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

Fessler, Pirmin[†] Kasy, Maximilian[‡]

May 18, 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. We characterize the geometry and the mean squared error of our estimator. One of our key theoretical results explicitly describes the risk-function of empirical Bayes under an asymptotic approximation. Simulations confirm our characterizations and the fact that our estimator uniformly dominates unrestricted estimation over a large space of parameter values.

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.

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

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

[*] We thank Alberto Abadie, Gary Chamberlain, Kei Hirano, and José Montiel Olea, as well as seminar participants at Duke, Harvard, UBC, and SFU 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.

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 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 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 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, 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.[1] (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]

We also provide novel insights on the behavior of empirical Bayes estimators. In section 4, we 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 4.3, which provides an explicit approximation of the risk function (mean squared error) of empirical Bayes based on an asymptotic argument for large dimensions. This asymptotic approximation is valid whenever the tuning parameter is estimated with small variance relative to the parameters of interest. 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.

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. We introduce our preferred estimator in section 2.4. This estimator shrinks a preliminary unrestricted estimator towards a structural model, to an extent which depends on how well the latter appears to fit the data. We then briefly explore the properties of our preferred estimator and propose a corresponding inference procedure. 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 explores the geometry of our proposed estimator and provides a theoretical characterization of its risk properties. Section 5 evaluates our estimation and

[1] It is possible to construct counterexamples, however, see section 4.3.

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 6 concludes. Appendix A 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 (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. D'Amuri and Peri (2015), studying European evidence like the present paper, 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 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. Goldin and Katz (2009) provide an extensive historical analysis of wage inequality in the US 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, 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). In section 4 we provide a theoretical characterization of the risk properties of our empirical Bayes procedure. This characterization relies on arguments similar to those invoked by Xie et al. (2012).

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 $i = 1, \ldots, n$.[2] 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 $Y_i = (Y_{1,i}, \ldots, Y_{J,i})$ and $X_i = (X_{1,i}, \ldots, X_{J,i})$. We are interested in the structural relationship between labor supply and wages, that is in the inverse demand function

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

where ϵ_i denotes a vector of unobserved demand shifters of unrestricted dimension.

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 in sections 2.1 and 2.2, and the general empirical Bayes approach in section 2.3. Section 2.4 presents our proposed empirical Bayes estimator, and section 2.5 discusses its advantages relative to structural and unrestricted estimation. We will initially focus on cross-sectional data with exogenous variation of labor supply; endogeneity, instruments and panel data are considered in section 2.6. Section 2.7 finally discusses the construction of empirical Bayes confidence sets. In section 4 we will further explore the geometry and the risk properties of our empirical Bayes estimator.

2.1 Structural estimation

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

Differenced estimates

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

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

[2] Card (2009) considers metropolitan statistical areas in the US. In our application we focus on NUTS 1 regions in Europe.

The coefficient β_0 in this regression is interpreted as the negative of the inverse elasticity of substitution between labor types j and j'.[3] The constant $\gamma_{j,j'}$ 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, 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.

Justification using production functions

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 of equation (1) can be justified based on the assumption that wages equal marginal productivity for some aggregate production function f,

$$w_{ij} = \frac{\partial f_i(N_{i1}, \ldots, N_{iJ})}{\partial N_{ij}}, \tag{2}$$

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

$$f_i(N_{i1}, \ldots, N_{iJ}) = \left(\sum_{j=1}^{J} \gamma_j N_{ij}^{\rho} \right)^{1/\rho}. \tag{3}$$

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

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

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

[3] The elasticity of substitution σ is defined as the relative change in the demand for different factors induced by a given change in their relative prices.

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:

$$Y_{j,i} = \alpha_i + \gamma_j + \sum_{j'} \beta_{j,j'} X_{j',i} + \epsilon_{j,i}, \tag{4}$$

$$\beta_{j,j'} = \beta_0 \cdot \begin{cases} \left(1 - \frac{1}{J}\right) & j = j' \\ -\frac{1}{J} & j \neq j' \end{cases} \tag{5}$$

Equation (5) can be written more compactly, in $J \times J$ matrix form, as

$$\beta = (\beta_{j,j'}) = \beta_0 \cdot \left(I_J - \tfrac{1}{J} E\right) = \beta_0 \cdot M_J, \tag{6}$$

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 (4) by Δ yields the differenced regression in matrix form,

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

2.2 Unrestricted least-squares estimation

Rather than imposing the 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

$$Y_{j,i} = \alpha_i + \gamma_j + \sum_{j'} \beta_{j,j'} X_{j',i} + \epsilon_{j,i}, \tag{7}$$

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 β 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:

$$\Delta \cdot Y_i = \Delta \cdot \gamma + \delta \cdot X_i + \Delta \cdot \epsilon_i. \tag{8}$$

$$\delta = \Delta \cdot \beta \tag{9}$$

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

We use the notation $\delta_\uparrow$ to denote the vectorized form of the $(J-1) \times J$ matrix δ, where the rows of δ have been stacked, and similarly for other such matrices. In this vectorized notation, we have $\delta \cdot X = (I_{J-1} \otimes X') \cdot \delta_\uparrow$, where $\otimes$ denotes the Kronecker product. We can thus write the OLS estimator for δ based on equation (8) as solution to the least-squares problem

$$\widehat{\delta} = \underset{d}{\operatorname{argmin}}\ E_n \left[\| \Delta Y - (I_{J-1} \otimes (X' - E_n[X'])) \cdot d_\uparrow \|^2 \right], \tag{10}$$

where the fixed effects γ have been taken care of by de-meaning X.

2.3 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. Both of these approaches have serious disadvantages, in theory as well as in practice, as we discuss in section 2.5 below. Estimation based on the structural model has a small variance, but yields to 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 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:[4]

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

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

where both f and π 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 η. Equation (12) describes a family of "prior distributions" for η, indexed by the hyperparameters θ.

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

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

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

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

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

Other estimators for θ are conceivable and commonly used, as well. In the second step, η is estimated as the "posterior expectation"[5] of η given Y and θ, substituting the estimate $\widehat{\theta}$ for the hyper-parameter θ,

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

The general empirical Bayes approach includes fully Bayesian estimation as a special case, if the family of priors π 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 θ is restricted to parameter values allowed by the structural model. We can think of such support restrictions as dogmatic imposition of prior beliefs, in contrast to non-dogmatic priors which have full support

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

2.4 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 an unrestricted estimator $\widehat{\delta}$ of the differenced model, as in equation (10), which might be obtained using OLS, IV, or some other method. This unrestricted estimator will then be mapped to an empirical Bayes estimator $\widehat{\delta}^{EB}$. To construct a family of priors for $\delta = \Delta \cdot \beta$, we assume that β is equal to a set of coefficients consistent with a structural model such as the one of equation (6), plus some noise of unknown variance.

[5] 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.

