How to use economic theory to improve estimators Supplementary Appendix

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

February 7, 2018

This appendix provides some additional discussion and results, supplementing the manuscript of "How to use economic theory to improve estimators."

In Section A of this appendix,we consider two additional applications of our proposed approach, to a general equilibrium model of financial markets, and to structural models of (consumer) preferences.

Our theoretical results suggest that the proposed empirical Bayes estimators should uniformly outperform unrestricted estimators and outperform structural (restricted) estimators for most parameter values. In Section C, we discuss some Monte Carlo simulations which do indeed confirm these predictions.

In section 3.1 of the manuscript we consider estimation of labor demand systems, shrinking toward the predictions of a CES production function model. In Section D of this appendix, we review CES production functions and derive from them the wage regressions considered in the manuscript.

In section 3.3 of the manuscript, we consider empirical Bayes estimators for choice probabilities, where estimation of the hyper-parameters involves maximization of a Dirichlet-multinomial likelihood subject to a set of linear inequality constraints. In Section E of this appendix, we discuss methods for the numerical solution of such maximization problems.

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

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

A Further applications

This section considers two additional applications of our proposed approach. Let us give a brief overview of both applications, before discussing them in greater detail.

Asset returns and the capital asset pricing model Various financial decisions such as capital budgeting and portfolio performance evaluation require precise estimates of the joint distribution of asset returns with market returns (see, e.g., Bossaerts 2013). The capital asset pricing model (CAPM) predicts that a regression of asset returns (in excess of the risk-free rate) on market returns (in excess of the risk-free rate) should have an intercept of 0 for each asset. Statistical tests of this prediction tend to reject (see, e.g., Jensen et al. 1972), but most intercepts of such regressions appear to be quite close to 0. We propose to construct empirical Bayes estimators shrinking toward this empirical prediction. Our proposed estimator remains valid in the presence of other factors explaining the cross-sectional distribution of returns and does not require the estimation of correlations between assets. We apply our estimator to monthly stock return data from the Center for Research in Security Prices (CRSP) covering the NYSE, AMEX, and NASDAQ. Our estimator achieves significant gains in out-of-sample predictive performance relative to both restricted and unrestricted OLS estimation.

Multinomial logit and mixed multinomial logit A workhorse model of discrete choice demand estimation is the multinomial logit model. A generalization of this model that allows for arbitrary patterns of substitutability is the mixed (or random coefficient) multinomial logit model, cf. Train (2009). Heterogeneity of preferences in the mixed multinomial logit model is plausibly identified when panel data of choices are available, but possibly imprecisely estimated. We propose estimation of flexible mixed multinomial logit models, shrinking the parameters governing coefficient heterogeneity towards no dispersion, as implied by the multinomial logit model.

Choice probabilities and economic decision theory Among the most general theories in economics are theories of decision making such as utility maximization, expected utility maximization, and exponential discounting. In considering such theories, we do want to allow for arbitrary preference heterogeneity across individuals. If choice sets are randomly assigned to individuals, these theories imply testable inequality restrictions on conditional choice probabilities (e.g., the stochastic axiom of revealed preference for utility maximization, cf. McFadden 2005). We provide a characterization of these restrictions for general theories of decision making, and construct an estimator of conditional choice probabilities shrinking toward these restrictions in a data-dependent way. This estimator is based on a family of Dirichlet priors centered on the simplex of conditional choice probabilities consistent with the theory of choice under consideration.

A.1 Asset returns and the capital asset pricing model

In this application we consider estimation of the joint distribution between the returns of individual financial assets and market returns. Estimation of these joint distributions is of key importance for financial decision making in various contexts, including capital budgeting and portfolio performance evaluation (see e.g. Bossaerts 2013). Estimation of these joint distributions involves high dimensional parameters of interest when we consider many different assets. In this application the "theory" that we propose shrinking to corresponds to restrictions on the joint first and second moments of financial assets implied by the capital asset pricing model (CAPM), a general equilibrium model of financial markets. These restrictions are discussed for instance in Jensen et al. (1972). Though CAPM is generally considered to be rejected by the data, it provides a useful approximation for decision making in practice. Our approach bears some resemblance to the Bayes and empirical Bayes methods in finance, reviewed in Jacquier et al. (2011). However, our approach is distinct in shrinking to the predictions of an economic theory, rather than some grand mean or similar object. A possible extension of our approach would be shrinkage toward the predictions of a multi-factor model of asset returns.

A.1.1 Setup

Consider a financial market on which assets $i \in \{1 \ldots N\}$ are traded, and assume that some risk-free asset exists on this market. Denote the market value of asset i at the beginning of period $t \in \{1, \ldots T\}$ by ω_{it}. Denote its realized return in period t, net of the risk-free rate of return, by R_{it}. Returns include both dividend payments and appreciation. Let R_t^M be the rate of return of the "market portfolio," net of the rate of return of the risk-free asset. The return of the market portfolio is the market value weighted average of the individual assets' returns. We shall assume further that returns are stationary over time. Define

$$\beta_i = \frac{\mathrm{Cov}(R_{it}, R_t^M)}{\mathrm{Var}(R_t^M)}.$$

This number β_i can be thought of as a measure of the non-diversifiable risk of asset i.

CAPM, structural and unrestricted estimation The CAPM relates the expected return of each asset i to its non-diversifiable risk. Under certain assumptions on investors' preferences, in the absence of transaction costs, and under the above restrictions, it can be shown that in general equilibrium the relationship

$$E[R_{it}] = \beta_i \cdot E[R_t^M] \tag{1}$$

holds for all assets i. This is a testable implication, and various tests have been proposed, including by Jensen (1968) and by Jensen et al. (1972). Since these early tests of CAPM, a large number of papers has appeared suggesting predictable cross-sectional variation in expected returns explained by observables or factors other than R_t^M; a comprehensive review is provided by Harvey et al. (2016). The potential presence of such predictable variation does not invalidate our approach as outlined below, and might be explicitly taken into account in extended versions of our estimator.

Consider the time series best linear predictor of R_{it} given R_t^M for each asset i separately,[1]

$$R_{it} = \alpha_i + \beta_i \cdot R_t^M + \epsilon_{it}, \tag{2}$$

where $\mathrm{Cov}(R_t^M, \epsilon_{it}) = 0$. The slope of this predictor is equal to β_i by definition, under the assumption of stationarity. Estimating the coefficients of the best linear predictor using OLS, we obtain unrestricted estimators $\left(\widehat{\alpha}_i, \widehat{\beta}_i\right)$, with estimated sampling variance $\widehat{V}_i$. Allowing for general heteroskedasticity and intertemporal dependence, we can use a heteroskedasticity and autocorrelation robust estimator for $\widehat{V}_i$. We do not need to impose any assumptions on cross-sectional dependence (across assets i) or intertemporal dependence (across t) so that our approach remains valid in the presence of further factors explaining some of the cross-sectional variation in returns.

Equation (1), which holds under the assumptions of CAPM, implies the restrictions $\alpha_i = 0$ for all i. We could obtain a restricted estimator of the parameters in Equation (2) that imposes this restriction by running a time series OLS regression of R_{it} on R_t^M with no intercept.

A.1.2 Empirical Bayes estimation, shrinking toward CAPM

In the spirit of the present paper, we do not want to test or impose the theoretical restrictions implied by CAPM. Instead we want to construct estimators of the α_i and β_i that perform particularly well if these restrictions are approximately true. The resulting estimates can then serve as inputs for financial decision making in capital budgeting, portfolio evaluation, etc.

Applying our general approach as introduced in Section 2 of the manuscript to the present setting, we propose to take the unrestricted OLS estimates $\left(\widehat{\alpha}_i, \widehat{\beta}_i\right)$ and $\widehat{V}_i$ as point of departure when constructing estimates which are shrunk toward the theory. We consider the family of priors

$$(\alpha_i, \beta_i) \sim^{iid} N\left((0, \beta^0), \Upsilon\right). \tag{3}$$

If Υ_{11} were set equal to 0, this prior would impose the restriction of Equation (1), as implied by CAPM. The parameter Υ_{11} thus takes the role that τ^2 had in the simplified setting of Section 2 of the manuscript.

