# Labor demand and wage inequality in Europe – an empirical Bayes approach*
# Preliminary and incomplete!

Fessler, Pirmin† Kasy, Maximilian‡

February 25, 2015

**Abstract**

To what extent can changes in the distribution of wages be explained by changes in labor supply of various groups (due to demographic change, migration, or expanded access to education), and to what extent are other factors (technical and institutional change) at work?

We develop a flexible methodology for answering this central question of labor economics, using an empirical Bayes approach, without imposing the restrictions on heterogeneity and on cross-elasticities of labor demand assumed by the literature. Our approach allows to reduce the variance of estimates by exploiting the information embodied in economic structural models, while avoiding the inconsistency and non-robustness of misspecified structural models. This approach also allows to overcome the issues associated with pretesting and the conventional duality of testing theories / imposing theories.

In our empirical application, we analyze changes since 2003 of the wage distribution in the countries of the European Union, using the EU-SILC data. We find ***

KEYWORDS: INEQUALITY, LABOR DEMAND, PRODUCTION FUNCTIONS, EMPIRICAL BAYES ESTIMATION, SHRINKAGE

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

## 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.

*We thank Alberto Abadie, Gary Chamberlain, and Kei Hirano for helpful discussions and comments.

†Economic Analysis Division, Österreichische Nationalbank; Address: Postbox 61, 1011 Vienna, Austria; e-Mail: pirmin.fessler@oenb.at.

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

Various explanations have been offered for this increase in wage inequality, involving factors such as 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 importance 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 such a structural model 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 very 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 considers 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 model 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 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 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. (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.

In addition to our methodological contribution, we provide new evidence on the evolution of wage inequality in Europe and the factors driving it. We use data from the EU Survey on Income and Living Conditions (EU-SILC). The EU-SILC is an annual survey conducted since 2003, which covers the "old" EU-15 member countries since 2004, and all of the EU-25, as well as some other countries, since 2005. The EU-SILC provides detailed evidence on earnings and labor supply as well as on a rich set of demographics for a representative sample of individuals from these countries.

[empirical results to be discussed here]

This paper is structured as follows: Section 1.1 provides a brief literature review. Section 2 discusses estimation methods, first reviewing structural and unrestricted estimation, discussing their drawbacks, and reviewing the general empirical Bayes approach. In section 2.5, our preferred estimator is introduced, which shrinks a preliminary unrestricted estimator towards a structural model, to an extent which depends on how well the latter appears to fit the data. The properties of our preferred estimator are then explored, and a corresponding inference procedure is proposed. Section 3.1 introduces the EU-SILC data used in this paper, and section 3.2 provides some preliminary empirical evidence, replicating the approaches taken in the literature (which mostly focuses on the United States) in the European context. Section 3.3 presents our main empirical results based on the empirical Bayes estimation procedure. Section 4 evaluates our estimation and inference procedure using a range of Monte Carlo simulations, both calibrated to the data and theoretically motivated, and evaluates the out-of-sample predictive performance of our procedure using the EU-SILC data. Section 5 concludes. Appendix A provides some additional details, and appendix B contains all proofs.

## 1.1 Related literature

This paper mainly builds on two distinct literatures: The literature on labor supply/demand and wage inequality in economics, and the literature on shrinkage and empirical Bayes estimation in statistics. 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 relevant labor literature encompasses various sub-literatures, concerned with different factors potentially affecting wage inequality (migration, 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.

Another, related, literature studies the impact of technical change on wage inequality, and in particular on the college premium. Autor et al. (1998) argue that technical change lead to a continuous rise of the relative demand for workers with college degrees, a rise which was offset partially 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. 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, thus leading to a polarization of the wage distribution.

The second literature relevant for us is the statistical literature on empirical Bayes methods and shrinkage. This literature 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 and 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 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 (Efron, 2010, chapter 1).

## 2 Estimation – structural, unrestricted, and an empirical Bayes alternative

Suppose there are $J$ types of workers, defined for instance by their level of education, age, and country of origin. Consider a cross-section of labor markets $i = 1, \ldots, n$; in our application we will focus on NUTS 2 regions in Europe. Let $Y_{j,i}$, $j = 1, \ldots, J$ be the average log wage for workers of type $j$ in labor market $i$, and let $X_{j,i}$ be the log labor supply of these same workers. Denote $X_i = (X_{1,i}, \ldots, X_{J,i})$. We are interested in the structural relationship between labor supply and wages, that is in the inverse demand function

$$Y_i = (Y_{1,i}, \ldots, Y_{J,i}) = y(X_{1,i}, \ldots, X_{J,i}, \epsilon_i),$$

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

There are various alternative ways to estimate this inverse demand function. One option, taken by the majority of contributions to the field, is to impose a tightly parametrized structural model, based on the assumptions of a parametric aggregate production function, a small number of labor-types, and wages which equal marginal productivity. Another option is to simply estimate a flexible regression model without any of the functional form restrictions imposed by the structural approach. We will argue that both approaches have serious shortcomings, and that a third option – empirical Bayes estimation, with details to be discussed below – combines some desirable features of both approaches, while avoiding their shortcomings.

We start by reviewing structural and unrestricted estimation and their shortcomings in sections 2.1 through 2.3, and the general empirical Bayes approach in section 2.4. Section 2.5 presents our proposed empirical Bayes estimator, and section 2.6 discusses its advantages. We will focus on cross-sectional data with exogenous variation of labor supply throughout; endogeneity, instruments and panel data are considered in section 2.7. Section 2.8 finally discusses the construction of empirical Bayes confidence sets.

### 2.1 Structural estimation

Let us start by reviewing the most common approach in the literature, structural estimation, and its conceptual justification.

#### Differenced estimates

Many papers in the literature run regressions of the following form; examples include Autor et al. (2008) and Card (2009).

$$Y_{j,i} - Y_{j',i} = \gamma_{j,j'} + \beta_0 \cdot (X_{j,i} - X_{j',i}) + \epsilon_{j,j',i}. \tag{1}$$

The coefficient $\beta_0$ in this regression is interpreted as the negative of the inverse elasticity of substitution between labor types $j$ and $j'$.[1] The constant

[1] The 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.

$\gamma_{j,j'}$ 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, 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, cf. appendix A.

### Justification using production function

Denote wages by $w$ and labor supply by $N$, so that $Y_{ij} = \log(w_{ij})$ and $X_{ij} = \log(N_{ij})$. The differenced regression specification can be justified based on the assumption 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 form,

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

These two assumptions together imply

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

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

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

Taking logs yields

$$Y_{j,i} - Y_{j',i} = \log(\gamma_j) - \log(\gamma_{j'}) + \beta_0 \cdot (X_{j,i} - X_{j',i}),$$

where $\beta_0 = \rho - 1$. This equation has the desired form.

### Equivalence to fixed effects regression with coefficient restrictions

There are various observationally and numerically equivalent ways to rewrite and estimate regression (1). Note first that equation (1) 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 equivalently be written in fixed effects form, including labor supply of all types $j'$ among the regressors, but imposing restrictions across coefficients:

$$Y_{j,i} = \alpha_i + \gamma_j + \sum_{j'} \beta_{j,j'} X_{j',i} + \epsilon_{j,i}, \tag{2}$$

$$\beta_{j,j'} = \beta_0 \cdot \begin{cases} \left(1 - \frac{1}{J}\right) & j = j' \\ -\frac{1}{J} & j \neq j' \end{cases} \tag{3}$$

Equation (3) can be written more compactly, in $J \times J$ matrix form, as

$$\beta = (\beta_{j,j'}) = \beta_0 \cdot \left(I_J - \tfrac{1}{J} E\right) = \beta_0 \cdot M_J, \tag{4}$$

where $I_J$ is the identity matrix, $E$ is a matrix of 1s, and $M_J$ is the demeaning-matrix, projecting $\mathbb{R}^J$ on the subspace of vectors of mean 0.

