restrictions on the local parameter. Under the framework of local asymptotic normality, the assumptions we impose are therefore without loss of generality. Rather than explicitly invoking the apparatus of limiting experiments and local asymptotic normality, the theoretical discussion in Section 4 below will impose the normal means / linear restrictions form of the limiting experiment directly.

2.2 General empirical Bayes estimation

Two approaches to estimation are commonly used in settings of this kind, one imposing the restrictions of the theoretical model, and one leaving the model unrestricted. Estimation based on the theoretical model has a small variance, but yields non-robust conclusions and estimates that are biased and inconsistent if the model is mis-specified. Estimation using the unrestricted model is unbiased and consistent, but leads to estimates of large variance.

The paradigm of empirical Bayes estimation allows one to cover a middle ground between these two approaches, and combines the advantages of both. An elegant exposition of this approach can be found in Morris (1983). The parametric empirical Bayes approach can be summarized as follows:

$$Y|\eta \sim f(Y|\eta) \tag{3}$$

$$\eta \sim \pi(\eta|\theta), \tag{4}$$

where Y are the observed data, both f and π describe parametric families of distributions, and where usually $\dim(\theta) \leq \dim(\eta) - 2$. Note that θ might include a subset of the parameters in η. Equation (3) describes the unrestricted model for the distribution of the data given the full set of parameters η. Equation (4) describes a family of "prior distributions" for η, indexed by the hyper-parameters θ.

Estimation in the empirical Bayes paradigm proceeds in two steps. First we obtain an estimator of θ. This can be done by considering the marginal likelihood of Y given θ, which is calculated by integrating over the distribution of the parameters η:

$$Y|\theta \sim g(Y|\theta) := \int f(Y|\eta)\pi(\eta|\theta)d\eta. \tag{5}$$

In models with suitable conjugacy properties, such as the one we will consider below, the marginal likelihood g can be calculated in closed form. A natural estimator for θ is obtained by maximum likelihood,

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\ g(Y|\theta). \tag{6}$$

Other estimators for θ are conceivable and commonly used, as well. In the second step of empirical Bayes estimation, η is estimated as the "posterior expectation"[3] of η given Y and θ, substituting the estimate $\widehat{\theta}$ for the hyper-parameter θ,

$$\widehat{\eta} = E\left[\eta|Y, \theta = \widehat{\theta}\right]. \tag{7}$$

The general empirical Bayes approach includes fully Bayesian estimation as a special case if the family of priors π contains just one distribution. This general approach also includes unrestricted frequentist estimation as a special case, when $\theta = \eta$. The general approach finally includes structural estimation when again $\theta = \eta$, and the support of θ is restricted to parameter values allowed by the structural model. We can think of such support restrictions as imposition of dogmatic prior beliefs, in contrast to non-dogmatic priors that have full support.

2.3 An empirical Bayes model for our setup

Let us now specialize the general empirical Bayes approach to the setting considered in this paper. We directly model the distribution of the unrestricted estimator $\widehat{\beta}$. This unrestricted estimator is then mapped to an empirical Bayes estimator $\widehat{\beta}^{EB}$. To construct a family of priors for β, we assume that β is equal to a vector of parameters consistent with the structural model plus noise of unknown variance.

Modeling $\widehat{\beta}$ We assume that the unrestricted estimator $\widehat{\beta}$ is normally distributed given the true coefficients, unbiased for the true coefficient vector β, and has a variance V,

$$\widehat{\beta}|\beta, V \sim N(\beta, V). \tag{8}$$

This assumption can be justified by conventional asymptotics, letting the number n of cross-sectional units go to infinity in many applications of interest (as in Hansen 2016).

[3] The quotation marks reflect the fact that this would only be a posterior expectation in the strict sense if $\widehat{\theta}$ had been chosen independently of the data, rather than estimated.

We emphasize again that normality of $\widehat{\beta}$ is only used for estimator construction, and is not imposed in our theoretical discussion of its properties in Section 4. We further assume that we have a consistent estimator $\widehat{V}$ of V, i.e.,

$$\widehat{V} \cdot V^{-1} \to^p I,$$

where $\to^p$ denotes convergence in probability.

Prior distributions We next need to specify a family of prior distributions. We model β as corresponding to the coefficients of the structural model plus some disturbances, that is

$$\begin{aligned} \beta &= \beta^0 + \zeta, \\ \zeta &\sim N(0, \tau^2 \cdot I), \\ \beta^0 &\in B^0. \end{aligned} \tag{9}$$

The term $\beta^0 \in B^0$ corresponds to a set of coefficients satisfying the structural model. The term ζ is equal to a random J-vector with variance $\mathrm{Var}(\zeta) = \tau^2 \cdot I$.

If we were to set $\tau^2 = 0$, the empirical Bayes approach would reduce to imposing the theoretical model. If we let τ^2 go to infinity, we effectively recover the unrestricted model. We consider τ^2 to be a parameter to be estimated, however, which measures how well the given theoretical model fits the data. Note that this choice of a family of priors is not "correct" or "incorrect" in an empirical setting; rather it is a device for the construction of an estimator.

Summarizing our setting in terms of the general notation introduced in Section 2.2, we get:

$$\begin{aligned} \eta &= (\beta, V) \\ \theta &= (\beta^0, \tau^2, V) \\ \widehat{\beta}|\eta &\sim N(\beta, V) \\ \beta|\theta &\sim N(\beta^0, \tau^2 \cdot I). \end{aligned} \tag{10}$$

Solving for the empirical Bayes estimator In order to obtain estimators of β^0 and τ^2, consider the marginal distribution of $\widehat{\beta}$ given θ. This marginal distribution is normal,

$$\widehat{\beta}|\theta \sim N(\beta^0, \Sigma(\tau^2, V)), \tag{11}$$

where (leaving the conditioning on θ implicit)

$$\begin{aligned}\Sigma(\tau^2, V) = \mathrm{Var}\left(\widehat{\beta}\right) &= \mathrm{Var}\left(E\left[\widehat{\beta}|\eta\right]\right) + E\left[\mathrm{Var}\left(\widehat{\beta}|\eta\right)\right] \\ &= \tau^2 \cdot I + V.\end{aligned}$$

Substituting the consistent estimator $\widehat{V}$ for V, we obtain the empirical Bayes estimators of β^0 and τ^2 as the solution to the maximum (marginal) likelihood problem

$$\begin{aligned}(\widehat{\beta}^0, \widehat{\tau}^2) = \underset{b^0 \in B^0,\, t^2 \geq 0}{\operatorname{argmin}} & \log\left(\det(\Sigma(t^2, \widehat{V}))\right) \\ & + (\widehat{\beta} - b^0)' \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot (\widehat{\beta} - b^0).\end{aligned} \tag{12}$$

If B^0 is only subject to equality restrictions, but no inequality restrictions, then we can simplify this optimization problem by concentrating out b^0. Let M be a matrix the columns of which form a basis of the orthocomplement of R_1, so that $R_1 \cdot M = 0$ and $\mathrm{rank}(M) + \mathrm{rank}(R) = J$. With this notation, and given t^2, the optimal b^0 takes the form of a GLS estimator and is equal to

$$\widehat{\beta}^0 = M \cdot (M \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot M')^{-1} \cdot M \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot \widehat{\beta}.$$

Substituting this expression into the objective function, we obtain a function of t^2 alone, which is easily optimized numerically.

Given the unrestricted estimates $\widehat{\beta}$, as well as the estimates $\widehat{\beta}^0$ and $\widehat{\tau}^2$, we can finally obtain the "posterior expectation" of β as

$$\widehat{\beta}^{EB} = \widehat{\beta}^0 + \left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot (\widehat{\beta} - \widehat{\beta}^0). \tag{13}$$

This is the empirical Bayes estimator of the coefficient vector of interest.

Discussion It is instructive to relate the proposed empirical Bayes procedure to restricted estimation, where the theoretical model is imposed. The empirical Bayes estimator $\widehat{\beta}^{EB}$ of β is not given by $\widehat{\beta}^0$. Instead we can think of it as an intermediate point between $\widehat{\beta}^0$ and the unrestricted estimator $\widehat{\beta}$. The relative weights of these two are determined by the matrices $\widehat{\tau}^2 \cdot I$ and $\widehat{V}$. When $\widehat{\tau}^2$ is close to 0, we get $\widehat{\beta}^{EB} \approx \widehat{\beta}^0$.

When $\widehat{\tau}^2$ is large, we get $\widehat{\beta}^{EB} \approx \widehat{\beta}$, cf. Equation (13).

Our construction of a family of priors thus implies the following: When the restricted model appears to describe the data well, then our estimate of β will be close to what is prescribed by the restricted model. When the restricted model fits poorly, then the estimator will essentially disregard it and provide estimates close to the unrestricted ones. A key point to note is that this is done in a data-dependent and smooth way, in contrast to the discontinuity of pre-testing estimators such as

$$\widehat{\beta}^{PT} = \widehat{\beta}^0 + \psi \cdot (\widehat{\beta} - \widehat{\beta}^0), \qquad \psi = \mathbf{1}\left((\widehat{\beta} - \widehat{\beta}^0)'\widehat{V}^{-1}(\widehat{\beta} - \widehat{\beta}^0) > \chi\right),$$

where χ is the $1 - \alpha$ critical value of the appropriate χ^2 distribution.

The estimator $\widehat{\beta}^0$ is very similar to the restricted estimator of β obtained by directly imposing the theoretical constraints when estimating β; in both cases we are considering an orthogonal projection of the unrestricted estimator $\widehat{\beta}$ onto the set B^0 of estimates consistent with the theory. The projection is with respect to different norms, however. When the restricted estimator is obtained by least squares regression of Y on X subject to linear constraints, the projection is with respect to the norm $\|b\|_\beta := (b' \cdot \mathrm{Var}(X) \cdot b)^{1/2}$. In the context of our empirical Bayes approach, the projection is with respect to the norm $\|b\|_{\beta,EB} = (b' \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot b)^{1/2}$. The two objective functions coincide (up to a multiplicative constant) if (i) $t^2 = 0$, so that the restricted model is assumed to be correct, and (ii) $\widehat{V}$ is estimated assuming homoskedasticity.

Our approach is based upon directly modeling the distribution of the unrestricted estimator $\widehat{\beta}$. If $\widehat{\beta}$ contains the coefficients of an OLS regression, there is a one-to-one mapping between (i) the dependent variables Y and (ii) the estimated coefficients and residuals of the unrestricted model. To the extent that $\widehat{\beta}$ is a sufficient statistic for β, our approach does not waste any information; this is true, in particular, for a standard parametric linear/normal model.

Using SURE to estimate τ^2 Our empirical Bayes estimator uses the marginal likelihood to estimate both τ^2 and β^0. The resulting estimator of τ^2 behaves well when the dimension of B^0 is small relative to the dimension of β. When the dimension of B^0 is not small, then maximum likelihood might choose a value of τ^2 that is too small. The resulting estimator $\widehat{\beta}^{EB}$ shrinks too aggressively toward B^0. This is the case for instance when B^0 is constrained only by inequality restrictions, as in our application to consumer demand.

An alternative to maximization of the marginal likelihood for choosing hyper-parameters is minimization of Stein's unbiased risk estimate. Denote $\widehat{\beta}^{EB}(\tau^2) = \widehat{\beta}^0 + \left(I + \frac{1}{\tau^2}\widehat{V}\right)^{-1} \cdot (\widehat{\beta} - \widehat{\beta}^0)$, where $\widehat{\beta}^0 = \text{argmin}_{\,b^0 \in B^0}(\widehat{\beta} - b^0)' \cdot \Sigma(\tau^2, \widehat{V})^{-1} \cdot (\widehat{\beta} - b^0)$ as before. Let

$$\begin{aligned} g(\widehat{\beta}) &= \widehat{\beta} - \widehat{\beta}^{EB}(\tau^2) \\ SURE(\tau^2) &= \text{trace}\left(\widehat{V}\right) + \left\|g(\widehat{\beta})\right\|^2 + 2 \cdot \text{trace}\left(\nabla g(\widehat{\beta}) \cdot \widehat{V}\right). \end{aligned} \tag{14}$$

It follows from Theorem 1 in Stein (1981) that $SURE(\tau^2)$ is an unbiased estimator of the mean squared error of the estimator $\widehat{\beta}^{EB}(\tau^2)$ when $\widehat{\beta} \sim N(\beta, \widehat{V})$. Choosing $\widehat{\tau}^2$ as the minimizer of $SURE(\tau^2)$ yields an estimator $\widehat{\beta}^{EB}(\widehat{\tau}^2)$ with small mean squared error. We use this approach in our application to consumer demand in Section 3.2.