In the second step of estimation we need to obtain estimates of the hyperparameters β^0 and Υ. Previously, we estimated hyperparameters via maximization of the

[1] In this section we stick with the standard finance notation of α_i and β_i, deviating slightly from our previous notation based on which we would subsume both of these, for all i, in a vector of interest β.

marginal likelihood. Such an approach is complicated in the present setting by the fact that the estimates $\left(\widehat{\alpha}_i, \widehat{\beta}_i\right)$ are correlated across i due to correlated returns across different assets, and that their covariances are hard to estimate. We can, however, easily construct method of moments estimators of β^0 and Υ that avoid the need to estimate these covariances. In particular, let

$$\widehat{\beta}^0 = \tfrac{1}{N} \sum_i \widehat{\beta}_i \tag{4}$$

and

$$\widehat{\Upsilon} = \tfrac{1}{N} \sum_i \left(\begin{pmatrix} \widehat{\alpha}_i \\ \widehat{\beta}_i - \widehat{\beta}^0 \end{pmatrix} \cdot \left(\widehat{\alpha}_i, \widehat{\beta}_i - \widehat{\beta}^0\right) - \widehat{V}_i \right). \tag{5}$$

Empirical Bayes estimates of (α_i, β_i) are then obtained in the final step via

$$(\widehat{\alpha}_i^{EB}, \widehat{\beta}_i^{EB}) = \left(0, \widehat{\beta}^0\right) + \widehat{\Upsilon} \cdot \left(\widehat{\Upsilon} + \widehat{V}_i\right)^{-1} \left(\widehat{\alpha}_i, \widehat{\beta}_i - \widehat{\beta}^0\right). \tag{6}$$

For comparison, we also consider a restricted empirical Bayes estimator. The restricted empirical Bayes estimator takes the restricted OLS estimates of the β_i as its point of departure – these already impose $\alpha_i = 0$ for all i – and is based on the family of priors

$$\beta_i \sim^{iid} N\left(\beta^0, \upsilon^2\right).$$

Restricted empirical Bayes otherwise proceeds like our preferred empirical Bayes estimator, shrinking toward the theory.

A.1.3 Empirical application

We apply this approach to data from the Center for Research in Security Prices (CRSP) NYSE/AMEX/NASDAQ monthly stock file, accessed through the Wharton Research Data Services (WRDS) web page. These data are available for the years 1926 to 2015. We consider two sub-samples, a recent 6 year sample for the years 2010 to 2015, and a sample for the years 1931-1965. The latter corresponds to the period considered by Jensen et al. (1972) and is included for comparability.

Following the literature, we use market excess returns R_t^M defined as the value-weighted return of all CRSP firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ and which have a CRSP share code of 10 or 11 at the beginning of month t, good shares and price data at the beginning of t, and good return data for t. We equate the risk-free rate, relative to which excess returns are defined, to the one-month Treasury bill rate.[2] We drop duplicates and all firms not existing for more than 2 months.

[2] For market excess return, cf. `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` accessed Sep 16 2016. Background on the CRSP data can be found at `http://www.crsp.com/products/research-products/crsp-us-stock-databases`, accessed Sep 16 2016. All data were downloaded from the UPenn Wharton Research Data Services web page, `https://wrds-web.wharton.upenn.edu/wrds/`, on Aug 9, 2016.

Table 1: One month ahead prediction MSE for asset returns, 2010-15

OLS	Restricted OLS	EB	Restricted EB
0.0255	0.0232	0.0219	0.0218

Note: This table shows the mean squared error of alternative predictors of excess asset returns R_{it} of the form $\widehat{\alpha}_i + \widehat{\beta}_i R_t^M$, where $(\widehat{\alpha}_i, \widehat{\beta}_i)$ are estimated using data for the 5-year windows $[t-60, t-1]$, starting in January 2010.

Predictive performance, 2010-15 In order to compare alternative estimators of the asset-specific parameters α_i and β_i, we consider their predictive performance. We calculate the mean squared error of alternative predictors of realized returns $R_{i,t+1}$ in period $t+1$ using market returns R_{t+1}^M and estimates of α_i, β_i, based on observations for the periods 1 through t. We repeatedly estimate α_i and β_i using 5-year windows of data and form predictions one month ahead. Thus, we predict returns for January 2015 using estimates based on returns for January 2010 to December 2014, then predict returns for February 2015 using estimates based on returns for February 2010 through January 2015, etc.

We compare four estimators, (i) unrestricted OLS, (ii) restricted OLS imposing an intercept of 0, (iii) our preferred empirical Bayes estimator shrinking to the theory, and (iv) empirical Bayes imposing an intercept of 0 and shrinking β_i to the grand mean.

Mean squared errors, averaged across assets and across time periods, are reported in Table 1. As can be seen from this table, using empirical Bayes estimators results in important reductions of prediction mean squared error both relative to unrestricted OLS and relative to restricted OLS imposing an intercept of 0 (as implied by CAPM). Our preferred estimator is the estimator shrinking to the theory, with MSE reported in column 3. This estimator essentially ties in terms of MSE with the restricted empirical Bayes estimator imposing the theory, in column 4.

Distribution of estimates for the period 2011-15 We next report estimates based on the last five years of data. For financial decision making, one would be interested in the actual asset-specific parameters α_i and β_i. For the purpose of this paper, and given the large number of assets i, we focus on estimating hyperparameters and on summarizing the distribution of alternative estimates for α_i and β_i.

Applying the method of moments estimators of equations (4) and (5) to the data for 2011-2015, we obtain estimates $\widehat{\beta}^0 = 0.96$ and

$$\widehat{\Upsilon} = \begin{pmatrix} 0.001 & -0.016 \\ -0.016 & 0.863 \end{pmatrix},$$

which implies a correlation between α and β across assets of -0.72. These estimates suggest that the predictions of CAPM are very accurate for this time period – the estimated mean square deviation $\widehat{\Upsilon}_{11}$ of α_i from 0 equals 0.0006. Recall that $\widehat{\Upsilon}_{11}$ corresponds the role of τ in the simplified setting considered in Section 2 of the manuscript and thus provides a measure of model fit.

We plot the distribution of estimates for α_i and β_i across assets i in Figure 1. In interpreting these figures, note the different scale of the axes between α and β. As suggested by the estimated $\widehat{\Upsilon}$, it appears that α has very small dispersion around 0, in line with the predictions of CAPM, and the same is true for estimators $\widehat{\alpha}_i$. Unsurprisingly empirical Bayes, our preferred estimator, as shown in the third row, delivers estimates that are less dispersed than the unrestricted OLS estimates, for both α and β. The median shrinkage factor of the OLS estimates of α toward 0 implied by the empirical Bayes estimator equals 0.82, while the median shrinkage factor of the OLS estimates of β toward $\widehat{\beta}^0$ equals 0.91. For purposes of comparison, the first row of Figure 1 shows the distribution of estimates of β for the restricted empirical Bayes estimator, which imposes $\alpha = 0$.

Figure 2 depicts the joint distribution of OLS and empirical Bayes estimates. The two are obviously positively correlated, and empirical Bayes estimates tend to be closer to the grand mean, but there is variability of the empirical Bayes estimates given the OLS estimates. This stands in contrast to component-wise linear shrinkage estimators such as the ones discussed by Hansen (2016). The bottom plot in Figure 2 shows that empirical Bayes and restricted empirical Bayes estimates are rather close to each other.

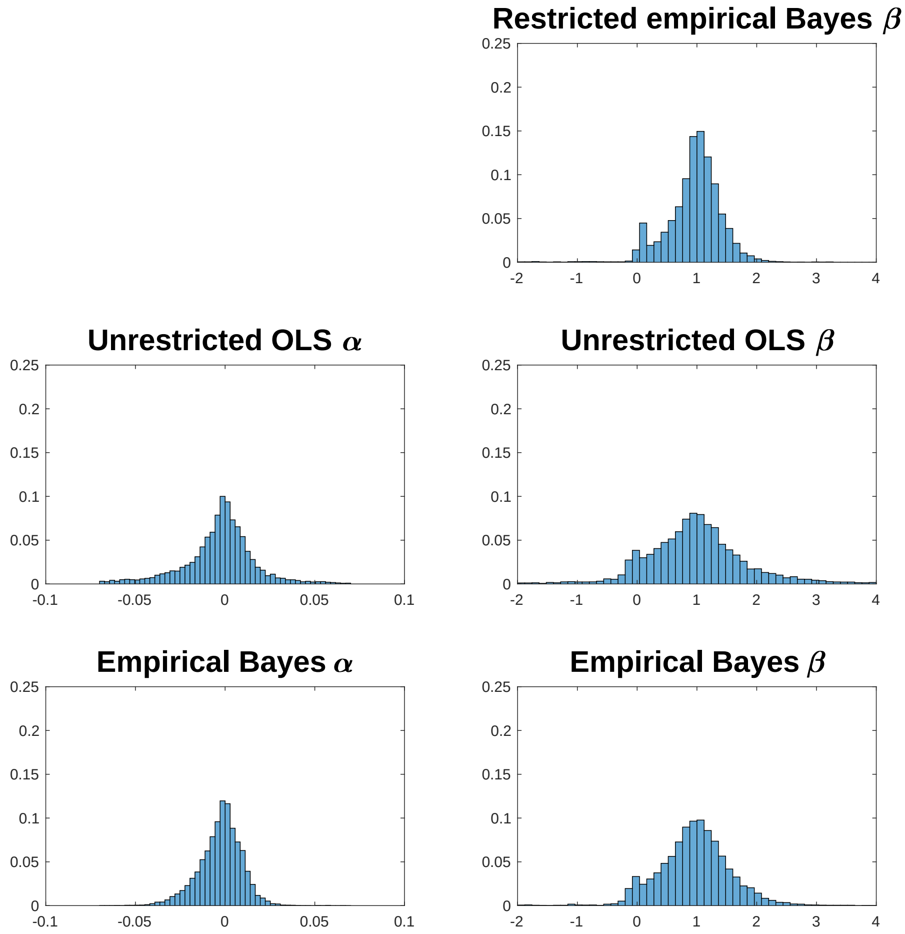
Results for the period 1931-1965 We next report similar results for the earlier period, which is the one studied by Jensen et al. (1972). For this period, the method of moments yields estimates $\widehat{\beta}^0 = 1.178$ and

$$\widehat{\Upsilon} = \begin{pmatrix} 0.000 & -0.000 \\ -0.000 & 0.197 \end{pmatrix},$$

which implies a correlation between α and β across assets of -0.05. These estimates again suggest that the predictions of CAPM are very accurate – the mean square deviation of α_i from 0 equals 0.0001.

