cemmap

centre for microdata methods and practice

Posterior average effects

Stéphane Bonhomme
Martin Weidner

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP43/19

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

An ESRC Research Centre

Posterior Average Effects*

Stéphane Bonhomme† Martin Weidner‡

September 6, 2019

Abstract

Economists are often interested in estimating averages with respect to distributions of unobservables. Examples are moments of individual fixed-effects, average effects in discrete choice models, or counterfactual simulations in structural models. For such quantities, we propose and study "posterior average effects", where the average is computed *conditional* on the sample, in the spirit of empirical Bayes and shrinkage methods. While the usefulness of shrinkage for prediction is well-understood, a justification of posterior conditioning to estimate population averages is currently lacking. We establish two robustness properties of posterior average effects under misspecification of the assumed distribution of unobservables: they are optimal in terms of *local* worst-case bias, and their *global* bias is at most twice the minimum worst-case bias within a large class of estimators. We establish related robustness results for posterior predictors. In addition, we suggest a simple measure of the information contained in the posterior conditioning. Lastly, we present two empirical illustrations, to estimate the distributions of neighborhood effects in the US, and of permanent and transitory components in a model of income dynamics.

JEL CODES: C13, C23.
KEYWORDS: model misspecification, robustness, sensitivity analysis, empirical Bayes, posterior conditioning, latent variables.

*We thank Manuel Arellano, Tim Armstrong, Raj Chetty, Tim Christensen, Nathan Hendren, Peter Hull, Max Kasy, Derek Neal, Jesse Shapiro, Danny Yagan, and audiences at various places for comments. Bonhomme acknowledges support from the NSF, Grant SES-1658920. Weidner acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA.

†University of Chicago. Email: sbonhomme@uchicago.edu
‡University College London. Email: m.weidner@ucl.ac.uk

1 Introduction

In many settings, applied researchers wish to estimate population averages with respect to a distribution of unobservables. For example, in a discrete choice model, one may want to estimate an average partial effect corresponding to a change in a covariate. In panel data, moments or other features of the distribution of individual fixed-effects are often of interest. In applications of structural methods, researchers often compute average welfare effects in counterfactual policy scenarios, which are expectations taken with respect to a joint distribution of shocks and individual heterogeneity.

The standard approach in applied work is to make parametric assumptions on the distribution of unobservables and to compute the average effect under that parametric distribution. For example, under this approach, in a binary choice model, the researcher may assume normality of the error term, and compute the average partial effect under normality. We refer to this approach as the "model-based" estimation of average effects.

In this paper, we consider a different type of estimator, where the average effect is computed *conditional on the observation sample*. We refer to such estimators as "posterior average effects". Posterior averaging is appealing for prediction purposes. It also plays a central role in Bayesian and empirical Bayes approaches (e.g., Berger, 1980, Morris, 1983). Here we focus on the estimation of population expectations. Our goal is twofold: to propose a novel class of estimators, and to provide a frequentist framework to understand for which purpose and in which circumstances posterior conditioning may be useful in estimation.

Posterior average effects are closely related to empirical Bayes (EB) estimators, which are increasingly popular in applied economics. Consider a fixed-effects model of teacher quality, which we use as our main example. When the number of observations per teacher is small, the dispersion of teacher fixed-effects is likely to overstate that of true teacher quality, since the teacher effects are estimated with noise. An alternative approach is to postulate a prior distribution for teacher quality – typically, a normal – and report posterior estimates. The hope is that such EB estimates, which are shrunk toward the prior, are less affected by noise than the teacher fixed-effects (e.g., Kane and Staiger, 2008, Chetty *et al.*, 2014, Angrist *et al.*, 2017). However, while EB estimates are well-justified predictors of the quality of individual teachers, it is not obvious how to aggregate EB estimates across teachers when the goal is to estimate a population average such as a moment or a distribution function.

As an example, suppose we wish to estimate the distribution of teacher quality evaluated at a point. Since this quantity is an average of indicator functions, the posterior average effect is simply an average of posterior means – that is, of empirical Bayes estimates – *of the indicator functions*. This estimator is available in closed form. However, note that it differs from the empirical distribution of the EB estimates of teacher effects. Indeed, while the variance of EB estimates is too small relative to that of latent teacher quality, the posterior average effect has the correct variance.