2.4 Related literature

The main contribution of the present paper is to bring together economic theory with the tools of the empirical Bayes paradigm, to leverage economic theory for improved estimation. Our approach builds on a long tradition of research on empirical Bayes methods in statistics, which has its roots in the seminal contributions of Robbins (1956), who first considered the empirical Bayes approach for constructing estimators, and James and Stein (1961), who demonstrated that the conventional estimator for the mean of a multivariate normal vector (of dimension greater than 2) is inadmissible and dominated by empirical Bayes estimators. Empirical Bayes approaches were developed further by Efron and Morris (1973) and Morris (1983). The latter introduced the parametric empirical Bayes framework on which we build. Practical implementation of non-parametric empirical Bayes has recently been discussed by Koenker and Mizera (2014). Inference in empirical Bayes settings was discussed by Laird and Louis (1987) and Carlin and Gelfand (1990), among others; a review can be found in Casella et al. 2012. A good introduction to empirical Bayes estimation can be found in Efron (2010), another review is provided by Zhang (2003). In the introduction, we contrasted our approach to the pre-testing approach. In this alternative approach, the theoretical predictions $\beta \in B^0$ are first tested and then imposed on a second-stage estimator if and only if they are not rejected. Such pre-testing approaches are known to perform poorly in terms of mean squared error for intermediate parameter values in a neighborhood of the set B^0. This is familiar from the literature on "Hodges' estimator"; see for instance the discussion in Leeb and Pötscher (2005).

In Section 4, we provide a theoretical characterization of the risk properties of our empirical Bayes procedure. The first characterization relies on asymptotic arguments related to those invoked by Xie et al. (2012), the second one on the characterization of risk by Stein (1981). Graham and Hirano (2011) propose an estimator for missing data models that shrinks nonparametric estimates towards the predictions of a parametric linear model, using an approach similar to ours. Ideas related to our approach, in a fully Bayesian setting, have also been used in the literature on macroeconomic forecasting, where theoretical DSGE models can be used to inform priors for the parameters of statistical VAR models fit to the data. Del Negro and Schorfheide (2004) and Del Negro et al. (2007), for instance, construct hierarchical Bayesian models for VARs, with a hyperparameter measuring the fit of the theoretical model.

In an elegant recent paper complementing our analysis, Hansen (2016) studies the asymptotic properties of component-wise linear shrinkage estimators in parametric models. Allowing for nonlinear models, smooth nonlinear restrictions, and smooth non-quadratic loss functions, Hansen (2016) uses local asymptotics to recover a setting with normally distributed estimators, linear restrictions, and quadratic loss functions. Hansen (2016) proposes a class of shrinkage estimators that component-wise linearly interpolate between an unrestricted and a restricted estimator and studies their risk properties using the asymptotic normal approximation.

The focus of the present paper differs from Hansen (2016). Our focus is on leveraging economic theory in empirical research, on providing guidelines for estimator construction, and on implementations in several economic settings. The models and estimators we consider also differ from those in Hansen (2016). We construct estimators based on economic theory using the empirical Bayes paradigm. To this end, we consider a family of priors centered on theoretical restrictions. The resulting estimators differ from the component-wise linear interpolation with constant shrinkage factors proposed by Hansen (2016); our estimators shrink components which are less precisely estimated more toward the theory. Our estimators therefore combine information from the data with extrapolations from other components of the parameters of interest, extrapolations that are implied by economic theory, as we discuss in Section 4.1.

The framework which we have discussed in Section 2 takes normality of unrestricted estimators and linearity of restrictions as given; this is asymptotically justified even in nonlinear models as Hansen (2016) demonstrates. Future research might more formally apply his asymptotic results to our settings. In the supplementary Appendix, we also implement our general approach in economic settings involving non-normal models.

3 Applications

We now turn to two applications of our proposed approach. These applications are chosen from the fields of (i) labor demand and wage inequality, and (ii) consumer demand and estimation of price and income elasticities. The supplementary appendix discusses additional applications to (iii) financial asset returns and the capital asset pricing model, (iv) multinomial logit and mixed multinomial logit models of discrete choice in panel data, and (v) economic choice and general theories of decision making, such as utility maximization.

Applications (i), (ii) and (iii) are covered by the normal-linear framework introduced in Section 2, up to some minor modifications. Applications (iv) and (v) demonstrate the possibility of extensions to nonlinear settings. In each of these settings we construct estimators shrinking toward an economic theory, where the meaning of "economic theory" differs across applications, ranging from parametric structural models of production or of preferences, to the theory of utility maximizing consumer choice, to general equilibrium models of financial markets, to abstract theories of decision making.

3.1 Labor demand and wage inequality

In our first application we consider estimation of labor demand systems. Such systems are commonly estimated in the literature on skill-biased technical change, e.g. Autor et al. (2008), and in the literature on the impact of immigration, e.g. Card (2009). Estimation of such demand systems involves high dimensional parameters to the extent that we want to allow for flexible interactions between the supply of many types of workers. In this application, the "theory" that we propose shrinking to corresponds to models of wage determination consistent with wages equal to marginal productivity where output is determined by a CES or nested CES production function.

3.1.1 Setup

Suppose there are J types of workers, $j = 1, \dots, J$, defined for instance by their level of education and their potential experience. Consider a cross-section of labor markets $i = 1, \dots, n$.[4] Let Y_{ij} be the average log wage for workers of type j in labor market i, and let X_{ij} be the log labor supply of these same workers. Denote $Y_i = (Y_{i1}, \dots, Y_{iJ})$

[4] We adopt cross-sectional notation for simplicity, similar arguments apply to time series or panel data.

and $X_i = (X_{i1}, \ldots, X_{iJ})$. We are interested in the structural relationship between labor supply and wages, that is, in the inverse demand function.

CES-production functions, structural and unrestricted estimation The majority of contributions to the field impose a structural model, based on the assumptions of a parametric aggregate production function of a CES or nested CES form, a small number of labor-types, and wages equal to marginal productivity.[5] These assumptions motivate regressions of the following form (see for instance Autor et al. 2008 and Card 2009):

$$Y_{ij} - Y_{ij'} = \gamma_{jj'} + \theta^0 \cdot (X_{ij} - X_{ij'}) + \epsilon_{ijj'}. \tag{15}$$

Equation (15) can be rewritten in a numerically equivalent way as a fixed effects regression with restrictions across coefficients:

$$\begin{aligned} Y_{ij} &= \alpha_i + \gamma_j + \sum_{j'} \beta_{jj'} X_{ij'} + \epsilon_{ij}, \\ \beta &\in B^0 = \{b : \ b = \theta^0 \cdot M\}, \qquad M = I - \tfrac{1}{J}E, \end{aligned} \tag{16}$$

where β is a $J \times J$ matrix of coefficients, I is the identity matrix, E is a matrix of 1s, and M is the demeaning-matrix, projecting $\mathbb{R}^J$ on the subspace of vectors of mean 0. To verify this equivalence, take the difference $Y_{ij} - Y_{ij'}$ based on Equation (16). This equivalence is familiar from difference-in-differences regressions, which can equivalently be written in fixed-effects form or in differenced form.

Rather than imposing the strong assumptions implied by the CES production function model or its generalizations, we could instead consider a linear specification with a large number of types J and unrestricted own- and cross-elasticities. That is, we could estimate (16), using least squares, without imposing any cross-restrictions on the parameters $\beta_{jj'}$. Relative to this model, the CES production function restricts the J^2-dimensional parameter β to lie in a 1-dimensional subspace B^0. Note, however, that Equation (16) is not identified without further restrictions. Given the presence of the fixed effects α_i, we cannot pin down the effect of labor supply on the overall level of wages. Adding an arbitrary vector to all rows of β, and adjusting the α_i accordingly, yields an observationally equivalent model. Differencing (16) across types j however yields a model which *is* identified. Let Δ be a $(J-1) \times J$ matrix which subtracts the

[5] The CES production function takes the form $f_i(N_{i1}, \ldots, N_{iJ}) = \left(\sum_{j=1}^{J} \gamma_j N_{ij}^{\theta^0+1}\right)^{1/(\theta^0+1)}$, where $N_{ij} = \exp(X_{ij})$. Details are reviewed in the supplementary online appendix.

first entry from each component of a J vector, $\Delta = (-e, I_{J-1})$, and define the differenced matrix of coefficients $\delta = \Delta \cdot \beta$. We will consider δ as our main object of interest, and estimate the unrestricted regression

$$\Delta \cdot Y_i = \Delta \cdot \gamma + \delta \cdot X_i + \Delta \cdot \epsilon_i. \tag{17}$$

There are $J \cdot (J-1)$ free slope parameters to be estimated in the matrix δ. Relative to this general linear fixed effects model, the CES production function imposes $\delta = \Delta \cdot (\theta^0 \cdot M) = \theta^0 \cdot \Delta$, which implies $J^2 - J - 1$ additional restrictions. The last equation holds because $\Delta \cdot M = \Delta$.

3.1.2 Empirical Bayes estimators

Empirical Bayes estimation, shrinking toward the J-type CES model We next adapt the general approach introduced in Section 2 to the estimation of labor demand. We discuss two cases. We first consider shrinkage toward the CES model for the *same* set of types over which the unrestricted model is estimated. This CES model is nested in the unrestricted model. We then discuss shrinkage of an unrestricted model with many types toward the CES model for only *two* types. When types are defined based on college / no college, this two-type model is the canonical model of the literature on skill-biased technical change, cf. Acemoglu and Autor (2011). Similar estimators are easily constructed for other models of production, such as the nested CES model advocated by Card (2009).

Some minor modifications of the approach introduced in Section 2 are necessary. In particular, the coefficients of interest δ that we now consider are in matrix form. We denote the vectorized version of δ, stacking the columns on top of each other, by $\delta_\uparrow = vec(\delta)$, and similarly for other matrices. Furthermore, a family of priors is most naturally specified for β while estimation is for $\delta = \Delta \cdot \beta$. We model the coefficient matrix β as corresponding to the coefficients of the structural CES model plus some disturbances, that is

$$\beta = \theta^0 \cdot M + \zeta, \qquad\qquad \zeta_\uparrow \sim N(0, \tau^2 I).$$

Differencing this model yields

$$\delta = \Delta \cdot \beta = \theta^0 \cdot \Delta + \Delta \cdot \zeta.$$

The variance of the second term, reflecting "prior uncertainty," is given by $\mathrm{Var}((\Delta\cdot\zeta)_\uparrow) = \tau^2 \cdot P \otimes I$, where $P := \Delta \cdot \Delta' = I_{J-1} + E_{J-1}$ and $\otimes$ is the Kronecker product of matrices. This implies a prior variance of the unrestricted OLS estimator $\widehat{\delta}_\uparrow$ equal to

$$\Sigma(\tau^2, V) = \mathrm{Var}\left(\widehat{\delta}_\uparrow\right) = \tau^2 \cdot P \otimes I + V.$$

Substituting a consistent estimator $\widehat{V}$ for V, we obtain the empirical Bayes estimators of θ^0 and τ^2 as solutions to the maximum (marginal) likelihood problem

$$(\widehat{\theta}^0, \widehat{\tau}^2) = \underset{h^0, t^2}{\operatorname{argmin}} \; \log\left(\det(\Sigma(t^2, \widehat{V}))\right) + (\widehat{\delta}_\uparrow - h^0 \cdot \Delta_\uparrow)' \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot (\widehat{\delta}_\uparrow - h^0 \cdot \Delta_\uparrow).$$

Given t^2, the optimal h^0 is equal to $\widehat{\theta}^0 = (\Delta \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot \Delta')^{-1} \cdot \Delta \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot \widehat{\delta}_\uparrow$. Substituting this expression into the objective function, we obtain a function of t^2 alone that we optimize numerically. Given the unrestricted estimates $\widehat{\delta}$, as well as the estimates $\widehat{\beta}^0$ and $\widehat{\tau}^2$, we obtain the empirical Bayes estimator of δ as

$$\widehat{\delta}_\uparrow^{EB} = \widehat{\theta}^0 \cdot \Delta_\uparrow + P \otimes I \cdot \left(P \otimes I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot (\widehat{\delta}_\uparrow - \widehat{\theta}^0 \cdot \Delta_\uparrow). \tag{18}$$

Empirical Bayes estimation, shrinking toward the 2-type CES model The approach just described assumes that the structural model that we are shrinking to is the CES model with types $j = 1, \ldots, J$. In practice, we might want to shrink towards a CES model with more aggregated types, such as the canonical model (cf. Acemoglu and Autor, 2011) with just two types k of workers, where $k = 1$ denotes those with some college or more, and $k = 2$ denotes those with high school or less.

To nest the 2-type model in a setting with J types, denote the aggregate type k corresponding to type j by k_j and denote the aggregate labor supply of this type by $\tilde{N}_{ik}$. Define $X_{ij} = \log(N_{ij}/\tilde{N}_{ik_j})$ and $\tilde{X}_{ik} = \log(\tilde{N}_{ik})$. Using this notation, we can nest the canonical CES model in the following regression specification, which includes regressors for both the dis-aggregated types j and the aggregated types k,

$$Y_{ij} - Y_{i1} = (\gamma_j - \gamma_1) + \sum_{j'} \delta_{jj'} X_{ij'} + \theta^0 \cdot (\tilde{X}_{ik_j} - \tilde{X}_{i1}) + (\epsilon_{ij} - \epsilon_{i1}). \tag{19}$$

In this setting the matrix δ captures the additional effect of labor supply on relative wages beyond the effect already taken care of by the term $\theta^0 \cdot (\tilde{X}_{ik_j} - \tilde{X}_{i1})$.

