How to use economic theory to improve estimators

Fessler, Pirmin* Kasy, Maximilian†

March 7, 2017

Abstract

We propose to use economic theories to construct estimators that perform well when the theories’ empirical implications are approximately correct, but are robust even if the theories are completely wrong. We describe a general construction of such estimators using the empirical Bayes paradigm. We implement this construction in various settings, including labor demand and wage inequality, asset pricing, economic decision theory, and structural discrete choice models. We provide theoretical characterizations of the behavior of the proposed estimators, and evaluate them using Monte Carlo simulations. Our approach is an alternative to the use of theory as something to be tested or to be imposed on estimates. Our approach complements uses of theory for identification and extrapolation.

Keywords: Empirical Bayes estimation, shrinkage, labor demand, asset pricing, decision theory, discrete choice
JEL codes: C11, C52, J23, J31, G12, D01, D12

*Economic Analysis Division, Oesterreichische Nationalbank; Address: Postbox 61, 1011 Vienna, Austria; e-mail: pirmin.fessler@oenb.at. Opinions expressed by the authors of studies do not necessarily reflect the official viewpoint of the Oesterreichische Nationalbank or of the Eurosystem.
†Associate Professor, Department of Economics, Harvard University; Address: 1805 Cambridge Street, Cambridge, MA 02138; e-mail: maximiliankasy@fas.harvard.edu.
1 Introduction

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

We propose a further, alternative use of economic theory in empirical research. We suggest a framework for the construction of estimators which perform particularly well when the empirical implications of a theory under consideration are approximately true. By “approximately true” we mean that deviations from the theory’s predictions are of the same order of magnitude as the standard errors of unrestricted estimates. Estimators constructed in the proposed way tend to outperform estimators ignoring the theory, no matter what the true data generating process is. Our approach provides an alternative to the testing and imposition of theories. We will argue that it is well suited for theories that are only approximately true, as might be the case for many theories in economics. If theories are only approximately true, their rejection by tests is only a matter of sample size, and their imposition might cause estimators to be biased and inconsistent. Our approach is complementary to the roles of theory in identification and in guiding the choice of research questions.

Our approach is based on estimators shrinking toward the theory in a data dependent way. Such estimators can be constructed as follows. Our construction uses the empirical Bayes paradigm, which requires a family of priors. We consider families of priors for the parameters of interest, where the priors are centered on the set of parameters consistent with the predictions of the theory. These priors are further governed by a parameter of dispersion, providing a measure for how well the theory appears to describe the data. A prior with a dispersion of zero would correspond to

---

1 We thank Alberto Abadie, Isaiah Andrews, Gary Chamberlain, Ellora Derenoncourt, Liran Einav, Kei Hirano, Michael Kummer, José Montiel Olea, Neil Shephard, several anonymous referees, as well as seminar participants at Duke, Harvard, UBC, SFU, Hebrew University, Tel Aviv University, University of Graz, JKU Linz, University of Rochester, and Bank of Canada for helpful discussions and comments.
imposing the theory, an infinite dispersion to an uninformative prior, ignoring the theory.

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

The main contribution of the present paper is to bring together economic theory with the tools of the empirical Bayes paradigm, in order to leverage economic theory for improved estimation in a way that contrasts with the “testing and imposition” approach. Empirical Bayes estimators were originally proposed by Robbins (1956); they are closely related to shrinkage estimators as introduced by James and Stein (1961). Parametric empirical Bayes was introduced by Morris (1983). The empirical Bayes estimators usually considered in the statistics literature shrink toward an arbitrary point in the parameter space, such as 0. We instead modify the construction to shrink toward parameter sets consistent with economic theories such as structural models of labor demand (as in Card 2009), general equilibrium models of asset markets (as in Jensen et al. 1972), abstract theories of economic decision making (as in McFadden 2005), or structural discrete choice models of consumer demand (as in Train 2009). In addition to proposing this new role for economic theory, and providing guidelines and examples for how to implement it, we develop some novel statistical theory results, characterizing the behavior of the proposed estimators. Our estimators are related to but different from the shrinkage estimators discussed in Hansen (2016); our results complement those of Hansen (2016) in a way discussed in greater detail below.

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

After discussing our approach in a stylized baseline setting in Section 2, we implement it in a diverse set of economic contexts in Section 3. These contexts are distinguished in particular by the type of “theory” considered, ranging from parametric structural models of production or of preferences, to general equilibrium models of financial markets, to abstract theories of decision making. Let us briefly sketch the settings considered.

**Labor demand and wage inequality**  Wage inequality has increased significantly in most industrial countries since the 1980s. There is considerable disagreement over the relative contribution to this increase of alternative factors, such as technical change (Autor et al., 2008), migration (Card, 2009), and institutional factors (Fortin and Lemieux, 1997). Some of these disagreements have methodological roots. The workhorse method of estimating structural models of labor demand yields results that depend on the specific model chosen and the implied substitutability patterns. Flexible (unrestricted) estimation, on the other hand, results in very noisy estimates. We propose instead estimating flexible systems of labor demand, shrinking toward the predictions of a canonical model such as the 2-type constant elasticity of substitution (CES) model. We apply this method to data from the Current Population Survey (CPS) and the American Community Survey (ACS). We generally find negative but small inverse elasticities of substitution. The explanatory power of changes in labor

---

2It is possible to construct counter-examples, however, see Section 4.3.
supply for changes in relative wages appears to be quite small, based on our estimates.

**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.

**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.

**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.

In Section 4 we provide characterizations of the behavior of the proposed estimators, focusing on the baseline setting introduced in Section 2. We show consistency, characterize the mapping from unrestricted estimates to empirical Bayes estimates, and provide visual representations. Theorem 1 in Section 4 provides a simple, tractable approximation of the risk function (mean squared error) of our estimator. This (asymptotic) approximation is valid whenever the dispersion parameter is estimated with small variance relative to the parameters of interest. Our characterization of risk demonstrates the favorable properties of our proposed estimator, relative to both restricted and unrestricted estimation. In contrast to classic derivations of risk for James-Stein shrinkage, Theorem 1 is only valid under this approximation, but it extends classic results to the practically relevant case where neither normality nor (more importantly) homoskedasticity are imposed. Monte Carlo simulations (in the supplementary online appendix) confirm the validity of these approximations for realistic specifications.

The rest of this paper is structured as follows: Section 2 discusses the empirical Bayes paradigm and introduces our proposed construction of estimators for a stylized baseline case. This section also reviews some related literature. Section 3 implements and adapts this construction to the economic settings described above. Section 4 develops some statistical theory for the estimators we consider, including consistency, and an exploration of their geometry. We also provide a theoretical characterization of their risk properties. Section 5 concludes. Appendix A describes a way to construct empirical Bayes confidence sets, based on the heuristic arguments of Laird and Louis (1987). Appendix B contains all proofs; figures and tables can be found in Appendix C. A supplementary online appendix contains some Monte Carlo simulations, as well as some additional discussion of labor demand systems, and a discussion of numerical methods for maximizing the marginal likelihood for our third application.

\footnote{Asymptotic normality need not hold under our large-dimension asymptotics, in contrast to large-sample asymptotics. Homoskedasticity is equivalent to having estimator variance and prior variance proportional and implies component-wise proportional shrinkage.}
2 Estimator construction in a simplified setting

In this section we consider a simplified setting to introduce our approach. We assume that there is a preliminary estimator \( \hat{\beta} \) of the parameter vector of interest, \( \beta \), where the preliminary estimator does not make use of restrictions implied by economic theory. Economic theory is assumed to provide overidentifying restrictions on \( \beta \); for simplicity of exposition, we focus on linear restrictions. We use these restrictions to construct an estimator \( \hat{\beta}^{EB} \) designed to outperform \( \hat{\beta} \) if the restrictions are approximately true, and to perform no worse than \( \hat{\beta} \) if they are not. In Section 3 we will then modify this simplified setting as needed for our applications, detailing in each case where the preliminary estimator and the theoretical restrictions are coming from.

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

2.1 Setup

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

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

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

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

\[
\beta = \beta_0 \cdot M
\]
for some known vector $M$ and unknown scalar $\beta_0$. In the applications of Section 3, we will see examples where our construction is extended to affine restrictions on $\beta$ of different dimension. More generally, shrinkage toward non-linear smooth restrictions (imposing that $\beta$ lies in some smooth manifold) could be considered. Theorem 1 in Hansen (2016) provides the type of result needed for such a generalization, using a local asymptotic framework; the theoretical discussion in the present paper will focus on the linear case, however.