In addition to the distribution of teacher quality, posterior average effects can be used to estimate coefficients in a regression of teacher quality on covariates, or more generally any quantity that involves an expectation with respect to the latent teacher effects. Other applications include the estimation of neighborhood or place effects (Chetty and Hendren, 2017, Finkelstein *et al.*, 2017) or hospital quality (Hull, 2018), for example.

Importantly, although posterior averages have desirable properties for predicting individual parameters, their usefulness for estimating population average quantities is not evident. For example, suppose that teacher quality is normally distributed. In this case, a model-based normal estimator of the distribution of teacher quality is consistent. Moreover, it is asymptotically efficient when means and variances are estimated by maximum likelihood. Hence, in the correctly specified case, there is no reason to deviate from the standard model-based approach and compute posterior estimators. The main insight of this paper is that, under misspecification (e.g., when teacher quality is not normally distributed), conditioning on the data can be beneficial.

To describe the behavior of estimators under misspecification, we focus on worst-case asymptotic bias in a nonparametric neighborhood of the reference parametric distribution (e.g., a normal). We consider neighborhoods based on ϕ-divergence, which is a family of distance measures often used to study misspecification. In the following, we often simply use *bias* to denote worst-case asymptotic bias in such a neighborhood. We establish two main properties.

Firstly, we show that the *local* bias of posterior average effects – calculated in an asymptotic where the size of the neighborhood tends to zero – is the smallest possible within a large class of estimators. Hence, posterior average effects are least sensitive to small departures from correct specification.

Secondly, we establish a *global* bound on the bias of posterior average effects, which holds irrespective of the neighborhood size. Specifically, we show that the worst-case bias of posterior average effects is at most twice as large as the smallest possible worst-case bias that can be achieved.

In addition, our analysis suggests a simple measure of "informativeness" of the posterior conditioning. As our examples highlight, the information contained in the posterior conditioning is setting-specific. Intuitively, posterior averages behave better when the realizations of outcome variables (such as test scores) are more informative about the values of the unobservables (such as the quality of a teacher). We show how the degree of posterior informativeness can be measured by an easily computable R^2 coefficient.

After those results for the *estimation* of population averages, we then also revisit the question of optimality of empirical Bayes estimates for *predicting* individual parameters. It is increasingly common in applications to report posterior predictors of neighborhood effects or teacher quality, for example. In a fixed-effects model, EB estimates minimize mean squared error when the normal reference model is correctly specified. However, when normality fails, the best predictor is a different posterior mean, which does not generally coincide with the EB estimate based on a normal prior. Intuitively, allowing for nonlinear functions of the data in the conditioning may improve prediction accuracy.

By extending our misspecification analysis from worst-case bias of sample averages to worst-case mean squared prediction error, we then derive a local optimality result for posterior predictors and show that EB estimators remain optimal, up to small-order terms, under local deviations from normality. In addition, we derive a bound for their worst-case asymptotic mean squared error that holds globally. Hence, our results for estimation and prediction highlight that, through the use of Bayes rule, conditioning on the data has robustness benefits when the prior is misspecified.

To illustrate the scope of posterior average effects for applications, we consider two empirical settings. In the first one, we study the estimation of neighborhood effects in the US. Chetty and Hendren (2017) report estimates of the variance of neighborhood effects, as well as empirical Bayes estimates of those effects. Our goal in the first illustration is to estimate the distribution of effects across neighborhoods.

We find that, when using a normal prior as in Chetty and Hendren (2017), our posterior estimator of the density of neighborhood effects across commuting zones is not normal. However, we also show through simulations and computation of our posterior informativeness measure that the signal-to-noise ratio in the data is not high enough to be confident about the exact shape of the distribution. Hence, in this setting, posterior average effects inform our knowledge of the density of neighborhood effects, and motivate future analyses using more flexible model specifications and individual-level data.

