How to use economic theory to improve estimators, with an application to labor demand and wage inequality

Fessler, Pirmin∗ Kasy, Maximilian†
March 12, 2016

Abstract

Economic theory, when it has empirical content, provides testable restrictions on empirically identified objects. These empirical objects might be estimated in an unrestricted way, leading to estimates of potentially large variance, or subject to the theoretical restrictions, leading to estimates of lower variance that are potentially biased, inconsistent, and non-robust.

We propose an alternative approach, based on the empirical Bayes paradigm, which avoids both large variance and large bias, and which performs particularly well when the theory is approximately correct. In our key theoretical contribution, we characterize the geometry and the risk-function (mean squared error) of the proposed estimator. Simulations confirm the good performance of our estimator relative to both unrestricted and structural estimation.

We apply our approach to models of labor demand which are used to analyze to what extent changes in the distribution of wages can be explained by changes in labor supply (due to demographic change, migration, or expanded access to education), as opposed to other factors (technical and institutional change). Taking our estimator to well known data from the United States (CPS and ACS), we find (i) that the composition of the workforce matters for relative wages beyond the effects captured by a 2-type model, but (ii) the effect of labor supply on wages overall is rather small.

Keywords: Empirical Bayes estimation, shrinkage, inequality, labor demand

JEL codes: C11, C52, D31, J23, J31

∗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.

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

Wage inequality has increased significantly in most industrial countries since the 1980s; see for instance Autor et al. (2008) for the case of the United States and Gottschalk and Smeeding (2000) for evidence on incomes in other rich countries. Various explanations have been offered for this increase in wage inequality, including the decline of minimum wages and unions, technical change, demographic change, migration, and international trade. Disentangling the relative contribution of these factors is important for assessing potential policy responses.

There is considerable disagreement regarding the contribution of these various factors; see for instance Autor et al. (2008) regarding technical change, and Card (2009) regarding migration. We argue that part of this disagreement has methodological roots. One of the workhorse methods of the literature on wage inequality is the estimation of models for labor demand. The models used are derived from a parametric specification of an aggregate production function. Qualitative conclusions, predictions and counterfactual analyses tend to be quite sensitive to specific choices of functional form for these production functions, as demonstrated by Card (2009) in his review of the literature on the impact of migration.

An alternative to the imposition of restrictions implied by structural models of labor demand would be the estimation of an unrestricted model of labor demand, allowing for a large number of types and unrestricted own- and cross-elasticities. The problem with such unrestricted models is that they require estimation of a large number of parameters using a potentially small number of observations, leading to estimates of high variance and possibly to lack of identification.

We propose to instead use an empirical Bayes approach for the construction of estimators avoiding the problems of both structural and unrestricted estimation. The empirical Bayes approach models parameters, such as own- and cross-elasticities, to be themselves drawn from some random distribution. This distribution is governed by hyper-parameters that have to be estimated. We model the elasticities (in a setting with many types of workers) as being equal to (i) the elasticities implied by a structural model plus (ii) random noise of unknown variance. This variance has to be estimated. If this variance is estimated to be zero, estimation of elasticities proceeds

---

1 We thank Alberto Abadie, Gary Chamberlain, Ellora Derenoncourt, Kei Hirano, Susanne Kimm, and José Montiel Olea, 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.
as under the structural model. If this variance is estimated to be infinite, estimation of elasticities proceeds as under the unrestricted model. In general, estimates will interpolate between these two extremes in an optimal, data-dependent way.

There are a number of advantages to our empirical Bayes approach: (i) The resulting elasticity estimates are consistent, i.e., converge to the truth as samples get large, for any parameter values, in contrast to structural estimation. (ii) The variance and mean squared error of the estimates is smaller than under unrestricted estimation. Simulations and asymptotic approximations suggest this is the case uniformly over most of the parameter space. (iii) In contrast to a fully Bayesian approach, no tuning parameters (features of the prior) have to be picked by the researcher. (iv) Counterfactual predictions and forecasts are driven by the data whenever the latter are informative. (v) The empirical Bayes approach avoids the irregularities of pre-testing (cf. Leeb and Pötscher, 2005) which are associated with testing structural models and imposing them if they are not rejected.

After proposing our general approach for using economic theory to construct empirical Bayes estimators in Section 2, we provide novel insights on the behavior of empirical Bayes estimators in Section 3. We show consistency, characterize the mapping from unrestricted estimates to empirical Bayes estimates, and provide visual representations. A key theoretical contribution of this paper is Theorem 1 in Section 3.2, which provides an explicit formula for an approximation of the risk function (mean squared error) of empirical Bayes. This asymptotic approximation is valid whenever the tuning parameter is estimated with small variance relative to the parameters of interest. Our characterization of risk provides formal backing for the assertions on the favorable properties of our estimator, relative to both structural and unrestricted estimation. In contrast to classic derivations of risk for James-Stein shrinkage, Theorem 1 is only valid under this approximation, but it extends classic results to the practically relevant case where neither normality nor (more importantly) homoskedasticity are imposed. Monte Carlo simulations confirm the validity of these approximations for realistic specifications.

In section 5 we apply our approach to study the evolution of wage inequality in the United States, using data from the Current Population Survey (CPS) and the American Community Survey (ACS). Across specifications, we generally find negative but small inverse elasticities of substitution, that is, small effects of relative labor supply on relative wages. The estimated elasticities are significantly smaller than esti-

\footnote{It is possible to construct counterexamples, however, see Section 3.2}
mates discussed in the literature, and not significantly different from 0. Specifications derived from a 2-type constant elasticity of substitution (CES) production function do not appear to provide a particularly good fit to the data – the composition of the workforce affects wages beyond the effect captured by a 2-type model. Overall, the explanatory power of changes in labor supply for changes in relative wages is rather small, however.

This paper is structured as follows: Section 1.1 provides a brief literature review. We introduce our approach for constructing estimators based on restrictions implied by economic theory in Section 2. The proposed estimators shrink a preliminary unrestricted estimator towards a structural model, to an extent that depends on how well the latter appears to fit the data. We propose a corresponding inference procedure in Section 2.4 and discuss the conceptual and methodological motivation for our approach in Section 2.5. Section 3 shows consistency, explores the geometry of our proposed estimator and provides a theoretical characterization of its risk properties.

Section 4 reviews structural estimation of labor demand based on CES production functions, and adapts our empirical Bayes approach to this setting. Section 5 applies our approach to data from the United States. Section 6 evaluates our estimation and inference procedure using a range of Monte Carlo simulations. Section 7 concludes. Appendix A contains all proofs; figures and tables can be found in Appendix B.

1.1 Related literature

This paper mainly builds on two distinct literatures: the literature on shrinkage and empirical Bayes estimation in statistics, and the literature on labor supply, labor demand, and wage inequality in economics. Both literatures are very large so that it is impossible to do full justice to either; we shall only discuss a few key references.

The statistical literature on empirical Bayes methods and shrinkage 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 the striking result that the conventional estimator for the mean of a multivariate normal vector with unit variance is inadmissible. It is dominated in terms of mean squared error by empirical Bayes estimators. This is true whenever the dimension of the vector is at least 3.

Empirical Bayes approaches were developed further by later contributions such
as Efron and Morris (1973). Morris (1983) was the first to discuss the parametric version of the empirical Bayes approach. Inference in empirical Bayes settings was discussed by Laird and Louis (1987) and Carlin and Gelfand (1990), among others. A good introduction to empirical Bayes estimation can be found in chapter 1 of Efron (2010), another review is to be found in Zhang (2003). In Section 3 we provide a theoretical characterization of the risk properties of our empirical Bayes procedure. This characterization relies on arguments related to those invoked by Xie et al. (2012).

The recently published Hansen (2016) complements our analysis. Rather than empirical Bayes estimation, as in the present paper, Hansen (2016) considers James-Stein-like shrinkage of all components of a preliminary estimator towards a restricted parameter space by the same factor, independent of their variance. Different shrinkage depending on estimation variance is key to our argument about combining theory and evidence, cf. Section 3 below. Hansen (2016) allows for non-linear settings and uses local asymptotics to retrieve linear settings like the one we shall consider. 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.

The labor literature relevant for us encompasses various sub-literatures, concerned with different factors potentially affecting wage inequality (in particular migration and technical change), but united by a common method based on estimating the parameters of a model for labor demand. The models used are justified by constant elasticity of substitution (CES) production functions or generalizations thereof.

The literature on the impact of migration on native wage inequality was pioneered by Card (1990), who studied the “natural experiment” of a large increase of the Cuban population in Miami, and did not find much of an effect on native wages or employment. Card (2001) studied the same question, but took a more structural approach based on production-function estimation, considering variation in immigration across metropolitan areas as predicted by a Bartik-type instrument. The approach based on cross-city comparisons has been criticized by Borjas et al. (1996), among others, who argue for considering the national economy rather than local labor markets, and who do find some effects of immigration on the wages of
native high-school dropouts. [Card (2009)] reviews this debate and argues that the divergent findings might be driven by different choices of functional form (number of groups in the CES specification) rather than the local versus national distinction. This lack of robustness to functional form choices motivates the methods proposed in this paper. Our methods aim to avoid such non-robustness. [Borjas et al. (2012)] similarly demonstrate the sensitivity of conclusions to details of the specification imposed. [D’Amuri and Peri (2015)], studying European evidence, even find a positive effect of migration on native wages, mediated through a process of job upgrading.

Another, related, literature studies the impact of technical change on wage inequality, and on the college premium in particular. [Autor et al. (1998)] argue that technical change leads to a continuous rise of the relative demand for workers with college degrees, a rise which was offset partially by an increase of relative supply in periods of expansion of college enrollment. They interpret the residual of a CES-regression specification as reflecting technical change. [Autor et al. (2008)] review and update this argument. [Goldin and Katz (2009)] provide an extensive historical analysis of wage inequality in the United States and how it was affected by changes in education. More recently, [Autor and Dorn (2013)] argue that technical change in recent decades has created substitutes for middle income and routine clerical work, while leaving unaffected low-wage service jobs and increasing the wages of highly educated workers. This lead to a polarization of the wage distribution. An extensive review and discussion is provided by [Acemoglu and Autor (2011)].

2 Using economic theory to construct empirical Bayes estimators

In this section we consider a simplified setting to introduce our approach. We assume that there is a preliminary estimator \( \hat{\beta} \) of the parameter vector of interest, \( \beta \), where the preliminary estimator does not make use of restrictions implied by economic theory. Economic theory is then supposed to provide overidentifying restrictions on \( \beta \); for simplicity of exposition we focus on linear restrictions. We use these restrictions to construct an estimator \( \hat{\beta}^{EB} \) designed to outperform \( \hat{\beta} \) if the restrictions are approximately true, and to perform no worse than \( \hat{\beta} \) if they are not.