Differencing this fixed-effects regression across different values of $j$ yields specification (1), with $\gamma_{j,j'} = \gamma_j - \gamma_{j'}$ and $\epsilon_{j,j',i} = \epsilon_{j,i} - \epsilon_{j',i}$. In matrix notation, let

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

be the $(J-1) \times J$ matrix which subtracts the first entry from each component of a $J$ vector. Differencing the matrix $M$ yields $\Delta \cdot M_J = \Delta$. Pre-multiplying equation (2) by $\Delta$ yields the differenced regression in matrix form,

$$\Delta \cdot Y_i = \Delta \cdot \gamma + \beta_0 \cdot \Delta \cdot X_i + \Delta \cdot \epsilon_i.$$

It is useful to discuss the economic content of the restrictions on $\beta$ imposed by equation (4):

1. $\beta \cdot e = 0$ 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.

2. $\beta_{j,j'} = \beta_{j,j''}$ 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 $\sigma = -1/\beta_0$.

3. $\beta_{j,j} = \beta_{j',j'}$:
   The elasticity of demand is the same for all types of labor.

   In combination, these restrictions 1-3 in fact imply the CES regression model.

4. The CES model additionally implicitly entails 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.

## 2.2 Unrestricted least-squares estimation

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

$$Y_{j,i} = \alpha_i + \gamma_j + \sum_{j'} \beta_{j,j'} X_{j',i} + \epsilon_{j,i}, \tag{5}$$

using least squares, without imposing any cross-restrictions on the parameters $\beta_{j,j'}$. 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. The data are informative about the effect of labor supply on relative wages:

$$\Delta \cdot Y_i = \Delta \cdot \gamma + \delta \cdot X_i + \Delta \cdot \epsilon_i. \tag{6}$$

$$\delta = \Delta \cdot \beta \tag{7}$$

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.

Throughout the rest of this section, we will use the notation $\delta_{\uparrow}$ to denote the vectorized form of the $(J-1) \times J$ matrix $\delta$, where the rows of $\delta$ have been stacked, and similarly for other such matrices. In this vectorized notation, we have $\delta \cdot X = (I_{J-1} \otimes X') \cdot \delta$, where $\otimes$ denotes the Kronecker product. We can thus write the OLS estimator for $\delta$ based on equation (6) as solution to the least-squares problem

$$\widehat{\delta} = \operatorname*{argmin}_{d} \; E_n \left[ \| \Delta Y - (I_{J-1} \otimes (X' - E_n[X'])) \cdot d_{\uparrow} \|^2 \right]. \tag{8}$$

## 2.3 Drawbacks

### Structural estimation

There are obvious drawbacks to an approach based on the strong restrictions implied by the CES model, or by its generalizations as reviewed in appendix A. The estimates will in particular be inconsistent if the model is misspecified. The following proposition provides an explicit characterization of misspecification bias.

**Proposition 1 (Misspecification)**

- *Suppose we observed i.i.d. draws of the $J$-vectors $X$ and $Y$. Suppose that these random vectors have finite joint second moments such that* $\det(\mathrm{Var}(X)) \neq 0$.

- *Let* $\widehat{\beta}_0$ *be the least squares estimator of the structural model in equation* (4)*, and* $\widehat{\delta}$ *the least squares estimator of the differenced unrestricted model in equation* (6)*.*

- *Let* $\beta_0$ *be the probability limit, as sample size goes to infinity, of* $\widehat{\beta}_0$*, and let* $\delta$ *be the probability limit of* $\widehat{\delta}$*.*

*Then we can write* $\beta_0$ *as*

$$\beta_0 = \underset{b_0}{\operatorname{argmin}} \ \|b_0 \cdot \Delta - \delta\|_\delta, \tag{9}$$

*where*

$$\|d\|_\delta := \left(d'_\uparrow \cdot (I_{J-1} \otimes \operatorname{Var}(X)) \cdot d_\uparrow\right)^{1/2}. \tag{10}$$

*In words,* $\beta_0 \cdot \Delta$ *is the orthogonal projection of* $\delta$ *onto the subspace of multiples of* $\Delta$ *with respect to the norm* $\|d\|_\delta$ *on* $\mathbb{R}^{(J-1)\times J}$*.*

The proof of this proposition can be found in appendix B. The result is easily generalized to other structural models, which impose for instance that $\beta = \beta_1 \cdot M_1 + \beta_2 \cdot M_2$ for some matrices $M_1$ and $M_2$. The matrix defining the norm $\|d\|_\delta$ is block-diagonal.

The bias induced by functional form choices in the structural model is not only a theoretical problem, but of practical importance in various contexts. This is reflected in non-robust findings, where qualitative conclusions depend on the specifics of the functional form assumptions imposed.

Card (2009, p5f) discusses an important example, the estimated impact of past migration on wage inequality in the US. 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[2] that (i) natives and migrants are imperfect substitutes, while (ii) high school dropouts and high school graduates are perfect substitutes.

We can interpret these diverging results in light of proposition 1. Suppose that types 1 and 2 (dropouts and high school graduates) are in fact perfect substitutes, and that the share of type 1 in the population is small. This implies a coefficient $\beta_{1,1}$ close to 0. Suppose that for other types $j$, the own-elasticity is negative, $\beta_{j,j} \ll 0$. The structural CES-model imposes all own-elasticities to be the same, so that $\widehat{\beta}_0 \ll 0$. An increase of the population of type 1 is then predicted to depress type 1's wages significantly, in contrast to what the correct, unrestricted model would have predicted.

[2] Card argues that these assumptions are justified by statistical tests.

### Unrestricted estimation

The key drawback of estimating an unrestricted model, on the other hand, is its large variance. Fitting the differenced model requires the estimation of $J^2$ parameters (including the fixed effects $\Delta \cdot \gamma$), using observations of only $n \cdot (J-1)$ outcomes. When the number $n$ of cross-sectional units is not much larger than the number $J$ of types, least squares will tend to over-fit, producing estimates with a very large variance. When the number of types exceeds the number of cross-sectional units, the model is actually not identified anymore. Presumably this is the main reason why the literature resorts to highly restrictive structural models, which reduce variance by heavily reducing the number of parameters to be estimated.

## 2.4 Empirical Bayes estimation

We have discussed two approaches to estimation, one imposing a lot of restrictions based on some structural model, and one leaving the model rather unrestricted. We have argued that both of these have serious disadvantages, in theory as well as in practice.