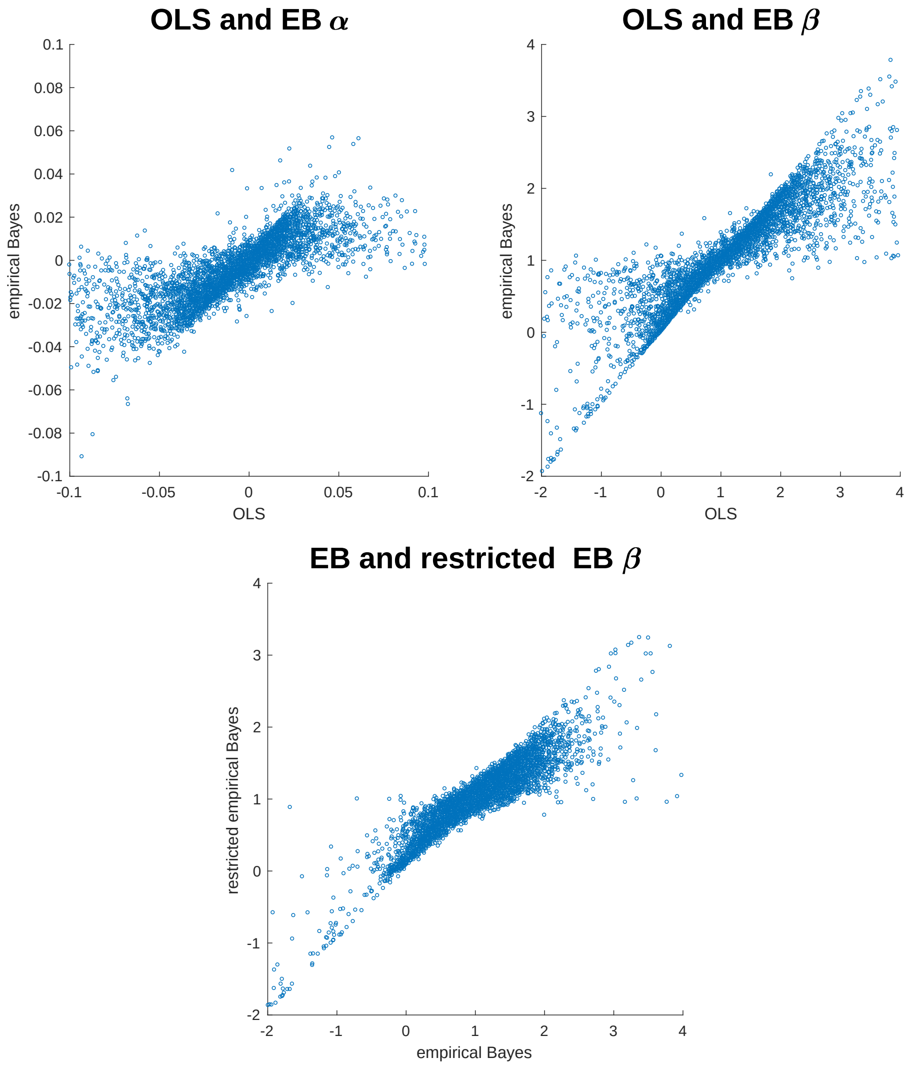
We plot the distribution of estimates for α_i and β_i across assets i in Figure 3. These plots reveal a pattern similar to the one discussed before, with more pronounced shrinkage for the empirical Bayes estimates than in the period 2011-15. The median shrinkage factor of the OLS estimates of α toward 0 implied by the empirical Bayes estimator equals 0.54, while the median shrinkage factor of the OLS estimates of β toward $\widehat{\beta}^0$ equals 0.75. Figure 4 depicts the joint distribution of OLS and empirical Bayes estimates. The pattern is again similar to the one discussed before, but shows more pronounced shrinkage.

Figure 1: Distribution of estimates of α and β across assets for the period 2011-15



Notes: These figures show histograms of the distribution of alternative estimators for α_i and β_i across assets i, as discussed in Section A.1.

Figure 2: OLS and empirical Bayes estimates of α and β for the period 2011-15



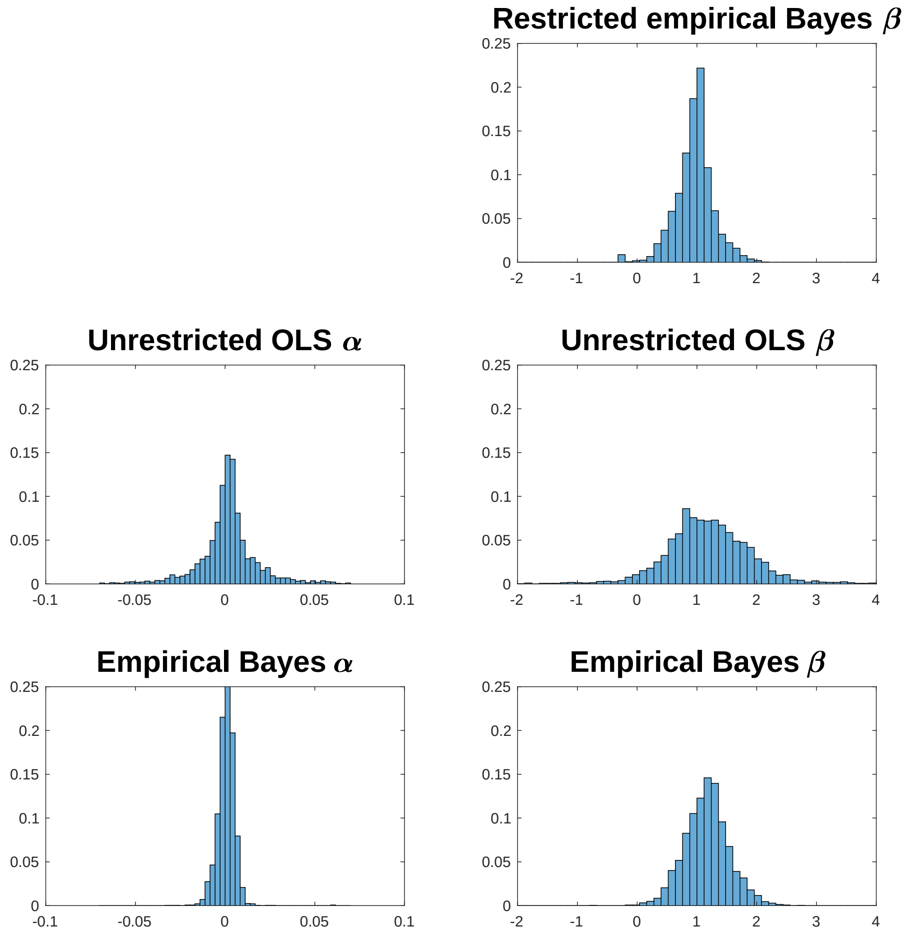
OLS and EB α

OLS and EB β

EB and restricted EB β

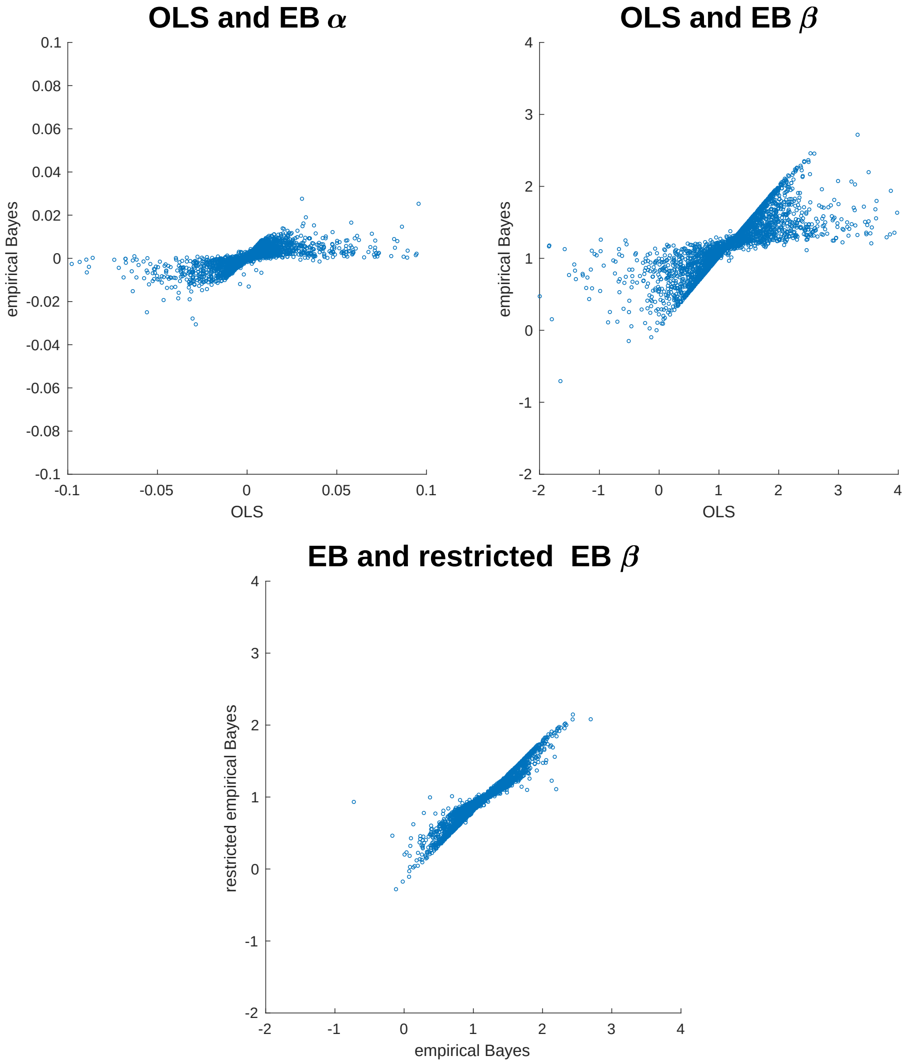
Notes: These figures show scatter plots of the joint distribution of alternative estimators for α_i and β_i across assets i, as discussed in Section A.1.