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 proposes a way to construct
empirical Bayes confidence sets. Section 2.5 discusses the conceptual and methodological motivation for our approach, and the role of economic theory in empirical research.

2.1 Setup

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

\[
\hat{\beta} \sim N(\beta, V),
\]  

(1)

of \( \beta \), with consistently estimable variance \( V \). This assumption implies that \( \beta \) 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; when studying the properties of these estimators we will not use normality. In the context of our application, asymptotically normal estimators \( \hat{\beta} \) will 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 simplest case, where a theoretical model implies that

\[
\beta = \beta_0 \cdot M
\]

(2)

for some known vector \( M \) and unknown scalar \( \beta_0 \). All of our discussion immediately extends to settings with linear or affine restrictions of arbitrary dimension on \( \beta \). The general argument also extends to arbitrary smooth restrictions (imposing that \( \beta \) lies in some smooth manifold), although some of the math has to be adapted for that case.

2.2 General empirical Bayes estimation

Two approaches to estimation are commonly used in settings of this kind, one imposing a lot of restrictions based on some structural model, and one leaving the model rather unrestricted. Estimation based on the structural 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 leads to estimates of large variance, but is (in principle) unbiased and consistent.

There is a paradigm in statistics, called empirical Bayes estimation, which provides a middle ground between these two approaches, and which 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:

\[
Y \mid \eta \sim f(Y \mid \eta) \tag{3}
\]

\[
\eta \sim \pi(\eta \mid \theta) \tag{4}
\]

where \( Y \) are the observed data, both \( f \) and \( \pi \) describe parametric families of distributions, and where usually \( \text{dim}(\theta) \leq \text{dim}(\eta) - 2 \). Equation (3) describes the unrestricted model for the distribution of the data given the full set of parameters \( \eta \). Equation (4) describes a family of “prior distributions” for \( \eta \), indexed by the hyper-parameters \( \theta \).

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

\[
Y \mid \theta \sim g(Y \mid \theta) := \int f(Y \mid \eta)\pi(\eta \mid \theta)\,d\eta. \tag{5}
\]

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 \( \theta \) is obtained by maximum likelihood,

\[
\hat{\theta} = \arg\max_{\theta} g(Y \mid \theta). \tag{6}
\]

Other estimators for \( \theta \) are conceivable and commonly used, as well. In the second step of empirical Bayes estimation, \( \eta \) is estimated as the “posterior expectation”\(^3\) of \( \eta \) given \( Y \) and \( \theta \), substituting the estimate \( \hat{\theta} \) for the hyper-parameter \( \theta \),

\[
\hat{\eta} = E[\eta \mid Y, \theta = \hat{\theta}]. \tag{7}
\]

The general empirical Bayes approach includes fully Bayesian estimation as a

\(^3\)The quotation marks reflect the fact that this would only be a posterior expectation in the strict sense if \( \hat{\theta} \) had been chosen independently of the data, rather than estimated.
special case, if the family of priors $\pi$ 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 $\theta$ is restricted to parameter values allowed by the structural model.

We can think of such support restrictions as a dogmatic imposition of 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 $\hat{\beta}$, which might be obtained using OLS, IV, or some other method. This unrestricted estimator is then mapped to an empirical Bayes estimator $\hat{\beta}^{EB}$. To construct a family of priors for $\beta$, we assume that $\beta$ is equal to a vector of parameters consistent with the structural model plus some noise of unknown variance.

**Modeling $\hat{\beta}$**

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

$$\hat{\beta}|\eta \sim N(\beta, V).$$  \hspace{1cm} (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. We further assume that we have a consistent estimator $\hat{V}$ of $V$, i.e.,

$$\hat{V} \cdot V^{-1} \rightarrow^p I,$$

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

**Prior distributions**

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

$$\beta = (\beta_j, j') = \beta_0 \cdot M + \zeta$$

$$\zeta_{j, j'} \sim^{iid} N(0, \tau^2);$$
The term $\beta_0 \cdot M$ corresponds to a set of coefficients satisfying the structural model. The term $\zeta$ is equal to a random $J$-vector with variance

$$\text{Var}(\zeta) = \tau^2 \cdot I.$$ 

If we were to set $\tau^2 = 0$, the empirical Bayes approach would reduce to estimation of the structural model. If we let $\tau^2$ go to infinity we effectively recover the unrestricted model. We consider $\tau^2$ to be a parameter to be estimated, however, which measures how well the given structural model fits the data.

Summarizing our model in terms of the general notation introduced in Section 2.2 with $\hat{\beta}$ taking the place of $Y$, we get:

$$\eta = (\beta, V)$$
$$\theta = (\beta_0, \tau^2, V)$$
$$\hat{\beta}|\eta \sim N(\beta, V)$$
$$\beta|\theta \sim N(\beta_0 \cdot M, \tau^2 \cdot I).$$ \hspace{1cm} (9)

**Solving for the empirical Bayes estimator**

In order to obtain estimators of $\beta_0$ and $\tau^2$, consider the marginal distribution of $\hat{\beta}$ given $\theta$. This marginal distribution is normal,

$$\hat{\beta}|\theta \sim N(\beta_0 \cdot M, \Sigma(\tau^2, V)), \hspace{1cm} (10)$$

where (leaving the conditioning on $\theta$ implicit)

$$\Sigma(\tau^2, V) = \text{Var} \left( \hat{\beta} \right) = \text{Var} \left( E \left[ \hat{\beta}|\eta \right] \right) + E \left[ \text{Var} \left( \hat{\beta}|\eta \right) \right]$$
$$\quad = \tau^2 \cdot I + V.$$

Substituting the consistent estimator $\hat{V}$ for $V$, we obtain the empirical Bayes estimators of $\beta_0$ and $\tau^2$ as solution to the maximum (marginal) likelihood problem

$$(\hat{\beta}_0, \hat{\tau}^2) = \arg\min_{b_0, t^2} \log \left( \det(\Sigma(t^2, \hat{V})) \right)$$
$$\quad + (\hat{\beta} - b_0 \cdot M)' \cdot \Sigma(t^2, \hat{V})^{-1} \cdot (\hat{\beta} - b_0 \cdot M).$$ \hspace{1cm} (11)
We can simplify this optimization problem by concentrating out $b_0$: given $t^2$, the optimal $b_0$ is equal to

$$\hat{\beta}_0 = (M \cdot \Sigma(t^2, \hat{V})^{-1} \cdot M')^{-1} \cdot M \cdot \Sigma(t^2, \hat{V})^{-1} \cdot \hat{\beta}. \quad (12)$$

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

Given the unrestricted estimates $\hat{\beta}$, as well as the estimates $\hat{\beta}_0$ and $\hat{\tau}^2$, we can finally obtain the “posterior expectation” of $\beta$ as

$$\hat{\beta}^{EB} = \hat{\beta}_0 \cdot M + \left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1} \cdot (\hat{\beta} - \hat{\beta}_0 \cdot M). \quad (13)$$

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

Discussion

It is instructive to relate the proposed empirical Bayes procedure to structural estimation. The empirical Bayes estimator $\hat{\beta}^{EB}$ of $\beta$ is not given by $\hat{\beta}_0 \cdot M$. Instead we can think of it as an intermediate point between $\hat{\beta}_0 \cdot M$ and the unrestricted estimator $\hat{\beta}$. The relative weights of these two are determined by the matrices $\hat{\tau}^2 \cdot I$ and $\hat{V}$. When $\hat{\tau}^2$ is close to 0, we get $\hat{\beta}^{EB} \approx \hat{\beta}_0 \cdot M$. When $\hat{\tau}^2$ is large, we get $\hat{\beta}^{EB} \approx \hat{\beta}$, cf. equation (13).

Our construction of a family of priors thus implies the following: When the structural model appears to describe the data well, then our estimate of $\beta$ will be close to what is prescribed by the structural model. When the structural 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, optimal, and smooth way, in contrast to the discontinuity of pre-testing procedures.

The estimator $\hat{\beta}_0 \cdot M$ is very similar to the structural estimator of $\beta$ obtained by directly imposing the theoretical constraints when estimating $\beta$; in both cases we are considering an orthogonal projection of the unrestricted estimator $\hat{\beta}$ onto the subspace of multiples of $M$. The projection is, in general, with respect to different norms, however. When the structural estimator is obtained by least squares regression of $Y$ on $X$ subject to linear constraints, the projection is with respect to the norm

$$\|b\|_\beta := (b' \cdot \text{Var}(X) \cdot b)^{1/2}.$$
In the context of our empirical Bayes approach, the projection is with respect to the norm
\[ \|b\|_{\beta,EB} = \left( b' \cdot \Sigma(t^2, \hat{V})^{-1} \cdot b \right)^{1/2}. \]

The two objective functions coincide (up to a multiplicative constant) if and only if (i) \( t^2 = 0 \), so that the structural model is assumed to be true, and (ii) \( \hat{V} \) is estimated assuming homoskedasticity.

Our approach is based upon directly modeling the distribution of the unrestricted estimator \( \hat{\beta} \). If \( \hat{\beta} \) are 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 residuals do not carry additional information about \( \beta \), our approach does not waste any information; this is true, in particular, for a standard parametric linear/normal model.

### 2.4 Inference

Inference in our setting is easily implemented, though conceptually somewhat subtle. We shall construct empirical Bayes confidence regions \( C \) for \( \beta \). Such confidence regions are required to satisfy
\[ P(\beta \in C|\theta) \geq 1 - \alpha, \tag{14} \]
and were first proposed by Morris (1983) and analyzed further by Laird and Louis (1987) and Carlin and Gelfand (1990). Definition (14) 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 (14) is slightly weaker than the requirement of frequentist coverage, \( P(\beta \in C|\eta) \geq 1 - \alpha \).

Our inference procedure builds on ideas of Laird and Louis (1987). We use standard frequentist inference to capture sampling variation of the estimates \( \hat{\beta}^{EB} \), and posterior inference to capture uncertainty about \( \beta \) given these estimates. The proposed procedure obtains a predictive distribution for \( \beta \) which is similar to a posterior distribution of the form
\[ P\left( \beta|\hat{\beta}, \hat{V} \right) = \int P\left( \beta|\hat{\beta}, \hat{V}, \theta \right) P\left( \theta|\hat{\beta}, \hat{V} \right) d\theta, \]
but replaces the posterior for the hyperparameter $\theta$ by the sampling distribution $Q_R$ for $\hat{\theta}$ obtained using standard frequentist inference, thus obtaining a mixture distribution

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

(15)

Our inference procedure can be summarized as follows:

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

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

3. Calculate the posterior mean $\hat{\beta}_{r}^{EB}$ and variance $V_r^{EB}$ for $\beta$ conditional on $\hat{\beta}_r$ and $\hat{\theta}_r$, using equation (13) and

$$
V_r^{EB} = \text{Var}(\beta | \hat{\beta} = \hat{\beta}_r, \theta = \hat{\theta}_r)
\quad = \hat{\tau}^2 \cdot I - (\hat{\tau}^2)^2 \cdot \left(\hat{\tau}^2 \cdot I + \hat{V}\right)^{-1}
\quad = \left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1} \cdot \hat{V}.
$$

4. Consider the mixture distribution

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

(16)

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

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 $P(\beta \in C | \hat{\beta}, \theta = \hat{\theta}) \geq 1 - \alpha$. Such sets are similar to Bayesian credible sets. Such sets ignore the fact that $\theta$ had to be estimated and therefore might undercover in the empirical Bayes sense. Second, one might estimate the sampling variation of $\hat{\beta}^{EB}$, 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 $\beta$ conditional on $\hat{\beta}$ and $\theta$.