Modeling $\widehat{\delta}$

We assume that the unrestricted estimator $\widehat{\delta}$ is normally distributed given the true coefficients, unbiased for the true coefficient matrix δ, 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 asymptotically holds for the panel data and instrumental variables models discussed below, as well. 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.[6]

Prior distributions

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

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

where, as before, $M_J = \left(I_J - \frac{1}{J}E\right)$. 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 β_0 times $\Delta \cdot M_J = \Delta$. 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 ζ pre-multiplied by Δ. The variance of this term is given by

$$\mathrm{Var}((\Delta \cdot \zeta)_{\uparrow}) = \tau^2 \cdot P \otimes I_J,$$

where $P := \Delta \cdot \Delta' = I_{J-1} + E$

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

[6] 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 $\widehat{\delta}_{\uparrow} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\Delta\mathbf{Y})$. We take $\widehat{V} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathrm{Var}}(\Delta\epsilon)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. where $\widehat{\mathrm{Var}}(\Delta\epsilon)_{ij,i'j'} = \begin{cases} 0 & i \neq i' \\ \mathbf{e}_{ij}\mathbf{e}_{i'j'} & i = i' \end{cases}$.

Summarizing our model in terms of the general notation introduced in section 2.3, with $\widehat{\delta}_{\uparrow}$ taking the place of Y, we get:

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

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 β_0^2 and τ^2, consider the marginal distribution of $\widehat{\delta}$ given θ. This marginal distribution is normal,

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

where (leaving the conditioning on θ implicit)

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

Substituting the consistent estimator $\widehat{V}$ for V, we obtain the empirical Bayes estimators of β_0 and τ^2 as solution to the maximum (marginal) likelihood problem

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

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 δ as

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

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

Discussion

- Our approach is based upon directly modeling the distribution of the unrestricted estimator $\widehat{\delta}$. If $\widehat{\delta}$ are the coefficients of an OLS regression, there is a one-to-one mapping between (i) Y and (ii) 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 δ, 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. The empirical Bayes estimator $\widehat{\delta}^{EB}$ of δ 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 $\widehat{\tau}^2 \cdot P \otimes I_J$ 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 $\widehat{\delta}^{EB} \approx \widehat{\delta}$.

- The estimator $\widehat{\beta}_0 \cdot \Delta$ is very similar to the structural estimator of δ discussed in section 2.1; in both cases we are considering an orthogonal projection of the unrestricted estimator $\widehat{\delta}$ onto the subspace of multiples of Δ. 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 below), 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.

- 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 δ 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.

2.5 Disadvantages of structural and unrestricted estimation, and advantages of the empirical Bayes approach

Structural estimation

It is useful to discuss the economic content of the restrictions on β imposed by the structural model and summarized by equation (6):

1. $\beta \cdot e = 0$ for $e = (1, \dots, 1)$:
   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 own-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.

There are obvious drawbacks to an approach based on the strong restrictions implied by the CES model, or by its generalizations. 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 (X_i, Y_i) are i.i.d. draws from the joint distribution of the J-vectors X and Y. Suppose that (X, Y) 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* (6)*, and $\widehat{\delta}$ the least squares estimator of the differenced unrestricted model in equation* (8)*.*

- *Let β_0 be the probability limit, as sample size goes to infinity, of $\widehat{\beta}_0$, and let δ be the probability limit of $\widehat{\delta}$.*

Then we can write β_0 as

$$\beta_0 = \operatorname*{argmin}_{b_0} \|b_0 \cdot \Delta - \delta\|_\delta, \tag{23}$$

where

$$\|d\|_\delta := \left(d'_\uparrow \cdot (I_{J-1} \otimes \mathrm{Var}(X)) \cdot d_\uparrow\right)^{1/2}. \tag{24}$$

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

The proof of this proposition, and of all other mathematical results, can be found in appendix A. 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 that (i) natives and migrants are imperfect substitutes, while (ii) high school dropouts and high school graduates are perfect substitutes.[7]

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.

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.

Advantages of empirical Bayes

The approach we propose has a number of advantages relative to structural and unrestricted estimation approaches, some of which we shall discuss next. An

[7] Card argues that the assumptions of such a specification are justified by statistical tests.

additional and more extensive analysis of the underlying geometric structure and risk properties of our empirical Bayes estimator can be found in section 4.

Consistency

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

Proposition 2 (Consistency)

- *Suppose (X_i, Y_i) are i.i.d. draws from the joint distribution of the J-vectors X and Y. Suppose that (X, Y) have finite joint second moments such that* $\det(\mathrm{Var}(X)) \neq 0$.
- *Let $\widehat{\delta}$ be the least squares estimator of the unrestricted model in equation* (7)*, and let δ be the probability limit of $\widehat{\delta}$.*
- *Let $\widehat{\delta}^{EB}$ be the empirical Bayes estimator of δ discussed in section 2.4.*

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 A.

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.

Now suppose that instead of $\mathrm{Var}(\widehat{\delta}) \approx 0$ we only have that the variance $\mathrm{Var}(\widehat{\delta} \cdot x')$ of the predicted value at some point x is small. The following argument shows that for such points x the predicted value $\widehat{y}$ using empirical Bayes is again close to the predicted value using unrestricted estimation – and thus also to the predicted value using the true coefficients δ, since the latter is estimated with small variance. 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 $\widehat{\delta}^{EB}$ of δ, equation (22). Rearranging this equation, we can write $\widehat{\delta}^{EB}_{\uparrow}$ as

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

Recall that $(I_{J-1} \otimes x') \cdot \delta_{\uparrow} = \delta \cdot x$. Consider a point x such that

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

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 point x we get

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

For what points x can we expect the condition $(I_{J-1} \otimes x') \cdot \widehat{V} \cdot (I_{J-1} \otimes x')' \approx 0$ to hold? For least squares estimation, 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 η, 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 η, 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 η is, as long as $\dim(Y) \geq 3$. We show in section 4.3, using an asymptotic approximation, that this dominance result generalizes subject to some qualifications.

We will also demonstrate numerically that dominance relative to both the unrestricted estimator and the structural estimator seems to hold for a wide range of values for η in our setting; see section 5 below.

2.6 Extensions

When we introduced our empirical Bayes estimator, we took as our point of departure some (asymptotically) normal unrestricted estimator $\widehat{\delta}$, in combination with some estimator $\widehat{V}$ of its variance. This point of departure could be justified by an assumption of exogenous cross-sectional variation of labor supply X, which implies 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 2.4 and 2.5 applies verbatim. After considering IV and panel data, we also discuss an extension of the CES model which is close in spirit to the nested CES specification.

Instrumental variables

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

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

If we imposed the assumption that the regressors X are exogenous, so that $\mathrm{Cov}(X_i, \Delta \cdot \epsilon_i) = 0$, then an unrestricted estimator of δ could be obtained by cross-sectional OLS. 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

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

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

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

This is just the conventional two-stage least squares formula for suitably expanded regressors and instruments. Under standard asymptotics, this gives an asymptotically normal estimator with a variance that can be consistently estimated by $\widehat{V} = (\mathbf{Z'X})^{-1}\mathbf{Z'}\widehat{\mathrm{Var}}(\Delta\epsilon)\mathbf{Z}(\mathbf{X'Z})^{-1}$, where $\mathbf{X}$ and $\mathbf{Z}$ are regressors and instruments stacked across observations. We are thus back to the setting imposed at the outset of section 2.4.