2.2 General empirical Bayes estimation

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

There is a paradigm in statistics called empirical Bayes estimation, which allows one to cover 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:

\begin{align}
Y|\eta &\sim f(Y|\eta) \quad (3) \\
\eta &\sim \pi(\eta|\theta), \quad (4)
\end{align}

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

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

\begin{equation}
Y|\theta \sim g(Y|\theta) := \int f(Y|\eta)\pi(\eta|\theta) d\eta. \quad (5)
\end{equation}

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

$$\hat{\theta} = \text{argmax}_\theta g(Y|\theta).$$

(6)

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

$$\hat{\eta} = E[\eta|Y, \theta = \hat{\theta}].$$

(7)

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

2.3 An empirical Bayes model for our setup

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

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

$$\hat{\beta}|\beta, V \sim N(\beta, V).$$

(8)

This assumption can be justified by conventional asymptotics, letting the number $n$ of
cross-sectional units go to infinity in many applications of interest (as in Hansen 2016).
We emphasize again that normality of $\hat{\beta}$ is only used for estimator construction, and

4The quotation marks reflect the fact that this would only be a posterior expectation in the strict
sense if $\hat{\theta}$ had been chosen independently of the data, rather than estimated.
is not imposed in our theoretical discussion of its properties in Section 4. We further assume that we have a consistent estimator $\hat{V}$ of $V$, i.e.,

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

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

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

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

$$\xi_{j,j'} \simiid N(0, \tau^2).$$

The term $\beta_0 \cdot M$ corresponds to a set of coefficients satisfying the structural model. The term $\zeta$ is equal to a random $J$-vector with variance

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

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

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

$$\eta = (\beta, V)$$
$$\theta = (\beta_0, \tau^2, V)$$
$$\tilde{\beta}|\eta \sim N(\beta, V)$$
$$\beta|\theta \sim N(\beta_0 \cdot M, \tau^2 \cdot I).$$

Solving for the empirical Bayes estimator In order to obtain estimators of $\beta_0$ and $\tau^2$, consider the marginal distribution of $\tilde{\beta}$ given $\theta$. This marginal distribution
is normal,
\[ \hat{\beta}|\theta \sim N(\beta_0 \cdot M, \Sigma(\tau^2, V)), \]  
(10)  
where (leaving the conditioning on \( \theta \) implicit)
\[ \Sigma(\tau^2, V) = \text{Var} \left( \hat{\beta} \right) \]
\[ = \tau^2 \cdot I + V. \]

Substituting the consistent estimator \( \hat{V} \) for \( V \), we obtain the empirical Bayes estimators of \( \beta_0 \) and \( \tau^2 \) as the solution to the maximum (marginal) likelihood problem
\[ (\hat{\beta}_0, \hat{\tau}^2) = \arg\min_{b_0, t^2} \log \left( \det(\Sigma(t^2, \hat{V})) \right) \]
\[ = (\hat{\beta} - b_0 \cdot M)' \cdot \Sigma(t^2, \hat{V})^{-1} \cdot (\hat{\beta} - b_0 \cdot M). \]  
(11)  
We can simplify this optimization problem by concentrating out \( b_0 \): given \( t^2 \), the optimal \( b_0 \) is equal to
\[ \hat{b}_0 = (M \cdot \Sigma(t^2, \hat{V})^{-1} \cdot M')^{-1} \cdot M \cdot \Sigma(t^2, \hat{V})^{-1} \cdot \hat{\beta}. \]  
(12)  
Substituting this expression into the objective function, we obtain a function of \( t^2 \) alone, which is easily optimized numerically.

Given the unrestricted estimates \( \hat{\beta} \), as well as the estimates \( \hat{\beta}_0 \) and \( \hat{\tau}^2 \), we can finally obtain the “posterior expectation” of \( \beta \) as
\[ \hat{\beta}^{EB} = \hat{\beta}_0 \cdot M + \left( I + \frac{1}{\hat{\tau}^2} \hat{V} \right)^{-1} \cdot (\hat{\beta} - \hat{\beta}_0 \cdot M). \]  
(13)  
This is the empirical Bayes estimator of the coefficient vector of interest.

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

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

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

$$\|b\|_{\beta} := (b' \cdot \text{Var}(X) \cdot b)^{1/2}.$$ 

In the context of our empirical Bayes approach, the projection is with respect to the norm

$$\|b\|_{\beta,EB} = \left(b' \cdot \Sigma(t^2, \hat{V})^{-1} \cdot b\right)^{1/2}.$$ 

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

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

2.4 Related literature

The main contribution of the present paper is to bring together economic theory with the tools of the empirical Bayes paradigm, to leverage economic theory for improved estimation in a novel way. Our approach builds on a long tradition of research on empirical Bayes methods in statistics, which has its roots in the seminal contributions of [Robbins (1956)], who first considered the empirical Bayes approach for constructing estimators, and [James and Stein (1961)], who demonstrated that the conventional estimator for the mean of a multivariate normal vector (of dimension greater than
2) is inadmissible and dominated by empirical Bayes estimators. Empirical Bayes approaches were developed further by Efron and Morris (1973) and Morris (1983). The latter introduced the parametric empirical Bayes framework on which we build. Inference in empirical Bayes settings was discussed by Laird and Louis (1987) and Carlin and Gelfand (1990), among others (a review can be found in Casella et al. 2012). A good introduction to empirical Bayes estimation can be found in Efron (2010), another review is provided by Zhang (2003). In Section 4, we provide a theoretical characterization of the risk properties of our empirical Bayes procedure. This characterization relies on arguments related to those invoked by Xie et al. (2012). Ideas related to our approach, in a fully Bayesian setting, have been used in the literature on macroeconomic forecasting, where theoretical DSGE models can be used to inform priors for the parameters of statistical VAR models fit to the data. Del Negro and Schorfheide (2004) and Del Negro et al. (2007), for instance, construct hierarchical Bayesian models for VARs, with a hyperparameter measuring the fit of the theoretical model.

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

The focus of the present paper differs from Hansen (2016), who focuses on general statistical theory. Our focus is on leveraging economic theory in empirical research, on providing guidelines for estimator construction, and on implementations in several economic settings. The models and estimators we consider also differ from those in Hansen (2016). We construct estimators based on economic theory using the empirical Bayes paradigm. To this end, we consider a family of priors centered on theoretical restrictions. The resulting estimators differ from the component-wise linear interpolation proposed by Hansen (2016). Our estimators combine information from the data with extrapolations from other components of the parameters of interest, extrapolations that are implied by economic theory, as we discuss in Section 4.2.

The expository example we have discussed in Section 2 takes normality of unre-
stricted estimators and linearity of restrictions as given; this is asymptotically justified even in nonlinear models as Hansen (2016) demonstrates. Future research might more formally apply his asymptotic results to our settings. In sections 3.3 and 3.4 we also implement our general approach in economic settings involving non-normal models.

3 Applications

We now turn to several applications of our proposed approach. These applications are chosen from a broad array of fields: (i) Labor demand and wage inequality, (ii) financial asset returns and the capital asset pricing model, (iii) economic choices and general theories of decision making, such as utility maximization, and (iv) multinomial logit and mixed multinomial logit models of discrete choice in panel data.

The first two of these applications fall within the confines of the normal-linear framework introduced in Section 2, up to some minor modifications. The latter two applications demonstrate the possibility of extensions to nonlinear settings. In each of these settings we construct estimators shrinking toward an economic theory, where the meaning of “economic theory” differs across applications, ranging from parametric structural models of production or of preferences, to general equilibrium models of financial markets, to abstract theories of decision making.

3.1 Labor demand and wage inequality

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

3.1.1 Setup

Suppose there are $J$ types of workers, defined for instance by their level of education and their potential experience. Consider a cross-section of labor markets $i = 1, \ldots, n$. We adopt cross-sectional notation for simplicity, similar arguments apply to time series or panel data. Let $Y_{ij}$, where $j = 1, \ldots, J$, be the average log wage for workers
of type \( j \) in labor market \( i \), and let \( X_{ij} \) be the log labor supply of these same workers. Denote \( Y_i = (Y_{i1}, \ldots, Y_{ij}) \) and \( X_i = (X_{i1}, \ldots, X_{ij}) \). We are interested in the structural relationship between labor supply and wages, that is in the inverse demand function \( Y_i = y(X_i, \epsilon_i) \), where \( \epsilon_i \) denotes a vector of unobserved demand shifters of unrestricted dimension.

**CES-production functions, structural and unrestricted estimation**

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

\[
Y_{ij} - Y_{ij'} = \gamma_{jj'} + \beta_0 \cdot (X_{ij} - X_{ij'}) + \epsilon_{ij'}.
\] (14)

Equation (14) can be rewritten in a numerically equivalent way as a fixed effects regression with restrictions across coefficients:

\[
Y_{ij} = \alpha_i + \gamma_j + \sum_{j'} \beta_{jj'} X_{ij'} + \epsilon_{ij},
\] (15)

\[
\beta_{jj'} = \beta_0 \cdot M_{jj'},
\] (16)

\[
M_{jj'} = (I - \frac{1}{J} E)_{jj'} = \begin{cases} 
(1 - \frac{1}{J}) & j = j' \\
-\frac{1}{J} & j \neq j'
\end{cases}.
\]

Here \( I \) is the identity matrix, \( E \) is a matrix of 1s, and \( M \) is the demeaning-matrix, projecting \( \mathbb{R}^J \) on the subspace of vectors of mean 0. To verify this equivalence, take the difference \( Y_{ij} - Y_{ij'} \) based on Equation (15).