In our second empirical illustration, our goal is to estimate the distributions of latent components in a permanent-transitory model of income dynamics (e.g., Hall and Mishkin, 1982, Blundell *et al.*, 2008). In this model, log-income is the sum of a permanent random-walk component and a transitory component that is independent over time. In the literature, researchers often estimate the covariance structure of the latent components in a first step. Then, in order to document distributional features of the permanent and transitory components, or to take the income process to a life-cycle model of consumption and saving, they often add parametric – Gaussian – assumptions. However, there is increasing evidence that income components are not Gaussian (e.g., Geweke and Keane, 2000, Bonhomme and Robin, 2010, Guvenen *et al.*, 2016).

In this empirical context, we show how posterior average effects can reveal the non-Gaussianity of permanent and transitory income components. We estimate posterior distribution functions and quantiles of the two income components, using recent waves from the Panel Study of Income Dynamics (PSID). The posterior estimates reveal that both income components are non-normal, especially the transitory one. The deviations from normality that we find are qualitatively similar to – though not as large as – those recently found by Arellano *et al.* (2017) using a flexible income process on the same data.

More generally, our results provide theoretical foundations for the use of posterior average effects in empirical work, beyond the settings that we analyze empirically. We analytically study several examples, including reduced-form and structural discrete choice models and censored regression models, and we provide illustrative simulations.

Related literature and outline. Posterior average effects are closely related to parametric empirical Bayes estimators, which were studied in a series of papers by Carl Morris and Bradley Efron; see for example Efron and Morris (1973) and Morris (1983). The literature on shrinkage methods, which dates back to James and Stein (1961), has found recent applications in econometrics (Hansen, 2016, Fessler and Kasy, 2018, Abadie and Kasy, 2018); see also Efron (2012) for a book-length treatment.

There is also a large literature on nonparametric empirical Bayes methods, which were pioneered by Herbert Robbins (e.g., Robbins, 1955, 1964), with recent contributions including Koenker and Mizera (2014) and Ignatiadis and Wager (2019). Unlike this line of work, and in contrast with deconvolution and other nonparametric approaches, in our framework, we allow for forms of misspecification under which the quantity of interest is not consistently estimable, and we search for estimators that have the smallest amount of bias.

Our measures of bias and posterior informativeness are related to analyses of sensitivity to the prior in Bayesian statistics, see Gustafson (2000) for a review. Mueller (2013) studies risk properties of Bayesian estimators under misspecification. Berger (1979) provides a gamma-minimax characterization of Bayes estimators in ϵ-contaminated neighborhoods.

Also related is the literature on random-effects and fixed-effects methods in panel data, in particular, Chamberlain (1984), Wooldridge (2010), and Arellano and Bonhomme (2009, 2012). Arellano and Bonhomme (2009) study the bias of random-effects estimators of averages of functions of covariates and individual effects as both dimensions n and T of the panel tend to infinity. They show that, when the distribution of individual effects is misspecified, whereas the other features of the model are correctly specified, posterior average effects are consistent as n and T tend to infinity. Moreover, they characterize the first-order contribution of the bias. By contrast, in our setup, only n tends to infinity, and misspecification may affect the entire joint distribution of unobservables.

Lastly, we borrow several concepts and techniques from the literature on robustness to model misspecification and sensitivity analysis, which is reviewed in Huber and Ronchetti (2009) and Hampel *et al.* (1986), for example. There is recent related work on robustness in econometrics, in particular, Andrews *et al.* (2017, 2018), Armstrong and Kolesár (2018), Bonhomme and Weidner (2018), and Christensen and Connault (2019). Here our aim is to propose and justify a class of estimators.

The plan of the paper is as follows. In Section 2 we motivate the analysis by considering a fixed-effects model of teacher quality. In Section 3 we describe the general form of model-based and posterior estimators. In Section 4 we establish the main theoretical results, and we discuss them in Section 5. In Section 6 we present results on prediction. In Section 7 we illustrate the use of posterior average effects in two empirical settings. In Section 8 we describe several additional examples. Finally, we conclude in Section 9.