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:

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

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 $\dim(Z) < \dim(X)$, if we allow some elements of $\widehat{V}$ to be infinite. This can be made formal by writing all expressions in terms of $\widehat{V}^{-1}$.

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

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

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

A generalization of the CES 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.

$$\begin{aligned}\beta = (\beta_{j,j'}) &= \beta_0 \cdot M_1 + \beta_1 \cdot M_2 + \zeta \\ \zeta_{j,j'} &\sim^{iid} N(0, \tau^2), \end{aligned} \tag{27}$$

where $M_1 = M$ as before, and

$$M_{2,j,j'} = \begin{cases} -\left(\frac{1}{k_j} - \frac{1}{J}\right) & j' \in B_j \\ \frac{1}{J} & \text{else} \end{cases}, \tag{28}$$

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.

2.7 Inference

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

$$P(\delta \in C|\theta) \geq 1 - \alpha, \tag{29}$$

and were first proposed by Morris (1983) and analyzed further by Laird and Louis (1987) and Carlin and Gelfand (1990). Definition (29) 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 pre-posterior inference. The requirement of definition (29) 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 δ given these estimates. The proposed procedure obtains a predictive distribution for δ 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 θ by the distribution Q_R for $\widehat{\theta}$ obtained using the bootstrap, thus obtaining a mixture distribution

$$M\left(\delta|\widehat{\delta},\widehat{V}\right) = \int P\left(\delta|\widehat{\delta},\widehat{V},\theta\right) Q_R\left(\theta|\widehat{\delta},\widehat{V}\right) d\theta. \tag{30}$$

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 (or IV, panel data with fixed effects,...),
   - $\widehat{V}_r$ using clustering-robust variance estimation,
   - and $\widehat{\theta}_r = (\widehat{\beta}_{0,r}, \widehat{\tau}^2_r)$ by maximizing the marginal likelihood,

   as discussed in section 2.4.
3. Calculate
   - the posterior mean $\widehat{\delta}^{EB}_r$ and variance V^{EB}_r for δ
   - conditional on $\widehat{\delta}_r$ and $\widehat{\theta}_r$,
   - using equation (22) and

   $$\begin{aligned} V^{EB}_r &= \mathrm{Var}(\delta|\widehat{\delta}=\widehat{\delta}_r, \theta=\widehat{\theta}_r) \\ &= \widehat{\tau}^2 \cdot P \otimes I_J - \widehat{\tau}^2 \cdot P \otimes I_J \cdot \left(\widehat{\tau}^2 \cdot P \otimes I_J + \widehat{V}\right)^{-1} \cdot \widehat{\tau}^2 \cdot P \otimes I_J \\ &= \widehat{\tau}^2 \cdot P \otimes I_J \cdot \left(\widehat{\tau}^2 \cdot P \otimes I_J + \widehat{V}\right)^{-1} \cdot \widehat{V}. \end{aligned}$$
4. Consider the mixture distribution

   $$M\left(\delta|\widehat{\delta},\widehat{V}\right) := \frac{1}{R}\sum_r N\left(\widehat{\delta}^{EB}_r, V^{EB}_r\right). \tag{31}$$
5. Obtain confidence intervals for components of δ 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 θ 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 δ conditional on $\widehat{\delta}$ and θ.

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

3 Empirical analysis – wage inequality in Europe

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[8], 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, and 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 existing literature on wage inequality, which mainly focuses on the United States and uses data from the US Current Population Survey (CPS) (Autor et al., 2008),[9] 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 1 regions; we perform additional analyses on the country- and EU-level, as well as on the level of NUTS 2 regions, however.

  Most NUTS 1 regions have between 3 million and 7 million inhabitants, most NUTS 2 regions between 800.000 and 3 million. Regional boundaries are defined based on existing administrative subdivisions; figure 1 shows maps of all these regions.

[8] EU-SILC homepage, accessed February 17 2015

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

- 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.

  For our preferred specifications, based on the empirical Bayes methodology proposed, we consider various richer sets of sub-types which classify workers additionally by age, work experience, and occupation, and use a model with super- and sub-types, as discussed toward the end of section ssec:ourempiricalBayes.

- 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 here]

3.2 Replication of Card (2009) for Europe

3.3 Main empirical results

4 The geometry of our empirical Bayes estimator and its risk function

In this section, we study the geometry of the empirical Bayes estimator proposed in section 2.4, as well as its risk properties. This estimator can be seen as providing a mapping from an unrestricted (preliminary) estimate $\widehat{\delta}$ to an empirical Bayes estimate $\widehat{\delta}^{EB}$. Understanding this mapping is key for understanding the behavior of our estimator.

To avoid dealing with distracting ancillary issues, we make the following simplifying assumptions:

1. We consider a vector of regression coefficients β for a regression that has *not* been differenced, with prior variance $\mathrm{Var}(\beta) = \tau^2 \cdot I$.
2. The variance V of the corresponding estimated coefficients $\widehat{\beta}$, conditional on β, is diagonal, $V = \mathrm{diag}(v)$.

Both of these assumptions can be achieved in more general settings through a change of basis. Under these assumptions, and assuming (solely for notational simplicity) that $\widehat{V} = V$, we get

$$\widehat{\beta}|\beta \sim N(\beta, \mathrm{diag}(v)) \tag{32}$$

$$\beta|\beta_0, \tau^2 \sim N(\beta_0 \cdot \mu, \tau^2 \cdot I), \tag{33}$$

where $\mu = M_{\uparrow}$ in the context of the structural model discussed in section 2.4. The implied marginal distribution of $\widehat{\beta}$ is given by

$$\widehat{\beta}|\beta_0, \tau^2 \sim N(\beta_0 \cdot \mu, \mathrm{diag}(v) + \tau^2 \cdot I).$$

Since both $\mathrm{Var}(\widehat{\beta}|\beta)$ and $\mathrm{Var}(\beta|\beta_0, \tau^2)$ are diagonal, we obtain the empirical Bayes estimator of β by component-wise shrinkage of $\widehat{\beta}$ toward $\widehat{\beta}_0 \cdot \mu$,

$$\widehat{\beta}^{EB} = \widehat{\beta}_0 \cdot \mu + \mathrm{diag}\left(\frac{\widehat{\tau}^2}{\widehat{\tau}^2 + v_k}\right) \cdot (\widehat{\beta} - \widehat{\beta}_0 \cdot \mu). \tag{34}$$

This expression equivalent equation (22) for an appropriate choice of coordinates.