The situation is analogous to the forecasting of outcomes using a linear regression. Forecast uncertainty involves uncertainty about regression slopes (analogous to $\theta$ 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.

2.5 Some conceptual remarks on the use of economic theory in empirical research

Before we proceed to analyze the properties of our proposed estimator, let us take a step back and discuss the conceptual motivation of our approach. The central underlying question is the following: How should economic theory be used in empirical research? This question is made particularly salient by settings such as the one discussed in Sections 4 and 5 below, where key qualitative conclusions are sensitive to the specifics of the theoretical assumptions imposed.

The standard prescriptions of the positivist paradigm of empirical research, dominant in economics, can be summarized as follows: Develop economic theories which have empirical content, that is, theories which yield falsifiable predictions (equivalently, overidentified models). Falsifiability requires theories to be universally true over some suitable domain (i.e., quantified by the universal quantifier “for all”). Confront your theories with data, using statistical tests. Maintain your theories as correct as long as you have not found evidence which falsifies them. Once you statistically reject a theory, discard it, and try to construct a new one consistent with all available evidence.

This paradigm is modeled on a highly stylized description of research in physics. We argue that this paradigm does not provide a very accurate or useful description of empirical research in economics. The main reason is that there are no theories in economics which are even approximately universally true over some domain. Consider the following theories, which are among the most general developed by economists: utility maximization, exponential discounting, maximization of expected utility under risk, Nash equilibrium in strategic situations, or Walrasian equilibrium of market participants acting as price takers. All of these can be and have been rejected numerous times in numerous contexts. And now consider the type of models imposed in much empirical research in economics: maximization of utility with additive EV1
errors, aggregate production following a CES production function with 3 inputs, etc. We know that such models have no claim at generalizable truth.

What to do then, if this premise is accepted? A first option would be to ignore this point, to keep nominally following the positivist ideal, and to argue that theories do not actually have to be true. This point of view is famously associated with Friedman (1953), and his arguments against realism. We believe that such a position amounts to an insulation of potentially ideological viewpoints from empirical evidence. A second option would be to forget about economic theory and to do solid data-driven research, using mostly harmless econometrics. We believe that this yields useful research. We also believe, however, that the insights of existing economic theory can and should be usefully employed, even absent claims for universal truth.

The estimators proposed in this paper are an attempt at providing a middle ground between imposing wrong theories and completely ignoring economic theory. We use economic theory to construct estimators which perform particularly well when the theory is approximately true (see Theorem 1 below), similarly to structural estimators imposing the theory. At the same time we avoid the inconsistency and non-robustness which follows from imposing wrong theories, similarly to estimators not using the theory. Our proposed estimators shrink estimates “toward the theory.” This could be done in a fully Bayesian way, by constructing priors putting low weight on parameter values deviating a lot from the theory. One disadvantage of such a fully Bayesian approach is that it requires applied researchers to pick tuning parameters such as our $\tau^2$, something empirical researchers are understandably reluctant to do. Another disadvantage is that fully Bayesian estimators in general perform poorly for parameter values in regions receiving low prior weight. In contrast, the empirical Bayes approach can achieve uniform improvements of risk (mean squared error) relative to unrestricted estimation, see again Theorem 1. This result is a generalization of the famous James-Stein result, demonstrating that unrestricted estimators are inadmissible for dimensions $J$ larger than two. Our empirical Bayes estimator can be thought of as a generalization of the James-Stein shrinkage estimator, shrinking toward a theory rather than toward an arbitrarily chosen origin.

Another alternative to our approach that might be seen as closely related is pre-testing: Test the theory, if you do not reject it, use structural estimation, otherwise use the unrestricted estimator. As is well known, and discussed for instance by Leeb and Pötscher (2005), such pre-testing procedures perform very poorly for intermediate parameter values, in sharp contrast to the dominance achieved by our approach.
3 Properties of the empirical Bayes estimator and its risk function

We shall next analyze various properties of the proposed empirical Bayes estimator. In Section 3.1, we show consistency of the estimator and show how counterfactual predictions combine theory and available evidence in a data-driven, intuitive way. Section 3.2 characterizes the risk properties of \( \hat{\beta}^{EB} \). Theorem 1 in particular, generalizes classic characterizations of the risk function of James-Stein shrinkage, using an asymptotic approximation to the behavior of \( \tau^2 \). This theorem forms the theoretical basis of our claims of dominance relative to unrestricted estimation. Section 3.3 explores the geometry of the mapping from the preliminary, unrestricted estimator \( \hat{\beta}^* \) to the empirical Bayes estimator \( \hat{\beta}^{EB} \).

3.1 Consistency and data-driven predictions

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

**Proposition 1** (Consistency)

Suppose that \( \hat{\beta} \to^p \beta \) and that \( \hat{V} \to^p 0 \). Let \( \hat{\beta}^{EB} \) be the empirical Bayes estimator of \( \beta \) discussed in Section 2.3. Then

\[
\hat{\beta}^{EB} \to^p \beta
\]

as sample size \( n \) goes to infinity.

The proof of this proposition can be found in appendix A. The proof of consistency relies on the fact that \( \hat{\beta}^{EB} \approx \hat{\beta} \) if \( \hat{V} \approx 0 \).

The formula for \( \hat{\beta}^{EB} \) given in equation (13) shows that the empirical Bayes estimator interpolates between the unrestricted estimator \( \hat{\beta} \) and the structural estimator \( \hat{\beta}^* = \hat{\beta}^0 \cdot M \). Suppose we are interested in making a prediction of the form \( \hat{y} = x \cdot \hat{\beta}^{EB} \). Heuristically, we would like our prediction to be based on the data alone (neglecting the structural model) whenever the data by themselves do allow 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 structural model. The
following proposition shows that this is exactly how the empirical Bayes estimator behaves.

**Proposition 2 (Counterfactual predictions)**

Let $\hat{\beta}^{EB}$ be given as in equation (13). Consider the prediction at $x$, $\hat{y} = x \cdot \hat{\beta}^{EB}$, and assume that $\hat{V}$ is non-singular. Then

\[
\begin{align*}
|\hat{y} - x \cdot \hat{\beta}| & \leq \frac{\sqrt{x \hat{V} x}}{\hat{\tau}} \cdot \|\hat{\beta}\| \\
|\hat{y} - x \cdot \hat{\beta}^s| & \leq \hat{\tau} \cdot \sqrt{x \hat{V}^{-1} x} \cdot \|\hat{\beta}\|
\end{align*}
\]

The first inequality of proposition 2 tells us that empirical Bayes predictions are close to unrestricted predictions whenever the standard deviation of the latter, $\sqrt{x \hat{V} x}$, is small relative to the measure of model fit $\hat{\tau}$. 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 (13),

\[
\hat{\beta}^{EB} = \hat{\beta} + \hat{V} \cdot \left( \hat{\tau}^2 \cdot I + \hat{V} \right)^{-1} \cdot (\hat{\beta}_0 \cdot M - \hat{\beta}).
\]

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

\[
x \cdot \hat{\beta}^{EB} = x \cdot \left[ \hat{\beta} + \hat{V} \cdot \left( \hat{\tau}^2 \cdot I + \hat{V} \right)^{-1} \cdot (\hat{\beta}_0 \cdot M - \hat{\beta}) \right] \approx x \cdot \hat{\beta}.
\]

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

### 3.2 The risk function of our estimator

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 the 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. The key argument in our characterization is that variability of $\hat{\tau}^2$ can be neglected for large $J$ when calculating the MSE. After rewriting the estimator in a canonical form, 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 $\tau^2$.

**Canonical form**

So far, we have allowed for arbitrary correlation of the unrestricted estimates $\hat{\beta}_j$ across $j$. For the rest of this section we simplify exposition by assuming that $V$ is diagonal and that $\hat{V}$ is equal to $V$, so that

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

(17)

The first assumption is without loss of generality, since we can always change coordinates by an orthogonal transformation to make $\hat{\beta}_j$ uncorrelated across $j$. The second assumption simplifies notation and could easily be dropped.

Under these assumptions, the empirical Bayes estimator of equation (13) is given by a component-wise weighted average of $\hat{\beta}_0 \cdot M$ and $\hat{\beta}$,

$$\hat{\beta}_{EB} = \text{diag} \left( \frac{v_j}{\hat{\tau}^2 + v_j} \right) \cdot \hat{\beta}_0 \cdot M + \text{diag} \left( \frac{\hat{\tau}^2}{\hat{\tau}^2 + v_j} \right) \cdot \hat{\beta}.$$  

(18)

The hyperparameters $\beta_0$ and $\tau^2$ are estimated by maximizing the marginal likelihood as in equation (11), which now simplifies to

$$(\hat{\beta}_0, \hat{\tau}^2) = \arg\min_{\beta_0, \tau^2} \frac{1}{J} \cdot \sum_j \left( \log(\tau^2 + v_j) + \frac{(\hat{\beta}_j - b_0 \cdot M_j)^2}{\tau^2 + v_j} \right).$$

(19)

**Asymptotic characterization of risk**

Our goal is to characterize the squared error

$$SE = \frac{1}{J} \cdot \sum_j \left( \hat{\beta}_{EB}^j - \beta_j \right)^2,$$
and the corresponding mean squared error \( \text{MSE} = E[SE] \) of the empirical Bayes estimator. In order to obtain our desired characterizations, we consider an asymptotic approximation where \( J \) becomes large, such that \( \hat{\beta}_0 \) and \( \hat{\tau}^2 \) converge in probability to constants. To do asymptotics, we need an assumption about what happens as the dimension \( J \) increases. We adopt a random effects framework for this purpose.