Figure 3: Distribution of estimates of α and β across assets for the period 1931-65



Notes: These figures show histograms of the distribution of alternative estimators for α_i and β_i across assets i, as discussed in Section A.1.

Figure 4: OLS and empirical Bayes estimates of α and β for the period 1931-65



Notes: These figures show scatter plots of the joint distribution of alternative estimators for α_i and β_i across assets i, as discussed in Section A.1.

Let us summarize our findings. A key prediction of CAPM – that all of the α_i are equal to 0 – does not appear to be exactly true (and has been rejected by statistical tests in the literature). This prediction however appears to be “approximately true,” in the sense of a small mean square deviation of α_i from 0. As a consequence, our preferred empirical Bayes estimator, which shrinks toward this theory, applies some non-negligible shrinkage relative to the unrestricted estimator. The out-of-sample predictive performance of our estimator exceeds that of competitors including both unrestricted and restricted estimation.

A.2 Choice probabilities and economic decision theory

In our second application we consider the problem of estimating the probability that an economic agent makes a certain choice when faced with a given choice set. Such estimation problems arise in many different economic settings. The data to estimate choice probabilities might for example come from lab experiments, or from household consumption surveys; see for instance Hoderlein and Stoye (2014) who use the the British Family Expenditure Survey to estimate bounds on the share of households violating the weak axiom of revealed preference. Choice probabilities and the restrictions on them implied by economic theory have been discussed in economic decision theory (see for instance McFadden 2005). Estimation of these choice probabilities involves high dimensional parameters to the extent that there are many possible choices and choice sets. In this application, the "theory" that we propose shrinking to corresponds to abstract theories of decision making, such as utility maximization, expected utility maximization, or exponential discounting.

Consider a set of individuals i who are randomly assigned to choice sets C. The individuals make choices using the choice functions d, which map choice sets into one of their elements. Suppose that all choices x belong to a finite set X of possible choices, and consider a collection $\mathscr{C}$ of subsets C of X. These are the possible choice sets faced by individuals. This setting is similar to the one considered by McFadden (2005).

In this setting, theories of decision making can be described by a collection $\mathscr{D}$ of choice functions d mapping each $C \in \mathscr{C}$ to one of its elements. A leading example of such a theory is maximization of strict preferences. This theory corresponds to the set of choice functions defined on $\mathscr{C}$ satisfying the strong axiom of revealed preference. Other examples of such theories are expected utility maximization (when the elements of X are lotteries), and exponential discounting (when the elements of X are time-paths of rewards).

We can identify choice functions with vectors as follows. For each combination of choice function d and choice x from choice set C, set d_{xC} equal to 1 if x is the element that d would choose from C. That is, $d_{xC} = 1$ iff $d(C) = x$ and $d_{xC} = 0$ otherwise. Once we stack the choice sets C and stack choices x within these sets, we are left with a vector d of 0s and 1s. Using this vector notation for choice functions, we can identify the collection of choice functions $\mathscr{D}$, which reflects the theory of decision making, with the matrix D containing all such column vectors d.

We want to allow for arbitrary heterogeneity, that is, arbitrary distributions of agents across the choice functions admitted by the theory. To this end, let $\pi \in \Delta$ be a probability distribution over choice functions. π_d is the probability that a random agent from the population of interest makes their choices according to the choice function $d \in \mathscr{D}$. Δ is the simplex of probability distributions on the elements of $\mathscr{D}$. Suppose now that agents (choice functions) are randomly assigned to choice sets.[3]

[3] This is a conceptual reference point and might be generalized in a number of ways. We could

Let $p = (p_{xC})$ be the vector of conditional choice probabilities for randomly assigned agents, where $p_{xC} = P(d(C) = x)$ is the probability that a random agent faced with choice set C will make choice $x \in C$. Our setup and notation now imply that

$$p = D \cdot \pi, \tag{7}$$

and thus, in particular,

$$p \in D \cdot \Delta. \tag{8}$$

The set $\mathscr{P} := D \cdot \Delta$ on the right hand side is in general a strict subset of the set of all conditional probability distributions for choices x given choice sets C. The statement $p \in \mathscr{P}$ is the empirical content of the theory of choice that imposes $d \in \mathscr{D}$ for all agents. When the theory considered is strict preference maximization, $D \cdot \Delta$ is the set of conditional choice probabilities satisfying the stochastic axiom of revealed preferences, as shown by McFadden (2005). By Farkas' Lemma we have $p \in \mathscr{P}$ if and only if there is no vector q such that

$$\begin{aligned} q' \cdot D &\geq 0 \\ q' \cdot p &< 0. \end{aligned}$$

Structural and unrestricted estimation Suppose now that we observe n i.i.d. draws (C_i, x_i), such that $x_i = d^i(C_i)$. Let n_{xC} be the number of observations such that $(C_i = C, x_i = x)$, and let n_C be the number of observations such that $C_i = C$. Once again we will consider three alternative approaches, this time for estimating the vector of conditional choice probabilities p. Unrestricted estimation simply estimates conditional probabilities by conditional frequencies, that is

$$\widehat{p}^u_{xC} = \frac{n_{xC}}{n_C}. \tag{9}$$

Restricted estimation estimates these probabilities while imposing that the vector p is consistent with our theory of choice, so that $p \in \mathscr{P} = D \cdot \Delta$. Structural estimation in this context thus imposes a set of linear inequality constraints on the vector p. The maximum likelihood estimator subject to this restriction can be written as

$$\widehat{p}^s = \underset{p \in \mathscr{P}}{\operatorname{argmax}} \sum_{C, x \in C} n_{xC} \cdot \log(p_{xC}). \tag{10}$$

Empirical Bayes estimation, shrinking toward the theory The third approach uses the empirical Bayes formalism to construct an estimator shrinking toward the theory. In the present context, we will deviate from the normal-normal setting considered thus far and instead consider multinomial sampling distributions

for instance replace independence by conditional independence given observed covariates.

and corresponding conjugate Dirichlet priors. Assume in particular for each choice set $C \in \mathscr{C}$ that

$$\begin{aligned}
(n_{xC})_{x\in C}|p &\sim MN((p_{xC})_{x\in C}, n_C) && (11)\\
(p_{xC})_{x\in C} &\sim Dir(\alpha \cdot (\overline{p}_{xC})_{x\in C}) && (12)\\
\overline{p} &\in \mathscr{P} \\
\alpha &\in \mathbb{R}^+.
\end{aligned}$$

We impose furthermore that independence of $((n_{xC})_{x\in C}, (p_{xC})_{x\in C})$ holds across different choice sets C. Equation (11) is implied by i.i.d. sampling. Equation (12) defines a family of priors, indexed by $\overline{p}$ and α. This family of priors will be used to construct the empirical Bayes estimator, where the hyperparameters α and $\overline{p}$ will be estimated using maximum marginal likelihood, and p itself will then be estimated using a Bayesian updating step.