4.1 Special case: $\mu = 0$

We shall first discuss the case where $\mu = 0$, so that we can ignore estimation of β_0. In this case, the expression for $\widehat{\beta}^{EB}$ simplifies further to

$$\widehat{\beta}^{EB} = \mathrm{diag}\left(\frac{\widehat{\tau}^2}{\widehat{\tau}^2 + v_k}\right) \cdot \widehat{\beta}.$$

This expression does not quite reveal the mapping from $\widehat{\beta}$ to $\widehat{\beta}^{EB}$, since $\widehat{\tau}^2$ itself is a function of $\widehat{\beta}$. This latter function is complicated; $\widehat{\tau}^2$ minimizes the negative log likelihood

$$\sum_k \log(t^2 + v_k) + \sum_k \frac{\widehat{\beta}_k^2}{t^2 + v_k},$$

and thus solves the first order condition

$$\sum_k \frac{1}{t^2 + v_k} = \sum_k \frac{\widehat{\beta}_k^2}{(t^2 + v_k)^2}.$$

Suppose that the minimizing value is given by $\widehat{\tau}^2$. The first order condition then implies that $\widehat{\beta}$ must be somewhere on the surface of an *ellipsoid* with semi-axes that have length

$$(\widehat{\tau}^2 + v_k) \cdot \sqrt{\sum_{k'} \frac{1}{\widehat{\tau}^2 + v_{k'}}} \tag{35}$$

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

$$\widehat{\tau}^2 \cdot \sqrt{\sum_{k'} \frac{1}{\widehat{\tau}^2 + v_{k'}}}. \tag{36}$$

Note that this value does not depend on $\widehat{\beta}$ beyond its effect on $\widehat{\tau}^2$. All estimates $\widehat{\beta}^{EB}$ corresponding to a given value of $\widehat{\tau}^2$ are on the surface of a *sphere* with this radius! Note finally that there is a natural lower boundary on $\widehat{\tau^2}$ of 0.[10] In particular, we have that $\widehat{\tau^2}$ is equal to 0 for any values of $\widehat{\beta}$ inside the ellipsoid with semi-axes of length

$$v_k \cdot \sqrt{\sum_{k'} \frac{1}{v_{k'}}}. \tag{37}$$

4.2 Visual representation

We can illustrate the mapping from $\widehat{\beta}$ to $\widehat{\tau}^2$ and $\widehat{\beta}^{EB}$ graphically when $\dim(\beta) = 2$. Suppose that $v_1 = 2$ and $v_2 = 1$. The top part of figure 2 shows $\widehat{\tau}^2$ as a function of $\widehat{\beta}$. This function is flat and equal to 0 inside the white ellipsoid; it rises smoothly and approaches a circular cone for large $\widehat{\beta}$. The bottom part of this same figure shows (i) $\widehat{\beta}^{EB} - \widehat{\beta}$ 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 3, in particular, plots, for

[10] Since we impose this boundary, our estimator resembles the positive-part James Stein estimator.

various values of $\widehat{\tau}^2$, (i) which values of $\widehat{\beta}$ would imply such values of $\widehat{\tau}^2$, and (ii) the corresponding estimates $\widehat{\beta}^{EB}$.

How can we interpret these figures? For small $\widehat{\beta}$, the estimator concludes that the "theory" is essentially correct, where the theory in this case reduces to the assumption $\beta = 0$. As $\widehat{\beta}$ gets larger, so does the estimated $\widehat{\tau}^2$ – the theory is considered less correct. Deviations from 0 in the direction of the first coordinate are weighted less heavily, since $\widehat{\beta}_1$ has a larger variance (is less precisely estimated). $\widehat{\beta}_1$ is shrunk most heavily if $\widehat{\beta}_2$ seems to confirm the theory while $\widehat{\beta}_1$ violates it moderately, as evident in the bottom right plot of figure 2. When $\widehat{\beta}$ is large, so is $\widehat{\tau}^2$, and the theory is essentially disregarded; $\widehat{\beta}^{EB}$ is basically equal to the unrestricted estimator, as evident in the bottom plots of figure 3.

[Figures 2 and 3 here]

4.3 Likelihood, loss, and risk

Our estimator is based on estimation of τ^2 using the marginal likelihood (MLLH) of $\widehat{\beta}$. Alternative ways of choosing τ^2 are conceivable, for instance using the method of moments (MOM).[11] We shall also consider the infeasible choice of τ^2 minimizing the loss (squared error) of the empirical Bayes estimator (SE-EB). The objective functions for these three alternatives are as follows, where we omit multiplicative constants and normalize by $1/K$.

$$\frac{1}{K} \cdot \sum_k \left(\log(\tau^2 + v_k) + \frac{\widehat{\beta}_k^2}{\tau^2 + v_k} \right) \qquad \text{(MLLH)} \qquad (38)$$

$$\frac{1}{K} \cdot \sum_k \left(\tau^2 + v_k - \widehat{\beta}_k^2 \right)^2 \qquad \text{(MOM)} \qquad (39)$$

$$\frac{1}{K} \cdot \sum_k \left(\frac{\tau^2}{\tau^2 + v_k} \widehat{\beta}_k - \beta_k \right)^2 \qquad \text{(SE-EB)} \qquad (40)$$

The minimizer of each of these objective functions satisfies the following first-order conditions.

$$\frac{1}{K} \cdot \sum_k \frac{1}{(\tau^2 + v_k)^2} \left(\tau^2 + v_k - \widehat{\beta}_k^2 \right) = 0 \qquad \text{(MLLH)} \qquad (41)$$

$$\frac{1}{K} \cdot \sum_k \left(\tau^2 + v_k - \widehat{\beta}_k^2 \right) = 0 \qquad \text{(MOM)} \qquad (42)$$

$$\frac{1}{K} \cdot \sum_k \frac{v_k^2}{(\tau^2 + v_k)^3} \cdot \left(\frac{\tau^2}{v_k} \left(\widehat{\beta}_k^2 - \beta_k \cdot \widehat{\beta}_k \right) - \beta_k \cdot \widehat{\beta}_k \right) = 0 \qquad \text{(SE-EB)} \qquad (43)$$

[11] Additional options are cross validation, generalized cross validation, and minimization of Stein's unbiased risk estimate. We will not discuss these here.

The frequentist expectation of each of these first order conditions, conditional on the true parameter β, is given as follows.

$$\frac{1}{K} \cdot \sum_k \frac{1}{(\tau^2 + v_k)^2} \left(\tau^2 - \beta_k^2\right) = 0 \qquad \text{(MLLH)} \qquad (44)$$

$$\frac{1}{K} \cdot \sum_k \left(\tau^2 - \beta_k^2\right) = 0 \qquad \text{(MOM)} \qquad (45)$$

$$\frac{1}{K} \cdot \sum_k \frac{v_k^2}{(\tau^2 + v_k)^3} \cdot \left(\tau^2 - \beta_k^2\right) = 0 \qquad \text{(SE-EB)} \qquad (46)$$

These expressions allow us (i) to relate our setting to the one by considered by James and Stein, and (ii) to gain a better understanding of the risk properties of our estimator. The James-Stein setting is a special case of our's, where James-Stein impose the additional restriction that all the v_k are the same (homoskedasticity). In that case maximum likelihood and method of moments yield the same estimator $\widehat{\tau}^2$, as is evident from the first order conditions (41) and (42). The original justification for the James-Stein estimator was in fact based on a method of moments argument. These expressions also give some intuition for the optimality of empirical Bayes in the homoskedastic case. In expectation, the first order condition for empirical Bayes, (44), and the first order condition for loss minimization, (46), are the same in this case.