**Assumption 1** (Random effects sequence)
The vectors \((\hat{\beta}_j, \beta_j, v_j, M_j)\) are i.i.d. draws from some distribution \( P \) with finite second moments such that

\[
E \left[ \hat{\beta}_j | \beta_j, v_j, M_j \right] = \beta_j, \\
\text{Var} \left( \hat{\beta}_j | \beta_j, v_j, M_j \right) = v_j.
\]

Let \( \hat{\beta}^{EB}(b_0, \tau^2) \) be the empirical Bayes estimator for given (non-random) hyperparameters \((b_0, \tau^2)\) and let \( SE(b_0, \tau^2) \) and \( MSE(b_0, \tau^2) \) be the corresponding squared error and mean squared error. The \( MSE \) given \( b_0 \) and \( \tau^2 \) can be written as a sum of variance and squared bias terms,

\[
MSE(b_0, \tau^2) = E \left[ \left( \frac{\tau^2}{\tau^2 + v_j} \right)^2 \cdot v_j + \left( \frac{v_j}{\tau^2 + v_j} \right) \cdot (\beta_j - b_0M_j)^2 \right]. \tag{20}
\]

Let \((\beta_0, \tau^{*2})\) be the maximizer of the expected log-likelihood given by the expectation of (19),

\[
(\beta_0, \tau^{*2}) = \arg\min_{b_0, \tau^2} E \left[ \log(\tau^2 + v_j) + \left( \frac{\hat{\beta}_j - b_0 \cdot M_j}{\tau^2 + v_j} \right)^2 \right].
\]

The following theorem shows that, as \( J \) becomes large, we can approximate the loss (squared error) of the empirical Bayes estimator \( \hat{\beta}^{EB} \) by the risk (mean squared error) of the infeasible estimator using the limiting pseudo-true values of \((\beta_0, \tau^{*2})\).

**Theorem 1**
Suppose that assumption \( \Box \) holds. Then \( SE(\hat{\beta}_0, \hat{\tau}^2) - MSE(\beta_0, \tau^{*2}) \to^p 0 \) as \( J \to \infty \).

**Discussion**
Recall that we obtain unrestricted estimation and structural estimation as limiting cases of our proposed estimator, where \( \tau^2 \to \infty \) corresponds to unrestricted esti-

---

4Actually, \( \hat{\beta}^{EB}(b_0, \tau^2) \) is the Bayes estimator for the prior \( \beta \sim N(b_0, \tau^2 I) \).
mation and $\tau^2 \to 0$ to restricted estimation. The mean squared error $MSE(b_0, \tau^2)$ for given values of $(b_0, \tau^2)$ is equal to the sum of a variance term and a squared bias term, cf. equation (20). The mean squared error of the unrestricted estimator only contains variance terms, $MSE(b_0, \infty) = E[v_j]$, and the mean squared error of the structural estimator converges to an average containing only bias terms, $\min_{b_0} MSE(b_0, 0) = \min_{b_0} E \left[ (\beta_j - b_0 M_j)^2 \right]$. Theorem 1 then immediately yields the following corollary.

**Corollary 1**

*Under assumption 2, and for large enough $J$, empirical Bayes has lower mean squared error than unrestricted estimation if*

$$MSE(\beta_0, \tau^{*2}) < E[v_j],$$

*and larger mean squared error if this inequality is reversed. It has lower mean squared error than restricted estimation for large $J$ if*

$$MSE(\beta_0, \tau^{*2}) < \min_{b_0} E \left[ (\beta_j - b_0 M_j)^2 \right],$$

*and larger mean squared error if this inequality is reversed.*

**The role of heteroskedasticity**

The infeasible oracle-optimal choice of $(b_0, \tau^2)$ would minimize $MSE(b_0, \tau^2)$ and automatically yield an estimator that dominates structural and unrestricted estimation uniformly. The first order condition for the optimal $\tau^{*2}$ minimizing the mean squared error is

$$E \left[ \frac{v_j^2}{(\tau^{*2} + v_j)^3} \cdot (\tau^{*2} - (\beta_j - \beta_0 M_j)^2) \right] = 0.$$

The empirical Bayes estimate $(\hat{\beta}_0, \hat{\tau}^{*2})$, by contrast, maximizes the marginal log likelihood, and its large $J$ limit $(\beta_0, \tau^{*2})$ maximizes the expected log likelihood. The first order condition characterizing $\tau^{*2}$ is

$$E \left[ \frac{1}{(\tau^{*2} + v_j)^2} \left( \tau^{*2} - (\beta_j - \beta_0 M_j)^2 \right) \right] = 0.$$

How does $\tau^{*2}$ relate to the optimal choice of $\tau^{*2}$? As can be seen from the first order conditions, both are weighted averages of $(\beta_j - \beta_0 M_j)^2$. 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 $\beta_j$ 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 $\beta_j$ and $v_j$, as the following corollary shows.

**Corollary 2**

*Under assumption $\mathcal{H}$, suppose $M \equiv 0$ and

$P(v_j = \beta_j = 0) = P(v_j = \beta_j = 2) = \frac{1}{2}.$

Then

$\tau^* = 0$

$MSE(\beta_0, \tau^*) = 2$

$MSE(b_0, \infty) = 1,$

so that unrestricted estimation has lower mean squared error than empirical Bayes for large samples.*

Restricted estimation dominates empirical Bayes for small enough samples if $\beta_j = 0$ with probability 1. Note, however, that in this case the two estimators become equivalent for large enough $J$ since $\hat{\tau}^2 \to^p 0$.

### 3.3 Geometry of our empirical Bayes estimator

In this section we study the geometry of the empirical Bayes estimator proposed in Section 2.3. This estimator can be seen as providing a mapping from an unrestricted (preliminary) estimate $\hat{\beta}$ to an empirical Bayes estimate $\hat{\beta}^{EB}$. Understanding this mapping is key for understanding the behavior of our estimator. We consider again the canonical setup of Section 3.2 where $\hat{V} = \text{diag}(v_i)$. 
Special case: $M = 0$

We shall first discuss the case where $M = 0$, so that we can ignore estimation of $\beta_0$. In this case, the expression for $\hat{\beta}^{EB}$ simplifies further to

$$\hat{\beta}^{EB} = \text{diag} \left( \frac{\tau^2}{\tau^2 + v_j} \right) \cdot \hat{\beta}.$$  

As $\tau$ varies, this equation describes a curve interpolating between the unrestricted estimate $\hat{\beta}$ and the “restricted estimate” $0$. All points along this curve are points of tangency between a sphere around 0 (corresponding to the prior variance), and an ellipsoid around $\hat{\beta}$ with axes of length proportional to $v_i$ (corresponding to estimator variance). This expression does not quite reveal the mapping from $\hat{\beta}$ to $\hat{\beta}^{EB}$ as $\tau^2$ itself is a function of $\hat{\beta}$, given by the solution to the first order condition

$$\sum_j \frac{1}{\tau^2 + v_j} = \sum_j \frac{\hat{\beta}_j^2}{(\tau^2 + v_j)^2}.$$  

Given $\tau^2$, this first order condition implies that $\hat{\beta}$ must be somewhere on the surface of an ellipsoid with semi-axes that have length

$$(\tau^2 + v_j) \cdot \sqrt{\sum_{j'} \frac{1}{\tau^2 + v_{j'}}} \quad (21)$$

along the $j$th dimension. This implies in turn that the length of $\hat{\beta}^{EB}$ is given by

$$\tau^2 \cdot \sqrt{\sum_{k'} \frac{1}{\tau^2 + v_{k'}}} \quad (22)$$

Note that this value does not depend on $\hat{\beta}$ beyond its effect on $\tau^2$. All estimates $\hat{\beta}^{EB}$ corresponding to a given value of $\tau^2$ are on the surface of a sphere with this radius. Note finally that there is a natural lower bound on $\tau^2$ of 0\(^5\). In particular, we have that $\tau^2$ is equal to 0 for any values of $\hat{\beta}$ inside the ellipsoid with semi-axes of length

$$v_j \cdot \sqrt{\sum_{k'} \frac{1}{v_{k'}}} \quad (23)$$

\(^5\)Since we impose this bound, our estimator resembles the positive-part James Stein estimator.
Visual representation

We can illustrate the mapping from $\hat{\beta}$ to $\hat{\tau}^2$ and $\hat{\beta}^{EB}$ graphically when $\dim(\beta) = 2$. Suppose that $v_1 = 2$, and $v_2 = 1$. The top part of Figure 1 shows $\hat{\tau}^2$ as a function of $\hat{\beta}$. This function is flat and equal to 0 inside the white ellipsoid; it rises smoothly and approaches a circular cone for large $\hat{\beta}$. The bottom part of this same figure shows (i) $\hat{\tau}^2 - \hat{\beta}$ as a vector field (arrows are proportional to, but smaller than, this difference) and (ii) a contour plot of the length of these vectors, that is, of the amount of shrinkage relative to the unrestricted estimator.

The structure of this mapping gets more transparent when considering the analytic characterizations we just derived. Figure 2, in particular, plots, for various values of $\hat{\tau}^2$, (i) which values of $\hat{\beta}$ would imply such values of $\hat{\tau}^2$ and (ii) the corresponding estimates $\hat{\beta}^{EB}$.

How can we interpret these figures? For small $\hat{\beta}$, the estimator concludes that the “theory” is essentially correct, where the theory in this case reduces to the assumption $\beta = 0$. As $\hat{\beta}$ gets larger, so does the estimated $\hat{\tau}^2$ – the theory is considered “less correct.” Deviations from 0 in the direction of the first coordinate are weighted less heavily as $\hat{\beta}_1$ has a larger variance (is less precisely estimated). $\hat{\beta}_1$ is shrunk most heavily if $\hat{\beta}_2$ seems to confirm the theory while $\hat{\beta}_1$ violates it moderately, as evident in the bottom right plot of Figure 1. When $\hat{\beta}$ is large, so is $\hat{\tau}^2$, and the theory is essentially disregarded; $\hat{\beta}^{EB}$ is basically equal to the unrestricted estimator, as evident in the bottom plots of Figure 2.

Geometry in the general case: $M \neq 0$

Let us now turn to the general case where $M \neq 0$ and estimation of $\beta_0$ has thus to be accounted for. This can be analyzed using the same “trick” as before, where we consider $\hat{\tau}^2$ and $\hat{\beta}_0$ to be given and derive the corresponding sets of $\hat{\beta}$ and $\hat{\beta}^{EB}$.