There is a paradigm in statistics, called empirical Bayes estimation, which can in many ways be seen as providing 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:[3]

$$Y|\eta \sim f(Y|\eta) \tag{11}$$

$$\eta \sim \pi(\eta|\theta), \tag{12}$$

where both $f$ and $\pi$ describe parametric families of distributions, and where usually $\dim(\theta) \leq \dim(\eta) - 2$. Equation (11) describes the unrestricted model for the distribution of the data given the full set of parameters $\eta$. Equation (12) 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 obtained by integrating out over the distribution of the parameters $\eta$:

$$Y|\theta \sim g(Y|\theta) := \int f(Y|\eta)\pi(\eta|\theta)d\eta. \tag{13}$$

In models with suitable conjugacy properties, such as the one we will consider below, this likelihood can be obtained in closed form. A natural estimator for $\theta$ is obtained by maximum likelihood,

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}}\ g(Y|\theta). \tag{14}$$

[3]All of the following probability statements are *conditional* on our regressors $X$

Other estimators for $\theta$ are conceivable and commonly used, as well. In the second step, $\eta$ is estimated as the "posterior expectation"[4] of $\eta$ given $Y$ and $\theta$, substituting the estimate $\widehat{\theta}$ for the hyper-parameter $\theta$,

$$\widehat{\eta} = E[\eta | Y, \theta = \widehat{\theta}]. \tag{15}$$

The general empirical Bayes approach includes fully Bayesian estimation as a special case, if the family of priors $\pi$ contains just one distribution. This general approach also includes unrestricted frequentist estimation, as in section 2.2, as a special case, when $\theta = \eta$. The general approach finally includes structural estimation, as in section 2.1, when again $\theta = \eta$, and the support of $\theta$ is restricted to parameter values allowed by the structural model.

The next section will specialize the empirical Bayes approach to our setting, the section thereafter will discuss the advantages of the empirical Bayes approach in this setting.

## 2.5 An empirical Bayes model for our problem

Let us now specialize the general empirical Bayes approach to the setting considered in this paper. Rather than providing a model for the distribution of the full data $Y$ given $X$, we directly model the distribution of the unrestricted OLS estimator $\widehat{\delta}$ of the differenced model, as in equation (8). This does not waste any information, allows us to easily extend our approach to panel data and instrumental variables below, and relies on a normal model which is justified asymptotically. To construct a family of priors for $\delta = \Delta \cdot \beta$, we assume that $\beta$ is equal to a set of coefficients consistent with a structural model such as the one of equation (4), plus some noise of unknown variance.

### Modeling $\widehat{\delta}$

We assume that the unrestricted OLS estimator $\widehat{\delta}$ is normally distributed given the true fixed effects and coefficients, unbiased for the true coefficient matrix $\delta$, and has a variance $V$:

$$\widehat{\delta}_{\uparrow} | \eta \sim N(\delta_{\uparrow}, V) \tag{16}$$

This assumption can be justified by conventional asymptotics, letting the number $n$ of cross-sectional units go to infinity. This assumption also holds for the panel data and instrumental variables models discussed below. We further assume that we have a consistent estimator $\widehat{V}$ of $V$, i.e.

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

We will use an estimator $\widehat{V}$ robust to clustering at the level of cross-sectional units $i$; appendix A provides a brief discussion.

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

**Prior distributions**

Let us now turn to specifying a family of "prior distributions." We model $\beta$ as corresponding to the coefficients of the structural CES model plus some disturbances, that is

$$\beta = (\beta_{j,j'}) = \beta_0 \cdot M_J + \zeta$$
$$\zeta_{j,j'} \sim^{iid} N(0, \tau^2),$$

where, as before, $M = \left(I_J - \frac{1}{J}E\right)$. Differencing this model yields

$$\delta = \Delta \cdot \beta = \beta_0 \cdot \Delta + \Delta \cdot \zeta \tag{17}$$

The term $\beta_0 \cdot \Delta$ is equal to a fixed scalar $\beta_0$ times $\Delta \cdot M_J = \Delta$, and 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 $\Delta$, reflecting the fact that we are estimating coefficients of the differenced model. Since $\Delta \cdot \Delta' = I_J + E =: P$, the variance of this term is given by

$$\text{Var}((\Delta \cdot \zeta)_\uparrow) = \tau^2 \cdot I_{J-1} \otimes P.$$

If we were to set $\tau^2 = 0$, the empirical Bayes approach would reduce to the structural CES 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 a CES model fits the data.

We can summarize our model as follows, using the same vectorized notation as before:

$$\begin{aligned}
\eta &= (\delta, V) \\
\theta &= (\beta_0, \tau^2, V) \\
\widehat{\delta}_\uparrow|\eta &\sim N(\delta_\uparrow, V) \\
\delta_\uparrow|\theta &\sim N(\beta_0 \cdot \Delta_\uparrow, \tau^2 \cdot I_{J-1} \otimes P)
\end{aligned} \tag{18}$$

The variance of $\delta_\uparrow$ in the last line is block-diagonal and equal to the variance of the vectorized matrix $(\Delta \cdot \zeta)_\uparrow$.

**Solving for the empirical Bayes estimator**

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

$$\widehat{\delta}_\uparrow|\theta \sim N(\beta_0 \cdot \Delta_\uparrow, \Sigma(\tau^2, V)), \tag{19}$$

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

$$\begin{aligned}
\Sigma(\tau^2, V) = \text{Var}\left(\widehat{\delta}_\uparrow\right) &= \text{Var}\left(E\left[\widehat{\delta}_\uparrow|\eta\right]\right) + E\left[\text{Var}\left(\widehat{\delta}_\uparrow|\eta\right)\right] \\
&= \tau^2 \cdot I_{J-1} \otimes P + V.
\end{aligned}$$

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

$$(\widehat{\beta}_0, \widehat{\tau}^2) = \underset{b_0, t^2}{\operatorname{argmin}} \log\left(\det(\Sigma(t^2, \widehat{V}))\right) + (\widehat{\delta}_\uparrow - b_0 \cdot \Delta_\uparrow)' \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot (\widehat{\delta}_\uparrow - b_0 \cdot \Delta_\uparrow). \tag{20}$$

We can simplify this optimization problem by "concentrating out" $b_0$. Given $t^2$, the optimal $b_0$ is easily seen to equal

$$\widehat{\beta}_0 = (\Delta \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot \Delta')^{-1} \cdot \Delta \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot \widehat{\delta}_\uparrow. \tag{21}$$

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

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

$$\widehat{\delta}_\uparrow^{EB} = \widehat{\beta}_0 \cdot \Delta_\uparrow + I_{J-1} \otimes P \cdot \left(I_{J-1} \otimes P + \frac{1}{\widehat{\tau}^2} \widehat{V}\right)^{-1} \cdot (\widehat{\delta}_\uparrow - \widehat{\beta}_0 \cdot \Delta_\uparrow) \tag{22}$$

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

**Discussion**

- Our approach is based upon directly modeling the distribution of the estimated OLS coefficients $\widehat{\delta}$. There is a one-to-one mapping between $Y$ and the estimated coefficients, fixed effects $\Delta \cdot \gamma$, and residuals of the unrestricted model. To the extent that residuals and fixed effects do not carry additional information about $\delta$, our approach does not waste any information; this is true, in particular, for a standard parametric linear/normal model .