These assumptions yield the following likelihoods corresponding to (i) sampling, (ii) the family of priors, (iii) the joint likelihood, and (iv) the marginal likelihood of the observed n_{xC} given the hyperparameters α and $\overline{p}$:

$$\begin{aligned}
P((n_{xC})|(p_{xC})) &= \prod_C \left[\left(\frac{n_C!}{\prod_{x\in C} n_{xC}!} \right) \times \prod_{x\in C} p_{xC}^{n_{xC}} \right] \\
P((p_{xC})) &= \prod_C \left[\left(\frac{\Gamma(\alpha)}{\prod_{x\in C} \Gamma(\alpha \cdot \overline{p}_{xC})} \right) \times \prod_{x\in C} p_{xC}^{\alpha \cdot \overline{p}_{xC}} \right] \\
P((n_{xC}), (p_{xC})) &= \prod_C \left[\left(\frac{\Gamma(\alpha) \cdot n_C!}{\prod_{x\in C} \Gamma(\alpha \cdot \overline{p}_{xC}) \cdot n_{xC}!} \right) \times \prod_{x\in C} p_{xC}^{\alpha \cdot \overline{p}_{xC} + n_{xC}} \right] \\
P((n_{xC})) &= \prod_C \left[\left(\frac{\Gamma(\alpha) \cdot n_C!}{\prod_{x\in C} \Gamma(\alpha \cdot \overline{p}_{xC}) \cdot n_{xC}!} \right) \times \frac{\prod_{x\in C} \Gamma(\alpha \cdot \overline{p}_{xC} + n_{xC})}{\Gamma(\alpha + n_C)} \right].
\end{aligned}$$

The marginal likelihood of the last equation is the product across choice sets C of so-called Dirichlet-multinomial distributions for each of the vectors $(n_{xC})_{x\in C}$. Conditional on the hyperparameters α and $\overline{p}_{xC}$ as well as the observed n_{xC}, the expectation of p_{xC} is given by

$$E[p_{xC}|\alpha, \overline{p}, (n_{xC})] = \frac{\alpha \cdot \overline{p}_{xC} + n_{xC}}{\alpha + n_C}. \qquad (13)$$

Plugging in estimates for α and $\overline{p}$, this expression gives the empirical Bayes estimates for p_{xC}^{EB}. These empirical Bayes estimates thus linearly interpolate between the unrestricted estimator and a structural estimator $\overline{p} \in \mathscr{P}$, as in the normal-normal setting considered before. Linear interpolation between structural and unrestricted estimators in fact will happen any time we are using conjugate priors for exponential families.

The empirical Bayes estimator of the hyperparameters α and $\overline{p}_{xC}$ maximizes the marginal likelihood, or equivalently, its logarithm:

$$(\alpha^{EB}, \overline{p}^{EB}) = \underset{\alpha,\overline{p}}{\operatorname{argmax}} \tag{14}$$

$$\sum_{C} \left(\log(\Gamma(\alpha)) - \log(\Gamma(\alpha + n_C)) + \sum_{x \in C} \Big(\log(\Gamma(\alpha \cdot \overline{p}_{xC} + n_{xC})) - \log(\Gamma(\alpha \cdot \overline{p}_{xC})) \Big) \right),$$

subject to $\overline{p} \in \mathscr{P}$ and $\alpha \in \mathbb{R}^{+}$. This optimization problem can be solved numerically; the supplementary appendix provides some discussion on numerical implementation of both structural estimation and maximizing the marginal likelihood in the present setting.

A.3 Multinomial logit and mixed multinomial logit

In our final application, we consider estimation of parametric structural models of discrete choice. Such models are used in many settings in applied microeconomics (Train 2009 provides a review). Estimation of these structural models of discrete choice might involve high dimensional parameters for several reasons. We might consider the influence of many characteristics of choices on choice probabilities as well as the influence of many characeristics of decision makers on choice probabilities. We might also wish to let either of these characteristics affect choice probabilities in a flexible way. In this application, the "theory" that we propose shrinking to corresponds to choice probabilities consistent with the multinomial logit model, which is nested in the more general mixed multinomial logit model. The multinomial logit model is arguably the most popular model of discrete choice, but imposes strong restrictions on demand. It imposes, in particular, the "independence of irrelevant alternatives" property. Tests for this property have been proposed by Hausman and McFadden (1984).

Consider a set of decision makers i who repeatedly, in periods t, choose between discrete alternatives j. Suppose that we observe these choices j, as well as a vector of observables x_{ijt} (with components x_{ijtk}) characterizing each of the available alternatives for decision maker i. Assume further that utility for these alternatives j is given by

$$u_{ijt} = x_{ijt} \cdot \beta_i + \epsilon_{ijt}, \tag{15}$$

where the ϵ_{ijt} are i.i.d. given x_i and follow an EV1 distribution, while the β_i are invariant across time and drawn from a distribution with density $f(\beta_i|\eta)$, i.i.d. across i. This is the setting considered in Train (2009) chapter 6.7, for instance. Availability of a panel, that is of repeated choices by the same decision makers, allows one to credibly identify heterogeneity of the preference parameters β_i across decision makers i.

Restricted and unrestricted estimation Under these assumptions, the probability of observing a sequence of choices $(j_1, \ldots, j_T)$ for any given decision maker i is equal to

$$P^{MML}(j_1, \ldots, j_T|x_{i..}) = \int \left(\prod_t \frac{\exp(x_{ij_t t} \cdot \beta)}{\sum_j \exp(x_{ijt} \cdot \beta)} \right) f(\beta|\eta) d\beta. \tag{16}$$

This is known as the mixed multinomial logit model. A special case of this model is the multinomial logit model, which imposes the additional restriction that there is no heterogeneity across i in terms of β, so that

$$P^{ML}(j_1, \ldots, j_T|x_{i..}) = \prod_t \frac{\exp(x_{ij_t t} \cdot \beta)}{\sum_j \exp(x_{ijt} \cdot \beta)}. \tag{17}$$

To fully parametrically specify the mixed multinomial logit model, we need to pick a family of distributions $f(\beta|\eta)$.

We assume that the vector β_i is normally distributed across i, that is

$$\beta_i|\eta \sim N(\mu, \Omega). \tag{18}$$

Under this assumption about the distribution of β_i, mixed multinomial logit reduces to multinomial logit in the boundary case $\Omega = 0$. Note that we allow for general correlations between the different components of β_i. This contrasts with the commonly imposed assumption that the different components of β_i are uncorrelated, as for instance in Train (2009) chapter 6.8. This increased flexibility allows for more realistic preference distributions, but requires estimation of a high-dimensional matrix Ω.

As before, we consider three alternative approaches for estimating these models and the implied choice probabilities. The first approach estimates the unrestricted mixed multinomial logit model, parametrized by (μ, Ω), using maximum likelihood. The second approach estimates the restricted multinomial logit model, parametrized by β, using maximum likelihood again.

Empirical Bayes estimation, shrinking toward the theory The third approach estimates the mixed multinomial logit model, shrinking it toward the multinomial logit model. We shall in particular consider the family of priors which imposes that the variance matrix Ω follows an Inverse-Wishart distribution with parameters $\left(\frac{1}{\tau} + p + 1\right)$ and 0,

$$\Omega \sim IW\left(\tfrac{1}{\tau} + p + 1, 0\right), \tag{19}$$

where $p = \dim(\beta_i)$. This parametrization is chosen to yield simple expressions for the conditional expectation of Ω below. We leave the mean vector μ unrestricted. In our general empirical Bayes notation, $\eta = (\mu, \Omega)$, a parameter of dimension $p + p\cdot(p+1)/2$, and $\theta = (\mu, \tau)$, a parameter of dimension $p + 1$.

This is a nonlinear model, and solutions have to be obtained using numerical methods. Empirical Bayes estimation of this model involves two steps. First we estimate the hyper-parameters μ and τ by maximizing the marginal likelihood. Then, we estimate the variance matrix Ω by its posterior mean, given μ and τ and given the observed data.

In order to evaluate the marginal likelihood, we propose to use a simulated likelihood approach. In order to calculate the posterior mean of Ω, we propose sampling from the posterior distributions of Ω and β_i, given the observed choices and given $\theta = (\mu, \tau)$, using Gibbs sampling.[4] For a detailed discussion of these numerical methods, the reader is pointed to chapters 10 and 12 in Train (2009) and chapters 11 and 12 in Gelman et al. (2014).

[4] Gibbs sampling is a Markov Chain Monte Carlo method designed to simulate draws from a distribution that decomposes in terms of several simpler conditional distributions.

Let us however briefly sketch some features of our model that simplify computation and shed some light on the behavior of the proposed empirical Bayes estimator. Given our modeling assumptions and given our family of priors, we have

$$\Omega|\mu, \tau, \beta_1, \ldots, \beta_n \sim IW\left(\tfrac{1}{\tau} + p + 1 + n, n \cdot \widehat{\Omega}\right), \tag{20}$$

where

$$\widehat{\Omega} = \tfrac{1}{n} \sum_i (\beta_i - \mu) \cdot (\beta_i - \mu)',$$

and thus

$$E[\Omega|\mu, \tau, \beta_1, \ldots, \beta_n] = \frac{n\tau}{1 + n\tau} \cdot \widehat{\Omega}.$$

If we hypothetically were to observe the β_i, then empirical Bayes estimation would involve linear shrinkage of the unrestricted variance estimator $\widehat{\Omega}$ toward 0. As the hyperparameter τ varies between 0 and ∞, the empirical Bayes estimator of Ω varies between 0 (so that we recover the restricted multinomial logit model) and the unrestricted maximum likelihood estimator $\widehat{\Omega}$ of Ω. If we observe many choices per agent so that T is large, while the number of observed agents n is not too large, this approximately describes the behavior of the empirical Bayes estimator where the β_i are unobserved.