Given $\hat{\tau}^2$, $\hat{\beta}_0$ minimizes the quadratic form

$$\sum_j \frac{(\hat{\beta}_j - \hat{\beta}_0 \cdot M_j)^2}{\hat{\tau}^2 + v_j},$$

where
so that
\[
\hat{\beta}_0 = \frac{\sum_j \hat{\beta}_j \cdot \frac{1}{\hat{\tau}^2 + v_j}}{\sum_j M_j \cdot \frac{1}{\hat{\tau}^2 + v_j}}.
\]

This equation defines a hyper-plane in the space of \( \hat{\beta} \). As before, the first order condition for \( \hat{\tau}^2 \) implies
\[
\sum_j \frac{1}{\hat{\tau}^2 + v_j} = \sum_j \frac{(\hat{\beta}_j - \hat{\beta}_0 \cdot M_j)^2}{(\hat{\tau}^2 + v_j)^2}.
\]

This equation describes an ellipsoid centered at \( \hat{\beta}_0 \cdot M \) with semi-axes of length \( v_j \cdot \sqrt{\sum_{k'} \frac{1}{v_{k'}}} \) along dimension \( k \). Given \( \hat{\tau}^2 \) and \( \hat{\beta}_0 \) we thus get that \( \hat{\beta} \) has to lie on the surfaces of this ellipsoid, intersected with a hyper-plane through the center of this ellipsoid. \( \hat{\beta}^{EB} \) is then obtained from \( \hat{\beta} \) by shrinking on the hyper-plane towards the center of the ellipsoid, where \( \hat{\beta}^{EB} \) again ends up on a sphere of radius \( \hat{\tau}^2 \cdot \sqrt{\sum_{k'} \frac{1}{v_{k'}}} \) around this center.

We can rephrase this argument by considering only \( \hat{\tau}^2 \) to be given. Conditional on \( \hat{\tau}^2 \), we get that \( \hat{\beta} \) has to lie on the surface of a hyper-cylinder with ellipsoid basis and axis going through the origin and pointing in the direction of the vector
\[
\left( \frac{1}{\hat{\tau}^2 + v_1}, \ldots, \frac{1}{\hat{\tau}^2 + v_J} \right).
\]

The corresponding estimates \( \hat{\beta}^{EB} \) are on the surface of a hyper-cylinder with spherical basis and the same axis. Note that the tilt of the axis depends on \( \hat{\tau}^2 \) and varies between (1, \ldots, 1) for large \( \hat{\tau}^2 \) and \( \left( \frac{1}{v_1}, \ldots, \frac{1}{v_J} \right) \) for \( \hat{\tau}^2 = 0 \).

4 Labor demand and wage inequality

We now turn to our motivating application, the estimation of labor demand. How to estimate labor demand is a central question for several literatures on wage inequality, including the literature on the impact of immigration, and the literature on skill-biased technical change. We start by giving a brief review of the most common structural estimation approach. We then adapt our general empirical Bayes approach to this setting. Section 5 below applies the resulting estimator to data from the United States.
Suppose there are $J$ types of workers, defined for instance by their level of education, potential experience, and whether they are migrants. Consider a cross-section of labor markets $i = 1, \ldots, n$. We adopt cross-sectional notation for simplicity, similar arguments apply to time series or panel data. Let $Y_{ij}, j = 1, \ldots, J$ 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 $Y_i = (Y_{i1}, \ldots, Y_{iJ})$ 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

$$Y_i = y(X_i, \epsilon_i),$$

where $\epsilon_i$ denotes a vector of unobserved demand shifters of unrestricted dimension.

### 4.1 CES-production functions and structural estimation

The majority of contributions to the field impose a tightly parametrized 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 which equal marginal productivity.

Denote wages by $w$ and labor supply by $N$, so that $Y_{ij} = \log(w_{ij})$ and $X_{ij} = \log(N_{ij})$. Assume that wages equal marginal productivity for some aggregate production function $f$,

$$w_{ij} = \frac{\partial f_i(N_{i1}, \ldots, N_{iJ})}{\partial N_{ij}},$$

and that the aggregate production function takes a constant elasticity of substitution (CES) form,

$$f_i(N_{i1}, \ldots, N_{iJ}) = \left( \sum_{j=1}^{J} \gamma_j N_{ij}^\rho \right)^{1/\rho}.$$  

These two assumptions together imply

$$w_{ij} = \frac{\partial f_i(N_{i1}, \ldots, N_{iJ})}{\partial N_{ij}} = \left( \sum_{j'=1}^{J} \gamma_{j'} N_{ij'}^\rho \right)^{1/\rho-1} \cdot \gamma_j \cdot N_{ij}^{\rho-1}.$$  

We get that the relative wage between groups $j$ and $j'$ is equal to

$$\frac{w_{ij}}{w_{ij'}} = \frac{\gamma_j}{\gamma_{j'}} \cdot \left( \frac{N_{ij}}{N_{ij'}} \right)^{\rho-1}.$$  

25
Taking logs yields
\[ Y_{ij} - Y_{ij'} = \log(\gamma_j) - \log(\gamma_{j'}) + \beta_0 \cdot (X_{ij} - X_{ij'}), \]
where \( \beta_0 = \rho - 1 \).

This result motivates regressions of the following form (see for instance Autor et al. 2008 and Card 2009):
\[ Y_{ij} - Y_{ij'} = \gamma_{jj'} + \beta_0 \cdot (X_{ij} - X_{ij'}) + \epsilon_{ijj'}. \] (28)

The coefficient \( \beta_0 \) in this regression is interpreted as the negative of the inverse elasticity of substitution between labor types \( j \) and \( j' \). The constant \( \gamma_{jj'} \) captures factors unaffected by labor supply which do affect relative wages. In practice, such regressions usually include additional controls for observables and/or time trends, as well as labor market fixed effects in panel data, and might be estimated using instrumental variables to account for the endogeneity of labor supply. More general specifications might also include additional terms for aggregate types of labor as motivated by nested CES models.

Equation (28) can be rewritten in a numerically equivalent way as a fixed effects regression with restrictions across coefficients:
\[ Y_{ij} = \alpha_i + \gamma_j + \sum_{j'} \beta_{jj'} X_{ij'} + \epsilon_{ij}, \] (29)
\[ \beta_{jj'} = \beta_0 \cdot M_{jj'}, \] (30)
\[ M_{jj'} = (I - \frac{1}{J} E)_{jj'} = \begin{cases} (1 - \frac{1}{J}) & j = j' \\ -\frac{1}{J} & j \neq j' \end{cases}. \]

Here \( I \) 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 see this, just take the difference \( Y_{ij} - Y_{ij'} \) based on equation (29).

To gain some intuition for the equivalence between equations (28) and (29), note that equation (28) has the form of a difference-in-differences regression, where differences are taken across types \( j \) of labor, as well as across cross-sectional units \( i \). Such difference-in-differences regressions can always equivalently be written in fixed-effects

---

6The elasticity of substitution \( \sigma \) is defined as the relative change in the demand for different factors induced by a given change in their relative prices.
form.

4.2 Unrestricted estimation

Rather than imposing the strong assumptions implied by the CES production function model or its generalizations, we could instead “let the data speak.” One natural way of doing so is to consider a linear specification with a large number of types $J$ and unrestricted own- and cross-elasticities. We could attempt to estimate the model

$$Y_{ij} = \alpha_i + \gamma_j + \sum_{j'} \beta_{jj'} X_{ij'} + \epsilon_{ij},$$

using least squares, without imposing any cross-restrictions on the parameters $\beta_{jj'}$. This is the same regression model as implied by the CES production function, except that the latter restricts the $J^2$-dimensional parameter $\beta$ to lie in a 1 dimensional subspace.

This general model is not identified. Differencing across types $j$ yields a model which is identified. Let $\Delta$ be a $(J - 1) \times J$ matrix which subtracts the first entry from each component of a $J$ vector,

$$\Delta = (-e, I_{J-1}).$$

The impact of labor supply on relative wages is identified, and thus so is the following model. Estimating this model yields an unrestricted estimator $\hat{\delta}$ of $\delta$.

$$\Delta \cdot Y_i = \Delta \cdot \gamma + \delta \cdot X_i + \Delta \cdot \epsilon_i,$$

$$\delta = \Delta \cdot \beta. \quad (31)$$

We thus have $J \cdot (J - 1)$ free slope parameters $\delta$ to be estimated. Relative to this general linear fixed effects model, the CES production function therefore implies $J^2 - J - 1$ additional restrictions.

Let us briefly discuss the economic content of the restrictions on $\beta$ imposed by the structural model relative to the the unrestricted regression $[31]$. First, $\beta \cdot e = 0$ for $e = (1, \ldots, 1)$. Proportionally increasing the labor supply of every group by the same factor does not affect wages. This is a restriction implied by constant returns to scale, if wages are assumed to correspond to marginal productivity based on an
aggregate production function. Second, $\beta_{jj'} = \beta_{jj''}$ for $j', j'' \neq j$. The elasticity of substitution between different groups is the same for all groups. The CES model imposes that there are only two possible degrees of substitutability between different workers – either they are perfect substitutes when they are the same type, or they have an elasticity of substitution of $\sigma = -1/\beta_0$. Third, $\beta_{jj} = \beta_{j'j'}$. The own-elasticity of demand is the same for all types of labor. In combination, these three restrictions in fact imply the CES regression model. The CES model additionally implies that changes in labor supply do not affect within-type inequality of wages. Given the small number of types usually imposed, this is a strong restriction.

4.3 Empirical Bayes estimation, shrinking toward the $J$-type CES model

We have just described a structural model for labor demand, and an unrestricted demand system which nests the structural model. In Section 2 above we introduced a general approach to construct estimators which optimally exploit the empirical implications of theoretical models, in order to reduce estimator variance, without suffering from the inconsistencies of mis-specified models. We shall now adapt this general approach to the estimation of labor demand.

We discuss two cases. In this section, we discuss 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. In Section 4.4, we 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).

Some minor modifications of the approach introduced in Section 2 are necessary. In particular, the coefficients of interest $\delta$ that we now consider are in matrix form. We denote the vectorized version of $\delta$, stacking the columns on top of each other, by $\delta^\uparrow$, and similarly for other matrices. Furthermore, a family of priors is most naturally specified for $\beta$ while estimation is for $\delta = \Delta \cdot \beta$.