- It is instructive to relate the proposed empirical Bayes procedure to structural estimation of the CES model. First, $\widehat{\beta}_0 \cdot \Delta$ is very similar to the structural estimator of $\delta$ discussed in section 2.1, in that in both cases we are considering an orthogonal projection of the unrestricted estimator $\widehat{\delta}$ onto the subspace of multiples of $\Delta$. The projection is with respect to different norms, however. In the case of section 2.1, the projection is with respect to the norm

  $$\|d\|_\delta := \left(d'_\uparrow \cdot (I_{J-1} \otimes \mathrm{Var}(X)) \cdot d_\uparrow\right)^{1/2}$$

  (compare proposition 1), in the context of our empirical Bayes approach the projection is with respect to the norm

  $$\|d\|_{\delta,EB} = \left(d'_\uparrow \cdot \Sigma(t^2, \widehat{V})^{-1} \cdot d_\uparrow\right)^{1/2}.$$

  The two objective functions coincide (up to a multiplicative constant) if and only if (i) $\tau^2 = 0$, and (ii) $\widehat{V}$ is estimated assuming homoskedasticity.

- Second, the empirical Bayes estimator $\widehat{\delta}^{EB}$ of $\delta$ is not given by $\widehat{\beta}_0 \cdot \Delta$. Instead we can think of it as an intermediate point between $\widehat{\beta}_0 \cdot \Delta$ and the unrestricted estimator $\widehat{\delta}$. The relative weights of these two are determined by the matrices $\widehat{\tau}^2 \cdot I_{J-1} \otimes P$ and $\widehat{V}$. When $\widehat{\tau}^2$ is close to 0, we get $\widehat{\delta}^{EB} \approx \widehat{\beta}_0 \cdot \Delta$. When $\widehat{\tau}^2$ is large, we get $\widehat{\delta}^{EB} \approx \widehat{\delta}$.

- Our construction of a family of priors thus implies that, when the structural model appears to describe the data well, then our estimate of $\delta$ 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 arbitrariness and discontinuity of pre-testing procedures.

**A slightly more general model**

In our application, we will consider specifications involving many types $j$. For such specifications, shrinking towards the CES model seems problematic. The CES model implies that all other types of labor are complements for a given type, with the same elasticity of substitution, including types very similar in their demographics to the given type.

In a spirit close to the nested CES models, our preferred specification will thus take the following, slightly more general form.

$$\begin{aligned} \beta = (\beta_{j,j'}) &= \beta_0 \cdot M_1 + \beta_1 \cdot M_2 + \zeta \\ \zeta_{j,j'} &\sim^{iid} N(0, \tau^2), \end{aligned} \tag{23}$$

where $M_1 = M$ as before, and

$$M_{2,j,j'} = \begin{cases} -\left(\frac{1}{k_j} - \frac{1}{J}\right) & j' \in B_j \\ \frac{1}{J} & \text{else} \end{cases}, \tag{24}$$

and where $B_j$ denotes a set of size $k_j$ of types $j'$ which are considered to be similar to $j$; analogous to the "nests" in the nested CES production function. All of our previous discussion immediately generalizes to this model.

## 2.6 Advantages of empirical Bayes estimation

The proposed approach has a number of advantages relative to structural and unrestricted estimation approaches. We now provide a formal discussion of some of these advantages.

**Consistency**

In contrast to structural estimation in the misspecified case, the empirical Bayes estimator of $\delta$ is consistent as sample size goes to infinity:

**Proposition 2 (Consistency)**

- *Suppose we observed i.i.d. draws of the J-vectors $X$ and $Y$, which have finite joint second moments, where* $\det(\mathrm{Var}(X)) \neq 0$.
- *Let $\widehat{\delta}$ be the least squares estimator of the unrestricted model in equation* (5)*, and let $\delta$ be the probability limit of $\widehat{\delta}$.*
- *Let $\widehat{\delta}^{EB}$ be the empirical Bayes estimator of $\delta$ discussed in section 2.5.*

*Then*

$$\widehat{\delta}^{EB} \to^p \delta$$

*as sample size $n$ goes to infinity.*

The proof of this proposition can again be found in appendix B.

**Data-driven predictions**

Our proof of consistency relies on the fact that the variance $V$ of $\widehat{\delta}$, es well as the corresponding estimate $\widehat{V}$, go to 0. In the limiting case, the empirical Bayes estimator becomes equal to the unrestricted estimator. We shall now discuss a variant of this argument which presumes not that $\mathrm{Var}(\widehat{\delta}) \approx 0$, but instead only that the variance of the predicted value at some point $x$, $\mathrm{Var}(x' \cdot \widehat{\delta})$ is small. The following argument shows that for such values the predicted value using empirical Bayes is again close to the predicted value using unrestricted estimation – and thus also to the predicted value using the true coefficients $\delta$. This insight is particularly valuable when considering historical counterfactuals ("how much did migration affect wage inequality?"), which might rely on variation which is actually observed in the data.

Consider again the formula for the empirical Bayes estimator of $\delta$, equation (22). Slightly rearranging the expression for $\widehat{\delta}^{EB}_{\uparrow}$ in this equation, we can write it as

$$\widehat{\delta}^{EB}_{\uparrow} = \widehat{\delta}_{\uparrow} + \widehat{V} \cdot \left(\widehat{\tau}^2 \cdot I_{J-1} \otimes P + \widehat{V}\right)^{-1} \cdot (\widehat{\beta}_0 \cdot \Delta_{\uparrow} - \widehat{\delta}_{\uparrow}).$$

Consider further a direction $x$ such that

$$(I_{J-1} \otimes x') \cdot \widehat{V} \cdot (I_{J-1} \otimes x')' \approx 0,$$

noting that $(I_{J-1} \otimes x') \cdot \delta_{\uparrow} = \delta \cdot x$. Because $\widehat{V}$ is a symmetric matrix, this condition holds if and only if $(I_{J-1} \otimes x') \cdot \widehat{V} \approx 0$. For this direction $x$ we get

$$\begin{aligned}
\widehat{\delta}^{EB} \cdot x &= (I_{J-1} \otimes x') \cdot \widehat{\delta}^{EB}_{\uparrow} \\
&= (I_{J-1} \otimes x') \cdot \left[\widehat{\delta}_{\uparrow} + \widehat{V} \cdot \left(\widehat{\tau}^2 \cdot I_{J-1} \otimes P + \widehat{V}\right)^{-1} \cdot (\widehat{\beta}_0 \cdot \Delta_{\uparrow} - \widehat{\delta}_{\uparrow})\right] \\
&\approx (I_{J-1} \otimes x') \cdot \widehat{\delta}_{\uparrow} = \widehat{\delta} \cdot x.
\end{aligned}$$

For what values $x$ can we expect the condition $(I_{J-1} \otimes x') \cdot \widehat{V} \cdot (I_{J-1} \otimes x')' \approx 0$ to hold? Given the form of $\widehat{V}$ for least squares estimation (cf. appendix A), this will happen whenever $x' \cdot \mathrm{Var}_n(X)^{-1} \cdot x \approx 0$.

#### James-Stein shrinkage and dominance

Empirical Bayes estimators are generalizations of the famous James-Stein shrinkage estimator; see for instance Efron and Morris (1973), Morris (1983), and Stigler (1990). James-Stein shrinkage applies to the setting where $Y_i|\eta \sim N(\eta_i, 1)$, the goal is to estimate $\eta$, and loss is evaluated in terms of mean squared error, summed across $i$. The empirical Bayes estimator in this setting, based on a family of normal i.i.d. priors for $\eta$, caused a great deal of surprise in statistics when it was demonstrated that it *uniformly dominates* the maximum likelihood estimator $\widehat{\eta} = Y$: The empirical Bayes estimator has smaller mean squared error, no matter what the true $\eta$ is, as long as $\dim(Y) \geq 3$. This dominance result is likely to generalize to many other settings (see for instance Xie et al. 2012), though it cannot be expected to hold for all empirical Bayes estimators.