Note that shrinkage happens for two distinct reasons in the mixed multinomial logit empirical Bayes setting, reflecting the fact that we have constructed a hierarchical model with three layers of parametrization. The first reason for shrinkage, present in conventional estimators of the MML model, is that we are estimating a discrete choice panel model with random coefficients β_i, where the random coefficients are considered to be drawn from a population distribution $f(\beta|\eta)$, so that we shrink β_i when interested in individual i's preferences. Conventional unrestricted estimators estimate η using maximum likelihood.

The second reason for shrinkage, which is specific to our approach, is that we are shrinking toward the multinomial logit model and its specific patterns of substitution. The multinomial logit model imposes independence of irrelevant alternatives, in particular, and depending on how well this assumption appears to apply in the available data, $\widehat{\tau}$ will be smaller or larger, so that the estimator of Ω shrinks more or less.

B Geometry of our empirical Bayes estimator

We conclude our analysis of the properties of $\widehat{\beta}^{EB}$ by studying its geometry. The proposed empirical Bayes estimator can be seen as providing a mapping from an unrestricted (preliminary) estimate $\widehat{\beta}$ to an empirical Bayes estimate $\widehat{\beta}^{EB}$. Understanding this mapping is key for understanding the behavior of our estimator. For this section, consider again the estimator of Assumption **??** in the canonical form Assumption **??**, where $\widehat{V} = \text{diag}(v_i)$.

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

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

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

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

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

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

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

$$\widehat{\tau}^2 \cdot \sqrt{\sum_{k'} \frac{1}{\widehat{\tau}^2 + v_{k'}}}. \tag{22}$$

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 bound on $\widehat{\tau^2}$ of 0.[5] 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

$$v_j \cdot \sqrt{\sum_{k'} \frac{1}{v_{k'}}}. \tag{23}$$

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 5 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

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

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

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 as $\widehat{\beta}_1$ 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 5. 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 6.

Geometry in the general case: $M \neq 0$ Let us now turn to the general case where $M \neq 0$, and where we must account for estimation of β^0. This can be analyzed using the same "trick" as before, 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

$$\sum_j \frac{(\widehat{\beta}_j - \widehat{\beta}^0 \cdot M_j)^2}{\widehat{\tau}^2 + v_j},$$

so that

$$\widehat{\beta}^0 = \frac{\sum_j \widehat{\beta}_j \cdot \frac{1}{\widehat{\tau}^2 + v_j}}{\sum_j M_j \cdot \frac{1}{\widehat{\tau}^2 + v_j}}. \tag{24}$$

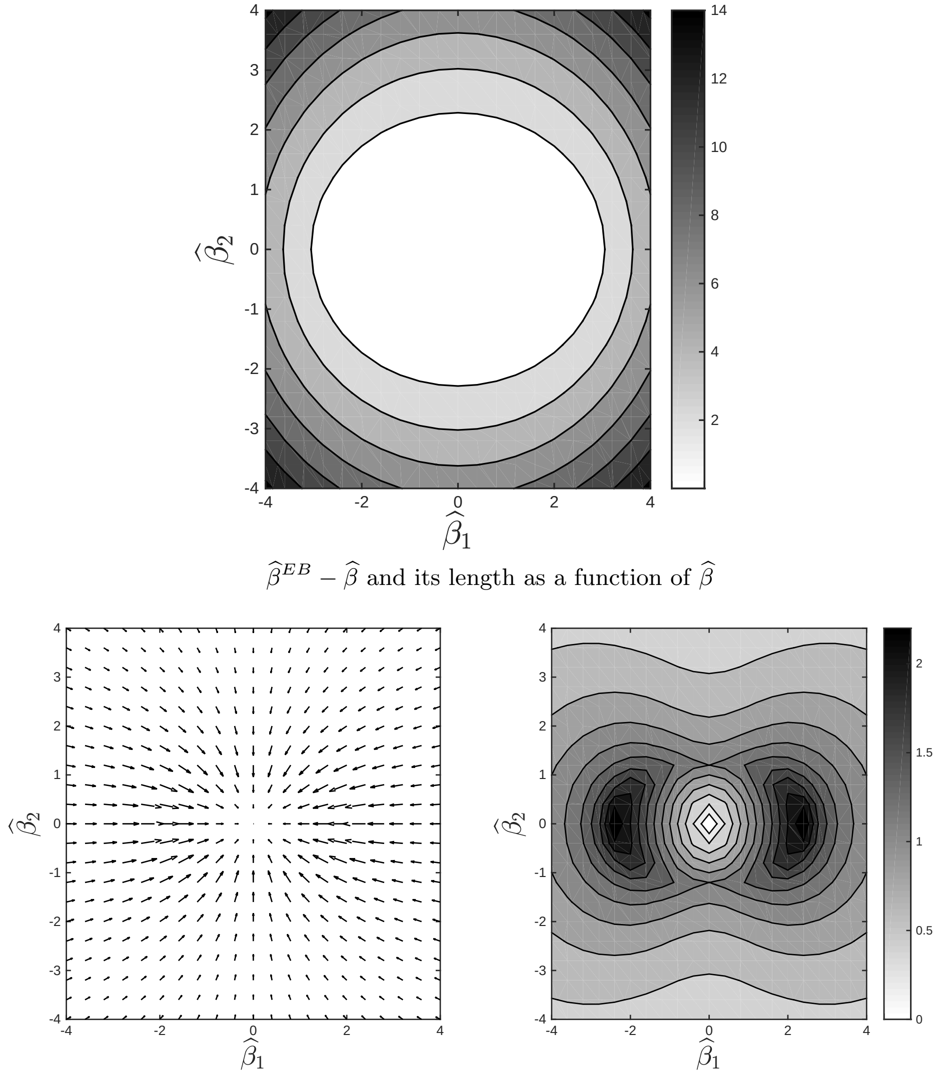
This equation defines a hyper-plane in the space of $\widehat{\beta}$. As before, the first order condition for $\widehat{\tau}^2$ implies

$$\sum_j \frac{1}{\widehat{\tau}^2 + v_j} = \sum_j \frac{(\widehat{\beta}_j - \widehat{\beta}^0 \cdot M_j)^2}{(\widehat{\tau}^2 + v_j)^2}.$$

This equation describes an ellipsoid centered at $\widehat{\beta}^0 \cdot M$ with semi-axes of length $v_j \cdot \sqrt{\sum_{k'} \frac{1}{v_{k'}}}$ 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 hyper-plane through the center of this ellipsoid. $\widehat{\beta}^{EB}$ is then obtained from $\widehat{\beta}$ by shrinking on the hyper-plane 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.

Figure 5: The mapping from $\widehat{\beta}$ to $\widehat{\tau}^2$ and $\widehat{\beta}^{EB}$

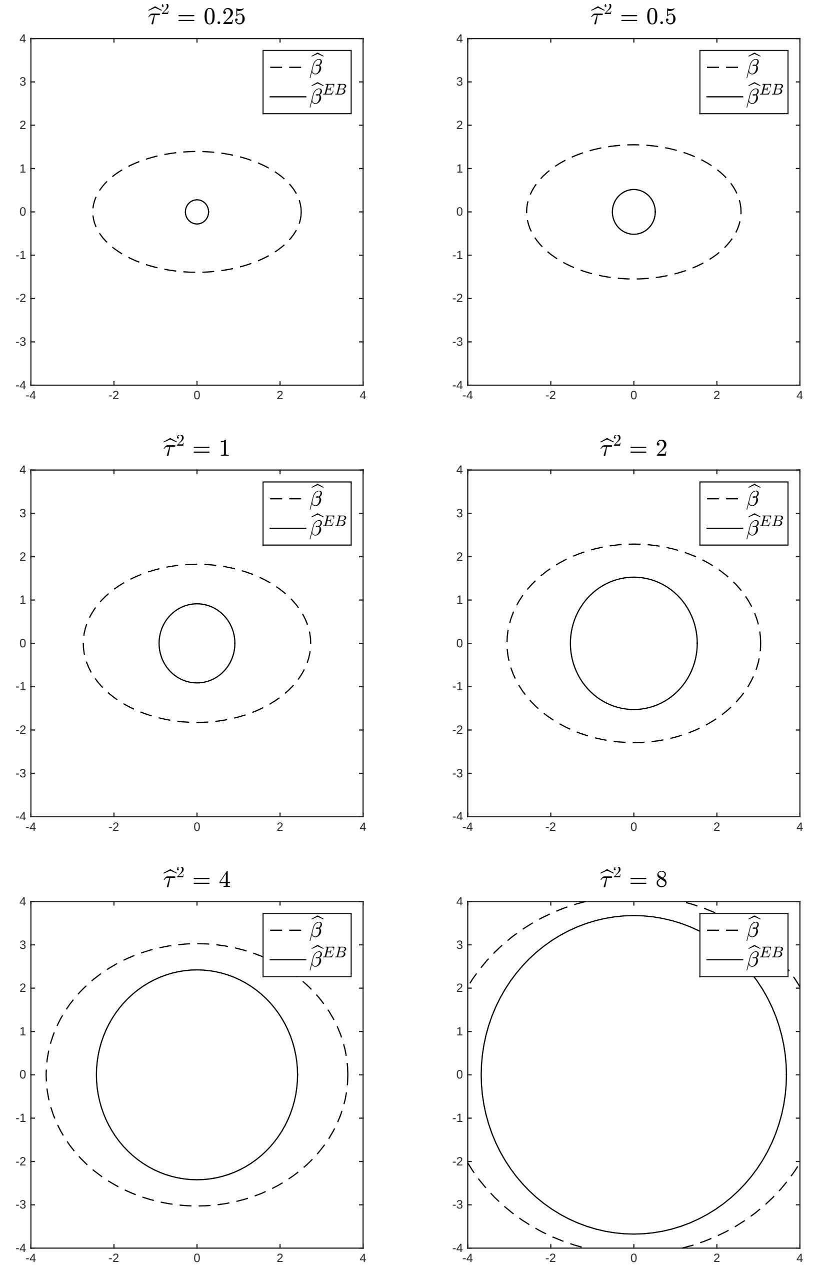
$\widehat{\tau}^2$ as a function of $\widehat{\beta}$