The canonical CES model implies the restriction $\delta = 0$. The unrestricted approach estimates versions of this equation with δ left fully flexible. Our empirical Bayes approach applied to this setting takes as its point of departure a first stage unrestricted estimator $(\widehat{\delta}, \widetilde{\theta}^0)$ of (δ, θ^0), with estimated covariance matrix $\widehat{V}$. We then consider the family of priors $\delta_\uparrow \sim N(0, \tau^2 \cdot P \otimes I)$, where, as before, θ^0 and τ^2 are hyperparameters. Denote the variance of the unrestricted estimators given θ^0 and τ^2 by

$$\Sigma(\tau^2, V) = \mathrm{Var}\left((\widehat{\delta}_\uparrow, \widetilde{\theta}^0)\right) = \begin{pmatrix} \tau^2 \cdot P \otimes I & 0 \\ 0 & 0 \end{pmatrix} + V;$$

the conditional mean is given by $(0, \theta^0)$. We obtain the empirical Bayes estimators of θ^0 and τ^2 as solutions to the maximum (marginal) likelihood problem

$$(\widehat{\theta}^0, \widehat{\tau}^2) = \operatorname*{argmin}_{h^0, t^2} \log\left(\det(\Sigma(t^2, \widehat{V}))\right) + (\widehat{\delta}_\uparrow, \widetilde{\theta}^0 - h^0)' \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot (\widehat{\delta}_\uparrow, \widetilde{\theta}^0 - h^0)'.$$

Given t^2, the optimal h^0 is equal to $\widehat{\theta}^0 = (e \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot e')^{-1} \cdot e \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot (\widehat{\delta}_\uparrow, \widetilde{\theta}^0)$, where $e = (0, \ldots, 0, 1)$. Substituting this expression into the objective function, we obtain a function of t^2 alone that we optimize numerically. We finally obtain the empirical Bayes estimator of δ as

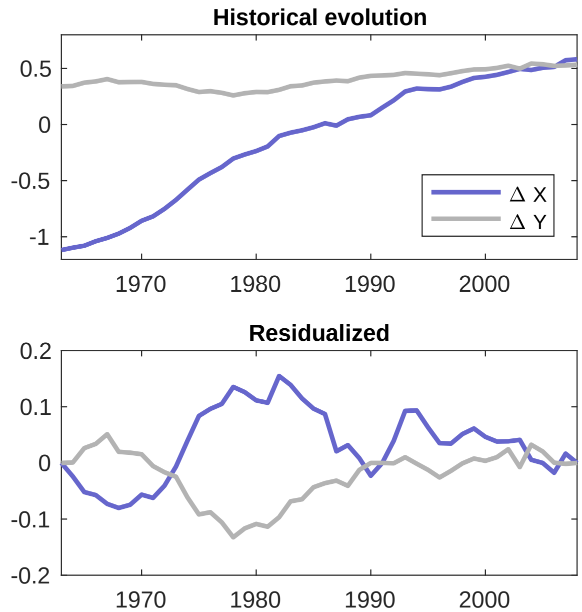
$$\widehat{\delta}_\uparrow^{EB} = (\widehat{\tau}^2 \cdot P \otimes I, 0) \cdot \Sigma(\widehat{\tau}^2, \widehat{V})^{-1} \cdot (\widehat{\delta}_\uparrow, \widetilde{\theta}^0 - h^0)'. \tag{20}$$

3.1.3 Empirical application

We now turn to our empirical application, studying labor demand in the United States. We use data that have been studied in the literatures on the impact of immigration on native wages and on the impact of skill-biased technical change; see for instance Card (2009), Autor et al. (2008), and Acemoglu and Autor (2011). We study the impact of historical changes of the labor force composition on relative wages. Rather than imposing one or the other of the models proposed in the literature (4-type CES, nested 2-type CES), we allow for arbitrary patterns of substitutability across a larger number of types, but use our empirical Bayes methodology to shrink to the canonical 2-type CES model.

Data Our analysis is based on the American Community Survey (ACS) data and Current Population Survey (CPS) data used in much of the literature. We build two aggregate data-sets. The first is a state-level panel for the years 1960, 1970, 1980, 1990, and 2000 using the CPS, and 2006 using the ACS. Our construction of this data-set

Figure 1: Log relative wages in the US – 2 types of workers



Note: The top graph of this figure shows the US time series of log relative wages and log relative labor supply between workers with more than a high school education, and those with high school or less. The bottom graph shows the same, after subtracting a linear trend in time with a kink-point in 1992. Calculations are based on the March CPS. For details, see Section 3.1.3. This figure replicates similar figures in Autor et al. (2008) and Acemoglu and Autor (2011).

builds on the specifications and the code provided by Borjas et al. (2012). The second data-set is a national annual time-series for the years 1963-2008 using the March CPS. Here we build on the specifications and code provided by Acemoglu and Autor (2011), including their pre-cleaning of the data.

For both these data-sets we restrict the sample to individuals aged between 25 and 64 years, and with less than 49 years of potential experience. We drop all self-employed or institutionalized workers. Labor supply for any given type of workers is defined as total hours worked. When calculating average log wages for any given type, we further restrict the sample to full-time workers (employed at least 40 weeks and working at least 35 hours per week) who are men. Our main analysis classifies workers into eight types, by education (high school dropouts, high school graduates, some college, and college graduates) and potential experience (less than 20 years and 20 years or more).

Results We first replicate results from the literature. The leading specification in the literature considers two types of workers, those with more than high school education and those with high school or less. Log relative wages of these two types are regressed on their log relative labor supply using national time series data for the US and controlling for a linear trend with a kink-point in 1992 (see Autor et al. 2008 and Acemoglu and Autor 2011). Running this regression, we replicate the estimate of -0.64 for the inverse elasticity of substitution reported by Acemoglu and Autor (2011). The corresponding time series are shown in Figure 1, where the first graph shows the actual series while the second graph shows the residualized series after controlling for a kinked time trend.

We next estimate the same parameter using our state-level decadal panel, and controlling for time and state fixed effects. Doing so, we find an elasticity of substitution of the same sign, but much smaller magnitude: -0.06, with a standard error of 0.04. We do not wish to take a stance on what causes this divergence of findings between the time-series and the state panel, but will proceed with obtaining our main estimates from the panel data. Using panel might be preferable to the extent that it allows us to control for business cycle variation and secular time trends using time fixed effects.

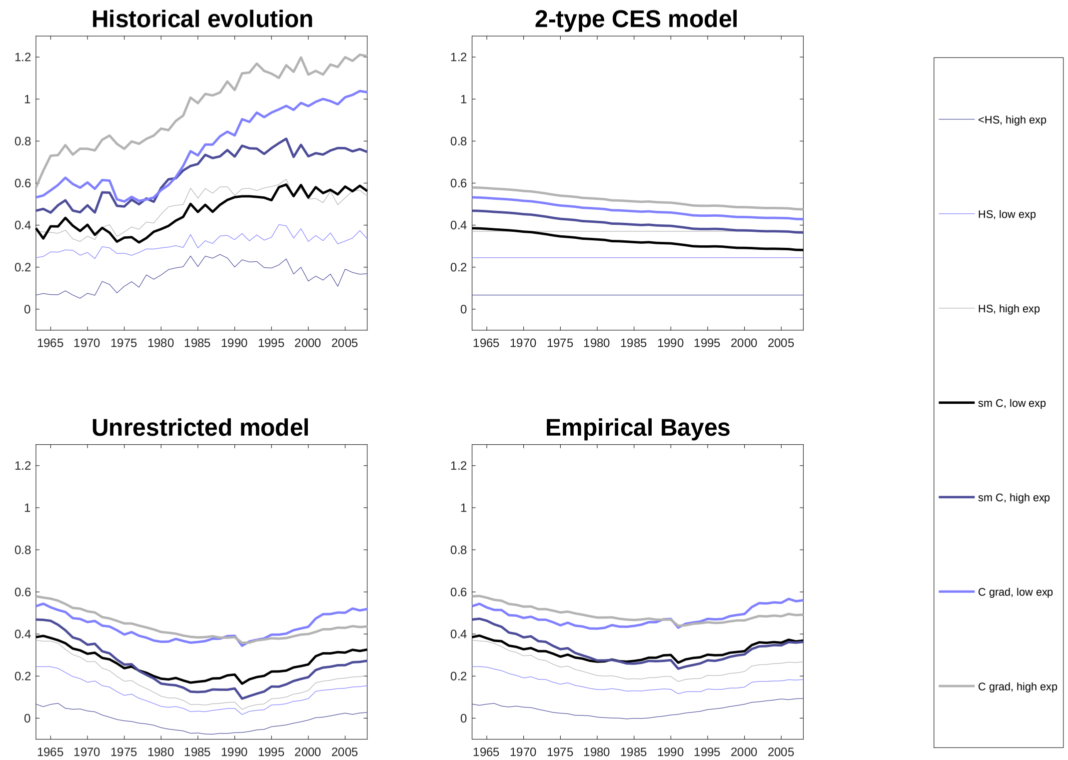
We now turn to our analysis using more disaggregated types of workers, classifying workers into 8 types by level of education and potential experience. The top left graph in Figure 2 shows the historical evolution of log wages of all types relative to the wage of high school dropouts with less than 20 years of potential experience. Clearly, there are patterns in the evolution of wages not captured by the classification into just 2 types. In particular, inequality across sub-types is rising over time, but in a non-linear manner.

The remaining graphs in this figure show the predicted (counterfactual) evolution of wages as implied by alternative estimates of labor demand (based on the state panel) and the historical evolution of labor supply (based on the national time series). Table 1 in the Appendix shows the corresponding coefficient estimates.

The top right graph of Figure 2 shows counterfactual wages as implied by the 2-type CES model. For this model, by construction, relative wages of sub-types remain fixed. The rising supply of college graduates, combined with the estimated inverse elasticity of -0.06, imply a modest compression of relative wages over time. The actually observed rising inequality would accordingly be due to demand factors.

The bottom left graph, and the second set of estimates in Table 1, are based on OLS estimation of the unrestricted model. These estimates suggest, as does the structural model, that changes of labor supply have induced a compression of wages over the initial three decades of our period. Some additional patterns emerge however. First, shifts in

Figure 2: Log relative wages in the US – actual evolution and counterfactual changes



Note: These figures show log wages of different types of workers relative to wages of high-school dropouts with less than 20 years of experience. The top left figure shows the actual historical evolution of relative wages, whereas the remaining figures show predicted counterfactual wages holding demand constant, based on the historical evolution of relative labor supply and alternative estimators of demand. Details are discussed in Section 3.1.3.

labor supply induced a widening of inequality over the most recent two decades. Second, these shifts also induced, over the initial three decades, a compression of wages between different workers with high school degrees or less and a widening between those with more than high school education. These effects appear to be reversed more recently.

The bottom right graph, and the final set of estimates in Table 1, are based on our preferred empirical Bayes estimator. As suggested by theory, and confirmed by visual inspection, these counterfactual predictions interpolate between those of the structural model and those of the unrestricted model. They are designed to balance bias and variance in a data-driven way. The predicted counterfactual changes of wages derived from these estimates are qualitatively similar to the unrestricted model, but of reduced magnitude.

The estimated $\widehat{\tau}^2$, our measure of model fit, is of a somewhat larger magnitude than the variance of the OLS coefficient estimates. This implies some, but not excessive, shrinkage towards the restricted estimates, thus leading to qualitatively similar conclusions of unrestricted and empirical Bayes predictions. This also suggests that the 2-type CES model does not provide a particularly good fit to our panel data.

3.2 Consumer demand and the Slutsky condition

In our second application we consider consumer choice and the restrictions on compensated demand implied by utility maximization. In this application we build on a rich literature on demand estimation, and more specifically on the recent contributions of Dette et al. (2016) and Blundell et al. (2017). Utility maximization by consumers implies that the matrix of slopes of compensated demand is symmetric and negative semi-definite. Consider quantile demand functions, where quantiles are across a population of consumers with arbitrary preference heterogeneity. Dette et al. (2016) have shown that negative semi-definiteness of compensated demand slopes holds not only for individual demand functions, but also for quantile demand functions.

This application involves high dimensional parameters to the extent that we are interested in estimating demand elasticities at many different price and income levels. In this application, the theoretical restrictions that we propose shrinking to implies negative semi-definiteness of compensated quantile demand elasticities, and in particular non-positive compensated own price elasticities.