We will demonstrate numerically that dominance relative to both the unrestricted estimator and the structural estimator seems to hold for a wide range of values for $\eta$ in our setting.

### 2.7 Extensions: instrumental variables, panel data

Recall that throughout our analysis of empirical Bayes estimation we took as our point of departure some (asymptotically) normal unrestricted estimator $\widehat{\delta}$, in combination with some estimator $\widehat{V}$ of its variance. We justified this point of departure by an assumption of exogenous cross-sectional variation of labor supply $X$, which implied that $\widehat{\delta}$ could be obtained using ordinary least squares.

In this section we consider two extensions, instrumental variables and panel data, which both yield unrestricted estimators $\widehat{\delta}$ and $\widehat{V}$ satisfying the same assumptions. Based on such unrestricted estimators, all our subsequent discussion in sections 2.5 and 2.6 applies verbatim.

#### Instrumental variables

Assume that we have data generated by the structural relationship considered in section 2.2, that is

$$\begin{aligned} \Delta \cdot Y_i &= \Delta \cdot \gamma + \delta \cdot X_i + \Delta \cdot \epsilon_i, \\ \delta &= \Delta \cdot \beta. \end{aligned}$$

Before, we imposed that the regressors $X$ are exogenous, so that $\mathrm{Cov}(X_i, \Delta \cdot \epsilon_i) = 0$. Assume now instead that there are instruments $Z$ at our disposition which satisfy

$$\mathrm{Cov}(Z_i, \Delta \cdot \epsilon_i) = 0. \tag{25}$$

This condition implies the estimating equation

$$E_n \left[ Z \cdot (\Delta Y - (I_{J-1} \otimes (X' - E_n[X'])) \cdot \widehat{\delta}_\uparrow)' \right] = 0.$$

If the model is just-identified given the available instruments, so that in particular $\dim(Z) = \dim(X)$, this implies that we can estimate $\delta$ by

$$\widehat{\delta}_\uparrow = E_n \left[ Z \cdot (I_{J-1} \otimes (X' - E_n[X']))' \right] \cdot E_n \left[ Z \cdot \Delta Y \right]. \tag{26}$$

Under standard asymptotics, this gives an asymptotically normal estimator with a variance that can be consistently estimated by

$$\widehat{V} = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}' \widehat{\mathrm{Var}}(\Delta\epsilon) \mathbf{Z} (\mathbf{X}'\mathbf{Z})^{-1},$$

cf. appendix A. We are thus back to the setting imposed at the outset of section 2.5.

An interesting case arises if some of the instruments appear to be weak. In that case there are values $x$ such that $x' \cdot E_n[Z \cdot (X' - E_n[X'])] \approx 0$, which in turn implies $(I_{J-1} \otimes X') \cdot \widehat{V}^{-1} \approx 0$. This is in some sense the reverse case of the one we discussed when considering data-driven predictions:

$$\begin{aligned}
\widehat{\delta}^{EB} \cdot x &= (I_{J-1} \otimes X') \cdot \widehat{\delta}_\uparrow^{EB} \\
&= (I_{J-1} \otimes X') \cdot \Bigg[ \widehat{\beta}_0 \cdot \Delta_\uparrow + \\
&\qquad \widehat{V}^{-1} \cdot I_{J-1} \otimes P \cdot \left( I_{J-1} \otimes P \cdot \widehat{V}^{-1} + \frac{1}{\widehat{\tau}^2} I \right)^{-1} \cdot (\widehat{\delta}_\uparrow - \widehat{\beta}_0 \cdot \Delta_\uparrow) \Bigg] \\
&\approx \widehat{\beta}_0 \cdot (I_{J-1} \otimes X') \cdot \Delta_\uparrow = \widehat{\beta}_0 \cdot \Delta \cdot x.
\end{aligned}$$

We thus get that for coefficients such that the variation in the data is uninformative, predictions are driven entirely by extrapolation from well-identified coefficients based on the structural model. This argument carries over to the limiting case of underidentified models, where $\dim(Z) < \dim(X)$.

**Panel data**

If panel data are available, we can allow for additional forms of endogenous unobserved heterogeneity, such as time-invariant market-level effects, and common time-trends across markets. We could for instance consider the model

$$\begin{aligned}
\Delta \cdot Y_{it} &= \gamma_t + \gamma_i + \delta \cdot X_i + \Delta \cdot \epsilon_{it}, \\
\delta &= \Delta \cdot \beta,
\end{aligned}$$

where

$$E[\epsilon_{it} | X] = 0.$$

As before, we can estimate this model by OLS with fixed effects, and will obtain an asymptotically normal estimator $\widehat{\delta}$ as well as a corresponding estimator $\widehat{V}$ of its variance.

## 2.8 Inference

Inference in our setting is easily implemented, though conceptually somewhat subtle. We shall construct empirical Bayes confidence regions $C$ for $\delta$. Such confidence regions are required to satisfy

$$P(\delta \in C|\theta) \geq 1 - \alpha, \tag{27}$$

and were first proposed by Morris (1983) and analyzed further by Laird and Louis (1987) and Carlin and Gelfand (1990). Definition (27) arguably captures the natural notion of inference corresponding to empirical Bayes estimation. Empirical Bayes confidence regions are intermediate between frequentist confidence sets (which satisfy $P(\delta \in C|\eta) \geq 1 - \alpha$), and Bayesian credible sets (which satisfy $P(\delta \in C|Y) \geq 1 - \alpha$). The requirement of definition (27) is slightly weaker than the requirement of frequentist coverage.

We follow Laird and Louis (1987) in constructing such an inference procedure, using the bootstrap to capture sampling variation of the estimates $\widehat{\delta}^{EB}$, and posterior inference to capture uncertainty about $\delta$ given these estimates. The proposed procedure obtains a predictive distribution for $\delta$ which is similar to a posterior distribution of the form

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

but replaces the posterior for the hyperparameter $\theta$ by the distribution for $\widehat{\theta}$ obtained using the bootstrap, thus obtaining a mixture distribution $M\left(\delta|\widehat{\delta}, \widehat{V}\right)$.

Our proposed procedure can be summarized as follows:

1. Draw $r = 1, \ldots, R$ i.i.d. bootstrap samples from the empirical distribution of $(Y_i, X_i)$.

2. For each of these $R$ samples, obtain estimates

   - $\widehat{\delta}_r$ using differenced OLS,
   - $\widehat{V}_r$ using clustering-robust variance estimation,
   - and $\widehat{\beta}_{0,r}$ and $\widehat{\tau}_r^2$ by maximizing the marginal likelihood,

   as discussed in section 2.5.

3. Calculate

   - the posterior mean $\widehat{\delta}_r^{EB}$ and variance $V_r^{EB}$ for $\delta$
   - conditional on $\widehat{\delta}_r$ and $\widehat{\theta}_r$,
   - using equation (22) and

$$\begin{aligned} V_r^{EB} &= \mathrm{Var}(\delta|\widehat{\delta} = \widehat{\delta}_r, \theta = \theta_r) \\ &= \frac{1}{\widehat{\tau}^2} I_{J-1} \otimes P \cdot \left(\frac{1}{\widehat{\tau}^2} I_{J-1} \otimes P + \widehat{V}\right)^{-1} \cdot \widehat{V}. \end{aligned}$$

4. Consider the mixture distribution

$$M\left(\delta|\widehat{\delta},\widehat{V}\right) := \frac{1}{R}\sum_{r} N\left(\widehat{\delta}_r^{EB}, V_r^{EB}\right). \tag{28}$$

5. Obtain confidence intervals for components of $\delta$ using the appropriate quantiles of the mixture distribution $M\left(\delta|\widehat{\delta},\widehat{V}\right)$.