Rather than imposing the strong assumptions implied by the CES production function model or its generalizations, we could instead “let the data speak,” considering a linear specification with a large number of types \( J \) and unrestricted own- and cross-elasticities. We could try to estimate (15), using least squares, without imposing any cross-restrictions on the parameters \( \beta_{jj'} \). Relative to this model, the CES production function restricts the \( J^2 \)-dimensional parameter \( \beta \) to lie in a 1-dimensional subspace. Note, however, that Equation (15) is not identified without further restric-

---

5The CES production function takes the form \( f_i(N_{i1}, \ldots, N_{ij}) = \left( \sum_{j=1}^{J} \gamma_j N_{ij}^{\beta_0+1} \right)^{1/(\beta_0+1)} \), where \( N_{ij} = \exp(X_{ij}) \). Details are reviewed in the supplementary online appendix.
tions. Given the presence of the fixed effects $\alpha_i$, we cannot pin down the effect of labor supply on the overall level of wages. Adding an arbitrary vector to all rows of $\beta$, and adjusting the $\alpha_i$ accordingly, yields an observationally equivalent model. Differencing (15) across types $j$ however yields a model which is identified. Let $\Delta$ be a $(J-1) \times J$ matrix which subtracts the first entry from each component of a $J$ vector, $\Delta = (-e, I_{J-1})$, and define the differenced matrix of coefficients $\delta = \Delta \cdot \beta$. We will consider $\delta$ as our main object of interest, and estimate the unrestricted regression

$$\Delta \cdot Y_i = \Delta \cdot \gamma + \delta \cdot X_i + \Delta \cdot \epsilon_i. \quad (17)$$

There are $J \cdot (J-1)$ free slope parameters to be estimated in the matrix $\delta$. Relative to this general linear fixed effects model, the CES production function imposes

$$\delta = \Delta \cdot (\beta_0 \cdot M) = \beta_0 \cdot \Delta,$$

which implies $J^2 - J - 1$ additional restrictions. The last equation holds because $\Delta \cdot M = \Delta$.

3.1.2 Empirical Bayes estimators

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

Some minor modifications of the approach introduced in Section 2 are necessary. In particular, the coefficients of interest $\delta$ that we now consider are in matrix form. We denote the vectorized version of $\delta$, stacking the columns on top of each other, by $\delta^\dagger$, and similarly for other matrices. Furthermore, a family of priors is most naturally specified for $\beta$ while estimation is for $\delta = \Delta \cdot \beta$. We model the coefficient matrix $\beta$ as corresponding to the coefficients of the structural CES model plus some disturbances,
that is
\[
\beta = (\beta_{jj'}) = \beta_0 \cdot M + \zeta
\]
\[
\zeta_{jj'} \sim iid N(0, \tau^2).
\]
Differencing this model yields
\[
\delta = \Delta \cdot \beta = \beta_0 \cdot \Delta + \Delta \cdot \zeta
\]
The variance of the second term, reflecting “prior uncertainty,” is given by
\[
\text{Var}((\Delta \cdot \zeta)') = \tau^2 \cdot P \otimes I,
\]
where
\[
P := \Delta \cdot \Delta' = I_{J-1} + E_{J-1}
\]
and \(\otimes\) is the Kronecker product of matrices. This implies a prior variance of the unrestricted OLS estimator \(\hat{\delta}_\tau\) equal to
\[
\Sigma(\tau^2, V) = \text{Var}(\hat{\delta}_\tau) = \tau^2 \cdot P \otimes I + V.
\]
Substituting a consistent estimator \(\hat{V}\) for \(V\), we obtain the empirical Bayes estimators of \(\beta_0\) and \(\tau^2\) as solutions to the maximum (marginal) likelihood problem
\[
(\hat{\beta}_0, \hat{\tau}^2) = \arg\min_{b_0, t^2} \log \left( \det(\Sigma(t^2, \hat{V})) \right)
\]
\[
+ (\hat{\delta}_\tau - b_0 \cdot \Delta_\tau)' \cdot \Sigma(t^2, \hat{V})^{-1} \cdot (\hat{\delta}_\tau - b_0 \cdot \Delta_\tau).
\]
(18)
Given \(t^2\), the optimal \(b_0\) is equal to
\[
\hat{b}_0 = (\Delta \cdot \Sigma(t^2, \hat{V})^{-1} \cdot \Delta')^{-1} \cdot \Delta \cdot \Sigma(t^2, \hat{V})^{-1} \cdot \hat{\delta}_\tau.
\]
Substituting this expression into the objective function, we obtain a function of \(t^2\) alone that we optimize numerically. Given the unrestricted estimates \(\hat{\delta}_\tau\), as well as the estimates \(\hat{\beta}_0\) and \(\hat{\tau}^2\), we obtain the empirical Bayes estimator of \(\delta\) as
\[
\hat{\delta}_\tau^{\text{EB}} = \hat{\beta}_0 \cdot \Delta_\tau + P \otimes I \cdot \left( P \otimes I + \frac{1}{\hat{\tau}^2} \hat{V} \right)^{-1} \cdot (\hat{\delta}_\tau - \hat{\beta}_0 \cdot \Delta_\tau).
\]
(19)

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

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

$$Y_{ij} - Y_{i1} = (\gamma_j - \gamma_1) + \sum_{j'} \delta_{jj'} X_{ij'} + \beta_0 \cdot (\tilde{X}_{ikj} - \tilde{X}_{i1}) + (\epsilon_{ij} - \epsilon_{i1}). \quad (20)$$

In this setting the matrix $\delta$ captures the additional effect of labor supply on relative wages beyond the effect already taken care of by the term $\beta_0 \cdot (\tilde{X}_{ikj} - \tilde{X}_{i1})$.

The canonical CES model implies the restriction $\delta = 0$. The unrestricted approach estimates versions of this equation with $\delta$ left fully flexible. Our empirical Bayes approach applied to this setting takes as its point of departure a first stage unrestricted estimator $(\hat{\delta}, \tilde{\beta}_0)$ of $(\delta, \beta_0)$, with estimated covariance matrix $\tilde{V}$. We then consider the family of priors

$$\delta \sim N(0, \tau^2 \cdot P \otimes I),$$

where, as before, $\beta_0$ and $\tau^2$ are hyperparameters. Denote the variance of the unrestricted estimators given $\beta_0$ and $\tau^2$ by

$$\Sigma(\tau^2, V) = \text{Var}
\begin{pmatrix}
\hat{\delta} \\
\tilde{\beta}_0
\end{pmatrix}
= \begin{pmatrix}
\tau^2 \cdot P \otimes I & 0 \\
0 & 0
\end{pmatrix} + V;$$

the conditional mean is given by $(0, \beta_0)$. We obtain the empirical Bayes estimators of $\beta_0$ and $\tau^2$ as solutions to the maximum (marginal) likelihood problem

$$(\hat{\beta}_0, \hat{\tau}^2) = \arg\min_{\beta_0, \tau^2} \log \left( \det(\Sigma(\tau^2, \tilde{V})) \right)
+ (\tilde{\delta}, \tilde{\beta}_0 - b_0)' \cdot \Sigma(\tau^2, \tilde{V})^{-1} \cdot (\tilde{\delta}, \tilde{\beta}_0 - b_0). \quad (21)$$
Given $t^2$, the optimal $b_0$ is equal to
\[
\hat{\beta}_0 = \left(e \cdot \Sigma(t^2, \hat{V})^{-1} \cdot e'\right)^{-1} \cdot e \cdot \Sigma(t^2, \hat{V})^{-1} \cdot (\hat{\delta}_\tau, \tilde{\beta}_0),
\]
where $e = (0, \ldots, 0, 1)$. Substituting this expression into the objective function, we obtain a function of $t^2$ alone that we optimize numerically. We finally obtain the empirical Bayes estimator of $\delta$ as
\[
\hat{\delta}^{\text{EB}}_1 = (\hat{\tau}^2 \cdot P \otimes I, 0) \cdot \Sigma(\hat{\tau}^2, \hat{V})^{-1} \cdot (\hat{\delta}_\tau, \tilde{\beta}_0 - b_0)^\prime.
\] (22)

### 3.1.3 Empirical application

We now turn to our empirical application, studying labor demand in the United States. We use data that have been studied in the literatures on the impact of immigration on native wages and on the impact of skill-biased technical change; see for instance Card (2009), Autor et al. (2008), and Acemoglu and Autor (2011).

We motivated our estimator arguing that (i) qualitative conclusions tend to be sensitive to auxiliary assumptions imposed in structural models, and (ii) estimates often have a high variance when not imposing a model. Card (2009, p5f) and Borjas et al. (2012) discuss an important example, the estimated impact of past migration on wage inequality in the US. 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.

In this section, we do not focus on the labor market impact of migration, but rather study the impact of historical changes of the labor force composition on relative wages in general. Rather than imposing one or the other of the models (4-type CES, nested 2-type CES), we allow for arbitrary patterns of substitutability across a larger number of types, but use our empirical Bayes methodology to shrink to the canonical 2-type CES model used in the literature.