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

Notes: These figures illustrate the mapping from preliminary estimates to empirical Bayes estimates when $\dim(\beta) = 2$, $\mathrm{Var}(\widehat{\beta}) = \mathrm{diag}(2, 1)$, and $M = 0$. The top figure shows how our measure of model fit $\widehat{\tau}^2$ varies with $\widehat{\beta}$, the bottom left figure shows the direction and magnitude of shrinkage from $\widehat{\beta}$ to $\widehat{\beta}^{EB}$, and the bottom right figure depicts just the magnitude of shrinkage. For details, see Section **??**.

Figure 6: The geometry of empirical Bayes



Notes: These figures illustrate the mapping from preliminary estimates for the same setting as in Figure 5. Each figure depicts, for a given value of $\widehat{\tau}^2$, which preliminary estimates $\widehat{\beta}$ yield this value and to what set of empirical Bayes estimates $\widehat{\beta}^{EB}$ these preliminary estimates are mapped. For details see Section **??**.

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

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

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

C Monte Carlo simulations

In this section, we present a series of simulations comparing the performance of our empirical Bayes procedure to its competitors, structural (restricted) estimation and unrestricted estimation. These simulations are modeled on the labor demand application of section 3.1 in the manuscript. Section C.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 C.2 presents simulations corresponding to the frequentist paradigm, fixing the parameter η and drawing from the implied distribution of the data Y. Section C.3 considers simulations similar to Sections C.1 and C.2, but governed by parameters calibrated to match our empirical application.

C.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 5,000 Monte Carlo draws for each design. Table 2 shows the results of these simulations. For each design we show the mean squared error, calculated as an average over Monte Carlo draws of 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, as it exploits additional correct information. And indeed we do find that structural estimation dominates empirical Bayes at the very extreme of the range of designs considered.

At the other extreme of the designs considered are those with large sample size, small number of regressors, small variance of residuals, and poor fit of the structural model (large τ^2). In these designs we would expect the unrestricted estimator to

Table 2: 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	Oracle E.B.
50	4	1.0	1.0	0.2	1.59	1.70	0.97
50	16	1.0	1.0	0.2	0.82	1.21	1.00
200	4	1.0	1.0	0.2	4.42	1.19	0.98
200	16	1.0	1.0	0.2	4.18	1.11	1.01
50	4	0.5	1.0	0.2	2.41	1.37	0.98
50	16	0.5	1.0	0.2	1.56	1.14	1.01
200	4	0.5	1.0	0.2	7.87	1.09	0.99
200	16	0.5	1.0	0.2	7.80	1.03	1.00
50	4	1.0	1.0	0.5	2.42	1.38	0.99
50	16	1.0	1.0	0.5	1.56	1.14	1.01
200	4	1.0	1.0	0.5	7.90	1.10	1.00
200	16	1.0	1.0	0.5	7.83	1.04	1.00
50	4	0.5	1.0	0.5	4.05	1.19	1.00
50	16	0.5	1.0	0.5	2.91	1.07	1.01
200	4	0.5	1.0	0.5	14.94	1.05	1.00
200	16	0.5	1.0	0.5	15.28	1.01	1.00
50	4	1.0	1.0	1.0	4.14	1.19	1.00
50	16	1.0	1.0	1.0	2.90	1.07	1.01
200	4	1.0	1.0	1.0	15.08	1.05	1.00
200	16	1.0	1.0	1.0	15.21	1.01	1.00
50	4	0.5	1.0	1.0	7.34	1.09	1.00
50	16	0.5	1.0	1.0	5.53	1.02	1.01
200	4	0.5	1.0	1.0	29.42	1.03	1.00
200	16	0.5	1.0	1.0	29.92	1.00	1.00

Notes: This table compares the performance of alternative estimators based on 5,000 Monte Carlo draws given θ. For details, see the description in Section C.1.

work well as it has a small variance and does not shrink toward the incorrect structural model. Nonetheless, we do find that unrestricted estimation never dominates empirical Bayes for any of the designs considered.

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 2 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 improvements of performance.

C.2 Monte Carlo results, fixing η, drawing from the distribution of Y

In Section C.1 we considered simulations where θ was fixed, but η was drawn repeatedly, an approach that 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 3 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 arbitrarily large when the model is exactly right, the number of parameters J is large, and estimates are noisy (small sample size n, large residual variance 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.

C.3 Calibrated Monte Carlo simulations

We conclude this section by presenting some simulations similar to those discussed before, but calibrated to our empirical results. We first estimate the model via

Table 3: Mean Squared Error of alternative estimators relative to empirical Bayes conditional on η

Sesign parameters						mean squared error	
n	J	σ^2	β_{00}	β_{01}	β_{02}	Structural	Unrestricted
50	4	1.0	1.0	1.0	1.0	0.25	2.13
50	16	1.0	1.0	1.0	1.0	0.02	1.32
200	4	1.0	1.0	1.0	1.0	0.18	1.47
200	16	1.0	1.0	1.0	1.0	0.04	2.30
50	4	0.5	1.0	1.0	1.0	0.19	1.70
50	16	0.5	1.0	1.0	1.0	0.02	1.32
200	4	0.5	1.0	1.0	1.0	0.16	1.27
200	16	0.5	1.0	1.0	1.0	0.09	5.39
50	4	1.0	1.0	1.0	6.0	3.86	1.13
50	16	1.0	1.0	1.0	6.0	0.61	1.20
200	4	1.0	1.0	1.0	6.0	14.69	1.04
200	16	1.0	1.0	1.0	6.0	3.10	1.12
50	4	0.5	1.0	1.0	6.0	7.09	1.05
50	16	0.5	1.0	1.0	6.0	1.14	1.13
200	4	0.5	1.0	1.0	6.0	28.66	1.02
200	16	0.5	1.0	1.0	6.0	5.81	1.05
50	4	1.0	0.0	1.0	6.0	4.60	1.05
50	16	1.0	0.0	1.0	6.0	0.81	1.18
200	4	1.0	0.0	1.0	6.0	18.74	1.00
200	16	1.0	0.0	1.0	6.0	4.07	1.08
50	4	0.5	0.0	1.0	6.0	8.76	1.00
50	16	0.5	0.0	1.0	6.0	1.51	1.11
200	4	0.5	0.0	1.0	6.0	37.35	1.01
200	16	0.5	0.0	1.0	6.0	7.73	1.03

Notes: This table compares the performance of alternative estimators based on 5,000 Monte Carlo draws given η. For details, see the description in Section C.2.

empirical Bayes as in our empirical application, using the US panel of states and controlling for time and state fixed effects, to obtain estimates $\widehat{\beta}^0$, $\widehat{\tau}^2$, and $\widehat{\delta}^{EB}$. We then perform two sets of simulations. For the first, we repeatedly draw values for δ conditional on $\widehat{\beta}^0$, $\widehat{\tau}^2$, and values of $\widehat{\delta}_r$ conditional on δ. These simulations are analogous to those of Section C.1, "conditional on θ." For the second set of simulations, we fix δ equal to the estimated $\widehat{\delta}^{EB}$, and draw values of $\widehat{\delta}_r$ from the corresponding sampling distribution. These simulations are analogous to those of Section C.2, "conditional on η." Each simulation is repeated 5000 times, and for each repetition we calculate structural, unrestricted, and empirical Bayes estimates based on $\widehat{\delta}_r$. We then calculate the mean squared errors of each of these and normalize them relative to the MSE of empirical Bayes estimation.

We do this for both of the following cases. First, we consider shrinkage toward the J-type CES model. Simulations conditional on θ and conditional on η are shown in Table 5. We second consider shrinkage toward the 2-type CES model of a demand system for 8 types of workers. Simulations conditional on θ and conditional on η are shown in Table 5.

These simulations show that our proposed empirical Bayes approach, in these empirical settings, performs consistently better than both unrestricted estimation and structural estimation. These simulation results support using our proposed estimator.