**Discussion**

Empirical Bayes confidence sets need to take into account two types of variation. This is best illustrated by first considering two invalid inference procedures, both of which ignore one of these two sources of variation. First, one might consider sets with the right coverage under the pseudo-posterior distribution, so that $P(\delta \in C|\widehat{\delta}, \theta = \widehat{\theta}) \geq 1 - \alpha$. 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 $\widehat{\delta}^{EB}$, 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 $\delta$ conditional on $\widehat{\delta}$ 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.

# 3 Empirical analysis

## 3.1 The EU-SILC data

Our empirical analysis uses the EU Survey of Income and Living Conditions (EU-SILC) data. These data are provided by Eurostat, the statistical agency of the European Union. Background on these data can be found on the website of EU-SILC[5], a very detailed description is available in Eurostat (2014). The EU-SILC project was launched in 2003 in six member states of the European Union (Belgium, Denmark, Greece, Ireland, Luxembourg and Austria) and Norway. Since 2004, the survey covers the old EU-15 member countries (except Germany, the Netherlands, the United Kingdom), as well as Estonia, Norway and Iceland. All countries of the EU-25 are covered since 2005.

The EU-SILC aims to collect comparable microdata on income, poverty, social exclusion and living conditions. EU-SILC participation is compulsory for all EU member states. The survey is based on a "common framework," defined by harmonised lists of variables, by a recommended design for implementing EU-SILC, by common requirements (for imputation, weighting, sampling errors calculation), common concepts (household and income) and classifications aiming at maximising comparability of the information produced.

The EU-SILC provides two types of annual data, cross-sectional data with variables on income, poverty, social exclusion and other living conditions, and longitudinal data pertaining to individual-level changes over time, observed periodically over a four year period. We only use the cross-sectional data. Social exclusion and housing condition information is collected mainly at the household level while labour, education and health information is obtained for all persons in the survey that are aged 16 and over. Income with detailed components is mainly collected at the personal level.

We use variables constructed in a way as close as possible to the literature, which mainly focuses on the United States and uses data from the US Current Population Survey (CPS) (Autor et al., 2008),[6] as well as from the US Census (Card, 2009). We map the variables available in the EU-SILC data to those of the models of labor supply considered in section 2 as follows:

- For our main analysis, the cross-sectional units $i$ considered are NUTS 2 regions; we perform additional analyses on the country- and EU-level, however. Most NUTS 2 regions have between 800.000 and 3 million inhabitants, regional boundaries are defined based on existing administrative subdivisions; figure 1 shows a map of all these regions.

- We employ various specifications for labor-types $j$. For our baseline results, which are replicating approaches from the literature, we classify workers by education (2 or 4 subgroups), and possibly by migrant/native status.

[5] EU-SILC home, accessed February 17 2015

[6] More specifically, the March CPS, May CPS, and Outgoing Rotation Group samples.

For our preferred specifications, based on the empirical Bayes methodology proposed, we consider various richer sets of types which classify workers additionally by age, work experience, and occupation.

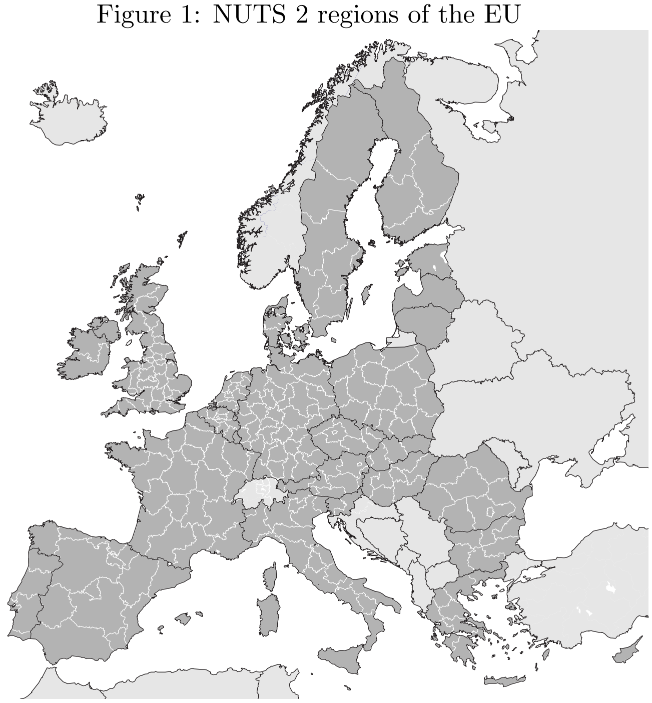
- Wages of each employed individual in the micro-data are calculated as $\frac{12}{52}$ times gross monthly earnings, divided by the number of hours usually worked per week in their main job.

  Type-specific wages $w_{j,i,t}$ are then calculated as averages (appropriately weighted using survey weights) for all individuals of a given type $j$ in region $i$ and year $t$. Outcomes $Y_{j,i,t}$ are defined as $Y_{j,i,t} = \log(w_{j,i,t})$.

- Following Card (2009), we take labor supply $N_{j,i,t}$ to equal the total hours worked per year for type $j$, region $i$, and year $t$. Regressors $X_{j,i,t}$ are defined as $X_{j,i,t} = \log(N_{j,i,t})$.

  As a robustness check, we alternatively define labor supply $N$ as the estimated total number of people of a given type in a given region and year.

Figure 1: NUTS 2 regions of the EU



**Note:** Map of the European Union NUTS 2 regions, 2007. Source: Wikipedia, January 2, 2015.

## 3.2 Replication of Card (2009) for Europe

## 3.3 Main empirical results

# 4 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 4.1 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 4.1 presents simulations corresponding to the frequentist paradigm, fixing the parameter $\eta$ and drawing from the implied distribution of the data $Y$.

We then discuss results based on our application. Section 4.3 considers simulations similar to section 4.2, but governed by parameters calibrated to match our empirical application. Section 4.4 implements split-sample exercises to evaluate the out-of-sample performance of alternative forecasting procedures.

## 4.1 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 fits 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 1 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, since 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 (large $\tau^2$). In these designs we would expect the unrestricted estimator to work well, since 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 the sense of average mean squared error given $\theta$.

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 1 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 significant improvements of performance.

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

The last subsection 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 2 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 $\beta_{02}$.

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 arbitrary 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.