3.2.1 Setup

Suppose that we have data on a set of consumers $i = 1, \ldots, n$, where we observe the log quantity Y_i of a good (gasoline, in our application) purchased by each consumer i. Consumer i is faced with a price P_i for the good under consideration and has income (total expenditures) W_i. Our goal is to estimate the price elasticity of demand for gasoline, at multiple price and income levels. To simplify our discussion, we shall assume that prices and incomes are statistically independent of unobserved preference heterogeneity across consumers. Controlling for covariates or using control functions as in Imbens and Newey (2009) would yield immediate extensions to the endogenous case.

Unrestricted estimator Let $q^\pi(p, w)$ be the π quantile of Y_i given $P_i = p, W_i = w$, where the quantile is across the distribution of consumers i and across the distribution of prices for other goods. Our goal is to estimate the (uncompensated) price elasticity β_j^p of the quantile demand function $q^\pi(p, w)$ at a series of price levels $p_1, \ldots, p_J$ and a given income level w, as well as the corresponding income elasticity β_j^w,

$$\beta_j^p = \frac{\partial \log q^\pi(p_j, w)}{\partial \log p}, \qquad \beta_j^w = \frac{\partial \log q^\pi(p_j, w)}{\partial \log w}.$$

We can get an unrestricted estimator of the price elasticity β_j^p and of the income elasticity β_j^w using local linear quantile regression,

$$\begin{aligned}(\widehat{\alpha}_j, \widehat{\beta}_j^p, \widehat{\beta}_j^w) = \underset{a, b^p, b^w}{\operatorname{argmin}} \sum_i & K_h(\log P_i - \log p_j, \log W_i - \log w) \\ & \cdot \rho_\pi(Y_i - a - b^p \cdot (\log P_i - \log p_j) - b^w \cdot (\log W_i - \log w)), \quad (21)\end{aligned}$$

where K_h is a kernel function of bandwidth h (we use the Epanechnikov kernel), and $\rho_\pi(e) = (\pi - \mathbf{1}(e < 0)) \cdot e$.[6] We estimate the variance V of $\widehat{\beta} = (\widehat{\beta}_j^p, \widehat{\beta}_j^w)_{j=1}^J$, jointly across all j, using the bootstrap. The variance of $\widehat{\alpha}_j$ is negligible relative to V under standard asymptotics, which is also true numerically in our application.

Negative compensated demand slopes Recall the classic consumer choice problem, as discussed in Chapter 3 of Mas-Colell et al. (1995). Let $\vec{X}_i$ be the k vector of goods demanded by consumer i, where $Y_i = \log X_{i1}$, and $\vec{P}_i$ the corresponding k vector

[6] For implementation of quantile regression, we use the code provided by Koenker (2005).

of prices. Denote $P_i = \vec{P}_{i1}$. Assume that $\vec{X}_i = \vec{X}_i(\vec{P}_i, W_i)$, where

$$\vec{X}_i(\vec{P}, w) = \underset{x}{\operatorname{argmax}}\ u_i(x) \quad s.t. \quad x \cdot \vec{P} \leq w,$$

where u_i is a continuous and locally nonsatiated utility function that represents a strictly convex preference relation. Define

$$S_i(\vec{P}, w) := \partial_{\vec{P}} \vec{X}_i(\vec{P}, w) + \partial_w \vec{X}_i(\vec{P}, w) \cdot \vec{X}_i(\vec{P}, w)'.$$

The $k \times k$ matrix $S_i(\vec{P}, w)$ collects the slopes of compensated (Hicksian) demand for consumer i. By Propositions 3.G.2 and 3.G.3 in Mas-Colell et al. (1995), the matrix $S_i(\vec{P}, w)$ is negative semi-definite, symmetric, and satisfies $S(\vec{P}, w)\vec{P} = 0$. Negative semi-definiteness implies, in particular, that the diagonal elements of $S_i(\vec{P}, w)$, corresponding to the compensated own-price elasticities of demand, are non-positive.

Absent restrictions on heterogeneity it is not possible to identify the slopes of any individual consumer's demand function. With exogeneous variation of P and w we can, however, identify quantiles $q^\pi(p, w)$ of demand for good 1 across consumers given the price of good 1 and given income. Under some regularity conditions, Theorem 1 in Dette et al. (2016) implies that

$$S^{\pi,1}(p, w) := \partial_p q^\pi(p, w) + \partial_w q_1^\pi(p, w) \cdot q^\pi(p, w) \leq 0.$$

Underlying this result is the fact that the slopes of of the quantile demand function $q^\pi(p, w)$ are equal to the average of individual demand slopes conditional on $X_{i1}(p, w) = q^\pi(p, w)$. Rewritten in terms of elasticities, we get the inequality

$$\partial_{\log p_1} \log q^\pi(p, w) + \partial_{\log w} \log q^\pi(p, w) \cdot q^\pi(p, w) p/w \leq 0. \tag{22}$$

This is the quantile analog of the condition that compensated own price elasticities are negative. Equation (22) is the key theoretical restriction which we use in this section in the construction of our empirical Bayes estimator.

3.2.2 Empirical Bayes estimation

Written in terms of the slope parameters of our quantile regressions, Equation (22) can be expressed as $\beta = (\beta_j^p, \beta_j^w)_{j=1}^J \in B^0$,

$$B^0 = \{b : \ b_j^p + b_j^w \cdot (\alpha_j p_j / w) \leq 0 \ \forall\, j\}. \tag{23}$$

We estimate β using the empirical Bayes estimator $\widehat{\beta}^{EB}$, which is constructed as follows. This estimator shrinks the unrestricted estimator $\widehat{\beta}$ toward $\widehat{\beta}^0 \in B^0$. The hyperparameter τ^2 is chosen to minimize Stein's unbiased risk estimate (SURE).

$$\begin{aligned}
\widehat{\beta}^0(\tau^2) &= \underset{b^0 \in B^0}{\operatorname{argmin}}\ (\widehat{\beta} - b^0)' \cdot (\tau^2 I + \widehat{V})^{-1} \cdot (\widehat{\beta} - b^0) \\
\widehat{\beta}^{EB}(\tau^2) &= \widehat{\beta}^0(\tau^2) + \left(I + \frac{1}{\tau^2}\widehat{V}\right)^{-1} \cdot (\widehat{\beta} - \widehat{\beta}^0(\tau^2)) \\
g(\widehat{\beta}) &= \widehat{\beta} - \widehat{\beta}^{EB}(\tau^2) \\
SURE(\tau^2) &= \left\| g(\widehat{\beta}) \right\|^2 + 2 \cdot \operatorname{trace}\left(\nabla g(\widehat{\beta}) \cdot \widehat{V}\right) \\
\widehat{\tau}^2 &= \operatorname{argmin}\ SURE(\tau^2) \\
\widehat{\beta}^{EB} &= \widehat{\beta}^{EB}(\widehat{\tau}^2).
\end{aligned} \tag{24}$$

3.2.3 Empirical application

We implement this approach in order to estimate the price and income elasticity of gasoline demand. We use the data and sample construction of Blundell et al. (2017); details can be found in their discussion, which we briefly summarize here. The data are from the 2001 National Household Travel Survey (NHTS). Heterogeneity is reduced by restricting the sample to households with a white respondent, two or more adults, at least one child under age 16, and at least one driver. Households in the most rural areas and in Hawaii are dropped, as are households with missing relevant variables or without a gasoline based vehicle. The resulting sample contains 3,640 observations.

We first present unrestricted estimates using local linear quantile regression, with bandwidths equal to the standard deviation of log price and log income, respectively, and with $\log p_j$ ranging over 80 gridpoints in the observed range of values for log price. The income levels w considered are at the .25, .5, and .75 quantiles of the income distribution in the sample. We focus on the median demand function, corresponding to $\pi = .5$. For the local linear quantile regressions, we choose the bandwidth for both $\log p$ and $\log w$ equal to their respective sample standard deviations. The joint variance of the resulting estimates is estimated using the bootstrap, resampling 1000 times.

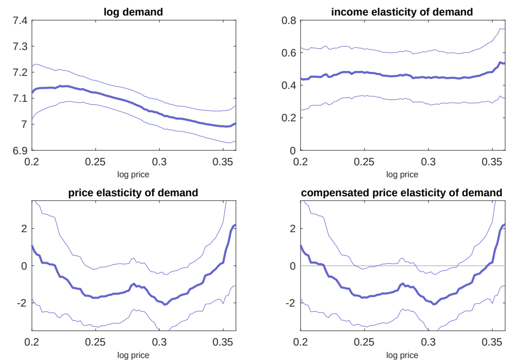
Figure 3 plots the resulting estimates for log gasoline demand, the budget share of gasoline, the price elasticity β^p, and the compensated price elasticity $\beta^p_j + \beta^w_j \cdot (\alpha_j p_j / w)$, each as a function of price p_j. The figure also shows 95% confidence bands, based on the bootstrapped standard errors. The estimates for log demand shown are similar to the unrestricted estimates of Figure 1 in Blundell et al. (2017), where they use spline regression instead of local linear regression.

The budget share $(\alpha_j p_j / w)$ of gasoline among total expenditures is fairly small for most price and income levels (the median budget share equals 0.026), and the income elasticity β_j^w is less than .5 in most cases, so that the compensated elasticity is quite close to the uncompensated elasticity β_j^p. The theoretical restriction implied by utility maximization is that the compensated elasticity is non-positive. This restriction is violated for our unrestricted estimates for low and high price levels, for all income levels. The restriction is satisfied for our unrestricted estimates at intermediate price levels.

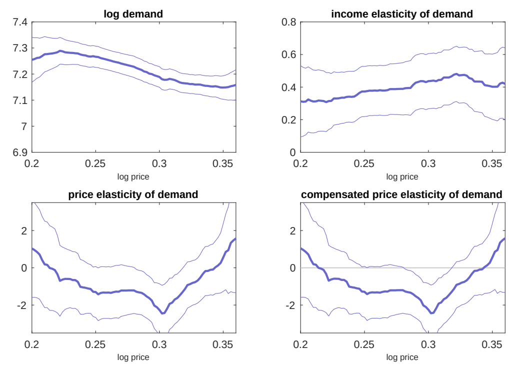
Figure 4 again plots estimates of price elasticities and income elasticities across price and income levels. This figure shows (i) unrestricted estimates $\widehat{\beta}$ (the same as in Figure 4), (ii) restricted estimates $\widehat{\beta}^0$, subject to the theoretical restriction on compensated price elasticities, and (iii) empirical Bayes estimates shrinking toward the theoretical restriction. The restricted estimates are equal to the unrestricted estimates for price levels p_j where the unrestricted estimates of compensated elasticities are already non-positive. The empirical Bayes estimates are intermediate between the unrestricted and restricted estimates. The optimal shrinkage parameter τ^2 is estimated using SURE in order to minimimize mean squared error. We obtain estimates of 0.48, 0.32, and 0.38 for low, middle, and high incomes, respectively. This results in estimates that are closer to the restricted estimates than to the unrestricted ones.