---

Card (2009) argues that the assumptions of such a specification are justified by statistical tests.
Data  Our analysis is based on the American Community Survey (ACS) data and Current Population Survey (CPS) data used in much of the literature. We build two aggregate data-sets. The first is a state-level panel for the years 1960, 1970, 1980, 1990, and 2000 using the CPS, and 2006 using the ACS. Our construction of this data-set builds on the specifications and the code provided by Borjas et al. (2012). The second data-set is a national annual time-series for the years 1963-2008 using the March CPS. Here we build on the specifications and code provided by Acemoglu and Autor (2011), including their pre-cleaning of the data.

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

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

We next aim to estimate the same parameter using our state-level decadal panel, and controlling for time and state fixed effects. Doing so, we find an elasticity of substitution of the same sign, but much smaller magnitude: -0.06, with a standard error of 0.04. We do not wish to take a stance on what causes this divergence of findings between the time-series and the state panel, but will proceed with obtaining our main estimates from the panel data. Using panel might be preferable to the extent that it allows us to control for business cycle variation and secular time trends using
We now turn to our analysis using more disaggregated types of workers, classifying workers into 8 types by level of education and potential experience. The top left graph in Figure 2 shows the historical evolution of log wages of all types relative to the wage of high school dropouts with less than 20 years of potential experience. Clearly, there are patterns in the evolution of wages not captured by the classification into just 2 types. In particular, inequality across sub-types is rising over time, but in a non-linear manner.

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

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

The bottom left graph, and the second set of estimates in Table 1, are based on OLS estimation of the unrestricted model. These estimates suggest, as does the structural model, that changes of labor supply have induced a compression of wages over the initial three decades of our period. Some additional patterns emerge however. First, shifts in labor supply induced a widening of inequality over the most recent two decades. Second, these shifts also induced a compression of wages between different workers with high school degrees or less and a widening between those with more than high school education.

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

The estimated \( \hat{\tau}^2 \), our measure of model fit, is of a somewhat larger magnitude than the variance of the OLS coefficient estimates. This implies some, but not excessive, shrinkage towards the restricted estimates, thus leading to qualitatively similar conclusions of unrestricted and empirical Bayes predictions. This also suggests that the 2-type CES model does not provide a particularly good fit to our panel data. Taken at face value, our estimates also suggest that past migration did not affect wage inequality between native workers much, in line with the conclusions of Card (2009).

3.2 Asset returns and the capital asset pricing model

In our second 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.

3.2.1 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 \( \omega_{it} \). 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

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

This number $\beta_i$ 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

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

holds for all assets $i$. This is a testable implication, and various tests have been proposed, including by [Jensen (1968)](1968) and by [Jensen et al. (1972)](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)](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:

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

where $\text{Cov}(R_t^M, \epsilon_{it}) = 0$. The slope of this predictor is equal to $\beta_i$ by definition, under the assumption of stationarity. Estimating the coefficients of the best linear predictor using OLS, we obtain unrestricted estimators $\left(\hat{\alpha}_i, \hat{\beta}_i\right)$, with estimated sampling variance $\hat{V}_i$. Allowing for general heteroskedasticity and intertemporal dependence, we can use a heteroskedasticity and autocorrelation robust estimator for $\hat{V}_i$. We do not need to impose any assumptions on cross-sectional dependence (across assets $i$)

---

7In this section we stick with the standard finance notation of $\alpha_i$ and $\beta_i$, deviating slightly from our previous notation based on which we would subsume both of these, for all $i$, in a vector of interest $\beta$.  

23
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 (23), 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 (24) that imposes this restriction by running a time series OLS regression of $R_{it}$ on $R_{tM}$ with no intercept.

### 3.2.2 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 $\alpha_i$ and $\beta_i$ 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 to the present setting, we propose to take the unrestricted OLS estimates $(\hat{\alpha}_i, \hat{\beta}_i)$ and $\hat{V}_i$ as point of departure when constructing estimates which are shrunk toward the theory. We consider the family of priors

$$(\alpha_i, \beta_i) \sim \text{iid } N((0, \beta_0), \Upsilon). (25)$$

If $\Upsilon_{11}$ were set equal to 0, this prior would impose the restriction of Equation (23), as implied by CAPM. The parameter $\Upsilon_{11}$ thus takes the role that $\tau^2$ had in the simplified setting of Section 2.

In the second step of estimation we need to obtain estimates of the hyperparameters $\beta_0$ and $\Upsilon$. Previously, we estimated hyperparameters via maximization of the marginal likelihood. Such an approach is complicated in the present setting by the fact that the estimates $(\tilde{\alpha}_i, \tilde{\beta}_i)$ 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 $\beta_0$ and $\Upsilon$ that avoid the need to estimate these covariances. In particular, let

$$\tilde{\beta}_0 = \frac{1}{N} \sum_i \tilde{\beta}_i \quad (26)$$

and

$$\tilde{\Upsilon} = \frac{1}{N} \sum_i \left( \left( \frac{\tilde{\alpha}_i}{\tilde{\beta}_i - \tilde{\beta}_0} \right) \cdot \left( \tilde{\alpha}_i, \tilde{\beta}_i - \tilde{\beta}_0 \right) - \tilde{V}_i \right). (27)$$
Empirical Bayes estimates of \((\alpha_i, \beta_i)\) are then obtained in the final step via

\[
(\hat{\alpha}^{EB}_i, \hat{\beta}^{EB}_i) = \left(0, \hat{\beta}_0 \right) + \hat{\Upsilon} \cdot \left(\hat{\Upsilon} + \hat{\Upsilon}_i\right)^{-1} \left(\hat{\alpha}_i, \hat{\beta}_i - \hat{\beta}_0\right).
\]

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

\[
\beta_i \sim \text{iid } N(\beta_0, \nu^2).
\]

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

### 3.2.3 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)](http://www.crsp.com/products/research-products/crsp-us-stock-databases) and is included for comparability.

Following the literature, we use market excess returns \(R_M^t\) 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.\(^8\) We drop duplicates and all firms not existing for more than 2 months.

**Predictive performance, 2010-15** In order to compare alternative estimators of the asset-specific parameters \(\alpha_i\) and \(\beta_i\), 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 \(R_{M,t+1}\) and estimates of \(\alpha_i, \beta_i\), based on observations

for the periods 1 through $t$. We repeatedly estimate $\alpha_i$ and $\beta_i$ 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 $\beta_i$ to the grand mean.

Table 2 here

Mean squared errors, averaged across assets and across time periods, are reported in Table 2. 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 $\alpha_i$ and $\beta_i$. 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 $\alpha_i$ and $\beta_i$.

Applying the method of moments estimators of equations (26) and (27) to the data for 2011-2015, we obtain estimates $\hat{\beta}_0 = 0.96$ and

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

which implies a correlation between $\alpha$ and $\beta$ 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 $\hat{\Upsilon}_{11}$ of $\alpha_i$ from 0 equals 0.0006. Recall that $\hat{\Upsilon}_{11}$ corresponds the role of $\tau$ in the simplified setting considered in Section 2 and thus provides a measure of model fit.
We plot the distribution of estimates for $\alpha_i$ and $\beta_i$ across assets $i$ in Figure 3. In interpreting these figures, note the different scale of the axes between $\alpha$ and $\beta$. As suggested by the estimated $\hat{\Upsilon}$, it appears that $\alpha$ has very small dispersion around 0, in line with the predictions of CAPM, and the same is true for estimators $\hat{\alpha}_i$. 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 $\alpha$ and $\beta$. The median shrinkage factor of the OLS estimates of $\alpha$ toward 0 implied by the empirical Bayes estimator equals 0.82, while the median shrinkage factor of the OLS estimates of $\beta$ toward $\hat{\beta}_0$ equals 0.91. For purposes of comparison, the first row of Figure 3 shows the distribution of estimates of $\beta$ for the restricted empirical Bayes estimator, which imposes $\alpha = 0$.

Figure 4 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 4 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 $\hat{\beta}_0 = 1.178$ and

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

which implies a correlation between $\alpha$ and $\beta$ across assets of $-0.05$. These estimates again suggest that the predictions of CAPM are very accurate – the mean square deviation of $\alpha_i$ from 0 equals 0.0001.
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 $\alpha$ toward 0 implied by the empirical Bayes estimator equals 0.54, while the median shrinkage factor of the OLS estimates of $\beta$ toward $\hat{\beta}_0$ equals 0.75. Figure 6 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.

Let us summarize our findings. A key prediction of CAPM – that all of the $\alpha_i$ 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 $\alpha_i$ 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.