Risk

Now consider the risk (expected squared error given β) of our empirical Bayes estimator. We motivated our empirical Bayes procedure by a comparison to two alternative estimators, the unrestricted one (corresponding to a choice of $\widehat{\tau}^2 = \infty$), and the one imposing the theory (corresponding to a choice of $\widehat{\tau}^2 = 0$.) More generally, we might compare our procedure to the class of estimators for β based on arbitrary first-stage estimators of τ^2. Clearly, none of these estimators can have lower risk than the "oracle" estimator based on a choice of τ^2 minimizing the loss (squared error) of empirical Bayes given by equation (40), and thus satisfying the first order condition (43).

The classic characterization of the risk-properties of the James-Stein estimator relies on an explicit derivation of its risk function, using the normality assumption as well as homoskedasticity (constant v_k). This derivation does not generalize to the case of heteroskedasticity (non-constant v_k).

In the more general case, where neither normality nor homoskedasticity is imposed, we can still obtain insightful characterizations of risk using the following asymptotic approximation:[12] For large dimension K and under mild regularity conditions the variability in $\widehat{\tau}^2$ can be ignored relative to the variability in $\widehat{\beta}$. This approximation relies on the consistency of maximum likelihood

[12] An elegant characterization of risk for estimators based on Stein's unbiased risk estimate using similar asymptotic arguments is discussed in Xie et al. (2012). Related intuitive arguments for the dominance of James-Stein can also be found in Stigler (1990).

(or alternative estimation approaches such as MOM) for the properly defined pseudo-true parameter $\overline{\tau}^2$. Ignoring the variability of $\widehat{\tau}^2$, we can evaluate the risk (mean squared error) of empirical Bayes at this pseudo-true parameter. Risk for non-stochastic τ^2 is given by the expectation of expression (40) given β and τ^2, as a sum of variance and squared bias,

$$MSE(\tau^2) := \sum_k \left[\left(\frac{\tau^2}{\tau^2 + v_k} \right)^2 \cdot v_k + \left(\frac{v_k}{\tau^2 + v_k} \right)^2 \cdot \beta_k^2 \right]. \tag{47}$$

For large K, our $\widehat{\tau}^2$ is (up to negligible error) equal to the maximizer of the expected log likelihood,

$$ELLH(\tau^2) := \frac{1}{K} \cdot \sum_k \left(\log(\tau^2 + v_k) + \frac{\beta_k^2 + v_k}{\tau^2 + v_k} \right) \tag{48}$$

and satisfies the first order condition (44).

The following theorem formalizes this argument and states sufficient regularity conditions which guarantee that the squared error at the ML estimate of τ^2, $SE(\widehat{\tau}^2)$, is asymptotically equivalent to the mean squared error evaluated at the pseudo true τ^2, $MSE(\overline{\tau}^2)$. The latter can be explicitly calculated and compared to the mean squared error of alternative procedures, in particular restricted and unrestricted estimation.

Theorem 1 (Asymptotic risk) *Let $\widehat{\tau}^2$ be the maximizer of the log likelihood in equation* (38)*, and $\overline{\tau}^2$ the maximizer of $ELLH(\tau^2)$. Let $SE(\tau^2)$ be the squared error in equation* (40)*.*

Suppose that $\widehat{\beta}_k$ has mean β_k and variance v_k such that $v_k \leq C_1$ and $|\beta_k| \leq C_2$ for all k, and that $E[\widehat{\beta}_k^4]/v_k^2 < C_3$. Suppose further that all $\widehat{\beta}_k$ are jointly independent. Assume that $\overline{\tau}^2 \to \tau^{*2}$ as $K \to \infty$.

Then

$$SE(\widehat{\tau}^2) - MSE(\overline{\tau}^2) \to 0, \tag{49}$$

in probability and in L^1.

Theorem 1 implies that empirical Bayes asymptotically dominates unrestricted and restricted estimation under conditions on on β and V which are easy to check algebraically:

Corollary 1 *Under the assumptions of theorem 1 and for large enough K, empirical Bayes has lower mean squared error than unrestricted estimation if*

$$MSE(\overline{\tau}^2) < MSE(\infty) = \frac{1}{K} \sum_k v_k,$$

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

$$MSE(\overline{\tau}^2) < MSE(0) = \frac{1}{K} \sum_k \beta_k^2,$$

and larger mean squared error if this inequality is reversed.

This theorem also suggests how dominance of empirical Bayes can be reversed, using the fact that the weighting of the asymptotic first order conditions (44) and (46) is different. The following corollary constructs an example.

Corollary 2 *Under the assumptions of theorem 1,*

$$v_k = \beta_k = \begin{cases} 0 & k \text{ even} \\ 2 & k \text{ odd} \end{cases} \tag{50}$$

Then

$$\begin{aligned} \widehat{\tau}^2 &\to^p 0 \\ E[SE(\widehat{\tau}^2)] &\to 2 \\ MSE(\infty) &\to 1, \end{aligned}$$

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

The preceding corollary constructs an example where *unrestricted* estimation has smaller mean squared error than empirical Bayes. The intuition behind this example is that the variation in variances v_k puts most of the weight for estimation of $\widehat{\tau}^2$ on those observations where β_k is small, leading to a small $\widehat{\tau}^2$ and large bias for those observations where β_k is large.

To construct an example where *restricted* estimation has lower mean squared error than empirical Bayes estimation, simply choose $\beta = 0$. Note however that, remarkably, this dominance does *not* hold for large K, where $\widehat{\tau}^2 \to^p 0$.

4.4 Geometry in the general case: $\mu \neq 0$

Let us now turn to the general case where $\mu \neq 0$ and estimation of β_0 has thus to be accounted for. This can be analyzed using the same "trick" as in section 4.1, where we consider $\widehat{\tau}^2$ and $\widehat{\beta}_0$ to be given and derive the corresponding sets of $\widehat{\beta}$ and $\widehat{\beta}^{EB}$.