Figure 3: Median gasoline demand

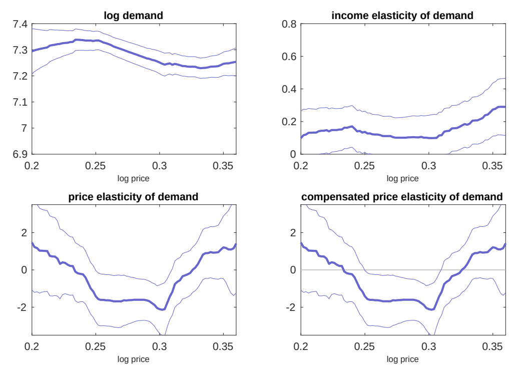
LOW INCOMES



MIDDLE INCOMES

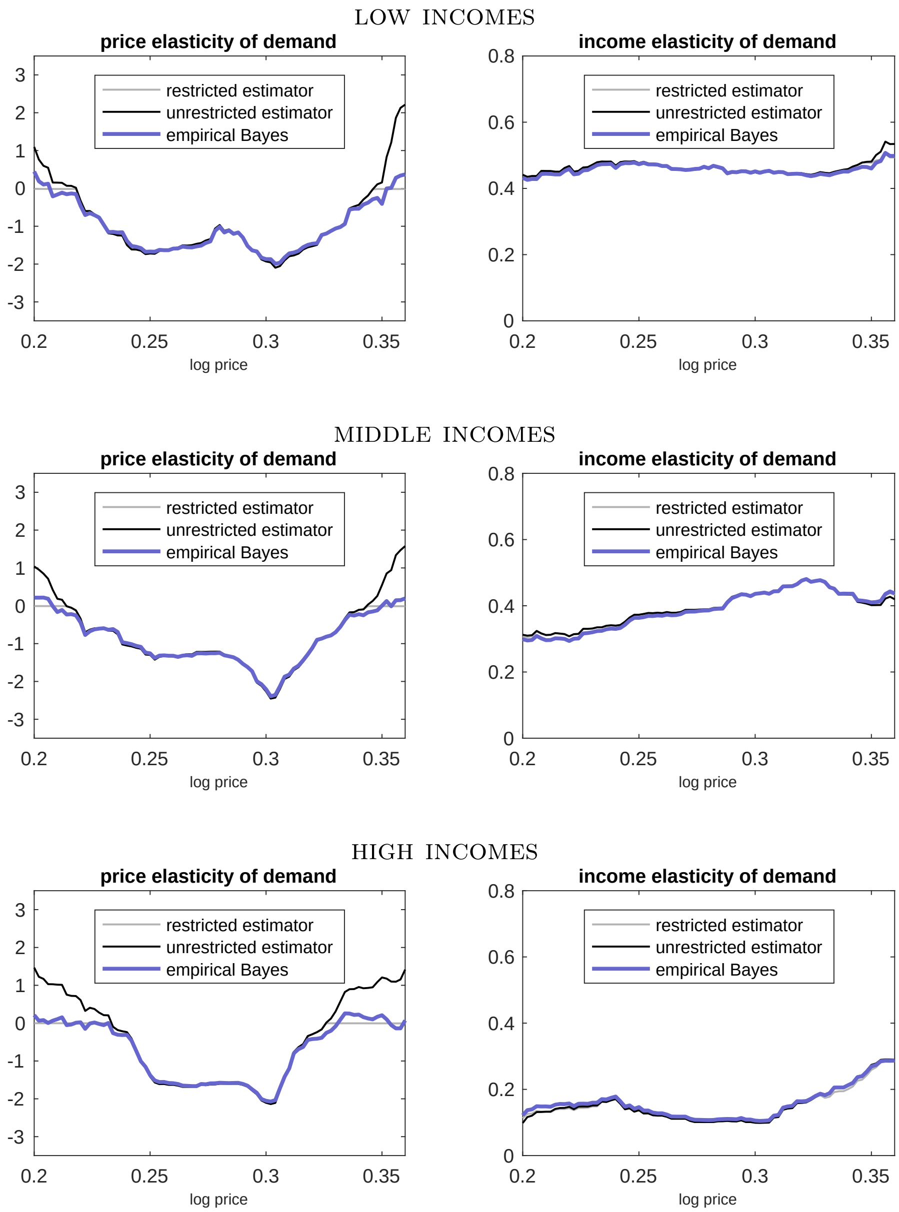


HIGH INCOMES



Note: This figure shows unrestricted estimates of gasoline demand and of the corresponding price and income elasticities across a range of price levels, based on local linear quantile regression. The thin lines are 95% confidence bands based on the bootstrap.

Figure 4: Empirical Bayes estimator of median gasoline demand elasticities



Note: This figure shows unrestricted, restricted, and empirical Bayes estimates of gasoline demand, where empirical Bayes shrinks unrestricted estimates toward the restriction of non-positive compensated price elasticities.

4 Behavior of the empirical Bayes estimator, its mean squared error, and uniform dominance

In this section, we characterize the behavior of the empirical Bayes estimator introduced in Section 2. We start with some basic properties. We show, in particular, consistency of the estimator, demonstrate how counterfactual predictions combine theory and available evidence in a data-driven, intuitive way, and rewrite the estimator in canonical coordinates. The rest of this Section is then dedicated to characterizing the risk function (mean squared error, MSE) of $\widehat{\beta}^{EB}$. Section B in the supplementary appendix explores the geometry of the mapping from the preliminary, unrestricted estimator $\widehat{\beta}$ to the empirical Bayes estimator $\widehat{\beta}^{EB}$.

The desirability of using our proposed estimator $\widehat{\beta}^{EB}$ hinges on the claim that it delivers more precise estimates (estimates with lower MSE) relative to the unrestricted estimator $\widehat{\beta}$. We justify this claim by characterizing the mean squared error of $\widehat{\beta}^{EB}$ using two complementary approaches. The first approach uses an asymptotic approximation, assuming that the dimension J of β is large relative to the dimension of the hyper-parameters. For such high-dimensional estimation problems, the variability of the estimated hyper-parameters $(\widehat{\beta}^0, \widehat{\tau}^2)$ is small relative to the variability of $\widehat{\beta}$. We can therefore approximate $(\widehat{\beta}^0, \widehat{\tau}^2)$, which maximize the marginal likelihood, by (β^0, τ^2), which maximize the expected marginal likelihood. With this approximation, $\widehat{\beta}^{EB}$ becomes a linear function of $\widehat{\beta}$, and we can write its MSE as a simple sum of variance and squared bias terms.

The second approach uses Stein's Unbiased Risk Estimate (SURE), to prove uniform dominance relative to $\widehat{\beta}$ for fixed J. This approach takes into account the variability of $(\widehat{\beta}^0, \widehat{\tau}^2)$. This approach extends the classic proof of uniform dominance of the James-Stein shrinkage estimator to the case of shrinkage toward more general linear equality and inequality restrictions.

These two approaches toward characterizing the MSE of our estimator are complementary and non-nested. While the first approach relies on a large-J approximation, it does not require normality of $\widehat{\beta}$, nor does it restrict the form of $V = \mathrm{Var}(\widehat{\beta})$ or the form of the theoretical restrictions B_0. The second approach, on the other hand, does not rely on approximations, but it requires normality of $\widehat{\beta}$ and restricts V to a canonical case.

Recall the form of the estimator introduced in Section 2,

$$
\begin{aligned}
(\widehat{\beta}^0, \widehat{\tau}^2) &= \underset{b^0 \in B^0, t^2}{\operatorname{argmin}} \log\left(\det(\Sigma(t^2, \widehat{V}))\right) + (\widehat{\beta} - b^0)' \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot (\widehat{\beta} - b^0), \\
B^0 &= \{b :\ R_1 \cdot b = 0,\ R_2 \cdot b \leq 0\}, \\
\widehat{\beta}^{EB} &= \widehat{\beta}^0 + \left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot (\widehat{\beta} - \widehat{\beta}^0),
\end{aligned} \tag{25}
$$

where $\widehat{\beta}$, $\widehat{V}$, and B^0 are known, and $\Sigma(\tau^2, \widehat{V}) = \tau^2 \cdot I + \widehat{V}$. We shall consider this estimator for the remainder of this section, and will assume throughout that $E_\beta[\widehat{\beta}] = \beta$ (the unrestricted estimator is unbiased) and $\mathrm{Var}_\beta(\widehat{\beta}) = V = \widehat{V}$ (the variance is known). We use subscript β to emphasize that expectation and variance are taken for a given, non-random β over the sampling distribution of $\widehat{\beta}$. We do not impose normality of $\widehat{\beta}$ until Section 4.3.

4.1 Consistency and data-driven predictions

In contrast to restricted estimation in the misspecified case, the empirical Bayes estimator of β is consistent as sample size n goes to infinity. If $\widehat{V} \to^p 0$, then $\widehat{\beta}^{EB}$ and $\widehat{\beta}$ become asymptotically equivalent. Consistency of $\widehat{\beta}^{EB}$ therefore follows immediately from consistency of unrestricted estimation.

Proposition 1 (Consistency)
Consider the empirical Bayes estimator defined in Equation (25). Assume that $\widehat{\beta} \to^p \beta$ and $\widehat{V} \to^p 0$ as $n \to \infty$. Then $\widehat{\beta}^{EB} \to^p \beta$ as n goes to infinity.

The proof of this proposition can be found in appendix B. The proof of consistency relies on the fact that $\widehat{\beta}^{EB} \approx \widehat{\beta}$ if $\widehat{V} \approx 0$.

The formula for $\widehat{\beta}^{EB}$ given in Equation (25) shows that the empirical Bayes estimator interpolates between the unrestricted estimator $\widehat{\beta}$ and the structural estimator $\widehat{\beta}^0$. Suppose we are interested in making a prediction of the form $\widehat{y} = x \cdot \widehat{\beta}^{EB}$. Heuristically, we would like our prediction to be based on the data alone (neglecting the structural model) whenever the data by themselves do allow us to make a precise prediction. When, on the other hand, a prediction of counterfactuals based on the data alone would be imprecise, we would like to leverage the theoretical model. The following proposition shows that this is exactly how the empirical Bayes estimator behaves.

Proposition 2 (Counterfactual predictions)
Consider the empirical Bayes estimator defined in Equation (25)*. Consider the prediction at* x*,* $\widehat{y} = x \cdot \widehat{\beta}^{EB}$*, and assume that* $\widehat{V}$ *is non-singular. Then*

$$\left|\widehat{y} - x \cdot \widehat{\beta}\right| \leq \frac{\sqrt{x\widehat{V}x}}{\widehat{\tau}} \cdot \|\widehat{\beta}\|,$$
$$\text{and } \left|\widehat{y} - x \cdot \widehat{\beta}^0\right| \leq \widehat{\tau} \cdot \sqrt{x\widehat{V}^{-1}x} \cdot \|\widehat{\beta}\|.$$

The first inequality of proposition 2 tells us that empirical Bayes predictions are close to unrestricted predictions whenever the standard deviation of the latter, $\sqrt{x\widehat{V}x}$, is small relative to the measure of model fit $\widehat{\tau}$. The second inequality tells us that empirical Bayes predictions are close to predictions using the structural model when the reverse situation holds. To gain intuition for this result, rearrange Equation (25),

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

Consider a point x such that $x \cdot \widehat{V} \cdot x' \approx 0$, which implies $x \cdot \widehat{V} \approx 0$. For such a point x, we get

$$x \cdot \widehat{\beta}^{EB} = x \cdot \left[\widehat{\beta} + \widehat{V} \cdot \left(\widehat{\tau}^2 \cdot I + \widehat{V}\right)^{-1} \cdot (\widehat{\beta}^0 - \widehat{\beta})\right] \approx x \cdot \widehat{\beta}.$$

This suggests that for points x with small variance of the unrestricted prediction $\widehat{y} = x \cdot \widehat{\beta}$, the predicted value $\widehat{y}$ using empirical Bayes is close to the predicted value using unrestricted estimation – and thus also close to the predicted value using the true coefficients β, as the latter is estimated with small variance. This insight is relevant in particular when considering historical counterfactuals ("how much did past changes in labor supply affect wage inequality?"), that might rely on variation which is actually observed in the data.

Canonical coordinates The variance matrix $\widehat{V}$ need not be diagonal in general. This matrix is, however, symmetric and positive semi-definite. We can therefore always find some orthonormal matrix O such that

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

Expressing both $\widehat{\beta}$ and $\widehat{\beta}^{EB}$ in terms of coordinates corresponding to the columns of O, it is then without loss of generality to assume $\widehat{V} = \text{diag}(v_j)$. Quadratic estimation error is invariant under such an orthonormal change of coordinates, as well.

Under the assumption that $\widehat{V} = \text{diag}(v_j)$, the empirical Bayes estimator is given by a component-wise weighted average of $\widehat{\beta}^0$ and $\widehat{\beta}$,

$$\widehat{\beta}_j^{EB} = \left(\frac{v_j}{\widehat{\tau}^2 + v_j}\right) \cdot \widehat{\beta}_j^0 + \left(\frac{\widehat{\tau}^2}{\widehat{\tau}^2 + v_j}\right) \cdot \widehat{\beta}_j. \tag{26}$$

The hyper-parameters β^0 and τ^2 are estimated by maximizing the marginal log likelihood, which now simplifies to

$$(\widehat{\beta}^0, \widehat{\tau}^2) = \underset{b^0 \in B^0, \tau^2}{\operatorname{argmin}} \, \tfrac{1}{J} \cdot \sum_j \left(\log(\tau^2 + v_j) + \frac{\left(\widehat{\beta}_j - b_j^0\right)^2}{\tau^2 + v_j} \right). \tag{27}$$

Writing $\widehat{\beta}^{EB}$ in canonical coordinates makes transparent how our estimator differs from the family of estimators considered by Hansen (2016), which (in our setting) take the form

$$\left(1 - \widehat{\lambda}\right) \cdot \widehat{\beta}^0 + \widehat{\lambda} \cdot \widehat{\beta},$$

so that each component of $\widehat{\beta}$ is shrunk by the same factor $\widehat{\lambda}$. Our estimator allows for a more flexible form of shrinkage, where precisely estimated components of $\widehat{\beta}$ (components with small v_j) are not shrunk by much, whereas imprecisely estimated components are shrunk substantially toward the predictions of the theoretical model.

4.2 Large J characterization of the MSE

One of the main arguments for using an empirical Bayes approach such as the one proposed in this paper is that it performs well in terms of risk (mean squared error, MSE). We might expect such favorable performance since our estimator is a close relative of the James-Stein shrinkage estimator, which is well known to uniformly dominate the unrestricted estimator for dimension $J \geq 3$.

We now proceed to characterize the risk of our estimator for large J. The key argument in our characterization is that variability of $(\widehat{\beta}^0, \widehat{\tau}^2)$ can be neglected for large J when calculating the MSE. We formalize this argument in Theorem 1. We then discuss the properties of the asymptotic approximation to risk obtained in this way and compare it to an oracle-optimal choice of (β^0, τ^2).