3.3 Choice probabilities and economic decision theory

In our third 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. 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 $\mathcal{C}$ of subsets $C$ of $X$. These are the possible choice sets faced by individuals. This setting is similar to the one considered by McFadden.
In this setting, theories of decision making can be described by a collection \( \mathcal{D} \) of choice functions \( d \) mapping each \( C \in \mathcal{C} \) 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 \( \mathcal{C} \) 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 \( \mathcal{D} \), 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. \( \pi_d \) is the probability that a random agent from the population of interest makes their choices according to the choice function \( d \in \mathcal{D} \). \( \Delta \) is the simplex of probability distributions on the elements of \( \mathcal{D} \). Suppose now that agents (choice functions) are randomly assigned to choice sets. 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

\[
p = D \cdot \pi, \tag{29}
\]

and thus, in particular,

\[
p \in D \cdot \Delta. \tag{30}
\]

The set \( \mathcal{P} := D \cdot \Delta \) 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 \( p \in \mathcal{P} \) is the empirical content of the theory of choice that imposes \( d \in \mathcal{D} \) for all agents. When the theory considered is strict preference maximization, \( D \cdot \Delta \) is

\footnote{This is a conceptual reference point and might be generalized in a number of ways. We could for instance replace independence by conditional independence given observed covariates.}
the set of conditional choice probabilities satisfying the stochastic axiom of revealed preferences, as shown by McFadden (2005). By Farkas’ Lemma we have $p \in \mathcal{P}$ if and only if there is no vector $q$ such that

$$q' \cdot D \geq 0$$
$$q' \cdot p < 0.$$

**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

$$\hat{p}^u_{xC} = \frac{n_{xC}}{n_C}. \quad (31)$$

Restricted estimation estimates these probabilities while imposing that the vector $p$ is consistent with our theory of choice, so that $p \in \mathcal{P} = D \cdot \Delta$. 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

$$\hat{p}^s = \arg\max_{p \in \mathcal{P}} \sum_{C, x \in C} n_{xC} \cdot \log(p_{xC}). \quad (32)$$

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

$$(n_{xC})_{x \in C} | p \sim MN((p_{xC})_{x \in C}, n_C) \quad (33)$$
$$(p_{xC})_{x \in C} \sim Dir(\alpha \cdot (\bar{p}_{xC})_{x \in C}) \quad (34)$$

$$\bar{p} \in \mathcal{P}$$
$$\alpha \in \mathbb{R}^+.$$
We impose furthermore that independence of \((n_{xC})_{x \in C}, (p_{xC})_{x \in C}\) holds across different choice sets \(C\). Equation (33) is implied by i.i.d. sampling. Equation (34) defines a family of priors, indexed by \(p\) and \(\alpha\). This family of priors will be used to construct the empirical Bayes estimator, where the hyperparameters \(\alpha\) 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 \(\alpha\) and \(\overline{p}\):

\[
P((n_{xC})|(p_{xC})) = \prod_C \left[ \left( \frac{n_C!}{\prod_{x \in C} n_{xC}!} \right) \times \prod_{x \in C} p_{xC}^{n_{xC}} \right]
\]

\[
P((p_{xC})) = \prod_C \left[ \left( \frac{\Gamma(\alpha)}{\prod_{x \in C} \Gamma(\alpha \cdot \overline{p}_{xC})} \right) \times \prod_{x \in C} p_{xC}^{\alpha \overline{p}_{xC}} \right]
\]

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

\[
P((n_{xC})) = \prod_C \left[ \left( \frac{\Gamma(\alpha) \cdot n_C!}{\prod_{x \in C} \Gamma(\alpha \cdot \overline{p}_{xC}) \cdot n_{xC}!} \right) \times \prod_{x \in C} \frac{\Gamma(\alpha \cdot \overline{p}_{xC} + n_{xC})}{\Gamma(\alpha + n_C)} \right].
\]

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 \(\alpha\) and \(\overline{p}_{xC}\) as well as the observed \(n_{xC}\), the expectation of \(p_{xC}\) is given by

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

(35)

Plugging in estimates for \(\alpha\) 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 \(\overline{p} \in \mathcal{P}\), 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 \(\alpha\) and \(\overline{p}_{xC}\) maximizes the marginal likelihood, or equivalently, its logarithm:

\[
(\alpha^{EB}, \overline{p}^{EB}) = \arg\max_{\alpha, \overline{p}}
\]  

(36)
\[
\sum_{C} \left( \log(\Gamma(\alpha)) - \log(\Gamma(\alpha + n_{C})) + \sum_{x \in C} \left( \log(\Gamma(\alpha \cdot \bar{p}_{xC} + n_{xC})) - \log(\Gamma(\alpha \cdot \bar{p}_{xC})) \right) \right),
\]
subject to $\bar{p} \in \mathcal{P}$ 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.

### 3.4 Multinomial logit and mixed multinomial logit

In our fourth 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 characteristics 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

\[ u_{ijt} = x_{ijt} \cdot \beta_i + \epsilon_{ijt}, \tag{37} \]

where the $\epsilon_{ijt}$ are i.i.d. given $x_i$ and follow an EV1 distribution, while the $\beta_i$ 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 $\beta_i$ across decision makers.
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

\[
P^{\text{MML}}(j_1, \ldots, j_T | x_{i..}) = \int \left( \prod_t \frac{\exp(x_{ij_t} \cdot \beta)}{\sum_j \exp(x_{ij_t} \cdot \beta)} \right) f(\beta | \eta) d\beta. \tag{38}
\]

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 \(\beta\), so that

\[
P^{\text{ML}}(j_1, \ldots, j_T | x_{i..}) = \prod_t \frac{\exp(x_{ij_t} \cdot \beta)}{\sum_j \exp(x_{ij_t} \cdot \beta)}. \tag{39}
\]

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 \(\beta_i\) is normally distributed across \(i\), that is

\[
\beta_i | \eta \sim N(\mu, \Omega). \tag{40}
\]

Under this assumption about the distribution of \(\beta_i\), 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 \(\beta_i\). This contrasts with the commonly imposed assumption that the different components of \(\beta_i\) 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 \(\Omega\).

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 \((\mu, \Omega)\), using maximum likelihood. The second approach estimates the restricted multinomial logit model, parametrized by \(\beta\), using maximum likelihood again.

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

$$\Omega \sim IW \left( \frac{1}{\tau} + p + 1, 0 \right), \quad (41)$$

where $p = \text{dim}(\beta_i)$. This parametrization is chosen to yield simple expressions for the conditional expectation of $\Omega$ below. We leave the mean vector $\mu$ 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 $\mu$ and $\tau$ by maximizing the marginal likelihood. Then, we estimate the variance matrix $\Omega$ by its posterior mean, given $\mu$ and $\tau$ 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 $\Omega$, we propose sampling from the posterior distributions of $\Omega$ and $\beta_i$, given the observed choices and given $\theta = (\mu, \tau)$, using Gibbs sampling\(^{10}\). 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).

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

$$\Omega | \mu, \tau, \beta_1, \ldots, \beta_n \sim IW \left( \frac{1}{\tau} + p + 1 + n, n \cdot \hat{\Omega} \right), \quad (42)$$

where

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

and thus

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

If we hypothetically were to observe the $\beta_i$, then empirical Bayes estimation would involve linear shrinkage of the unrestricted variance estimator $\hat{\Omega}$ toward 0. As the

\(^{10}\)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.
hyperparameter $\tau$ varies between 0 and $\infty$, the empirical Bayes estimator of $\Omega$ varies between 0 (so that we recover the restricted multinomial logit model) and the unrestricted maximum likelihood estimator $\hat{\Omega}$ of $\Omega$. 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 $\beta_i$ 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 $\beta_i$, where the random coefficients are considered to be drawn from a population distribution $f(\beta|\eta)$, so that we shrink $\beta_i$ when interested in individual $i$’s preferences. Conventional unrestricted estimators estimate $\eta$ 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, $\hat{\tau}$ will be smaller or larger, so that the estimator of $\Omega$ shrinks more or less.

4 Properties of the empirical Bayes estimator and its risk function

We next analyze some properties of the proposed empirical Bayes estimator. Throughout this section we consider the simplified setting of Section 2. Section 4.1 lists the assumptions we will impose in this section. In Section 4.2 we show consistency of the estimator and how counterfactual predictions combine theory and available evidence in a data-driven, intuitive way. Section 4.3 characterizes the risk properties of $\hat{\beta}_{EB}$. Theorem 1, in particular, characterizes the risk function using an asymptotic approximation to the behavior of $\hat{\tau}^2$. This theorem forms the theoretical basis of our claims of dominance relative to unrestricted estimation. This result is related to the classic characterization of the risk of the James-Stein shrinkage estimator. Section 4.4 explores the geometry of the mapping from the preliminary, unrestricted estimator $\hat{\beta}$.
to the empirical Bayes estimator $\hat{\beta}^{EB}$.

### 4.1 Assumptions

In order to concisely state the following results, we list all the assumptions that we impose at some point in this section. Throughout this section we consider the empirical Bayes estimator of Assumption 1. This is the estimator we constructed for the simplified setting of Section 2. When studying the risk properties of $\hat{\beta}^{EB}$ and the geometry of the mapping from $\hat{\beta}$ to $\hat{\beta}^{EB}$, we use canonical coordinates as in Assumption 2 which imposes that $V$ is diagonal. We will consider two types of asymptotics. Standard large sample asymptotics as in Assumption 3 yield consistency of $\hat{\beta}^{EB}$ for $\beta$. Large dimension asymptotics as in Assumption 4, applicable whenever $\dim(\beta)$ is large, allow us to characterize the risk function (mean squared error) of $\hat{\beta}^{EB}$ in an intuitive way.