2 Motivating example: a fixed-effects model

To motivate our results, we start by considering the following model

$$Y_{ij} = \alpha_i + \varepsilon_{ij}, \qquad i = 1, ..., n, \qquad j = 1, ..., J. \tag{1}$$

To fix ideas, we will think of Y_{ij} as an average test score of teacher i in classroom j, α_i as the quality of teacher i, and ε_{ij} as a classroom-specific shock. For simplicity, we abstract away from covariates (such as students' past test scores), but those will be present in the framework we will introduce in the next section. Although here we focus on teacher effects, this model is of interest in other settings, such as the study of neighborhood effects, school effectiveness, or hospital quality, for example.

Suppose we wish to estimate a feature of the distribution of teacher quality α. As an example, here we consider the distribution function of α at a particular point a,

$$F_\alpha(a) = \mathbb{E}\left[\mathbf{1}\{\alpha \leq a\}\right],$$

which is the percentage of teachers whose quality is below a. When estimated at all points a, the distribution function can be inverted or differentiated to compute the quantiles of teacher quality or its density.

A first estimator is the empirical distribution of the fixed-effects estimates $\widehat{\alpha}_i = \overline{Y}_i = \frac{1}{J}\sum_{j=1}^J Y_{ij}$, for all teachers $i = 1, ..., n$; that is,

$$\widehat{F}_\alpha^{\text{FE}}(a) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\overline{Y}_i \leq a\},$$

where FE stands for "fixed-effects". An obvious issue with this estimator is that $\overline{Y}_i = \alpha_i + \overline{\varepsilon}_i$ is a noisy estimate of α_i, where $\overline{\varepsilon}_i = \frac{1}{J}\sum_{j=1}^J \varepsilon_{ij}$. Under mild conditions, as J tends to infinity $\overline{Y}_i$ is consistent for α_i, and $\widehat{F}_\alpha^{\text{FE}}(a)$ is consistent for $F_\alpha(a)$. However, due to the presence of noise, for small J the distribution $\widehat{F}_\alpha^{\text{FE}}$ tends to be *too dispersed* relative to F_α.[1]

A different strategy is to model the joint distribution of $\alpha, \varepsilon_1, ..., \varepsilon_J$. A simple specification is a multivariate normal distribution with means μ_α and $\mu_\varepsilon = 0$, and variances σ_α^2 and σ_ε^2. This specification can easily be made more flexible by allowing for different $\sigma_{\varepsilon_j}^2$'s across j, for correlation between the different ε_j's, or for means and variances being functions of covariates, for example. Under the assumption that all components are uncorrelated, μ_α, σ_α^2

[1] The large-J leading order bias of $\widehat{F}_\alpha^{\text{FE}}(a)$ is worked out in in Jochmans and Weidner (2018), and for the kernel-smoothed version in Okui and Yanagi (2018).

and σ_ε^2 can be consistently estimated using quasi-maximum likelihood or minimum distance based on mean and covariance restrictions.[2]

Given estimates $\widehat{\mu}_\alpha$, $\widehat{\sigma}_\alpha^2$, $\widehat{\sigma}_\varepsilon^2$, we can compute empirical Bayes estimates (Morris, 1983) of the α_i as

$$\mathbb{E}\left[\alpha \,|\, Y = Y_i\right] = \widehat{\mu}_\alpha + \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha), \quad i = 1, ..., n, \tag{2}$$

where the expectation is taken with respect to the posterior distribution of α given $Y = Y_i$ for $\widehat{\mu}_\alpha$, $\widehat{\sigma}_\alpha^2$, $\widehat{\sigma}_\varepsilon^2$ fixed, and $\widehat{\rho} = \frac{\widehat{\sigma}_\alpha^2}{\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2/J}$ is a shrinkage factor. Here, Y_i are vectors containing all Y_{ij}, $j = 1, ..., J$. The empirical Bayes estimates in (2) are well-justified as predictors of the α_i, since – when treating $\widehat{\mu}_\alpha$, $\widehat{\sigma}_\alpha^2$, $\widehat{\sigma}_\varepsilon^2$ as fixed – $\widehat{\mu}_\alpha + \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha)$ is the minimum mean squared error predictor of α_i under normality.