The analogy to the famous result of James-Stein (that empirical Bayes dominates unrestricted estimation in the "many means" setting) would lead one to

Table 1: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on $\theta$

| design parameters | | | | | MSE relative to empirical Bayes estimation | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $J$ | $\sigma^2$ | $\beta_0$ | $\tau^2$ | structural | unrestricted | emp. Bayes | oracle e.B. |
| 50 | 4 | 1.0 | 1.0 | 0.2 | 1.53 | 1.62 | 1.00 | 0.94 |
| 50 | 16 | 1.0 | 1.0 | 0.2 | 0.82 | 1.21 | 1.00 | 1.00 |
| 200 | 4 | 1.0 | 1.0 | 0.2 | 3.69 | 1.01 | 1.00 | 0.84 |
| 200 | 16 | 1.0 | 1.0 | 0.2 | 4.17 | 1.11 | 1.00 | 1.01 |
| 50 | 4 | 0.5 | 1.0 | 0.2 | 2.28 | 1.32 | 1.00 | 0.94 |
| 50 | 16 | 0.5 | 1.0 | 0.2 | 1.54 | 1.14 | 1.00 | 1.01 |
| 200 | 4 | 0.5 | 1.0 | 0.2 | 5.28 | 0.74 | 1.00 | 0.67 |
| 200 | 16 | 0.5 | 1.0 | 0.2 | 7.86 | 1.04 | 1.00 | 1.00 |
| 50 | 4 | 1.0 | 1.0 | 0.5 | 2.39 | 1.31 | 1.00 | 0.97 |
| 50 | 16 | 1.0 | 1.0 | 0.5 | 1.56 | 1.15 | 1.00 | 1.01 |
| 200 | 4 | 1.0 | 1.0 | 0.5 | 7.78 | 1.08 | 1.00 | 0.97 |
| 200 | 16 | 1.0 | 1.0 | 0.5 | 7.89 | 1.03 | 1.00 | 1.00 |
| 50 | 4 | 0.5 | 1.0 | 0.5 | 3.95 | 1.14 | 1.00 | 0.97 |
| 50 | 16 | 0.5 | 1.0 | 0.5 | 2.91 | 1.07 | 1.00 | 1.01 |
| 200 | 4 | 0.5 | 1.0 | 0.5 | 14.87 | 1.04 | 1.00 | 0.99 |
| 200 | 16 | 0.5 | 1.0 | 0.5 | 15.15 | 1.01 | 1.00 | 1.00 |
| 50 | 4 | 1.0 | 1.0 | 1.0 | 4.06 | 1.19 | 1.00 | 0.99 |
| 50 | 16 | 1.0 | 1.0 | 1.0 | 2.94 | 1.06 | 1.00 | 1.01 |
| 200 | 4 | 1.0 | 1.0 | 1.0 | 15.47 | 1.05 | 1.00 | 1.00 |
| 200 | 16 | 1.0 | 1.0 | 1.0 | 14.94 | 1.01 | 1.00 | 1.00 |
| 50 | 4 | 0.5 | 1.0 | 1.0 | 7.22 | 1.08 | 1.00 | 0.99 |
| 50 | 16 | 0.5 | 1.0 | 1.0 | 5.53 | 1.02 | 1.00 | 1.01 |
| 200 | 4 | 0.5 | 1.0 | 1.0 | 30.13 | 1.02 | 1.00 | 1.00 |
| 200 | 16 | 0.5 | 1.0 | 1.0 | 29.97 | 1.00 | 1.00 | 1.00 |

**Notes:** This table compares the performance of alternative estimators based on 1.000 Monte Carlo draws given $\theta$. For details, see description in section 4.1.

conjecture that empirical Bayes might dominate unrestricted estimation in the present setting, as well. This seems to be the case over a very wide range of parameter values, but not uniformly so, as can be seen in table 2. Empirical Bayes compares favorably for most of the parameter space nonetheless; further exploration will be necessary to better understand its risk properties.

Table 2: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on $\eta$

| design parameters | | | | | | mean squared error | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $J$ | $\sigma^2$ | $\beta_{00}$ | $\beta_{01}$ | $\beta_{02}$ | structural | unrestricted | emp. Bayes |
| 50 | 4 | 1.0 | 1.0 | 1.0 | 1.0 | 0.41 | 3.48 | 1.00 |
| 50 | 16 | 1.0 | 1.0 | 1.0 | 1.0 | 0.02 | 1.35 | 1.00 |
| 200 | 4 | 1.0 | 1.0 | 1.0 | 1.0 | 0.75 | 6.37 | 1.00 |
| 200 | 16 | 1.0 | 1.0 | 1.0 | 1.0 | 0.09 | 5.36 | 1.00 |
| 50 | 4 | 0.5 | 1.0 | 1.0 | 1.0 | 0.54 | 4.61 | 1.00 |
| 50 | 16 | 0.5 | 1.0 | 1.0 | 1.0 | 0.02 | 1.36 | 1.00 |
| 200 | 4 | 0.5 | 1.0 | 1.0 | 1.0 | 1.33 | 11.17 | 1.00 |
| 200 | 16 | 0.5 | 1.0 | 1.0 | 1.0 | 0.09 | 5.08 | 1.00 |
| 50 | 4 | 1.0 | 1.0 | 1.0 | 6.0 | 3.91 | 1.16 | 1.00 |
| 50 | 16 | 1.0 | 1.0 | 1.0 | 6.0 | 0.61 | 1.21 | 1.00 |
| 200 | 4 | 1.0 | 1.0 | 1.0 | 6.0 | 14.65 | 1.04 | 1.00 |
| 200 | 16 | 1.0 | 1.0 | 1.0 | 6.0 | 1.30 | 0.47 | 1.00 |
| 50 | 4 | 0.5 | 1.0 | 1.0 | 6.0 | 7.15 | 1.07 | 1.00 |
| 50 | 16 | 0.5 | 1.0 | 1.0 | 6.0 | 1.14 | 1.13 | 1.00 |
| 200 | 4 | 0.5 | 1.0 | 1.0 | 6.0 | 28.82 | 1.02 | 1.00 |
| 200 | 16 | 0.5 | 1.0 | 1.0 | 6.0 | 1.18 | 0.21 | 1.00 |
| 50 | 4 | 1.0 | 0.0 | 1.0 | 6.0 | 4.67 | 1.07 | 1.00 |
| 50 | 16 | 1.0 | 0.0 | 1.0 | 6.0 | 0.80 | 1.17 | 1.00 |
| 200 | 4 | 1.0 | 0.0 | 1.0 | 6.0 | 18.92 | 1.01 | 1.00 |
| 200 | 16 | 1.0 | 0.0 | 1.0 | 6.0 | 3.83 | 1.01 | 1.00 |
| 50 | 4 | 0.5 | 0.0 | 1.0 | 6.0 | 8.90 | 1.01 | 1.00 |
| 50 | 16 | 0.5 | 0.0 | 1.0 | 6.0 | 1.50 | 1.11 | 1.00 |
| 200 | 4 | 0.5 | 0.0 | 1.0 | 6.0 | 37.52 | 1.00 | 1.00 |
| 200 | 16 | 0.5 | 0.0 | 1.0 | 6.0 | 5.02 | 0.67 | 1.00 |

**Notes:** This table compares the performance of alternative estimators based on 1.000 Monte Carlo draws given $\eta$. For details, see description in section 4.2.

### 4.3 Calibrated Monte Carlo simulations

### 4.4 Split sample results

## 5 Conclusion

# A Some additional details

### Generalizations of the CES approach

One straightforward generalization of the structural regression approach motivated by CES production functions replaces labor types $j$ by aggregated labor types, including various subtypes which are assumed to be perfectly substitutable. Such a generalization replaces log labor supply $X_j$ by

$$\tilde{X}_k = \log\left(\sum_{j:k_j=k} \theta_j N_j\right),$$

where the aggregate labor type $k$ includes subtypes $j$, which are aggregated linearly using productivity weights $\theta_j$.