**Assumption 1** (Empirical Bayes estimator)

Suppose we observe $\hat{\beta}$, $\hat{V}$, and $M$, where $E[\hat{\beta}] = \beta$ and $\text{Var}(\hat{\beta}) = V$. Let $\Sigma(\tau^2, \hat{V}) = \tau^2 \cdot I + \hat{V}$ and define the empirical Bayes estimator $\hat{\beta}^{EB}$ as follows.

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

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

**Assumption 2** (Canonical form)

$\hat{V} = V$ and $V$ is diagonal,

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

**Assumption 3** (Large sample asymptotics)

The distribution of $\hat{\beta}$ and $\hat{V}$ is a function of sample size $n$ such that

$$\hat{\beta} \xrightarrow{p} \beta,$$

$$\hat{V} \xrightarrow{p} 0$$

as $n \to \infty$.

**Assumption 4** (Large dimension asymptotics, random coefficients)

Under assumptions 1 and 2, let $\hat{\beta}, \beta, \text{diag}(V)$, and $M$ be $J$ vectors with components
which are i.i.d. draws from some distribution $P$ with finite second moments, such that

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

Consider limits as $J \to \infty$, where $P$ does not depend on $J$.

**Discussion** Assumption 1 is as before. Assumptions 2 through 4 deserve some justification. The canonical form of Assumption 2 has two parts. First, we impose that the variance $V$ of the unrestricted estimator is known in order to simplify our exposition. This is for convenience only. Asymptotic justifications are easily obtained (see for instance Hansen 2016, section 4). Second, we impose canonical coordinates. Use of such coordinates is without loss of generality, since $V$ can always be diagonalized using an orthogonal transformation. The family of priors considered, $\beta - \beta_0 M \sim N(0, \tau^2 I)$, and the squared error loss function, $\| \hat{\beta}^{EB} - \beta \|^2$, are both invariant under such transformations.

The large sample asymptotics of assumption 3 require only standard consistency for $\hat{\beta}$ and $\hat{V}$. Such standard consistency is satisfied for many conceivable unrestricted estimators. We will use it to show consistency of $\hat{\beta}^{EB}$ in Proposition 1.

The large dimension asymptotics of Assumption 4 are less familiar. These asymptotics are designed to approximate the risk of our estimator for large $J$, and are used in Theorem 1. For large $J$, variability in $\hat{\tau}^2$ becomes negligible relative to variability in $\hat{\beta}$, and the squared error $\| \hat{\beta}^{EB} - \beta \|^2$ approaches the mean squared error by a law of large numbers across components $j$. To formalize this idea, we need to spell out what happens to the components of $\beta$, $V$, and $M$ as $J$ increases. The easiest way to do this is in terms of the random coefficient setup of Assumption 4. An alternative to the random coefficient setup would be to consider deterministic sequences $(\beta_j, v_j, M_j)$ and impose constraints on their behavior. This is the approach taken by Xie et al. (2012), for instance. Further, Abadie and Kasy (2015) elaborate in detail on random coefficient large dimension asymptotics and on how the random coefficient (empirical Bayes) perspective relates to the compound risk perspective. We use random coefficient large dimension asymptotics to show uniform risk consistency results for a variety of estimators (including ridge, lasso, and pre-testing) and a variety of ways to pick the tuning parameter $\tau$ (including Stein’s unbiased risk estimate and cross
validation).

Finally, note that we are not imposing normality of \( \hat{\beta} \) at any point in this section. We used normality in Section 2 to motivate the construction of our estimator; we do not require normality to characterize its behavior. Under large dimension asymptotics, \( \hat{\beta} \) need not be asymptotically normal, either.

4.2 Consistency and data-driven predictions

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

**Proposition 1** (Consistency)

Consider the estimator of assumption \( \text{I} \) under large sample asymptotics as in assumption \( \text{III} \). Then

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

as \( n \) goes to infinity.

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

The formula for \( \hat{\beta}^{EB} \) given in Equation \( \text{(13)} \) shows that the empirical Bayes estimator interpolates between the unrestricted estimator \( \hat{\beta} \) and the structural estimator \( \hat{\beta}^s = \hat{\beta}^0 \cdot M \). Suppose we are interested in making a prediction of the form \( \hat{y} = x \cdot \hat{\beta}^{EB} \).

Heuristically, we would like our prediction to be based on the data alone (neglecting the structural model) whenever the data by themselves do allow us to make a precise prediction. When, on the other hand, a prediction of counterfactuals based on the data alone would be imprecise, we would like to leverage the theoretical model. The following proposition shows that this is exactly how the empirical Bayes estimator behaves.

**Proposition 2** (Counterfactual predictions)

Consider the estimator of assumption \( \text{I} \). Consider the prediction at \( x \), \( \hat{y} = x \cdot \hat{\beta}^{EB} \), and assume that \( \hat{V} \) is non-singular. Then

\[
\left| \hat{y} - x \cdot \hat{\beta} \right| \leq \frac{\sqrt{x^t \hat{V} x}}{\tau} \cdot \| \hat{\beta} \|,
\]

38
and $|\hat{y} - x \cdot \hat{\beta}^s| \leq \hat{\tau} \cdot \sqrt{\hat{V}^{-1} x \cdot \|\hat{\beta}\|}$.

The first inequality of proposition 2 tells us that empirical Bayes predictions are close to unrestricted predictions whenever the standard deviation of the latter, $\sqrt{\hat{V} x}$, is small relative to the measure of model fit $\hat{\tau}$. The second inequality tells us that empirical Bayes predictions are close to predictions using the structural model when the reverse situation holds. To gain intuition for this result, rearrange Equation (13),

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

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

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

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

### 4.3 The risk function of our estimator

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

We now proceed to characterize the risk of our estimator. The key argument in our characterization is that variability of $\hat{\tau}^2$ can be neglected for large $J$ when calculating the MSE. After rewriting the estimator in a canonical form, we formalize this argument in Theorem 1. We then discuss the properties of the asymptotic approximation to risk obtained in this way and compare it to an oracle-optimal choice of $\tau^2$. 

39
Canonical form  Consider the estimator of Assumption 1 in its canonical form, as imposed in assumption 2, so that \( \hat{V} = \text{diag}(v_j) \). Under this assumption, the empirical Bayes estimator is given by a component-wise weighted average of \( \hat{\beta}_0 \cdot M \) and \( \hat{\beta} \),

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

(43)

The hyperparameters \( \beta_0 \) and \( \tau^2 \) are estimated by maximizing the marginal log likelihood, which now simplifies to

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

(44)

The canonical form makes transparent how our estimator differs from the family of estimators considered by Hansen (2016), which (in our setting) take the form

\[
\left( 1 - \hat{\lambda} \right) \cdot \hat{\beta}_0 \cdot M + \hat{\lambda} \cdot \hat{\beta},
\]

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

Asymptotic characterization of risk  Our goal in this subsection is to characterize the squared error

\[
SE = \frac{1}{J} \cdot \sum_j \left( \hat{\beta}_j^{EB} - \beta_j \right)^2,
\]

and the corresponding mean squared error \( MSE = E[SE] \) of the empirical Bayes estimator. In order to obtain our desired characterizations, we consider an asymptotic approximation where \( J \) becomes large, such that \( \hat{\beta}_0 \) and \( \hat{\tau}^2 \) converge in probability to constants. This is accomplished by the large dimension asymptotics of Assumption 4.

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

\footnotetext{11}{Actually, \( \hat{\beta}^{EB}(b_0, \tau^2) \) is the Bayes estimator for the prior \( \beta \sim N(b_0 M, \tau^2 I) \).}

40
of variance and squared bias terms,

\[
MSE(b_0, \tau^2) = E \left[ \left( \frac{\tau^2}{\tau^2 + v_j} \right)^2 v_j + \left( \frac{v_j}{\tau^2 + v_j} \right)^2 \left( \beta_j - b_0 M_j \right)^2 \right]. \tag{45}
\]

Let \((\beta_0, \tau^*)\) be the maximizer of the expected log-likelihood given by the expectation of (44),

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

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

**Theorem 1**

Consider the estimator of Assumption \(I\) in the canonical form of Assumption \(I\) under large dimension asymptotics as in assumption \(I\). Then

\[
SE(\hat{\beta}_0, \tau^2) - MSE(\beta_0, \tau^2) \rightarrow^p 0
\]

as \(J \rightarrow \infty\).