We model the coefficient matrix $\beta$ as corresponding to the coefficients of the structural CES model plus some disturbances, that is

$$\beta = (\beta_{jj'}) = \beta_0 \cdot M + \zeta$$

$$\zeta_{jj'} \sim^{iid} N(0, \tau^2),$$
where, as before, \( M = (I - \frac{1}{J}E) \). Differencing this model yields

\[
\delta = \Delta \cdot \beta = \beta_0 \cdot \Delta + \Delta \cdot \zeta
\]

The term \( \beta_0 \cdot \Delta \) is equal to a fixed scalar \( \beta_0 \) times \( \Delta \cdot M = \Delta \). This term corresponds to a set of coefficients satisfying the CES-production function model. The term \( \Delta \cdot \zeta \) is equal to a random \( J \times J \) matrix \( \zeta \) pre-multiplied by \( \Delta \). The variance of this term is given by

\[
\text{Var}(\Delta \cdot \zeta) = \tau^2 \cdot P \otimes I,
\]

where \( P := \Delta \cdot \Delta' = I_{J-1} + E_{J-1} \). This implies a prior variance of \( \hat{\delta}^\perp \) equal to

\[
\Sigma(\tau^2, V) = \text{Var}(\hat{\delta}^\perp) = \tau^2 \cdot P \otimes I + V.
\]

Substituting the consistent estimator \( \hat{V} \) for \( V \), we obtain the empirical Bayes estimators of \( \beta_0 \) and \( \tau^2 \) as solutions to the maximum (marginal) likelihood problem

\[
(\hat{\beta}_0, \hat{\tau}^2) = \arg \min_{b_0, \tau^2} \log \left( \det(\Sigma(\tau^2, \hat{V})) \right)
+ (\hat{\delta}^\perp - b_0 \cdot \Delta^\perp)' \cdot \Sigma(\tau^2, \hat{V})^{-1} \cdot (\hat{\delta}^\perp - b_0 \cdot \Delta^\perp).
\]

(32)

Given \( t^2 \), the optimal \( b_0 \) is equal to

\[
\hat{\beta}_0 = (\Delta \cdot \Sigma(t^2, \hat{V})^{-1} \cdot \Delta')^{-1} \cdot \Delta \cdot \Sigma(t^2, \hat{V})^{-1} \cdot \hat{\delta}^\perp.
\]

Substituting this expression into the objective function, we obtain a function of \( t^2 \) alone that we optimize numerically. Given the unrestricted estimates \( \hat{\delta} \), as well as the estimates \( \hat{\beta}_0 \) and \( \hat{\tau}^2 \), we obtain the empirical Bayes estimator of \( \delta \) as

\[
\hat{\delta}^{EB} = \hat{\beta}_0 \cdot \Delta^\perp + P \otimes I \cdot \left( P \otimes I + \frac{1}{\tau^2} \hat{V} \right)^{-1} \cdot (\hat{\delta}^\perp - \hat{\beta}_0 \cdot \Delta^\perp).
\]

(33)

4.4 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 \( k_j \). Assume that the production function takes the form
\[
f_i(N_{i1}, \ldots, N_{iJ}) = \left( \sum_{k=1}^{2} \tilde{N}_{ik}^\rho \right)^{1/\rho}. \\
\tilde{N}_{ik} = \sum_{k_j = k} \gamma_j N_{ij}.
\]
Then
\[
w_{ij} = \frac{\partial f_i}{\partial N_{ij}} = \left( \frac{2}{\sum_{k=1}^{2} \tilde{N}_{ik}^\rho} \right)^{1/\rho-1} \cdot \gamma_j \cdot \tilde{N}_{ik}^{\rho-1},
\]
and thus, using the same notation as before,
\[
\frac{w_{ij}}{w_{ij'}} = \frac{\gamma_j}{\gamma_{j'}} + \left( \frac{\tilde{N}_{ikj}}{\tilde{N}_{ikj'}} \right)^{\rho-1}.
\]
With \( X_{ij} = \log(N_{ij}/\tilde{N}_{ikj}) \) and \( \tilde{X}_{ik} = \log(\tilde{N}_{ik}) \), we can thus nest the canonical CES model in the following regression specification, which includes regressors for both the dis-aggregated types \( j \) and the aggregated types \( k \),
\[
Y_{ij} - Y_{i1} = (\gamma_j - \gamma_1) + \sum_{j'} \delta_{jj'} X_{ij'} + \beta_0 \cdot (\tilde{X}_{ikj} - \tilde{X}_{i1}) + (\epsilon_{ij} - \epsilon_{i1}). \tag{34}
\]
In this setting the matrix \( \delta \) captures the additional effect of labor supply on relative wages beyond the effect already taken care of by the term \( \beta_0 \cdot (\tilde{X}_{ikj} - \tilde{X}_{i1}) \).

The canonical CES model implies the restriction \( \delta = 0 \). The unrestricted approach estimates versions of this equation with \( \delta \) left fully flexible. Our empirical Bayes approach applied to this setting takes as its point of departure a first stage unrestricted estimator \((\hat{\delta}, \hat{\beta}_1)\) of \((\delta, \beta_0)\), with estimated covariance matrix \( \hat{V} \). We then consider the family of priors
\[
\delta \sim \mathcal{N}(0, \tau^2 \cdot P \otimes I),
\]
where, as before, \( \beta_0 \) and \( \tau^2 \) are hyperparameters. Denote the variance of the unre-
stricted estimators given $\beta_0$ and $\tau^2$ by
\[
\Sigma(\tau^2, V) = \text{Var} \left( \left( \hat{\delta}_{\tau}, \hat{\beta}_1 \right) \right) = \begin{pmatrix} \tau^2 \cdot P \otimes I & 0 \\ 0 & 0 \end{pmatrix} + V;
\]
the conditional mean is given by $(0, \beta_0)$. We obtain the empirical Bayes estimators of $\beta_0$ and $\tau^2$ as solutions to the maximum (marginal) likelihood problem
\[
(\hat{\beta}_0, \hat{\tau}^2) = \arg\min_{b_0, t^2} \log \left( \det(\Sigma(t^2, \hat{V})) \right) + (\hat{\delta}_{\tau}, \hat{\beta}_1 - b_0)' \cdot \Sigma(t^2, \hat{V})^{-1} \cdot (\hat{\delta}_{\tau}, \hat{\beta}_1 - b_0)'.
\] (35)
Given $t^2$, the optimal $b_0$ is equal to
\[
\hat{b}_0 = \left( e \cdot \Sigma(t^2, \hat{V})^{-1} \cdot e' \right)^{-1} \cdot e \cdot \Sigma(t^2, \hat{V})^{-1} \cdot (\hat{\delta}_{\tau}, \hat{\beta}_1),
\]
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 $\delta$ as
\[
\hat{\delta}_{\tau}^{EB} = (\hat{\tau}^2 \cdot P \otimes I, 0) \cdot \Sigma(\hat{\tau}^2, \hat{V})^{-1} \cdot (\hat{\delta}_{\tau}, \hat{\beta}_1 - b_0)'.
\] (36)

5 Empirical analysis – United States

We now turn to our empirical application, studying labor demand in the United States. We use data which 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 motivated our estimator arguing that (i) qualitative conclusions tend to be sensitive to auxiliary assumptions imposed in structural models, and (ii) estimates often have a high variance when not imposing a model. Card (2009, p5f) and Borjas et al. (2012) discuss an important example, the estimated impact of past migration on wage inequality in the US, to which this argument applies. One side of the literature on this question argues that there were large effects. Their CES specifications assume (i) migrants and natives are perfect substitutes in the labor market while (ii) the elasticity of substitution between high school dropouts and high school graduates is the same as between either of those and college graduates or those with a postgraduate
degree. The other side of this literature argues that there were negligibly small effects. Their CES specifications assume that (i) natives and migrants are imperfect substitutes while (ii) high school dropouts and high school graduates are perfect substitutes.\footnote{Card argues that the assumptions of such a specification are justified by statistical tests.}

In this section, we do not focus on the labor market impact of migration, but rather study the impact of historical changes of the labor force composition on relative wages in general. Rather than imposing one or the other of the models (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 used in the literature.

5.1 Data

Our analysis for the United States 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 builds on the specifications and the code provided by \cite{BorjasEtAl2012}. 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 \cite{AcemogluAutor2011}, 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 male full-time workers (employed at least 40 weeks and working at least 35 hours per week). 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).

5.2 Results

As a first step, we seek to replicate results from the literature. The leading specification in the literature considers two types of workers, those with more than high school
education relative to 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 3, where the first graph shows the actual series whereas the second graph shows the residualized series after controlling for a kinked time-trend.

We next aim to 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.

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 4 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 shows the corresponding coefficient estimates, which are based on the models discussed in Section 4.4.

The top right graph of Figure 4 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 labor supply induced a widening of inequality over the most recent two decades. Second, these shifts also induced a compression of wages between different workers with high school degrees or less and a widening between those with more than high-school education.

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 optimally balance bias and variance. The predicted counterfactual changes of wages derived from these estimates are qualitatively similar to the unrestricted model, but of reduced magnitude.

The estimated $\hat{\tau}^2$, 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 of the unrestricted 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. Taken at face value, our estimates also suggest that past migration did not affect wage inequality between native workers much, in line with the conclusions of Card (2009).

6 Monte Carlo simulations – demonstrating the performance of the empirical Bayes estimator

In this section we present a series of simulation and evaluation exercises comparing the performance of our empirical Bayes procedure to its competitors, structural estimation and unrestricted estimation. Section 6.1 presents simulations corresponding to the empirical Bayes paradigm, fixing the hyperparameter $\theta$ and drawing from the implied distributions of the parameters $\eta$ and data $Y$. Section 6.2 presents simulations corresponding to the frequentist paradigm, fixing the parameter $\eta$ and drawing
Section 6.3 considers simulations similar to Section 6.1 and 6.2, but governed by parameters calibrated to match our empirical application.

### 6.1 Monte Carlo results, fixing \( \theta \), drawing from the distribution of \( \eta \) and \( Y \)

Corresponding to the different paradigms of statistical inference (Bayesian, frequentist, empirical Bayes), there are different notions of the performance of an estimator. The Bayesian perspective considers expected loss averaged over possible values of both \( \theta \) and \( \eta \). The frequentist perspective considers expected loss conditional on \( \eta \), averaging just over repeated draws of the data. The empirical Bayes perspective considers expected loss averaging over \( \eta \) but conditional on \( \theta \). Let us first consider simulations based on the empirical Bayes perspective, where we repeatedly draw values for \( \eta \) (in particular own- and cross-elasticities \( \beta \)) and data generated by the parameter \( \eta \).

In our simulations, we vary the sample size \( n \), the number of regressors \( J \), the residual variance \( \sigma^2 \), and the parameter \( \tau^2 \), which measures how well the structural model describes the data generating process. For all simulations, the regressors \( X_{ij} \) are i.i.d. draws from the uniform distribution on \([0, 1]\), and the regression residuals are normally distributed with variance \( \sigma^2 \). Results are based on 1,000 Monte Carlo draws for each design. Table 2 shows the results of these simulations. For each design we show the mean squared error, calculated as an average over Monte Carlo draws of \( \beta \) and \( Y \), for four alternative estimation procedures, relative to the proposed empirical Bayes procedure.

At one extreme of the designs considered are those with a small sample size, a large number of regressors, a high variance of residuals, and a good fit of the structural model (small \( \tau^2 \)). In these designs we would expect the structural model to work well and to potentially outperform the empirical Bayes procedure, as it exploits additional correct information. And indeed we do find that structural estimation dominates empirical Bayes at the very extreme of the range of designs considered.