Given their rationale for prediction purposes, it is appealing to try and aggregate the empirical Bayes estimates in some way in order to estimate our target quantity $F_\alpha(a)$. A possible estimator is

$$\widehat{F}_\alpha^{\text{PM}}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\widehat{\mu}_\alpha + \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha) \leq a\right\}, \tag{3}$$

where PM stands for "posterior means". Like $\widehat{F}_\alpha^{\text{FE}}(a)$, $\widehat{F}_\alpha^{\text{PM}}(a)$ is consistent as J tends to infinity under mild conditions, since the shrinkage factor $\widehat{\rho}$ tends to one. However, for small J the empirical Bayes estimates tend to be *less dispersed* than the true α_i, and $\widehat{F}_\alpha^{\text{PM}}(a)$ is biased. Indeed, while in large samples the variance of the fixed-effects estimates is $\rho^{-1}\sigma_\alpha^2 > \sigma_\alpha^2$, the variance of the empirical Bayes estimates is $\rho\sigma_\alpha^2 < \sigma_\alpha^2$, where $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2/J}$.

Instead of computing the distribution of empirical Bayes estimates as in (3), a related idea is to compute the posterior distribution estimator

$$\widehat{F}_\alpha^{\text{P}}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}\{\alpha \leq a\} \,|\, Y = Y_i\right],$$

where P stands for "posterior". Using the normality assumption, we obtain

$$\widehat{F}_\alpha^{\text{P}}(a) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{a - \widehat{\mu}_\alpha - \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha)}{\widehat{\sigma}_\alpha\sqrt{1-\widehat{\rho}}}\right), \tag{4}$$

where Φ denotes the distribution function of the standard normal.

$\widehat{F}_\alpha^{\text{P}}(a)$ is an example of a *posterior average effect*. One can check that it is consistent for any fixed J when the distribution of $\alpha, \varepsilon_1, ..., \varepsilon_J$ is normal. Under non-normality,

[2] A set of restrictions is $\mathbb{E}[\varepsilon_j] = 0$, $\mathbb{E}[\varepsilon_j^2] = \sigma_\varepsilon^2$, $\mathbb{E}[\alpha] = \mu_\alpha$, $\mathbb{E}[(\alpha - \mu_\alpha)^2] = \sigma_\alpha^2$, $\mathbb{E}[\varepsilon_j\alpha] = 0$, and $\mathbb{E}[\varepsilon_j\varepsilon_{j'}] = 0$ for all $j \neq j'$.

$\widehat{F}_\alpha^{\mathrm{P}}(a)$ is consistent as J tends to infinity with n, although it is generally biased for small J.[3] Moreover, the mean and variance of $\widehat{F}_\alpha^{\mathrm{P}}$ are $(1-\widehat{\rho})\widehat{\mu}_\alpha + \widehat{\rho}\frac{1}{n}\sum_{i=1}^n \overline{Y}_i$ and $(1-\widehat{\rho})\widehat{\sigma}_\alpha^2 + \widehat{\rho}^2\left[\frac{1}{n}\sum_{i=1}^n \overline{Y}_i^2 - (\frac{1}{n}\sum_{i=1}^n \overline{Y}_i)^2\right]$, respectively, which are consistent for μ_α and σ_α^2 for any J.

The last estimator we consider here is directly based on the normal specification for α,

$$\widehat{F}_\alpha^{\mathrm{M}}(a) = \Phi\left(\frac{a-\widehat{\mu}_\alpha}{\widehat{\sigma}_\alpha}\right), \tag{5}$$

where M stands for "model". This estimator enjoys attractive properties when the distribution of $\alpha, \varepsilon_1, ..., \varepsilon_J$ is indeed normal. In this case, $\widehat{F}_\alpha^{\mathrm{M}}(a)$ is consistent for any fixed J, and it is efficient when $\widehat{\mu}_\alpha$ and $\widehat{\sigma}_\alpha^2$ are maximum likelihood estimates. Moreover, the mean and variance of $\widehat{F}_\alpha^{\mathrm{M}}$ are $\widehat{\mu}_\alpha$ and $\widehat{\sigma}_\alpha^2$, which are consistent irrespective of normality. However, in contrast to the other estimators above, when $\alpha, \varepsilon_1, ..., \varepsilon_J$ is *not* normally distributed $\widehat{F}_\alpha^{\mathrm{M}}(a)$ is generally inconsistent for $F_\alpha(a)$ as J tends to infinity. The inconsistency arises from the fact that $\widehat{F}_\alpha^{\mathrm{M}}(a)$ only depends on the data through the mean $\widehat{\mu}_\alpha$ and the variance $\widehat{\sigma}_\alpha^2$. In particular, $\widehat{F}_\alpha^{\mathrm{M}}$ is always normal, even when the data show clear evidence of non-normality.