**Discussion**  Recall that we obtain unrestricted estimation and structural estimation as limiting cases of our proposed estimator, where \(\tau^2 \rightarrow \infty\) corresponds to unrestricted estimation and \(\tau^2 \rightarrow 0\) to restricted estimation. The mean squared error \(MSE(b_0, \tau^2)\) for given values of \((b_0, \tau^2)\) is equal to the sum of a variance term and a squared bias term, cf. Equation (45). The mean squared error of the unrestricted estimator contains only variance terms, \(MSE(b_0, \infty) = E[v_j]\), and the mean squared error of the structural estimator converges to an average containing only bias terms, \(\min_{b_0} MSE(b_0, 0) = \min_{b_0} E \left[ (\beta_j - b_0 M_j)^2 \right]\). Theorem 1 then immediately yields the following corollary.

**Corollary 1**

Under the assumptions of Theorem 1, and for large enough \(J\), empirical Bayes has lower mean squared error than unrestricted estimation if

\[
MSE(\beta_0, \tau^2) < E[v_j],
\]
and larger mean squared error if this inequality is reversed. It has lower mean squared error than restricted estimation for large $J$ if

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

and larger mean squared error if this inequality is reversed.

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

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

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

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

How does $\tau^*^2$ relate to the optimal choice of $\tau^x^2$? As can be seen from the first order conditions, both are weighted averages of $(\beta_j - \beta_0 M_j)^2$. The weights differ slightly, however. Minimization of the mean squared error assigns a slightly larger weight to draws $j$ with smaller values of $v_j$, relative to to maximization of the expected log likelihood. For homoskedastic settings ($v_j$ constant), or settings where $v_j$ and $\beta_j$ are independent across $j$, the two objectives do in fact coincide. In these cases it is immediate that our empirical Bayes estimator dominates both unrestricted and restricted estimation for large enough $J$. It is also possible to reverse the dominance of empirical Bayes relative to unrestricted estimation, however, by introducing strong correlation across $j$ between $\beta_j$ and $v_j$, as the following corollary shows.

**Corollary 2**

Under the assumptions of Theorem 1, suppose $M \equiv 0$ and

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

\[ \tau^* = 0 \]
\[ MSE(\beta_0, \tau^*) = 2 \]
\[ MSE(b_0, \infty) = 1, \]

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

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

4.4 Geometry of our empirical Bayes estimator

We conclude our analysis of the properties of \( \hat{\beta}^{EB} \) by studying its geometry. The proposed empirical Bayes estimator can be seen as providing a mapping from an unrestricted (preliminary) estimate \( \hat{\beta} \) to an empirical Bayes estimate \( \hat{\beta}^{EB} \). Understanding this mapping is key for understanding the behavior of our estimator. For this section, consider again the estimator of Assumption 1 in the canonical form Assumption 2, where \( \hat{V} = \text{diag}(v_i) \).

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

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

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

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

43
Given $\hat{\tau}^2$, this first order condition implies that $\hat{\beta}$ must be somewhere on the surface of an ellipsoid with semi-axes that have length

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

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

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

(47)

Note that this value does not depend on $\hat{\beta}$ beyond its effect on $\hat{\tau}^2$. All estimates $\hat{\beta}_{EB}$ corresponding to a given value of $\hat{\tau}^2$ are on the surface of a sphere with this radius. Note finally that there is a natural lower bound on $\hat{\tau}^2$ of 0.\footnote{Since we impose this bound, our estimator resembles the positive-part James-Stein estimator.}

In particular, we have that $\hat{\tau}^2$ is equal to 0 for any values of $\hat{\beta}$ inside the ellipsoid with semi-axes of length

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

(48)

**Visual representation** We can illustrate the mapping from $\hat{\beta}$ to $\hat{\tau}^2$ and $\hat{\beta}_{EB}$ graphically when $\dim(\beta) = 2$. Suppose that $v_1 = 2$, and $v_2 = 1$. The top part of Figure 7 shows $\hat{\tau}^2$ as a function of $\hat{\beta}$. This function is flat and equal to 0 inside the white ellipsoid; it rises smoothly and approaches a circular cone for large $\hat{\beta}$. The bottom part of this same figure shows (i) $\hat{\beta}_{EB} - \hat{\beta}$ as a vector field (arrows are proportional to, but smaller than, this difference) and (ii) a contour plot of the length of these vectors, that is, of the amount of shrinkage relative to the unrestricted estimator.

The structure of this mapping gets more transparent when considering the analytic characterizations we just derived. Figure 8, in particular, plots (i) which values of $\hat{\beta}$ would imply such values of $\hat{\tau}^2$ and (ii) the corresponding estimates $\hat{\beta}_{EB}$, for various values of $\hat{\tau}^2$.

[Figures 7 and 8 here]

How can we interpret these figures? For small $\hat{\beta}$, the estimator concludes that the
“theory” is essentially correct, where the theory in this case reduces to the assumption $\hat{\beta} = 0$. As $\hat{\beta}$ gets larger, so does the estimated $\hat{\tau}^2$ – the theory is considered “less correct.” Deviations from 0 in the direction of the first coordinate are weighted less heavily as $\hat{\beta}_1$ has a larger variance (is less precisely estimated). $\hat{\beta}_1$ is shrunk most heavily if $\hat{\beta}_2$ seems to confirm the theory while $\hat{\beta}_1$ violates it moderately, as evident in the bottom right plot of Figure 7. When $\hat{\beta}$ is large, so is $\hat{\tau}^2$, and the theory is essentially disregarded; $\hat{\beta}^{EB}$ is basically equal to the unrestricted estimator, as evident in the bottom plots of Figure 8.

**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 $\beta_0$. This can be analyzed using the same “trick” as before, where we consider $\hat{\tau}^2$ and $\hat{\beta}_0$ to be given and derive the corresponding sets of $\hat{\beta}$ and $\hat{\beta}^{EB}$.

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

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

so that

$$\hat{\beta}_0 = \frac{\sum_j \hat{\beta}_j \cdot \frac{1}{\hat{\tau}^2 + v_j}}{\sum_j M_j \cdot \frac{1}{\hat{\tau}^2 + v_j}}. \quad (49)$$

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

$$\sum_j \frac{1}{\hat{\tau}^2 + v_j} = \sum_j \frac{(\hat{\beta}_j - \hat{\beta}_0 \cdot M_j)^2}{(\hat{\tau}^2 + v_j)^2}.$$ 

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

We can rephrase this argument by considering only $\hat{\tau}^2$ to be given. Conditional on $\hat{\tau}^2$, we get that $\hat{\beta}$ has to lie on the surface of a hyper-cylinder with ellipsoid basis
and axis going through the origin and pointing in the direction of the vector
\[
\left( \frac{1}{\hat{\tau}^2 + v_1}, \ldots, \frac{1}{\hat{\tau}^2 + v_J} \right).
\]
The corresponding estimates \( \hat{\beta}_{EB} \) are on the surface of a hyper-cylinder with spherical basis and the same axis. Note that the tilt of the axis depends on \( \hat{\tau}^2 \) and varies between \((1, \ldots, 1)\) for large \( \hat{\tau}^2 \) and \( \left( \frac{1}{v_1}, \ldots, \frac{1}{v_J} \right) \) for \( \hat{\tau}^2 = 0 \).

## 5 Conclusion

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

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

In a normal-normal setting with linear restrictions implied by economic theory, our approach leads to particularly tractable and interpretable estimators. We provide several theoretical results for this case. One of our key results, Theorem 1, provides a characterization of the risk function of our estimator. The theorem is based on an asymptotic approximation that implies that the variability of the estimated hy-
perparameters is negligible relative to variability of the estimates of interest. This assumption is justified as long as the dimension of the parameters of interest is large relative to the dimension of the hyperparameters.
A Inference

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

\[
P(\beta \in C | \theta) \geq 1 - \alpha
\]

and were first proposed by [Morris (1983)] and analyzed further by [Laird and Louis (1987) and Carlin and Gelfand (1990)]. Definition (50) arguably captures the natural notion of inference corresponding to empirical Bayes estimation. Empirical Bayes confidence regions are intermediate between frequentist confidence sets and Bayesian pre-posterior inference. The requirement of definition (50) is weaker than the requirement of frequentist coverage, \( P(\beta \in C | \eta) \geq 1 - \alpha \).

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

\[
P(\beta | \hat{\beta}, \hat{V}) = \int P(\beta | \hat{\beta}, \hat{V}, \theta) P(\theta | \hat{\beta}, \hat{V}) d\theta,
\]

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

\[
M(\beta | \hat{\beta}, \hat{V}) = \int P(\beta | \hat{\beta}, \hat{V}, \theta) Q_R(\theta | \hat{\beta}, \hat{V}) d\theta. \tag{51}
\]

Our inference procedure can be summarized as follows:

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

2. For each of these \( R \) draws, obtain estimates \( \hat{\theta}_r = (\hat{\beta}_0, r_2) \) by maximizing the marginal likelihood, as discussed in Section 2.3.