At the other extreme of the designs considered are those with large sample size, small number of regressors, small variance of residuals, and poor fit of the structural model.
model (large $\tau^2$). In these designs we would expect the unrestricted estimator to work well, as it has a small variance and does not shrink toward the incorrect structural model. Nonetheless, we do find that unrestricted estimation never dominates empirical Bayes for any of the designs considered. It does seem like unrestricted estimation is uniformly dominated by empirical Bayes, in terms of MSE given $\theta$, confirming the conclusions we drew from theorem 1 in Section 3.

Over almost the entire range of the simulations considered, empirical Bayes performs very well and better than either of the alternatives: structural / unrestricted estimation. For designs where $\tau^2$ is large, estimation based on the structural model yields estimates that perform very poorly relative to empirical Bayes, as to be expected. And for all designs considered, the variance reduction achieved by empirical Bayes implies that empirical Bayes performs better than unrestricted estimation, sometimes significantly so.

The last column of Table 2 shows, for purposes of comparison, the infeasible oracle empirical Bayes estimator, where $\tau^2$ is assumed to be known rather than estimated. As this column shows, knowledge of $\tau^2$ does not appear to result in improvements of performance.

6.2 Monte Carlo results, fixing $\eta$, drawing from the distribution of $Y$

In Section 6.1 we considered simulations where $\theta$ was fixed, but $\eta$ was drawn repeatedly, an approach which corresponds to the empirical Bayes paradigm. We shall now turn to simulations in the spirit of the frequentist paradigm, where $\eta$ is fixed, and we repeatedly sample from the distribution of $Y$.

Specifically, we are considering coefficient matrices of the form

$$\beta = \beta_{00} \cdot M_{J0} + \beta_{01} \cdot M_{J1} + \beta_{02} \cdot M_{J2},$$

where $M_{J0}$ is equal to $M_J$ in the first $J/4$ columns, and zero elsewhere, $M_{J2}$ is equal to $M_J$ in the last $J/4$ columns, and zero elsewhere, and $M_{J1}$ is equal to $M_J$ in the middle $J/2$ columns, and zero elsewhere. This design implies that the structural model is correct if and only if $\beta_{00} = \beta_{01} = \beta_{02}$. Table 3 shows the results of these simulations. The values for $n$, $J$, and $\sigma^2$ are the same as considered before, as are the distributions of $X_{ij}$ and of the residuals. For each combination of these values, we consider different combinations of $\beta_{00}$, $\beta_{01}$, and $\beta_{02}$. 36
Structural estimation dominates empirical Bayes when the structural model is correctly specified, that is when $\beta_{00} = \beta_{01} = \beta_{02}$. Not very surprisingly, the reduction in MSE by imposing the structural model relative to empirical Bayes estimation can be made arbitrarily large when the model is exactly right, the number of parameters $J$ is large, and estimates are noisy (small sample size $n$, large residual variance $\sigma^2$). On the other hand, structural estimation performs significantly worse when the structural model is violated, and the variance of unrestricted estimation is not too large. These simulation results once again confirm the predictions regarding the favorable performance of empirical Bayes made by theorem 1 and its corollaries.

### 6.3 Calibrated Monte Carlo simulations

We conclude this section by presenting some simulations similar to those discussed before, but calibrated to our empirical results. We first estimate the model via empirical Bayes as in our empirical application, using the US panel of states and controlling for time and state fixed effects, to obtain estimates $\hat{\beta}^0$, $\hat{\tau}^2$, and $\hat{\delta}_{EB}$. We then perform two sets of simulations. For the first, we repeatedly draw values for $\delta$ conditional on $\hat{\beta}^0$, $\hat{\tau}^2$, and values of $\hat{\delta}_r$ conditional on $\delta$. These simulations are analogous to those of Section 6.1, “conditional on $\theta$.” For the second set of simulations, we fix $\delta$ equal to the estimated $\hat{\delta}_{EB}$, and draw values of $\hat{\delta}_r$ from the corresponding sampling distribution. These simulations are analogous to those of Section 6.2, “conditional on $\eta$.” Each simulation is repeated 1000 times, and for each repetition we calculate structural, unrestricted, and empirical Bayes estimates based on $\hat{\delta}_r$. We then calculate the mean squared errors of each of these and normalize them relative to the MSE of empirical Bayes estimation.

We do this for both of the following cases. First, we consider shrinkage toward the $J$-type CES model, as in Section 4.3. Simulations conditional on $\theta$ and conditional on $\eta$ are shown in Table 5. We second consider shrinkage toward the 2-type CES model of a demand system for 8 types of workers, as in Section 4.4. Simulations conditional on $\theta$ and conditional on $\eta$ are shown in Table 5.
These simulations show that our proposed empirical Bayes approach, in these empirical settings, performs consistently much better than both unrestricted estimation and structural estimation. These simulation results strongly support using our proposed estimator.

7 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 which 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.

In a normal-normal setting with linear restrictions implied by the theory, this approach leads to particularly tractable and interpretable estimators. We provide several theoretical results for this case. One of our key results, Theorem 1, provides a characterization of the risk function of our estimator. The theorem is based on an asymptotic approximation which 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.

We apply our approach to estimation of labor demand. A workhorse-method of labor economics is the estimation of demand systems for labor, derived from aggregate production functions having a CES-form, or something similar. Many qualitative conclusions of this literature, concerning for instance the impact of immigration on native wages, are quite sensitive to functional form choices in the theoretical model.
imposed. Rather than imposing such a model on the estimates, we propose to use it as a reference point toward which unrestricted estimates are shrunk in a data-dependent way.

We implement this approach to study labor demand and wage inequality in the United States using the CPS and ACS data. Using panel data for states and alternative estimation approaches, we find negative but small inverse elasticities of substitution. The estimated elasticities are significantly smaller than comparable estimates based on national time series for the United States, and not significantly different from 0. Specifications derived from a 2-type constant elasticity of substitution (CES) production function do not appear to provide a particularly good fit to the data. Overall, changes in labor supply don’t appear to be major drivers of relative wages.
A  Proofs

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

$$\hat{\beta}^{EB} = \hat{\beta} + \frac{1}{\hat{\tau}^2} \hat{V} \cdot \left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1} \cdot \left(\hat{\beta}_0 \cdot M - \hat{\beta}\right).$$

By assumption, $\hat{\beta} \to^p \beta$. Our claim follows, by Slutsky’s theorem, if we can show that $\frac{1}{\hat{\tau}^2} \hat{V} \to^p 0$, and $\hat{\beta}_0 = O_p(1)$. Since $\hat{V} \to^p 0$, this holds if $(\hat{\beta}_0, \hat{\tau}^2)$ 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 hyperparameters,

$$(\hat{\beta}_0, \hat{\tau}^2) \to^p \arg\min_{b_0, t^2} J \cdot \log (t^2) + \frac{1}{t^2} \|\beta - b_0 \cdot M\|^2$$

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

Proof of proposition 2: By equation (13),

$$\hat{y} - x \cdot \hat{\beta} = x \cdot \frac{1}{\hat{\tau}^2} \hat{V} \cdot \left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1} \cdot (\hat{\beta}_0 - \hat{\beta})$$

$$= \left(x \cdot \frac{1}{\hat{\tau}} \hat{V}^{1/2}\right) \cdot \left(\hat{\tau} \hat{V}^{-1/2} + \frac{1}{\hat{\tau}} \hat{V}^{1/2}\right)^{-1/2} \cdot \left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1/2} \cdot (\hat{\beta}_0 - \hat{\beta}),$$

and thus

$$|\hat{y} - x \cdot \hat{\beta}| \leq \left|x \cdot \frac{1}{\hat{\tau}} \hat{V}^{1/2}\right| \cdot \left(\hat{\tau} \hat{V}^{-1/2} + \frac{1}{\hat{\tau}} \hat{V}^{1/2}\right)^{-1/2} \cdot \left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1/2} \cdot (\hat{\beta}_0 - \hat{\beta}) \cdot .$$

By equation (11),

$$\left|\left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1/2} \cdot (\hat{\beta}_0 - \hat{\beta})\right| = \min_{\beta_0} \left|\left(I + \frac{1}{\hat{\tau}^2} \hat{V}\right)^{-1/2} \cdot (\beta_0 M - \hat{\beta})\right|.$$
\[
\leq \left\| \left( I + \frac{1}{\hat{\tau}^2} \hat{V} \right)^{-1/2} \cdot \hat{\beta} \right\| \leq \| \hat{\beta} \|.
\]

where the last inequality holds by positive definiteness of \( \hat{V} \), which also implies

\[
\left\| \left( \hat{\tau} \hat{V}^{-\frac{1}{2}} + \frac{1}{\hat{\tau}} \hat{V}^{1/2} \right)^{-1/2} \right\| \leq 1.
\]

The first inequality claimed in proposition 2 follows. The proof for \( \hat{y} - x \cdot \hat{\beta}^a \) proceeds analogously. □

The following simple lemma gives a sufficient condition which allows us to approximate the squared error for the estimator using the estimated \((\hat{\beta}_0, \hat{\tau}^2)\) by the mean squared error of the infeasible estimator using the non-random limits \((\beta_0, \tau^{*2})\). This lemma is used in the proof of theorem 1.

**Lemma 1**

Suppose that MSE is continuous at \((\beta_0, \tau^{*2})\), that \((\hat{\beta}_0, \hat{\tau}^2) \to_P (\beta_0, \tau^{*2})\), and that

\[
\sup_{(b_0, \tau^2) \in U} |SE(b_0, \tau^2) - MSE(b_0, \tau^2)| \to_P 0,
\]

where \( U \) is some neighborhood of \((\beta_0, \tau^{*2})\).

Then \( SE(\hat{\beta}_0, \hat{\tau}^2) - MSE(\beta_0, \tau^{*2}) \to_P 0. \)

**Proof of lemma 1**

This is immediate from

\[
|SE(\hat{\beta}_0, \hat{\tau}^2) - MSE(\beta_0, \tau^{*2})| \leq |SE(\hat{\beta}_0, \hat{\tau}^2) - MSE(\hat{\beta}_0, \hat{\tau}^2)| + |MSE(\hat{\beta}_0, \hat{\tau}^2) - MSE(\beta_0, \tau^{*2})|.
\]

□

**Proof of theorem 1**

We need to show that the sufficient conditions of lemma 1 are satisfied. Convergence
of \((\hat{\beta}_0, \hat{\tau}^2)\) to the pseudo-true parameters
\[
(\beta_0, \tau^{*2}) = \arg\min_{b_0, \tau^2} E \left[ \log(\tau^2 + v_j) + \frac{\left(\hat{\beta}_j - b_0 \cdot M_j\right)^2}{\tau^2 + v_j} \right]
\]
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 show uniform convergence, in a neighborhood \(U\) of \((\beta_0, \tau^{*2})\), of

\[
SE(b_0, \tau^2) - MSE(b_0, \tau^2) = (E_J - E) \left[ \left(\hat{\beta}_j^{EB}(b_0, \tau^2) - \beta_j\right)^2 \right],
\]

where \(E_J\) denotes the average from \(j = 1, \ldots, J\). Such uniform convergence follows if we can show that the family of mappings

\[
(\hat{\beta}_j, \beta_j, v_j, M_j) \rightarrow \left(\hat{\beta}_j^{EB}(b_0, \tau^2) - \beta_j\right)^2,
\]

indexed by \((b_0, \tau^2) \in U\), 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 since it is a special case of example 19.8, p.272 in van der Vaart (2000):

(i) Continuity of \(\left(\hat{\beta}_j^{EB}(b_0, \tau^2) - \beta_j\right)^2\) in \((b_0, \tau^2)\) is immediate.