Another generalization, motivated by nested CES production functions, is based on aggregate labor types as well, but assumes that subtypes are imperfectly substitutable. Such an approach would estimate regressions of the form

$$Y_{j,i} - Y_{j',i} = \gamma_{j,j'} + \beta \cdot (X_{j,i} - X_{j',i}) + \gamma \cdot (\tilde{X}_{k_j,i} - \tilde{X}_{k_{j'},i}) + \epsilon_{j,j',i},$$

where

$$\tilde{X}_k = \log\left(\left(\sum_{j:k_j=k} \theta_j N_j^{\rho_2}\right)^{1/\rho_2}\right),$$

and $\rho_2 = 1 + \beta$. This regression can be estimated in two steps, first considering comparisons of types $j$ of the same aggregate type $k_j$ to get $\beta$ and $\rho_2$, then considering comparisons of types $j$ of different aggregate type.

### Estimating the variance of $\widehat{\delta}$

Let $V$ be the variance of the OLS estimator of $\widehat{\delta}$ given $\eta$. Denote by $\mathbf{X}$ the matrix stacking $(I_{J-1} \otimes (X_i' - E_n[X']))$ across cross-sectional units, and $\Delta\mathbf{Y}$ the correspondingly stacked differenced outcomes $\Delta \cdot Y_i$ so that $\widehat{\delta}_\uparrow = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\Delta\mathbf{Y})$. Then $V$ is given by

$$V = \mathrm{Var}(\widehat{\delta}|\eta) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \,\mathrm{Var}(\Delta\epsilon)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

With homoskedastic errors in the differenced regression, this simplifies to $V = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$.

If errors are uncorrelated, we can estimate $V$ by the usual heteroskedasticity robust variance estimator

$$\widehat{V} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathrm{Var}}(\Delta\epsilon)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.$$

where

$$\widehat{\mathrm{Var}}(\Delta\epsilon) = \mathrm{diag}(\mathbf{e}_{ij}^2).$$

If clustering is a concern, as is the case in our setting, we can use the blog-diagonal

$$\widehat{\mathrm{Var}}(\Delta\epsilon)_{ij,i'j'} = \begin{cases} 0 & i \neq i' \\ \mathbf{e}_{ij}\mathbf{e}_{i'j'} & i = i' \end{cases}.$$

If we were willing to impose homoskedasticity, we could take

$$\widehat{V} = \mathbf{e}'\mathbf{e}/df \cdot (\mathbf{X}'\mathbf{X})^{-1}.$$

where $df = (nJ - J^2)$.

# B Proofs

**Proof of proposition 1:**

- As discussed in section 2.2, we can rewrite either estimator as solution to a least-squares problem after projecting out location means (i.e., the fixed effects $\alpha$) and regressor means (to take care of the fixed effects $\gamma$) for each location $i$, that is, we can write

$$\widehat{\delta} = \underset{d}{\operatorname{argmin}}\ E_n\left[\|\Delta Y - (I_{J-1} \otimes (X' - E_n[X'])) \cdot d_{\uparrow}\|^2\right]$$

  and

$$\widehat{\beta}_0 = \underset{b_0}{\operatorname{argmin}}\ E_n\left[\|\Delta Y - b_0 \cdot (I_{J-1} \otimes (X' - E_n[X'])) \cdot \Delta_{\uparrow}\|^2\right],$$

  where $E_n$ denotes sample averages.

- The usual arguments for consistency of m-estimators (cf. van der Vaart 2000, chapter 3) yield probability limits of

$$\delta = \underset{b}{\operatorname{argmin}}\ E\left[\|\Delta Y - (I_{J-1} \otimes (X' - E[X'])) \cdot d_{\uparrow}\|^2\right]$$

  and

$$\beta_0 = \underset{b_0}{\operatorname{argmin}}\ E\left[\|Y - b_0 \cdot (I_{J-1} \otimes (X' - E[X'])) \cdot \Delta_{\uparrow}\|^2\right].$$

- Both probability limits are orthogonal projections. The estimand $\beta_0$ results from an orthogonal projection on a linear subspace of the space projected onto for the unrestricted estimator $\delta$. The law of iterated projections thus yields

$$\beta_0 = \underset{b_0}{\operatorname{argmin}}\ E\left[\|(I_{J-1} \otimes (X' - E[X']) \cdot (\delta_{\uparrow} - b_0 \cdot \Delta_{\uparrow})\|^2\right],$$

  which shows that our claim holds for

$$\|d\|_{\delta}^2 = d'_{\uparrow} \cdot E\left[(I_{J-1} \otimes (X' - E[X']))' \cdot (I_{J-1} \otimes (X' - E[X']))\right] \cdot d_{\uparrow}.$$

- Algebraic manipulation of this expression finally yields

$$\|d\|_\delta^2 := d_\uparrow' \cdot (I_{J-1} \otimes \mathrm{Var}(X)) \cdot d_\uparrow.$$

□

**Proof of proposition 2:**

- By definition of $\delta$ we have $\widehat{\delta} \to^p \delta$. For the usual reasons, we have $V = \mathrm{Var}(\widehat{\beta}) = \frac{1}{n}V_1$, and thus $\widehat{V} = O_p(1/n)$.
- By the standard arguments for consistency of m-estimators van der Vaart (2000, chapter 3), we get convergence of the hyperparameters,

$$(\widehat{\beta}_0, \widehat{\tau}^2) \to^p \underset{b_0, t^2}{\operatorname{argmin}} \log\left(\det(\Sigma(t^2, 0))\right) \\ + (\widehat{\delta}_\uparrow - b_0 \cdot \Delta_\uparrow)' \cdot \Sigma(t^2, 0)^{-1} \cdot (\widehat{\delta}_\uparrow - b_0 \cdot \Delta_\uparrow).$$

  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 $\widehat{\beta}$ and $\widehat{V}$.
- Combining these results ($p\lim \widehat{\tau}^2 > 0$, $p\lim \widehat{V} = 0$, and $p\lim \widehat{\delta} = \delta$), the claim follows from

$$\widehat{\delta}_\uparrow^{EB} = \widehat{\beta}_0 \cdot \Delta_\uparrow + I_{J-1} \otimes P \cdot \left(I_{J-1} \otimes P + \frac{1}{\widehat{\tau}^2}\widehat{V}\right)^{-1} \cdot (\widehat{\delta}_\uparrow - \widehat{\beta}_0 \cdot \Delta_\uparrow)$$

  by Slutsky's theorem.

□

# References

Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Autor, D., Katz, L., et al. (1998). Computing inequality: Have computers changed the labor market? *Quarterly Journal of Economics*, 113(4):1169–1213.

Autor, D. H. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–97.

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

Borjas, G. J., Freeman, R. B., and Katz, L. F. (1996). Searching for the effect of immigration on the labor market. *The American Economic Review*, 86(2):pp. 246–251.

Card, D. (1990). The impact of the Mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review*, 43(2):245–257.

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics*, 19(1):22–64.

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

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

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

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

Eurostat (2014). Working paper with the description of the 'income and living conditions dataset'. Technical report, Eurostat.

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

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

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

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

Robbins, H. (1956). An empirical bayes approach to statistics.

Stigler, S. M. (1990). The 1988 neyman memorial lecture: a galtonian perspective on shrinkage estimators. *Statistical Science*, pages 147–155.

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

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