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

$$\sum_k \frac{(\widehat{\beta}_k - \widehat{\beta}_0 \cdot \mu_k)^2}{\widehat{\tau}^2 + v_k},$$

so that

$$\widehat{\beta}_0 = \frac{\sum_k \widehat{\beta}_k \cdot \frac{1}{\widehat{\tau}^2 + v_k}}{\sum_k \mu_k \cdot \frac{1}{\widehat{\tau}^2 + v_k}}. \tag{51}$$

This equation defines a *hyperplane* in the space of $\widehat{\beta}$. As before, the first order condition for $\widehat{\tau}^2$ implies

$$\sum_k \frac{1}{\widehat{\tau}^2 + v_k} = \sum_k \frac{(\widehat{\beta}_k - \widehat{\beta}_0 \cdot \mu_k)^2}{(\widehat{\tau}^2 + v_k)^2}.$$

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

We can rephrase this argument by considering only $\widehat{\tau}^2$ to be given. Conditional on $\widehat{\tau}^2$, we get that $\widehat{\beta}$ 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

$$\left(\frac{1}{\widehat{\tau}^2 + v_1}, \ldots, \frac{1}{\widehat{\tau}^2 + v_K} \right).$$

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

5 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 5.1 presents simulations corresponding to the empirical Bayes paradigm, fixing the hyperparameter θ and drawing from the implied distributions of the parameters η and data Y. Section 5.2 presents simulations corresponding to the frequentist paradigm, fixing the parameter η and drawing from the implied distribution of the data Y.

We then discuss results based on our application. Section 5.3 considers simulations similar to section 5.2, but governed by parameters calibrated to match our empirical application. Section 5.4 implements split-sample exercises to evaluate the out-of-sample performance of alternative forecasting procedures.

5.1 Monte Carlo results, fixing θ, drawing from the distribution of η 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 θ and η. The frequentist perspective considers expected loss conditional on η, averaging just over repeated draws of the data. The empirical Bayes perspective considers expected loss averaging over η but conditional on θ. Let us first consider simulations based on the empirical Bayes perspective, where we repeatedly draw values for η (in particular, own- and cross-elasticities β), and data generated by the parameter η.

In our simulations, we vary the sample size n, the number of regressors J, the residual variance σ^2, and the parameter τ^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 σ^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 β 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 τ^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 τ^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 θ.

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 τ^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 τ^2 is assumed to be known rather than estimated. As this column shows, knowledge of τ^2 does not appear to result in any improvements of performance.

[Table 1 here]

5.2 Monte Carlo results, fixing η, drawing from the distribution of Y

The last subsection considered simulations where θ was fixed but η 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 η 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 σ^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 β_{00}, β_{01}, and β_{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 σ^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) naturally leads to the conjecture that empirical Bayes might dominate unrestricted estimation in the present setting, as well. The results in table 2 suggest that this is indeed the case; over the range of parameters considered empirical Bayes seems to uniformly dominate unrestricted estimation. These numerical results are quite encouraging. Further theoretical exploration will be necessary to see whether uniform dominance indeed holds over all parameters.

[Table 2 here]

5.3 Calibrated Monte Carlo simulations

5.4 Split sample results

6 Conclusion

A 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 α) and regressor means (to take care of the fixed effects γ) 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 β_0 results from an orthogonal projection on a linear subspace of the space projected onto for the unrestricted estimator δ. 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

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

□

Proof of proposition 2:

- By definition of δ we have $\widehat{\delta} \to^p \delta$. For the usual reasons, we have $V = \text{Var}(\widehat{\beta}) = \frac{1}{n}V_1$, 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,

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

  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

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

  by Slutsky's theorem.

□

Proof of theorem 1: We can bound

$$|SE(\widehat{\tau}^2) - MSE(\overline{\tau}^2)| \leq |SE(\widehat{\tau}^2) - MSE(\widehat{\tau}^2)| \\ + |MSE(\widehat{\tau}^2) - MSE(\overline{\tau}^2)|.$$

This decomposition implies that convergence in probability follows if we can show that

1. $\widehat{\tau}^2 - \overline{\tau}^2 \to^p 0$, which in turn (by a slight modification of theorem 5.7 in in van der Vaart (2000)) follows from
   (a) $\sup_{\tau^2} |LLH(\tau^2) - ELLH(\tau^2)| \to^p 0$
   (b) $\sup_{|\tau^2 - \overline{\tau}^2| > \epsilon} \limsup_K -ELLH(\tau^2) < -ELLH(\overline{\tau}^2)$.[13]
2. $|SE(\tau^{*2}) - MSE(\tau^{*2})| \to^p 0$
3. $MSE(.)$ is continuous at τ^{*2}.

Let us now proof these claims in turn.

[13] The minus sign is necessary since we dropped constants earlier.

1. (a) We need to show that

$$D_1 := \sup_{\tau^2} \left| \frac{1}{K} \cdot \sum_k \left(\frac{v_k}{\tau^2 + v_k} \cdot \frac{\widehat{\beta}_k^2 - E[\widehat{\beta}_k^2]}{v_k} \right) \right| \to^p 0$$

as $K \to \infty$. For a given K, assume without loss of generality that $v_1 \geq v_2 \ldots \geq v_K$, so that

$$D_1 \leq \sup_{1 \geq c_1 \geq c_2 \geq \ldots \geq c_K \geq 0} \left| \frac{1}{K} \cdot \sum_k \left(c_k \cdot \frac{\widehat{\beta}_k^2 - E[\widehat{\beta}_k^2]}{v_k} \right) \right|.$$

Lemma 2.1 of Li (1986) (or basic linear programming), implies that the supremum on the right hand side is equal to

$$\max_{\tilde{K} \leq K} \left| \frac{1}{K} \cdot \sum_{k=1}^{\tilde{K}} \left(\frac{\widehat{\beta}_k^2 - E[\widehat{\beta}_k^2]}{v_k} \right) \right|.$$

This maximum is taken over the values of a martingale, so that Doob's martingale inequality (equivalently: Kolmogorov's inequality) applies, which yields

$$\begin{aligned} P(D_1 \geq \epsilon) &\leq \frac{1}{\epsilon^2} \cdot E\left[\left(\frac{1}{K} \cdot \sum_{k=1}^{\tilde{K}} \left(\frac{\widehat{\beta}_k^2 - E[\widehat{\beta}_k^2]}{v_k} \right) \right)^2 \right] \\ &= \frac{1}{(K\epsilon)^2} \cdot \sum_k E\left[\left(\frac{\widehat{\beta}_k^2 - E[\widehat{\beta}_k^2]}{v_k} \right)^2 \right]. \end{aligned}$$

Our assumed bounds on the moments of $\widehat{\beta}_k$ now immediately yield the claim.