Table 4: Mean Squared Error of alternative estimators relative to empirical Bayes, calibrated specifications, 8-type CES model

Specification	Mean squared error	
Def of supply	Structural	Unrestricted
Given η	3.06	3.14
Given θ	644.72	590.35

Notes: This table compares the performance of alternative estimators based on 5,000 Monte Carlo draws, based on specifications calibrated to our empirical results. For details, see description in Section C.3.

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

Specification	Mean squared error	
Def of supply	Structural	Unrestricted
Given η	5.21	3.21
Given θ	2.71	3.60

Notes: This table compares the performance of alternative estimators based on 5,000 Monte Carlo draws, based on specifications calibrated to our empirical results. For details, see the description in Section C.3.

D Labor demand and CES production functions

We next review the structural models of labor demand justifying the restricted wage regressions considered in section 3.1 of the manuscript. Assume that wages equal marginal productivity for some aggregate production function f,

$$w_{ij} = \frac{\partial f_i(N_{i1}, \ldots, N_{iJ})}{\partial N_{ij}}, \tag{25}$$

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

$$f_i(N_{i1}, \ldots, N_{iJ}) = \left(\sum_{j=1}^{J} \gamma_j N_{ij}^{\rho} \right)^{1/\rho}. \tag{26}$$

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

$$Y_{ij} - Y_{ij'} = \log(\gamma_j) - \log(\gamma_{j'}) + \beta^0 \cdot (X_{ij} - X_{ij'}),$$

where $\beta^0 = \rho - 1$.

This result motivates regressions of the following form (see for instance Autor et al. 2008 and Card 2009):

$$Y_{ij} - Y_{ij'} = \gamma_{jj'} + \beta^0 \cdot (X_{ij} - X_{ij'}) + \epsilon_{ijj'}. \tag{27}$$

The coefficient β^0 in this regression is interpreted as the negative of the inverse elasticity of substitution between labor types j and j'.[6] 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

[6] 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.

specifications might also include additional terms for aggregate types of labor as motivated by nested CES models.

Let us briefly discuss the economic content of the restrictions on β imposed by the structural model relative to an unrestricted regression of the form

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

First, $\beta \cdot e = 0$ 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. Second, $\beta_{jj'} = \beta_{jj''}$ for $j', j'' \neq j$. The elasticity of substitution between different groups is the same for all groups. The CES model imposes that there are only two possible degrees of substitutability between different workers – either they are perfect substitutes when they are the same type, or they have an elasticity of substitution of $\sigma = -1/\beta^0$. Third, $\beta_{jj} = \beta_{j'j'}$. The own-elasticity of demand is the same for all types of labor. In combination, these three restrictions in fact imply the CES regression model. The CES model additionally implies that changes in labor supply do not affect within-type inequality of wages. Given the small number of types usually imposed, this is a strong restriction.

D.1 2-type CES

In the manuscript, we also consider the canonical 2-type CES production function model. Assume that the production function takes the form

$$f_i(N_{i1}, \dots, N_{iJ}) = \left(\sum_{k=1}^{2} \tilde{N}_{ik}^{\rho} \right)^{1/\rho}.$$

$$\tilde{N}_{ik} = \sum_{k_j = k} \gamma_j N_{ij}.$$

Then

$$w_{ij} = \frac{\partial f_i}{\partial N_{ij}} = \left(\sum_{k=1}^{2} \tilde{N}_{ik}^{\rho} \right)^{1/\rho - 1} \cdot \gamma_j \cdot \tilde{N}_{ik_j}^{\rho-1},$$

and thus, using the same notation as before,

$$\frac{w_{ij}}{w_{ij'}} = \frac{\gamma_j}{\gamma_{j'}} + \left(\frac{\tilde{N}_{ik_j}}{\tilde{N}_{ik_{j'}}} \right)^{\rho-1}.$$

Taking logs suggests the regression specification.

$$Y_{ij} - Y_{i1} = (\gamma_j - \gamma_1) + \beta^0 \cdot (\tilde{X}_{ik_j} - \tilde{X}_{i1}) + (\epsilon_{ij} - \epsilon_{i1}).$$

The 2 type CES model thus imposes the restriction $\delta = 0$ on the unrestricted model for labor demand given by

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

E Numerical implementation for marginal maximum likelihood in the Multinomial-Dirichlet setting

In section 3.3 of the manuscript, we consider estimation of conditional choice probabilities, shrinking toward the prediction of general theories of choice. In order to calculate the structural maximum likelihood estimator proposed in this section, we need to solve the constrained optimization problem

$$\begin{aligned} \widehat{p}^s &= \underset{p}{\operatorname{argmax}} \sum_{C, x \in C} n_{xC} \cdot \log(p_{xC}) \\ s.t.\ p &= D \cdot \pi \\ \pi &\in \Delta. \end{aligned}$$

Note that π itself is in general *not* identified when the theory is correct ($p \in \mathscr{P}$), but the pseudo-true π *is* generically unique and given by a corner of $\mathscr{P}$ otherwise. This corner solution is the projection of the true p onto $\mathscr{P}$ with respect to the Kulbach-Leibler divergence; similarly, the structural estimator is given by a projection of the unrestricted estimator onto $\mathscr{P}$. We can solve this problem by searching for a (potentially non-unique) maximizer in Δ. Since this is a convex optimization problem,numerical maximization is fairly straightforward, using for instance the barrier method, and any solution satisfies

$$\begin{aligned} \widehat{p}^s &= D \cdot \widehat{\pi} \\ \widehat{\pi} &= \underset{\pi: \sum_d \pi_d = 1}{\operatorname{argmax}} \sum_{C, x \in C} n_{xC} \cdot \log((D \cdot \pi)_{xC}) + \lambda \cdot \pi \\ \lambda \cdot \widehat{\pi} &= 0, \end{aligned}$$

for Lagrange multipliers $\lambda \geq 0$.

Let us now turn to the empirical Bayes estimator. The difficult step in solving for the empirical Bayes estimator is estimating the hyper-parameters α and $\overline{p}$. Our proposed algorithm builds on the iterative procedure suggested by Minka (2000), section 4. This iterative procedure alternates maximizing the likelihood with respect to α and with respect to $\overline{p}$. The first order condition for maximization with respect to α is given by

$$\sum_C \left(\Psi(\alpha) - \Psi(\alpha + n_C) + \sum_{x \in C} \Big(\Psi(\alpha \cdot \overline{p}_{xC} + n_{xC})) - \Psi(\alpha \cdot \overline{p}_{xC}) \Big) \right) = 0, \qquad (28)$$

where $\Psi := \partial_x \log(\Gamma(x))$ is the so-called digamma function. As shown by Minka (2000), the solution to this FOC can be found by fixed-point iteration of the form

$$\alpha^{new} = \alpha \cdot \frac{\sum_C \sum_{x \in C} \Big(\Psi(\alpha \cdot \overline{p}_{xC} + n_{xC}) - \Psi(\alpha \cdot \overline{p}_{xC}) \Big)}{\sum_C \Psi(\alpha) - \Psi(\alpha + n_C)}.$$

Faster second-order methods are available, as well.

The likelihood for $\overline{p}$ given α and the data is proportional to

$$\prod_C \prod_{x \in C} \frac{\Gamma(\alpha \cdot \overline{p}_{xC} + n_{xC})}{\Gamma(\alpha \cdot \overline{p}_{xC})}. \tag{29}$$

Using the fact that $\Psi(x+1) = \Psi(x) + 1/x$, the derivative of the logarithm of this expression with respect to $\overline{p}_{xC}$ can be written as

$$\alpha \cdot \Big(\Psi(\alpha \cdot \overline{p}_{xC} + n_{xC})) - \Psi(\alpha \cdot \overline{p}_{xC}) \Big) = \sum_{j=0}^{n_{xC}-1} \frac{1}{\overline{p}_{xC} + j/\alpha}.$$

Define the notation

$$\nu(m, q) := q \cdot \Big(\Psi(q + m)) - \Psi(q) \Big) = \sum_{j=0}^{m-1} \frac{1}{1 + j/q}.$$

Were there no constraint on $\overline{p}$ then we could find the maximum likelihood estimator by a simple fixed point iteration, setting $\overline{p}_{xC}^{new}$ proportional to $\nu(n_{xC}, \alpha \cdot \overline{p}_{xC})$, appropriately normalized to ensure summing to 1 for each C, cf. Minka (2000), section 5. Imposing the constraints of theory, we again obtain a convex optimization problem which might be solved using the barrier method.

References

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.

Bossaerts, P. (2013). *The paradox of asset pricing.* Princeton university press.

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Taylor & Francis.

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

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.

Hausman, J. and McFadden, D. (1984). Specification tests for the multinomial logit model. *Econometrica*, 52(5):1219–1240.

Hoderlein, S. and Stoye, J. (2014). Revealed preferences in a heterogeneous population. *The Review of Economics and Statistics*, 96(2):197–213.

Jacquier, E., Polson, N., Geweke, J., Koop, G., and Van Dijk, H. (2011). Bayesian methods in finance. *Handbook of Bayesian econometrics. Oxford University Press, Oxford*, pages 439–512.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of finance*, 23(2):389–416.

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

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

Minka, T. (2000). Estimating a Dirichlet distribution.

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