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

$$V_{rEB} = \text{Var}(\beta | \hat{\beta} = \hat{\beta}_r, \theta = \hat{\theta}_r)$$

$$= \tilde{\tau}^2 \cdot I - (\tilde{\tau}^2)^2 \cdot (\tilde{\tau}^2 \cdot I + \hat{\mathcal{V}})^{-1}$$

$$= \left( I + \frac{1}{\tilde{\tau}^2} \hat{\mathcal{V}} \right)^{-1} \cdot \hat{\mathcal{V}}.$$

4. Consider the mixture distribution

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

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

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

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

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

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

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

By the standard arguments for consistency of m-estimators (see for instance van der Vaart [2000] chapter 3), we get convergence of these hyperparameters, $(\hat{\beta}_0, \hat{\tau}^2) \xrightarrow{p} \arg\min \log \left( \det(\Sigma(t^2, 0)) \right) + (\beta - b_0 \cdot M)' \cdot \Sigma(t^2, 0)^{-1} \cdot (\beta - b_0 \cdot M)$

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

Proof of proposition 2: By Assumption 1,

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

and thus

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

By Equation 11,

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

where the last inequality holds by positive definiteness of \(\hat{V}\), which also implies
\[
\|\left(\frac{\hat{\tau}V^{-1/2} + \frac{1}{\tau}V^{1/2}}{\tau} \right)^{-1/2}\| \leq 1.
\]

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

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

**Lemma 1**

Suppose that \(MSE\) is continuous at \((\beta_0, \tau^*2)\), that \((\hat{\beta}_0, \hat{\tau}^2) \to p (\beta_0, \tau^*2)\), and that
\[
\sup_{(b_0, \tau^2) \in U} |SE(b_0, \tau^2) - MSE(b_0, \tau^2)| \to p 0,
\]

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

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

**Proof of lemma**

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

□

**Proof of Theorem**

We need to show that the sufficient conditions of lemma 1 are satisfied. Convergence
of $(\hat{\beta}_0, \hat{\tau}^2)$ to the pseudo-true parameters

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

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

It remains to be shown that uniform convergence of

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

holds in a neighborhood $U$ of $(\beta_0, \tau^*^2)$, where $E_J$ denotes the average from $j = 1, \ldots, J$. Such uniform convergence follows if we can show that the family of mappings

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

indexed by $(b_0, \tau^2) \in U$, is a Glivenko-Cantelli class, cf. van der Vaart (2000) chapter 19.2.

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

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

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

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

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

This envelope is integrable because Assumption 1 imposed finite second moments, given the form of $\hat{\beta}_j^{EB}(b_0, \tau^2)$. □

Proof of corollary 1: Immediate from Theorem 1. □
Proof of corollary 2: Under the given assumptions, evaluating the asymptotic first order condition for maximizing the likelihood yields

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

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

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

Note: The top graph of this figure shows the US time series of log relative wages and log relative labor supply between workers with more than a high school education, and those with high school or less. The bottom graph shows the same, after subtracting a linear trend in time with a kink-point in 1992. Calculations are based on the March CPS. For details, see Section 3.1.3. This figure replicates similar figures in Autor et al. (2008) and Acemoglu and Autor (2011).
Figure 2: Log relative wages in the US – actual evolution and counterfactual changes

Note: These figures show log wages of different types of workers relative to wages of high-school dropouts with less than 20 years of experience. The top left figure shows the actual historical evolution of relative wages, whereas the remaining figures show predicted counterfactual wages holding demand constant, based on the historical evolution of relative labor supply and alternative estimators of demand. Details are discussed in Section 3.1.3.
Table 1: Estimated effects of labor supply on wage inequality, panel of US states

<table>
<thead>
<tr>
<th>Supply of type</th>
<th>Struct</th>
<th>Unrest</th>
<th>EB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2-1</td>
<td>3-1</td>
</tr>
<tr>
<td>1</td>
<td>-0.022</td>
<td>0.063</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td>(0.034)</td>
<td>(0.027)</td>
</tr>
<tr>
<td>2</td>
<td>-0.002</td>
<td>-0.085</td>
<td>-0.080</td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
<td>(0.049)</td>
<td>(0.045)</td>
</tr>
<tr>
<td>3</td>
<td>-0.182</td>
<td>-0.320</td>
<td>-0.398</td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.071)</td>
<td>(0.068)</td>
</tr>
<tr>
<td>4</td>
<td>-0.023</td>
<td>-0.044</td>
<td>-0.107</td>
</tr>
<tr>
<td></td>
<td>(0.047)</td>
<td>(0.048)</td>
<td>(0.043)</td>
</tr>
<tr>
<td>5</td>
<td>-0.105</td>
<td>-0.032</td>
<td>-0.074</td>
</tr>
<tr>
<td></td>
<td>(0.072)</td>
<td>(0.080)</td>
<td>(0.086)</td>
</tr>
<tr>
<td>6</td>
<td>-0.011</td>
<td>-0.001</td>
<td>-0.028</td>
</tr>
<tr>
<td></td>
<td>(0.086)</td>
<td>(0.085)</td>
<td>(0.082)</td>
</tr>
<tr>
<td>7</td>
<td>-0.206</td>
<td>0.093</td>
<td>-0.003</td>
</tr>
<tr>
<td></td>
<td>(0.126)</td>
<td>(0.113)</td>
<td>(0.117)</td>
</tr>
<tr>
<td>8</td>
<td>-0.119</td>
<td>0.077</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>(0.074)</td>
<td>(0.081)</td>
<td>(0.082)</td>
</tr>
</tbody>
</table>

| $\beta_0$  | -0.061 | 0.051 |     |     |     |     |     | 0.052 |     |     |     |     |     |     |
|            | (0.058) | (0.065) |     |     |     |     |     | (0.042) |     |     |     |     |     |     |
| $\tau^2$  |     |     |     |     |     |     |     | 0.050 |     |     |     |     |     |     |
| Time FE  | YES | YES |     |     |     |     |     | YES |     |     |     |     |     |     |
| State FE | YES | YES |     |     |     |     |     | YES |     |     |     |     |     |     |
| N        | 306 | 306 |     |     |     |     |     | 306 |     |     |     |     |     |     |

Notes: This table shows three alternative estimates of labor demand using (i) the structural model based on the 2-type CES production function, (ii) unrestricted OLS regression using 8-types of the model nesting 2-type CES, and (iii) empirical Bayes estimation of the same model. Regressions control for time and state fixed effects. Standard errors are clustered across types of workers. Standard errors for empirical Bayes are calculated as discussed in Appendix A. For details, see Section 3.1.3.
Table 2: One month ahead prediction MSE for asset returns, 2010-15

<table>
<thead>
<tr>
<th>OLS</th>
<th>Restricted OLS</th>
<th>EB</th>
<th>Restricted EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0255</td>
<td>0.0232</td>
<td>0.0219</td>
<td>0.0218</td>
</tr>
</tbody>
</table>

Note: This table shows the mean squared error of alternative predictors of excess asset returns $R_{it}$ of the form $\hat{\alpha}_i + \hat{\beta}_i R^M_t$, where $(\hat{\alpha}_i, \hat{\beta}_i)$ are estimated using data for the 5-year windows $[t - 60, t - 1]$, starting in January 2010.
Figure 3: Distribution of estimates of $\alpha$ and $\beta$ across assets for the period 2011-15

Notes: These figures show histograms of the distribution of alternative estimators for $\alpha_i$ and $\beta_i$ across assets $i$, as discussed in Section 3.2.
Figure 4: OLS and empirical Bayes estimates of $\alpha$ and $\beta$ for the period 2011-15

**Notes:** These figures show scatter plots of the joint distribution of alternative estimators for $\alpha_i$ and $\beta_i$ across assets $i$, as discussed in Section 3.2.
Figure 5: Distribution of estimates of $\alpha$ and $\beta$ across assets for the period 1931-65

Notes: These figures show histograms of the distribution of alternative estimators for $\alpha_i$ and $\beta_i$ across assets $i$, as discussed in Section 3.2.
Figure 6: OLS and empirical Bayes estimates of $\alpha$ and $\beta$ for the period 1931-65

Notes: These figures show scatter plots of the joint distribution of alternative estimators for $\alpha_i$ and $\beta_i$ across assets $i$, as discussed in Section 3.2.
Figure 7: The mapping from $\hat{\beta}$ to $\hat{\tau}^2$ and $\hat{\beta}^{EB}$

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

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

Notes: These figures illustrate the mapping from preliminary estimates to empirical Bayes estimates when $\dim(\beta) = 2$, $\var(\hat{\beta}) = \text{diag}(2, 1)$, and $M = 0$. The top figure shows how our measure of model fit $\hat{\tau}^2$ varies with $\hat{\beta}$, the bottom left figure shows the direction and magnitude of shrinkage from $\hat{\beta}$ to $\hat{\beta}^{EB}$, and the bottom right figure depicts just the magnitude of shrinkage. For details, see Section 4.4.
Figure 8: The geometry of empirical Bayes

Notes: These figures illustrate the mapping from preliminary estimates for the same setting as in Figure 7. Each figure depicts, for a given value of $\hat{\tau}^2$, which preliminary estimates $\hat{\beta}$ yield this value and to what set of empirical Bayes estimates $\hat{\beta}^{EB}$ these preliminary estimates are mapped. For details see Section 4.4.
References


64