(b) ***

2. That $SE(\tau^2) - MSE(\tau^{2*}) \to 0$ in L^2 is immediate by our bounds on the momets of $\widehat{\beta}_k$.

3. Continuity of MSE is immediate from equation (40).

Convergence in probability then implies convergence in L^1 by theorem 2.20 in van der Vaart (2000) since SE can be bounded as follows:

$$\begin{aligned} SE(\widehat{\tau}^2) &= \frac{1}{K} \cdot \sum_k \left(\frac{\tau^2}{\tau^2 + v_k} \widehat{\beta}_k - \beta_k \right)^2 \\ &\leq \frac{1}{K} \cdot \sum_k \max\left(\left(\widehat{\beta}_k - \beta_k \right)^2, \beta_k^2 \right) \\ &\leq \frac{1}{K} \cdot \sum_k \left(\left(\widehat{\beta}_k - \beta_k \right)^2 + \beta_k^2 \right). \end{aligned}$$

This last expression has expectation $MSE(0) + MSE(\infty)$ and is uniformly integrable by our assumptions bounding the moments of $\widehat{\beta}_k$. □

Proof of corollary 1: Immediate from theorem 1. □

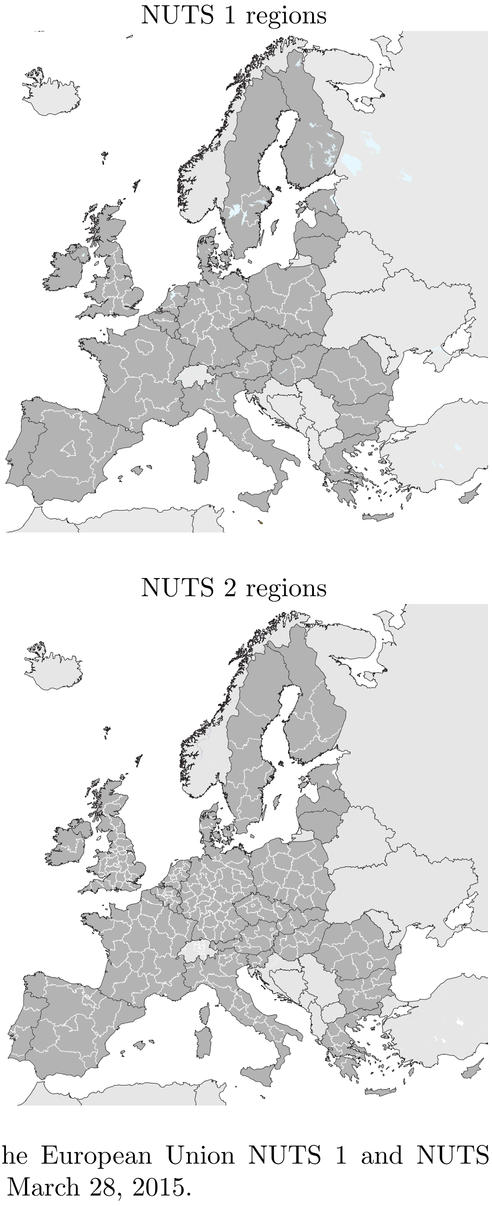
Proof of corollary 2: For even K, we have

$$\begin{aligned} \overline{\tau}^2 &= 0 \\ MSE(\overline{\tau}^2) &= \frac{1}{2}(\beta_1^2 + \beta_2^2) = 2 \\ MSE(\infty) &= \frac{1}{2}(v_1 + v_2) = 1. \end{aligned}$$

For odd K the same equations hold up to a remainder of order $1/K$. The claim now follows immediately from theorem 1. □

B Figures and tables

Figure 1: NUTS regions of the EU

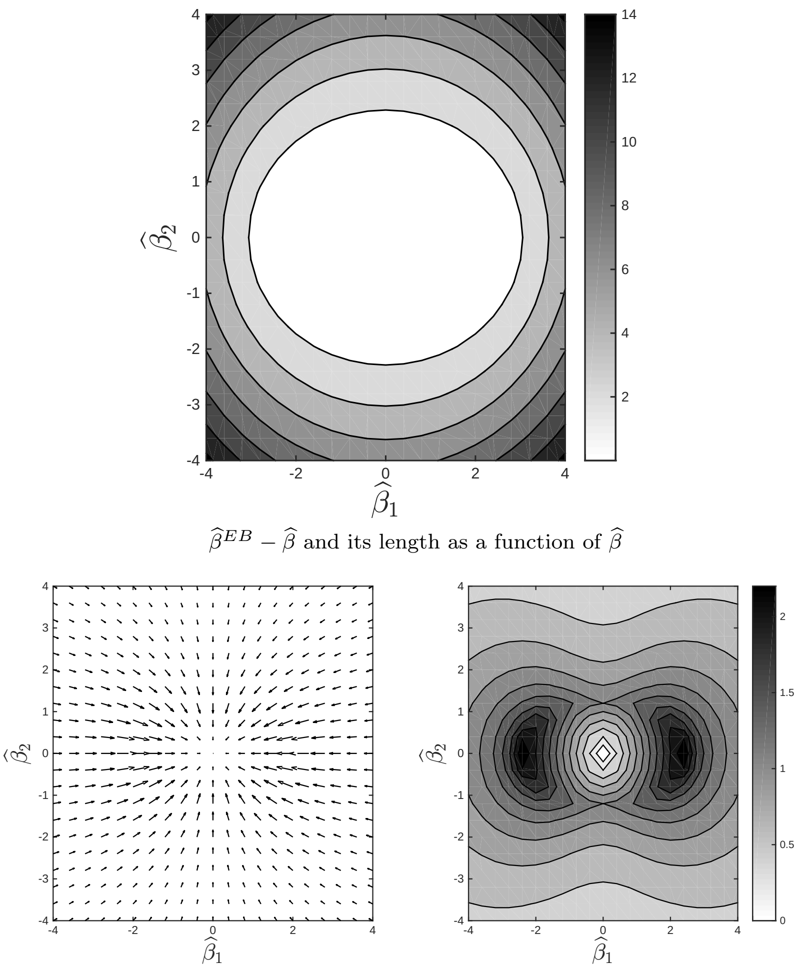


NUTS 1 regions

NUTS 2 regions

Note: Map of the European Union NUTS 1 and NUTS 2 regions, 2007. Source: Wikipedia, March 28, 2015.

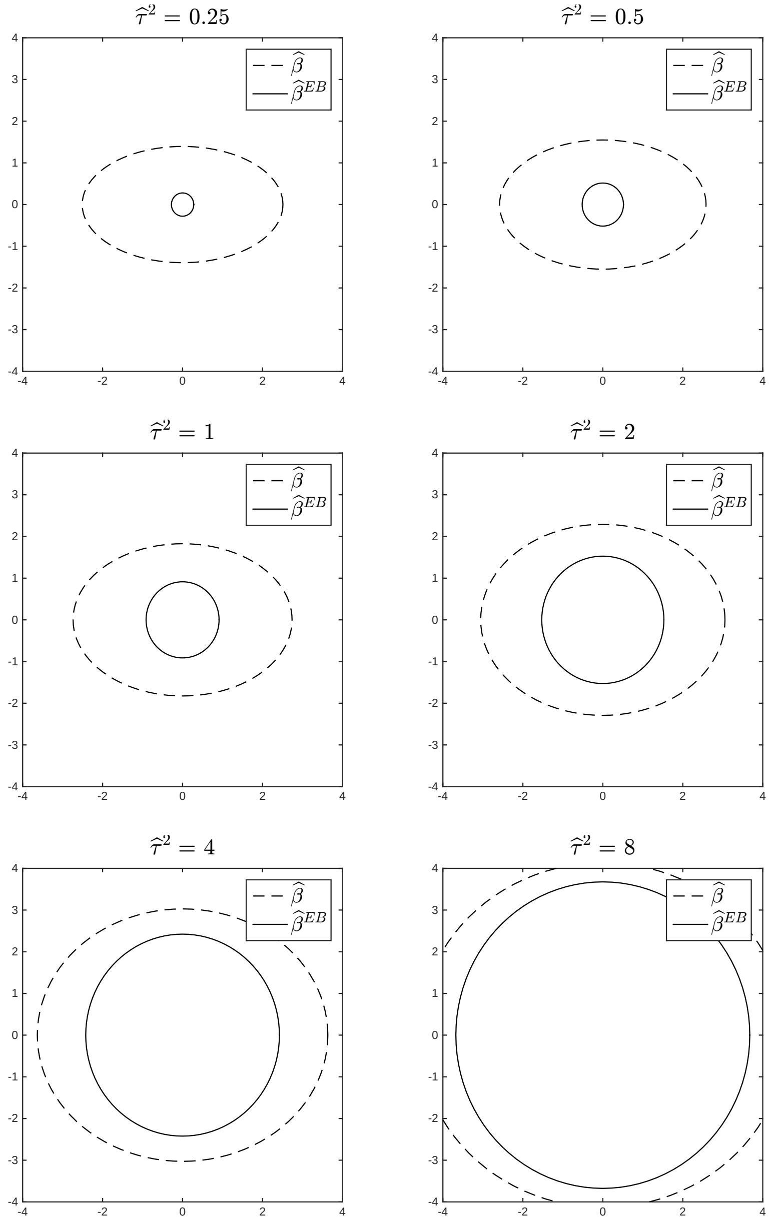
Figure 2: The mapping from $\widehat{\beta}$ to $\widehat{\tau}^2$ and $\widehat{\beta}^{EB}$