(ii) Compactness of the neighborhood \(U\) of \((\beta_0, \tau^{*2})\) to be considered can be imposed without loss of generality.

(iii) It remains to show the existence of an integrable envelope function. Suppose w.l.o.g. that the neighborhood \(U\) is of the form \([b, \tilde{b}] \times \left[t^2, \tilde{t}^2\right]\). Then \(\left(\hat{\beta}_j^{EB}(b_0, \tau^2) - \beta_j\right)^2\) always attains its maximum at one the corners of \(U\). This holds by monotonicity of \(\hat{\beta}_j^{EB}(b_0, \tau^2)\) in its arguments and the convexity of squaring. An envelope is therefore given by

\[
\max_{(b_0, \tau^2) \in \{b, \tilde{b}\} \times \{t^2, \tilde{t}^2\}} \left(\hat{\beta}_j^{EB}(b_0, \tau^2) - \beta_j\right)^2.
\]

This envelope is integrable since assumption \([1]\) imposed finite second moments, and given the form of \(\hat{\beta}_j^{EB}(b_0, \tau^2)\). □

**Proof of corollary** \([1]\) Immediate from Theorem \([1]\) □
Proof of corollary 2: Under the given assumptions, evaluating the asymptotic first order condition for maximizing the likelihood yields

\[ E \left[ \frac{1}{(\tau^2 + v_j)^2} (\tau^2 - \beta_j^2) \right] > 0, \]

for any \( \tau^2 > 0 \), which implies \( \tau^{*2} = 0 \). The other claims are immediate. \( \square \)
B Figures and tables

Figure 1: The mapping from \( \hat{\beta} \) to \( \hat{\tau}^2 \) and \( \hat{\beta}^{EB} \)

\( \hat{\tau}^2 \) as a function of \( \hat{\beta} \)

\( \hat{\beta}^{EB} - \hat{\beta} \) and its length as a function of \( \hat{\beta} \)

Notes: These figures illustrate the mapping from preliminary estimates to empirical Bayes estimates when \( \dim(\beta) = 2, \ Var(\hat{\beta}) = \text{diag}(2, 1) \), and \( M = 2 \). The top figure shows how our measure of model fit \( \hat{\tau}^2 \) varies with \( \hat{\beta} \), the bottom left figure shows the direction and magnitude of shrinkage from \( \hat{\beta} \) to \( \hat{\beta}^{EB} \), and the bottom right figure depicts just the magnitude of shrinkage. For details see section 3.3.
Figure 2: The geometry of empirical Bayes

Notes: These figures illustrate the mapping from preliminary estimates for the same setting as in Figure [1]. Each figure depicts, for a given value of $\hat{\tau}^2$, which preliminary estimates $\hat{\beta}$ yield this value, and to what set of empirical Bayes estimates $\hat{\beta}_{EB}$ these preliminary estimates are mapped. For details see section 3.3.
Figure 3: Log relative wages in the US – 2 types of workers

Note: The top graph of this figure shows US time series of log relative wages and log relative labor supply between workers with more than high school, and those with high school or less. The bottom graph shows the same, after controlling for a linear trend in time with a kink-point in 1992. Calculations are based on the March CPS. For details, see section 5. This figure replicates similar figures in Autor et al. (2008) and Acemoglu and Autor (2011).
Figure 4: 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 5.2.
Table 1: Estimated effects of labor supply on wage inequality, panel of US states

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

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-types 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 Section 2.4. For details, see section 5.2.
Table 2: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on $\theta$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$J$</th>
<th>$\sigma^2$</th>
<th>$\beta_0$</th>
<th>$\tau^2$</th>
<th>MSE relative to empirical Bayes estimation</th>
<th>structural</th>
<th>unrestricted</th>
<th>emp. Bayes</th>
<th>oracle e.B.</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>0.2</td>
<td>1.66</td>
<td>1.70</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>1.0</td>
<td>1.0</td>
<td>0.2</td>
<td>0.83</td>
<td>1.20</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>0.2</td>
<td>4.37</td>
<td>1.19</td>
<td>1.00</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>16</td>
<td>1.0</td>
<td>1.0</td>
<td>0.2</td>
<td>4.15</td>
<td>1.11</td>
<td>1.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>2.39</td>
<td>1.35</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>1.55</td>
<td>1.15</td>
<td>1.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>8.16</td>
<td>1.09</td>
<td>1.00</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>16</td>
<td>0.5</td>
<td>1.0</td>
<td>0.2</td>
<td>7.76</td>
<td>1.04</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>2.42</td>
<td>1.39</td>
<td>1.00</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>1.55</td>
<td>1.13</td>
<td>1.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>7.92</td>
<td>1.10</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>16</td>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>7.93</td>
<td>1.04</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>4.14</td>
<td>1.18</td>
<td>1.00</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>2.91</td>
<td>1.06</td>
<td>1.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>15.43</td>
<td>1.05</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>16</td>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>14.94</td>
<td>1.01</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>4.03</td>
<td>1.19</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.91</td>
<td>1.07</td>
<td>1.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>15.33</td>
<td>1.05</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>16</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>15.49</td>
<td>1.01</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>7.47</td>
<td>1.08</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>16</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>5.59</td>
<td>1.02</td>
<td>1.00</td>
<td>1.01</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>4</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>30.54</td>
<td>1.03</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>16</td>
<td>0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>30.04</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table compares the performance of alternative estimators based on 1,000 Monte Carlo draws given $\theta$. For details, see description in Section 6.1.
Table 3: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on $\eta$

<table>
<thead>
<tr>
<th>design parameters</th>
<th>mean squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>structural</td>
</tr>
<tr>
<td>$n$, $J$, $\sigma^2$, $\beta_{00}$, $\beta_{01}$, $\beta_{02}$</td>
<td></td>
</tr>
<tr>
<td>50, 4, 1.0, 1.0, 1.0, 1.0</td>
<td>0.24</td>
</tr>
<tr>
<td>50, 16, 1.0, 1.0, 1.0, 1.0</td>
<td>0.02</td>
</tr>
<tr>
<td>200, 4, 1.0, 1.0, 1.0, 1.0</td>
<td>0.18</td>
</tr>
<tr>
<td>200, 16, 1.0, 1.0, 1.0, 1.0</td>
<td>0.04</td>
</tr>
<tr>
<td>50, 4, 0.5, 1.0, 1.0, 1.0</td>
<td>0.20</td>
</tr>
<tr>
<td>50, 16, 0.5, 1.0, 1.0, 1.0</td>
<td>0.02</td>
</tr>
<tr>
<td>200, 4, 0.5, 1.0, 1.0, 1.0</td>
<td>0.16</td>
</tr>
<tr>
<td>200, 16, 0.5, 1.0, 1.0, 1.0</td>
<td>0.09</td>
</tr>
<tr>
<td>50, 4, 1.0, 1.0, 1.0, 6.0</td>
<td>3.83</td>
</tr>
<tr>
<td>50, 16, 1.0, 1.0, 1.0, 6.0</td>
<td>0.61</td>
</tr>
<tr>
<td>200, 4, 1.0, 1.0, 1.0, 6.0</td>
<td>15.04</td>
</tr>
<tr>
<td>200, 16, 1.0, 1.0, 1.0, 6.0</td>
<td>3.11</td>
</tr>
<tr>
<td>50, 4, 0.5, 1.0, 1.0, 6.0</td>
<td>6.89</td>
</tr>
<tr>
<td>50, 16, 0.5, 1.0, 1.0, 6.0</td>
<td>1.15</td>
</tr>
<tr>
<td>200, 4, 0.5, 1.0, 1.0, 6.0</td>
<td>28.41</td>
</tr>
<tr>
<td>200, 16, 0.5, 1.0, 1.0, 6.0</td>
<td>5.84</td>
</tr>
<tr>
<td>50, 4, 1.0, 0.0, 1.0, 6.0</td>
<td>4.61</td>
</tr>
<tr>
<td>50, 16, 1.0, 0.0, 1.0, 6.0</td>
<td>0.81</td>
</tr>
<tr>
<td>200, 4, 1.0, 0.0, 1.0, 6.0</td>
<td>19.37</td>
</tr>
<tr>
<td>200, 16, 1.0, 0.0, 1.0, 6.0</td>
<td>4.08</td>
</tr>
<tr>
<td>50, 4, 0.5, 0.0, 1.0, 6.0</td>
<td>9.06</td>
</tr>
<tr>
<td>50, 16, 0.5, 0.0, 1.0, 6.0</td>
<td>1.51</td>
</tr>
<tr>
<td>200, 4, 0.5, 0.0, 1.0, 6.0</td>
<td>37.64</td>
</tr>
<tr>
<td>200, 16, 0.5, 0.0, 1.0, 6.0</td>
<td>7.77</td>
</tr>
</tbody>
</table>

Notes: This table compares the performance of alternative estimators based on 1,000 Monte Carlo draws given $\eta$. For details, see description in Section 6.2.
Table 4: Mean Squared Error of alternative estimators relative to empirical Bayes, calibrated specifications, 8-type CES model

<table>
<thead>
<tr>
<th>specification def of supply</th>
<th>mean squared error</th>
<th>structural</th>
<th>unrestricted</th>
<th>emp. Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>given η</td>
<td>3.09</td>
<td>3.04</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>given θ</td>
<td>42.34</td>
<td>39.68</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table compares the performance of alternative estimators based on 1,000 Monte Carlo draws, based on specifications calibrated to our empirical results. For details, see description in Section 6.3.

Table 5: Mean Squared Error of alternative estimators relative to empirical Bayes, calibrated specifications, 2-type CES model

<table>
<thead>
<tr>
<th>specification def of supply</th>
<th>mean squared error</th>
<th>structural</th>
<th>unrestricted</th>
<th>emp. Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>given η</td>
<td>5.17</td>
<td>3.17</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>given θ</td>
<td>2.71</td>
<td>3.66</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table compares the performance of alternative estimators based on 1,000 Monte Carlo draws, based on specifications calibrated to our empirical results. For details, see description in Section 6.3.
References