Asymptotic characterization of risk Our goal is to characterize the squared error of the empirical Bayes estimator, $SE(\widehat{\beta}^{EB}, \beta) = \frac{1}{J}\|\widehat{\beta}^{EB} - \beta\|^2$, and the corresponding mean squared error (MSE) given β,

$$MSE(\widehat{\beta}^{EB}, \beta) = E_\beta\left[\frac{1}{J}\|\widehat{\beta}^{EB} - \beta\|^2\right] = E_\beta\left[\frac{1}{J} \cdot \sum_{j=1}^{J} \left(\widehat{\beta}_j^{EB} - \beta_j\right)^2\right].$$

The mean squared error is the most common criterion for evaluating the performance of estimators in the theory of point estimation; see for instance chapter 7 in Casella and Berger (2001). The MSE is equal to the variance of the estimator plus the square of its bias. Estimators with good performance in terms of MSE trade off bias and variance. This is familiar from non-parametric estimation in econometrics, and central to the more recent literature on machine learning. Depending on context, other loss functions might sometimes be appropriate.

In order to obtain our desired characterizations, we consider an asymptotic approximation where J becomes large, such that $\widehat{\beta}^0$ and $\widehat{\tau}^2$ converge in probability. Let $\widehat{\beta}^{EB}(b^0, \tau^2)$ be the empirical Bayes estimator for given (non-random) hyper-parameters (b^0, τ^2),[7] and let $MSE(\widehat{\beta}^{EB}(b^0, \tau^2), \beta)$ be the corresponding mean squared error. Recalling our assumption that $\widehat{V} = V = \mathrm{diag}(v_j)$, the MSE given b^0 and τ^2 can be written as a sum of variance and squared bias terms,

$$MSE(\widehat{\beta}^{EB}(b^0, \tau^2), \beta) = \frac{1}{J} \cdot \sum_{j=1}^{J} \left[\left(\frac{\tau^2}{\tau^2 + v_j}\right)^2 \cdot v_j + \left(\frac{v_j}{\tau^2 + v_j}\right)^2 \cdot (\beta_j - b_j^0)^2\right]. \quad (28)$$

Define (β^0, τ^{*2}) to be the maximizer of the expected marginal log-likelihood, or equivalently the minimizer of the expectation of (27),

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

The following theorem shows that as J becomes large, we can approximate the loss (squared error) of the empirical Bayes estimator $\widehat{\beta}^{EB}$ by the risk (mean squared error) of the infeasible estimator using the limiting pseudo-true values of (β^0, τ^{*2}). The theorem relies on an assumption regarding the behavior of β, V and B^0 when J goes to infinity, which we shall discuss immediately after stating our result.

[7] Actually, $\widehat{\beta}^{EB}(b^0, \tau^2)$ is the Bayes estimator for the prior $\beta \sim N(b^0, \tau^2 I)$.

Theorem 1
Consider the empirical Bayes estimator of Equations (26) *and* (27). *Under Assumption 1,*

$$SE(\widehat{\beta}^{EB}, \beta) - MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) \to^p 0$$

as $J \to \infty$.

Assumption 1 (Random coefficient sequence)
For the estimator defined by Equations (26) *and* (27), *assume that* B^0 *is of the form*

$$B^0 = \{b^0 : \; b^0 = M' \cdot c, \; S \cdot c \leq 0\},$$

where c *is of dimension* k, S *does not depend on* J, *and* $M = (M_1, \ldots, M_J)$. *The components* $(\widehat{\beta}_j, \beta_j, v_j, M_j)$ *of* $(\widehat{\beta}, \beta, \mathrm{diag}(v), M)$ *are i.i.d. draws from some distribution* P, *where* P *does not depend on* J, *and where* $\|(\beta_j, v_j, M_j)\| < C$ *for some fixed constant* C *with probability* 1. *As before,* $E[\widehat{\beta}|\beta, V, M] = \beta$ *and* $\mathrm{Var}(\widehat{\beta}|\beta, V, M) = V$.

Discussion Our goal in this section is to give a simple characterization of the MSE of our proposed estimator, based on the variance and bias squared formula of Equation (28), and based on the expected first order conditions for the hyper-parameters. Theorem 1 states that the conditions of Assumption 1 are sufficient to allow us to do so, as long as J is large.

In order to state results of this form, we need to spell out what happens to the components of β, V, and B_0 as J increases. The easiest way to do this is in terms of the random coefficient setup of Assumption 1. Note that this assumption also implies that the dimension k of the set B^0 stays constant as J increases. This is achieved by reparametrizing, and writing $\beta = M'c$ for fixed c of dimension k and M a sequence of random vectors. An alternative to the random coefficient setup would be to consider deterministic sequences (β_j, v_j, M_j), and to impose constraints on their behavior. This is the approach taken by Xie et al. (2012), for instance.

In related work (Abadie and Kasy, 2017) we provide stronger results of the form of Theorem 1, proving uniform risk consistency of tuning parameter choice using criteria such as cross validation or Stein's unbiased risk estimate. The uniform risk consistency is toward estimators using an infeasible oracle optimal choice of tuning parameters. Theorem 1, by contrast, shows point-wise risk consistency toward the pseudo-true choice of (β^0, τ^{*2}), which need not be optimal.

Recall that we obtain unrestricted estimation and structural estimation as limiting cases of our proposed estimator, where $\tau^2 \to \infty$ corresponds to unrestricted estimation and $\tau^2 \to 0$ to restricted estimation. The mean squared error $MSE(\widehat{\beta}^{EB}(b^0, \tau^2), \beta)$ for given values of (b^0, τ^2) is equal to the sum of a variance term and a squared bias term, cf. Equation (28). The mean squared error of the unrestricted estimator contains only variance terms, $MSE(b^0, \infty) = \frac{1}{J}\sum_j v_j$, and the mean squared error of the structural estimator converges to an average containing only bias terms, $\min_{b^0 \in B^0} MSE(b^0, 0) - \min_{b^0 \in B^0} \frac{1}{J}\sum_j (\beta_j - b_j^0)^2 \to^p = 0$.

Under the assumptions of Theorem 1 it then follows immediately that, for large enough J, our estimator has lower mean squared error than the unrestricted estimator if

$$MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) < \tfrac{1}{J} \cdot \sum_{j=1}^{J} v_j,$$

and larger mean squared error if this inequality is reversed. Our estimator has lower mean squared error than the restricted estimator for large J if

$$MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) < \min_{b^0 \in B^0} \tfrac{1}{J} \cdot \sum_{j=1}^{J} (\beta_j - b_j^0)^2,$$

and larger mean squared error if this inequality is reversed.

The role of heteroskedasticity The infeasible oracle-optimal choice of (b^0, τ^2) would minimize $MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta)$ and automatically yield an estimator that dominates structural and unrestricted estimation uniformly. The first order condition for the optimal $\tau^{\times 2}$ that minimizes the mean squared error is

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

The empirical Bayes estimate $(\widehat{\beta}^0, \widehat{\tau}^2)$, by contrast, maximizes the marginal log likelihood, and for large J (β^0, τ^{*2}) approximately maximizes the expected log likelihood. The first order condition characterizing τ^{*2} is

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

How does τ^{*2} relate to the optimal choice of $\tau^{\times 2}$? As can be seen from the first order conditions, both are weighted averages of $(\beta_j - \beta_j^0)^2$. The weights differ slightly, however. Minimization of the mean squared error assigns a slightly larger weight to draws j with smaller values of v_j, relative to to maximization of the expected log likelihood. For homoskedastic settings (v_j constant), or settings where v_j and β_j are independent across j, the two objectives do in fact coincide. In these cases it is immediate that our empirical Bayes estimator dominates both unrestricted and restricted estimation for large enough J. It is also possible to reverse the dominance of empirical Bayes relative to unrestricted estimation, however, by introducing strong correlation across j between β_j and v_j. Suppose in particular that J is even, that $B^0 = \{0\}$, and that

$$v_j = \beta_j = 0 \text{ for j odd,}$$
$$v_j = \beta_j = 2 \text{ for j even.}$$

Then $\tau^{*2} = 0$ and $MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) = 2$ while $MSE(\beta^0, \infty) = 1$ so that unrestricted estimation has lower mean squared error than empirical Bayes for large samples. Restricted estimation, on the other hand, dominates empirical Bayes for small enough samples if $\beta \in B^0$. Note, however, that in this case the two estimators become equivalent for large enough J since $\widehat{\tau}^2 \to^p 0$.

4.3 Fixed J characterization of the MSE

The last section characterized the mean squared error of $\widehat{\beta}^{EB}$ under the assumption that sampling variability of the hyper-parameters $(\widehat{\beta}^0, \widehat{\tau}^2)$ is negligible relative to the variance of $\widehat{\beta}$. We formally showed that this is a valid assumption for the case for large J under a random coefficient sequence. The advantage of this characterization is that it yields simple and easily interpreted expressions for the MSE. The disadvantage is that it relies on an approximation that might be misleading when J is too small.

In this section, we characterize the MSE taking into account the sampling variability of the hyper-parameters $(\widehat{\beta}^0, \widehat{\tau}^2)$, but restrict our attention to the homoskedastic and normally distributed case with canonical coordinates for the restrictions imposed by B^0. In this section J is fixed and β is non-random. The results in this section generalize a classic proof by Stein (1981), of the uniform dominance of James-Stein shrinkage, to our estimator.

Homoskedastic case, canonical coordinates We assume for the rest of this section that $V = I$, and that the restrictions imposed by B^0 take the canonical form

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

These assumptions are restrictive. Homoskedasticity eliminates the weighting issues discussed in the previous section. The form of the equality restrictions in the definition of B^0 is without loss of generality. The assumed form of the inequality restrictions is restrictive whenever $L - K > 1$, but our derivation easily generalizes to more general sets B_0. Denote

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

Under these assumptions, our empirical Bayes estimator is given as follows.

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

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

$$\widehat{\beta}_j^{EB} = \begin{cases} \frac{\widehat{\tau}^2}{\widehat{\tau}^2+1} \cdot \widehat{\beta}_j & j = 1, \dots K \\ & \quad \textit{or } j = K+1, \dots, L \textit{ and } \widehat{\beta}_j > 0, \\ \widehat{\beta}_j & \textit{else.} \end{cases} \tag{29}$$

Stein's Unbiased Risk Estimate (SURE) A celebrated result by Stein (1981) provides a characterization of the mean squared error of arbitrary estimators of the form $\widetilde{\beta} = \widehat{\beta} + g(\widehat{\beta})$, whenever $\widehat{\beta} \sim N(\beta, I)$ and g is almost differentiable.[8] By Theorem 1 of Stein (1981), the risk function (mean squared error) of $\widetilde{\beta}$ as an estimator of β is given by

$$MSE(\widetilde{\beta}, \beta) = 1 + \tfrac{1}{J} E_\beta \left[\|g(\widehat{\beta})\|^2 + 2\nabla \cdot g(\widehat{\beta}) \right], \tag{30}$$

where $\nabla \cdot g = \sum_j \partial_j g_j$ is the divergence of g and the expectation is taken for a fixed β. The risk function of $\widehat{\beta}$ itself as an estimator of β is given by $MSE(\widehat{\beta}, \beta) = 1$. Stein's

[8] g is almost differentiable if there exists a function $\nabla g = (\partial_1 g, \dots, \partial_J g)$ such that we can write $g(b'') - g(b') = \int_{b'}^{b''} \nabla g(b) db$ for all b', b'' and arbitrary paths of integration between these two points.

result immediately implies that $\widetilde{\beta}$ uniformly dominates $\widehat{\beta}$ in terms of MSE if

$$\|g(\widehat{\beta})\|^2 + 2\nabla \cdot g(\widehat{\beta}) < 0 \tag{31}$$

for all $\widehat{\beta}$. Note that this is a sufficient but not necessary condition for uniform dominance.

SURE for our estimator We now apply Stein's general result to our estimator, as defined in Equation (29).

Theorem 2
Assume that $\widehat{\beta} \sim N(\beta, I)$ and consider the estimator $\widehat{\beta}^{EB}$ as defined in Equation (29). Then $MSE(\widehat{\beta}^{EB}, \beta) = 1 + E_\beta[\Delta]$, where

$$\Delta = \begin{cases} \frac{1}{R} \cdot [J + 4 - 2J^*] & R > J \\ \frac{1}{J} \cdot [R - 2J^*] & else, \end{cases} \tag{32}$$

$R = \sum_{j=1}^{K} \widehat{\beta}_j^2 + \sum_{j=K+1}^{L} \max(\widehat{\beta}_j, 0)^2$, and $J^* = K + \sum_{j=K+1}^{L} \mathbf{1}(\widehat{\beta}_j > 0)$.

By Stein's result, $\widehat{\beta}^{EB}$ uniformly dominates $\widehat{\beta}$ in terms of MSE if $\|g(\widehat{\beta})\|^2 + 2\nabla \cdot g(\widehat{\beta}) < 0$ for all $\widehat{\beta}$. By Equation (32), the empirical Bayes estimator therefore has uniformly lower risk than the unrestricted estimator for all β if

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

Since $J^* \geq K$, this holds automatically if $K > J/2 + 2$, that is, if there are enough equality restrictions. Note, however, that the inequality restrictions also contribute to reducing risk.