$\widehat{\tau}^2$ as a function of $\widehat{\beta}$

$\widehat{\beta}^{EB} - \widehat{\beta}$ and its length as a function of $\widehat{\beta}$

Figure 3: The geometry of empirical Bayes



$\widehat{\tau}^2 = 0.25$

$\widehat{\tau}^2 = 0.5$

$\widehat{\tau}^2 = 1$

$\widehat{\tau}^2 = 2$

$\widehat{\tau}^2 = 4$

$\widehat{\tau}^2 = 8$

Table 1: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on θ

design parameters					MSE relative to empirical Bayes estimation			
n	J	σ^2	β_0	τ^2	structural	unrestricted	emp. Bayes	oracle e.B.
50	4	1.0	1.0	0.2	1.66	1.70	1.00	0.98
50	16	1.0	1.0	0.2	0.83	1.20	1.00	1.00
200	4	1.0	1.0	0.2	4.37	1.19	1.00	0.99
200	16	1.0	1.0	0.2	4.15	1.11	1.00	1.01
50	4	0.5	1.0	0.2	2.39	1.35	1.00	0.98
50	16	0.5	1.0	0.2	1.55	1.15	1.00	1.01
200	4	0.5	1.0	0.2	8.16	1.09	1.00	0.98
200	16	0.5	1.0	0.2	7.76	1.04	1.00	1.00
50	4	1.0	1.0	0.5	2.42	1.39	1.00	0.99
50	16	1.0	1.0	0.5	1.55	1.13	1.00	1.01
200	4	1.0	1.0	0.5	7.92	1.10	1.00	1.00
200	16	1.0	1.0	0.5	7.93	1.04	1.00	1.00
50	4	0.5	1.0	0.5	4.14	1.18	1.00	0.99
50	16	0.5	1.0	0.5	2.91	1.06	1.00	1.01
200	4	0.5	1.0	0.5	15.43	1.05	1.00	1.00
200	16	0.5	1.0	0.5	14.94	1.01	1.00	1.00
50	4	1.0	1.0	1.0	4.03	1.19	1.00	1.00
50	16	1.0	1.0	1.0	2.91	1.07	1.00	1.01
200	4	1.0	1.0	1.0	15.33	1.05	1.00	1.00
200	16	1.0	1.0	1.0	15.49	1.01	1.00	1.00
50	4	0.5	1.0	1.0	7.47	1.08	1.00	1.00
50	16	0.5	1.0	1.0	5.59	1.02	1.00	1.01
200	4	0.5	1.0	1.0	30.54	1.03	1.00	1.00
200	16	0.5	1.0	1.0	30.04	1.00	1.00	1.00

Notes: This table compares the performance of alternative estimators based on 1.000 Monte Carlo draws given θ. For details, see description in section 5.1.

Table 2: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on η

design parameters						mean squared error		
n	J	σ^2	β_{00}	β_{01}	β_{02}	structural	unrestricted	emp. Bayes
50	4	1.0	1.0	1.0	1.0	0.24	2.11	1.00
50	16	1.0	1.0	1.0	1.0	0.02	1.32	1.00
200	4	1.0	1.0	1.0	1.0	0.18	1.47	1.00
200	16	1.0	1.0	1.0	1.0	0.04	2.30	1.00
50	4	0.5	1.0	1.0	1.0	0.20	1.71	1.00
50	16	0.5	1.0	1.0	1.0	0.02	1.32	1.00
200	4	0.5	1.0	1.0	1.0	0.16	1.27	1.00
200	16	0.5	1.0	1.0	1.0	0.09	5.09	1.00
50	4	1.0	1.0	1.0	6.0	3.83	1.15	1.00
50	16	1.0	1.0	1.0	6.0	0.61	1.20	1.00
200	4	1.0	1.0	1.0	6.0	15.04	1.03	1.00
200	16	1.0	1.0	1.0	6.0	3.11	1.12	1.00
50	4	0.5	1.0	1.0	6.0	6.89	1.05	1.00
50	16	0.5	1.0	1.0	6.0	1.15	1.13	1.00
200	4	0.5	1.0	1.0	6.0	28.41	1.02	1.00
200	16	0.5	1.0	1.0	6.0	5.84	1.05	1.00
50	4	1.0	0.0	1.0	6.0	4.61	1.04	1.00
50	16	1.0	0.0	1.0	6.0	0.81	1.18	1.00
200	4	1.0	0.0	1.0	6.0	19.37	1.01	1.00
200	16	1.0	0.0	1.0	6.0	4.08	1.08	1.00
50	4	0.5	0.0	1.0	6.0	9.06	0.99	1.00
50	16	0.5	0.0	1.0	6.0	1.51	1.11	1.00
200	4	0.5	0.0	1.0	6.0	37.64	1.01	1.00
200	16	0.5	0.0	1.0	6.0	7.77	1.03	1.00

Notes: This table compares the performance of alternative estimators based on 1.000 Monte Carlo draws given η. For details, see description in section 5.2.

References

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.

D'Amuri, F. and Peri, G. (2015). Immigration, jobs and labor market institutions: Evidence from europe. *Journal of European Economic Association*, forthcoming.

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.

Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology.* Harvard University Press.

Gottschalk, P. and Smeeding, T. M. (2000). Empirical evidence on income inequality in industrialized countries. volume 1 of *Handbook of Income Distribution*, chapter 5, pages 261 – 307. Elsevier.

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.

Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):pp. 1101–1112.

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.