Corrected degrees of freedom We can improve risk uniformly by applying a degree of freedom corrections to the estimation of $\widehat{\tau}^2$. Replacing, in particular, J by $L - 2$ in the denominator of the expression defining $\widehat{\tau}^2$, we get

$$\widehat{\tau}^2 = \max\left(\tfrac{1}{L-2} R - 1, 0\right),$$

$$\|g(\widehat{\beta})\|^2 + 2\nabla \cdot g(\widehat{\beta}) = \begin{cases} \frac{L-2}{R} \cdot [L + 2 - 2J^*] & R > L - 2 \\ R - 2J^* & else, \end{cases}$$

which is uniformly more negative than the corresponding expression for our empirical Bayes estimator. To see this, note that the quadratic expression $x[x + 4 - 2J^*]$ is minimized at $x = J^* - 2 \geq L - 2$. This estimator with corrected degrees of freedom uniformly dominates the maximum likelihood estimator if $J^* > \frac{L-2}{2}$, which holds automatically if $2K > L - 2$. This estimator shrinks less aggressively toward the set B^0 relative to the empirical Bayes estimator discussed before.

5 Conclusion

We have proposed a general purpose approach for using economic theory in order to construct estimators. These estimators perform particularly well when the empirical predictions of the theory are approximately correct, but are robust to moderate or large violations of the theoretical predictions.

Our approach can be summarized as follows: (i) Obtain a first-stage estimate of the parameters of interest that neglects the theoretical predictions. This first-stage estimate will often have a large variance. (ii) Assume that the true parameter values are equal to parameter values conforming to the theoretical predictions (the structural model), plus some noise of unknown variance. This assumption yields a family of priors for the parameters of interest. The priors are indexed by hyperparameters, namely the variance of noise and the parameters of the structural model. (iii) Use the marginal likelihood of the data given the hyperparameters to obtain estimates of the latter. The estimated variance of noise, in particular, provides a measure of model fit. (iv) Use Bayesian updating conditional on the estimated hyperparameters and the data in order to obtain estimates of the parameters of interest. We demonstrate how to implement this approach in a variety of settings, constructing estimators that shrink toward parameter sets consistent with economic theories, such as structural models of labor demand, consumer demand satisfying Slutsky conditions, general equilibrium models of asset markets, abstract theories of economic decision making, or structural discrete choice models.

In a normal-normal setting with linear equality and inequality restrictions implied by economic theory, our approach leads to particularly tractable and interpretable estimators. Theorems 1 and 2 provide characterizations of the risk function of our estimator. Theorem 1 is based on an asymptotic approximation that implies that the variability of the estimated hyperparameters is negligible relative to variability of the estimates of interest. This assumption is justified as long as the dimension of the parameters of interest is large relative to the dimension of the hyperparameters. Theorems 2 uses Stein's unbiased risk estimate to provide a characterization and proof of uniform dominance that does not rely on this asymptotic approximation.

A Inference

This paper does not contribute to the theory of shrinkage inference. For empirical applications we adapt the heuristic approach introduced by Laird and Louis (1987) to our setting. Inference in our setting is easily implemented, though conceptually somewhat subtle. We construct empirical Bayes confidence regions C for β. Such confidence regions must satisfy

$$P(\beta \in C|\theta) \geq 1 - \alpha \tag{33}$$

and were first proposed by Morris (1983) and analyzed further by Laird and Louis (1987) and Carlin and Gelfand (1990). Definition (33) arguably captures the natural notion of inference corresponding to empirical Bayes estimation. Empirical Bayes confidence regions are intermediate between frequentist confidence sets and Bayesian pre-posterior inference. The requirement of definition (33) is weaker than the requirement of frequentist coverage, $P(\beta \in C|\eta) \geq 1 - \alpha$.

We use standard frequentist inference to capture sampling variation of the estimates $\widehat{\beta}^{EB}$ and posterior inference to capture uncertainty about β given these estimates. The proposed procedure obtains a predictive distribution for β that is similar to a posterior distribution of the form

$$P\left(\beta|\widehat{\beta}, \widehat{V}\right) = \int P\left(\beta|\widehat{\beta}, \widehat{V}, \theta\right) P\left(\theta|\widehat{\beta}, \widehat{V}\right) d\theta,$$

but replaces the posterior for the hyperparameter θ by the sampling distribution Q_R for $\widehat{\theta}$ obtained using standard frequentist inference, thus obtaining a mixture distribution

$$M\left(\beta|\widehat{\beta}, \widehat{V}\right) = \int P\left(\beta|\widehat{\beta}, \widehat{V}, \theta\right) Q_R\left(\theta|\widehat{\beta}, \widehat{V}\right) d\theta. \tag{34}$$

Our inference procedure can be summarized as follows:

1. Obtain $r = 1, \ldots, R$ i.i.d. draws $\widehat{\beta}_r$ from the distribution $N\left(\widehat{\beta}, \widehat{V}\right)$.

2. For each of these R draws, obtain estimates $\widehat{\theta}_r = (\widehat{\beta}_{0,r}, \widehat{\tau}_r^2)$ by maximizing the marginal likelihood, as discussed in Section 2.3.

3. Calculate the posterior mean $\widehat{\beta}_r^{EB}$ and variance V_r^{EB} for β conditional on $\widehat{\beta}_r$ and

$\widehat{\theta}_r$, using Equation (13) and

$$\begin{aligned} V_r^{EB} &= \mathrm{Var}(\beta|\widehat{\beta} = \widehat{\beta}_r, \theta = \widehat{\theta}_r) \\ &= \widehat{\tau}^2 \cdot I - \left(\widehat{\tau}^2\right)^2 \cdot \left(\widehat{\tau}^2 \cdot I + \widehat{V}\right)^{-1} \\ &= \left(I + \tfrac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot \widehat{V}. \end{aligned}$$

4. Consider the mixture distribution

$$M\left(\beta|\widehat{\beta}, \widehat{V}\right) := \frac{1}{R}\sum_r N\left(\widehat{\beta}_r^{EB}, V_r^{EB}\right). \tag{35}$$

5. Obtain standard errors based on the variance of the mixture distribution, and confidence intervals for components of β using the appropriate quantiles of the mixture distribution $M\left(\beta|\widehat{\beta}, \widehat{V}\right)$.

Discussion Empirical Bayes confidence sets need to take into account two types of variation. This is best illustrated by first considering two invalid inference procedures, both of which ignore one of these two sources of variation. First, one might consider sets with the right coverage under the pseudo-posterior distribution, so that $P(\beta \in C|\widehat{\beta}, \theta = \widehat{\theta}) \geq 1 - \alpha$. These sets are similar to Bayesian credible sets. Such sets ignore the fact that θ had to be estimated and therefore might undercover in the empirical Bayes sense. Second, one might estimate the sampling variation of $\widehat{\beta}^{EB}$, for instance using the bootstrap. Confidence sets obtained in this way are similar to frequentist confidence sets, but ignore the fact that there is residual uncertainty about β conditional on $\widehat{\beta}$ and θ.

The situation is analogous to the forecasting of outcomes using a linear regression. Forecast uncertainty involves uncertainty about regression slopes (analogous to θ in our case, and captured by the bootstrap), and uncertainty about the outcome around its conditional expectation (analogous to the pseudo-posterior distribution in our setting). A correct inference procedure combines both aspects.

B Proofs

Proof of Proposition 1: Rearranging our expression for the empirical Bayes estimator, we can write

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

By assumption, $\widehat{\beta} \to^p \beta$. Our claim follows, by Slutsky's theorem, if we can show that $\frac{1}{\widehat{\tau}^2}\widehat{V} \to^p 0$, and $\widehat{\beta}^0 = O_p(1)$. Since $\widehat{V} \to^p 0$, this holds if $(\widehat{\beta}^0, \widehat{\tau}^2)$ converge in probability.

By the standard arguments for consistency of m-estimators (see for instance van der Vaart 2000, chapter 3), we get convergence of these hyper-parameters,

$$\begin{aligned}(\widehat{\beta}^0, \widehat{\tau}^2) \to^p & \operatorname*{argmin}_{b^0, t^2} \ \log\left(\det(\Sigma(t^2, 0))\right) + (\beta - b^0)' \cdot \Sigma(t^2, 0)^{-1} \cdot (\beta - b^0) \\ = & \operatorname*{argmin}_{b^0, t^2} \ J \cdot \log\left(t^2\right) + \frac{1}{t^2}\|\beta - b^0\|^2.\end{aligned}$$

The required conditions for applicability of this general consistency result are uniform consistency of the objective function and well-separatedness of the maximum. Both are easily verified given convergence of $\widehat{\beta}$ and $\widehat{V}$. $\square$

Proof of proposition 2: By Equation (25),

$$\begin{aligned}\widehat{y} - x \cdot \widehat{\beta} &= x \cdot \frac{1}{\widehat{\tau}^2}\widehat{V} \cdot \left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot (\widehat{\beta}^0 - \widehat{\beta}) \\ &= \left(x \cdot \frac{1}{\widehat{\tau}}\widehat{V}^{1/2}\right) \cdot \left(\widehat{\tau}\widehat{V}^{-1/2} + \frac{1}{\widehat{\tau}}\widehat{V}^{1/2}\right)^{-1/2} \cdot \left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1/2} \cdot (\widehat{\beta}^0 - \widehat{\beta}),\end{aligned}$$

and thus

$$\left|\widehat{y} - x \cdot \widehat{\beta}\right| \le \left\|x \cdot \frac{1}{\widehat{\tau}}\widehat{V}^{1/2}\right\| \cdot \left\|\left(\widehat{\tau}\widehat{V}^{-1/2} + \frac{1}{\widehat{\tau}}\widehat{V}^{1/2}\right)^{-1/2}\right\| \cdot \left\|\left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1/2} \cdot (\widehat{\beta}^0 - \widehat{\beta})\right\|.$$

By Equation (25), again,

$$\begin{aligned}\left\|\left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1/2} \cdot (\widehat{\beta}^0 - \widehat{\beta})\right\| &= \min_{\beta^0 \in B^0} \left\|\left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1/2} \cdot (\beta^0 - \widehat{\beta})\right\| \\ &\le \left\|\left(I + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1/2} \cdot \widehat{\beta}\right\| \le \left\|\widehat{\beta}\right\|,\end{aligned}$$

where the last inequality holds by positive definiteness of $\widehat{V}$, which also implies

$$\left\| \left(\widehat{\tau} \widehat{V}^{-1/2} + \frac{1}{\widehat{\tau}} \widehat{V}^{1/2} \right)^{-1/2} \right\| \leq 1.$$

The first inequality claimed in proposition 2 follows. The proof for $\widehat{y} - x \cdot \widehat{\beta}^0$ proceeds analogously. $\square$

The following simple lemma gives a sufficient condition which allows us to approximate the squared error for the estimator using the estimated $(\widehat{\beta}^0, \widehat{\tau}^2)$ by the mean squared error of the infeasible estimator using (β^0, τ^{*2}). This lemma is used in the proof of Theorem 1. Let $\widehat{c}^0$ and c^0 be such that $\widehat{\beta}^0 = M' \cdot \widehat{c}^0$ and $\beta^0 = M' \cdot c^0$, where M is as in Assumption 1.

Lemma 1
Suppose that Assumption 1 holds, that $(\widehat{c}^0, \widehat{\tau}^2) - (c^0, \tau^{*2}) \to^p 0$, and that

$$\sup_{(c,\tau^2) \in U} \left| SE(\widehat{\beta}^{EB}(M' \cdot c, \tau^2), \beta) - MSE(\widehat{\beta}^{EB}(M' \cdot c, \tau^2), \beta) \right| \to^p 0,$$

where U is some neighborhood of $\operatorname{plim}(c^0, \tau^{*2})$. *Then*

$$SE(\widehat{\beta}^{EB}, \beta) - MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) \to^p 0.$$

Proof of lemma 1:
This is immediate from

$$\begin{aligned} \left| SE(\widehat{\beta}^{EB}, \beta) - MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) \right| \leq & \left| SE(\widehat{\beta}^{EB}, \beta) - MSE(\widehat{\beta}^{EB}, \beta) \right| \\ & + \left| MSE(\widehat{\beta}^{EB}, \beta) - MSE(\widehat{\beta}^{EB}(\beta^0, \tau^{*2}), \beta) \right|, \end{aligned}$$

once we note that $MSE(\widehat{\beta}^{EB}(M' \cdot c, \tau^2)$ is uniformly continuous as a function of (c, τ^2) by boundedness of the sequence (β_j, v_j, M_j). $\square$

Proof of Theorem 1:
We need to show that the sufficient conditions of lemma 1 are satisfied. Convergence of

$(\widehat{c}^0, \widehat{\tau}^2)$ and (c^0, τ^{*2}) to the limiting pseudo-true parameters

$$\operatorname{plim}(c^0, \tau^{*2}) = \underset{\tau^2, c: S\cdot c \leq 0}{\operatorname{argmin}} \; E\left[\log(\tau^2 + v_j) + \frac{\left(\widehat{\beta}_j - M_j c^0\right)^2}{\tau^2 + v_j}\right]$$

follows from standard results on the consistency of maximum likelihood estimators, cf. van der Vaart (2000), chapters 5.2 and 5.5.

It remains to be shown that uniform convergence of

$$SE(\widehat{\beta}^{EB}(M'c, \tau^2), \beta) - MSE(\widehat{\beta}^{EB}(M'c, \tau^2), \beta) = (E_J - E)\left[\left(\widehat{\beta}_j^{EB}(M'c, \tau^2) - \beta_j\right)^2\right]$$

holds in a neighborhood U of $\operatorname{plim}(c^0, \tau^{*2})$, where E_J denotes the average over $j = 1, \ldots, J$. Such uniform convergence follows if we can show that the family of mappings

$$(\widehat{\beta}_j, \beta_j, v_j, M_j) \to \left(\widehat{\beta}_j^{EB}(M'c^0, \tau^2) - \beta_j\right)^2,$$

indexed by $(c^0, \tau^2) \in U$, is a Glivenko-Cantelli class, cf. van der Vaart (2000) chapter 19.2.

That this family of mappings is in fact a Glivenko-Cantelli class follows because it is a special case of example 19.8, p.272 in van der Vaart (2000):

(i) Continuity of $\left(\widehat{\beta}^{EB}(M'c, \tau^2) - \beta_j\right)^2$ in (c, τ^2) is immediate.

(ii) Compactness of the neighborhood U of $\operatorname{plim}(c^0, \tau^{*2})$ can be imposed without loss of generality.

(iii) It remains to be shown that an integrable envelope function exists on U. Suppose w.l.o.g. that the neighborhood U is of the form $[\underline{c}, \overline{c}] \times [\underline{t^2}, \overline{t^2}]$. Then $\left(\widehat{\beta}_j^{EB}(M'c, \tau^2) - \beta_j\right)^2$ always attains its maximum at one of the corners of U. This holds by monotonicity of $\widehat{\beta}_j^{EB}(M'c, \tau^2)$ in its arguments and the convexity of squaring. An envelope is therefore given by

$$\max_{(c, \tau^2) \in \{\underline{c}, \overline{c}\} \times \{\underline{t^2}, \overline{t^2}\}} \left(\widehat{\beta}_j^{EB}(M'c, \tau^2) - \beta_j\right)^2.$$

This envelope is integrable since we assumed finite second moments, given the form of $\widehat{\beta}_j^{EB}(M'c, \tau^2)$. $\square$

Proof of Theorem 2:

By Equation (29) and the definition of g, we have

$$g_j(\widehat{\beta}) = \widehat{\beta}_j^{EB} - \widehat{\beta}_j = \begin{cases} -\frac{1}{\widehat{\tau}^2+1} \cdot \widehat{\beta}_j & j = 1, \dots K \\ & \quad or\ j = K+1, \dots, L\ and\ \widehat{\beta}_j > 0, \\ 0 & else. \end{cases}$$

Noting that $(\widehat{\tau}^2 + 1)^2 = \frac{1}{J} \max(R, J) > 0$, the squared norm of g is given by

$$\|g(\widehat{\beta})\|^2 = \frac{R}{(\widehat{\tau}^2+1)^2} = \frac{J^2 R}{\max(R,J)^2} = \begin{cases} J^2/R & R > J \\ R & else. \end{cases}$$

The function g is almost differentiable. It has kink points at values of $\widehat{\beta}$ where either $R = J$ or $j = K+1, \dots, L$ and $\widehat{\beta}_j = 0$. The derivatives $\partial_j \widehat{\tau}^2$ and $\partial_j g_j$ away from these kink points are given by

$$\partial_j \widehat{\tau}^2 = \mathbf{1}(R > J) \cdot \begin{cases} 2\widehat{\beta}_j / J & j = 1, \dots K \\ & \quad or\ j = K+1, \dots, L\ and\ \widehat{\beta}_j > 0, \\ 0 & else. \end{cases}$$

and

$$\begin{aligned} \partial_j g_j(\widehat{\beta}) &= \begin{cases} -\frac{1}{\widehat{\tau}^2+1} + \frac{\widehat{\beta}_j}{(\widehat{\tau}^2+1)^2} \cdot \partial_j \widehat{\tau}^2 & j = 1, \dots K \\ & \quad or\ j = K+1, \dots, L\ and\ \widehat{\beta}_j > 0, \\ 0 & else \end{cases} \\ &= \left[-\frac{J}{\max(R,J)} + 2J \frac{\mathbf{1}(R>J)\widehat{\beta}_j^2}{\max(R,J)^2} \right] \cdot \mathbf{1}(j = 1, \dots K\ or\ j = K+1, \dots, L\ and\ \widehat{\beta}_j > 0) \\ &= \min(J/R, 1) \cdot \left[-1 + 2\frac{\mathbf{1}(R>J)\widehat{\beta}_j^2}{R} \right] \cdot \mathbf{1}(j = 1, \dots K\ or\ j = K+1, \dots, L\ and\ \widehat{\beta}_j > 0). \end{aligned}$$

Recalling our notation $J^* = K + \sum_{j=K+1}^{L} \mathbf{1}(\widehat{\beta}_j > 0)$, and summing across j, we get

$$\begin{aligned} \nabla \cdot g(\widehat{\beta}) = \sum_j \partial_j g_j(\widehat{\beta}) &= \min(J/R, 1) \cdot [-J^* + 2 \cdot \mathbf{1}(R > J)] \\ &= \begin{cases} \frac{J}{R} \cdot [2 - J^*] & R > J \\ -J^* & else. \end{cases} \end{aligned}$$

Collecting terms yields

$$\begin{aligned}\|g(\widehat{\beta})\|^2 + 2\nabla \cdot g(\widehat{\beta}) &= \min(R, J^2/R) + 2[-J^* + 2] \cdot \min(J/R, 1) \\ &= \begin{cases} \frac{J}{R} \cdot [J + 4 - 2J^*] & R > J \\ R - 2J^* & else. \end{cases}\end{aligned}$$

The claim now follows.□

Table 1: Estimated effects of labor supply on wage inequality, panel of US states

Supply of type	Struct	Unrest							EB						
		2-1	3-1	4-1	5-1	6-1	7-1	8-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1
1		-0.022	0.063	0.056	0.124	0.066	0.146	0.050	-0.009	0.063	0.062	0.127	0.081	0.155	0.071
		(0.028)	(0.034)	(0.027)	(0.026)	(0.029)	(0.033)	(0.036)	(0.028)	(0.035)	(0.030)	(0.028)	(0.032)	(0.036)	(0.039)
2		-0.002	-0.085	-0.080	-0.098	-0.019	-0.100	0.024	-0.033	-0.088	-0.095	-0.095	-0.033	-0.110	0.020
		(0.049)	(0.049)	(0.045)	(0.043)	(0.054)	(0.052)	(0.058)	(0.051)	(0.050)	(0.046)	(0.041)	(0.050)	(0.053)	(0.054)
3		-0.182	-0.320	-0.398	-0.306	-0.369	-0.238	-0.149	-0.103	-0.199	-0.257	-0.182	-0.226	-0.152	-0.087
		(0.072)	(0.071)	(0.068)	(0.077)	(0.088)	(0.072)	(0.092)	(0.064)	(0.061)	(0.061)	(0.065)	(0.070)	(0.060)	(0.071)
4		0.023	-0.044	-0.107	-0.115	-0.175	-0.107	-0.167	0.024	-0.069	-0.108	-0.118	-0.154	-0.100	-0.117
		(0.047)	(0.048)	(0.043)	(0.049)	(0.057)	(0.048)	(0.064)	(0.042)	(0.045)	(0.043)	(0.047)	(0.053)	(0.045)	(0.055)
5		-0.105	-0.032	-0.074	0.006	-0.048	-0.217	-0.162	-0.050	-0.065	-0.095	-0.056	-0.097	-0.192	-0.101
		(0.072)	(0.080)	(0.086)	(0.084)	(0.104)	(0.080)	(0.112)	(0.057)	(0.059)	(0.058)	(0.057)	(0.068)	(0.063)	(0.069)
6		-0.011	-0.001	-0.028	-0.012	0.087	-0.183	-0.088	0.040	-0.044	-0.043	-0.056	0.004	-0.160	-0.070
		(0.086)	(0.085)	(0.082)	(0.083)	(0.105)	(0.080)	(0.131)	(0.057)	(0.055)	(0.054)	(0.056)	(0.064)	(0.057)	(0.069)
7		-0.206	0.093	-0.003	0.127	0.207	-0.082	-0.191	-0.065	0.045	-0.013	0.081	0.134	-0.051	-0.038
		(0.126)	(0.113)	(0.117)	(0.114)	(0.127)	(0.120)	(0.162)	(0.071)	(0.067)	(0.070)	(0.067)	(0.071)	(0.069)	(0.077)
8		-0.119	0.077	0.052	0.215	0.159	0.108	-0.039	-0.025	0.012	0.001	0.139	0.119	0.096	0.023
		(0.074)	(0.081)	(0.082)	(0.081)	(0.092)	(0.080)	(0.111)	(0.059)	(0.062)	(0.061)	(0.057)	(0.062)	(0.056)	(0.066)
$\widehat{\beta}_0$	-0.061	0.051							0.052						
	(0.058)	(0.065)							(0.042)						
$\widehat{\tau}^2$									0.0050						
Time FE	YES	YES							YES						
State FE	YES	YES							YES						
N	306	306							306						

Notes: This table shows three alternative estimates of labor demand using (i) the structural model based on the 2-type CES production function, (ii) unrestricted OLS regression using 8-types of the model nesting 2-type CES, and (iii) empirical Bayes estimation of the same model. Regressions control for time and state fixed effects. Standard errors are clustered across types of workers. Standard errors for empirical Bayes are calculated as discussed in Appendix A. For details, see Section 3.1.3.

References

Abadie, A. and Kasy, M. (2017). The risk of machine learning. *Working Paper, Harvard University.*

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of labor economics*, 4:1043–1171.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in US wage inequality: Revising the revisionists. *The Review of Economics and Statistics*, 90(2):300–323.

Blundell, R., Horowitz, J., and Parey, M. (2017). Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *Review of Economics and Statistics*, 99(2):291–304.

Borjas, G. J., Grogger, J., and Hanson, G. H. (2012). Comment: On estimating elasticities of substition. *Journal of the European Economic Association*, 10(1):198–210.

Card, D. (2009). Immigration and inequality. *The American Economic Review*, 99(2):1–21.

Carlin, B. P. and Gelfand, A. E. (1990). Approaches for empirical Bayes confidence intervals. *Journal of the American Statistical Association*, 85(409):105–114.

Casella, G. and Berger, R. (2001). *Statistical inference.* Duxbury Press.

Casella, G., Hwang, J. G., et al. (2012). Shrinkage confidence procedures. *Statistical Science*, 27(1):51–60.

Del Negro, M. and Schorfheide, F. (2004). Priors from general equilibrium models for VARS. *International Economic Review*, 45(2):643–673.

Del Negro, M., Schorfheide, F., Smets, F., and Wouters, R. (2007). On the fit of new Keynesian models. *Journal of Business & Economic Statistics*, 25(2):123–143.

Dette, H., Hoderlein, S., and Neumeyer, N. (2016). Testing multivariate economic restrictions using quantiles: the example of slutsky negative semidefiniteness. *Journal of Econometrics*, 191(1):129–144.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Fortin, N. M. and Lemieux, T. (1997). Institutional changes and rising wage inequality: Is there a linkage? *The Journal of Economic Perspectives*, 11(2):pp. 75–96.

Graham, B. S. and Hirano, K. (2011). Robustness to parametric assumptions in missing data models. *American Economic Review*, 101(3):538–43.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Imbens, G. W. and Newey, W. (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, 77:1481–1512.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Jensen, M. C., Black, F., and Scholes, M. S. (1972). The capital asset pricing model: Some empirical tests. *Studies in the theory of capital markets*.

Koenker, R. (2005). Quantile Regression. `http://www.econ.uiuc.edu/~roger/research/rq/rq.html`.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82(399):739–750.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59.

Mas-Colell, A., Whinston, M., and Green, J. (1995). *Microeconomic theory*. Oxford University Press.

McFadden, D. L. (2005). Revealed stochastic preference: a synthesis. *Economic Theory*, 26(2):245–264.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):pp. 47–55.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

van der Vaart, A. (2000). *Asymptotic statistics*. Cambridge University Press.

Xie, X., Kou, S., and Brown, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association*, 107(500):1465–1479.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods: invited paper. *The Annals of Statistics*, 31(2):379–390.