The question we ask in this paper is which one of these estimators one should use. The answer is not obvious since they are all biased for small J in general. We provide optimality results under misspecification of the normal distribution for $\alpha, \varepsilon_1, ..., \varepsilon_J$. We show that the posterior average effect $\widehat{F}_\alpha^{\mathrm{P}}(a)$ is bias-optimal under local misspecification, and that it is near bias-optimal under global misspecification. To our knowledge, unlike the other three estimators above, posterior estimators of distributions are novel to practitioners. They are also straightforward to implement. Our results provide a justification for reporting them in applications.

Note that one may wish to relax the normality in (5) by making the specification of α, and possibly ε_j, more flexible. Deconvolution and nonparametric maximum likelihood estimators are often used for this purpose (e.g., Delaigle *et al.*, 2008, Bonhomme and Robin, 2010, Koenker and Mizera, 2014). These estimators will be consistent even when α is not normal, under suitable conditions. For example, the assumptions in Kotlarski (1967) require that $\alpha, \varepsilon_1, ..., \varepsilon_J$ be mutually independent. However, these conditions may fail to hold, and we allow for such general forms of misspecification in our framework. Like $\widehat{F}_\alpha^{\mathrm{M}}$, deconvolution and nonparametric maximum likelihood estimators are thus "model-based", even though they are based on a more flexible model than the normal.

[3] Consistency of $\widehat{F}_\alpha^{\mathrm{P}}(a)$ as J tends to infinity comes from the fact that $\widehat{\mu}_\alpha + \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha)$ approaches α_i, and $\widehat{\rho}$ approaches one, so $\Phi\left(\frac{a-\widehat{\mu}_\alpha-\widehat{\rho}(\overline{Y}_i-\widehat{\mu}_\alpha)}{\widehat{\sigma}_\alpha\sqrt{1-\widehat{\rho}}}\right)$ becomes increasingly concentrated around $\mathbf{1}\{\alpha_i \leq a\}$.

In model (1), the researcher may be interested in estimating a different quantity. As an example, consider the coefficient in the population regression of teacher quality α on a vector of covariates W; that is,

$$\overline{\delta} = \left(\mathbb{E}[WW']\right)^{-1} \mathbb{E}[W\alpha]. \tag{6}$$

In applications, it is common to regress fixed-effects estimates on covariates to help interpret them (as in Dobbie and Fryer, 2013, among many others), and to compute

$$\widehat{\delta}^{\mathrm{FE}} = \left(\sum_{i=1}^{n} W_i W_i'\right)^{-1} \sum_{i=1}^{n} W_i \overline{Y}_i. \tag{7}$$

Alternatively, one may regress the empirical Bayes estimates of α_i, as given by (2), on covariates (as in Angrist *et al.*, 2017, and Hull, 2018, for example), and compute

$$\widehat{\delta}^{\mathrm{P}} = \left(\sum_{i=1}^{n} W_i W_i'\right)^{-1} \sum_{i=1}^{n} W_i \left(\widehat{\mu}_\alpha + \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha)\right), \tag{8}$$

which is a posterior average effect based on a normal reference specification for α. We will see that, in our framework, the rationale for reporting $\widehat{\delta}^{\mathrm{P}}$ or $\widehat{\delta}^{\mathrm{FE}}$ depends on the form of misspecification that the researcher is concerned about.

The framework we describe next applies to the estimation of different quantities in a variety of settings. For example, the permanent-transitory model of income dynamics (e.g., Hall and Mishkin, 1982) has a similar structure as model (1). In Section 7 we will report posterior estimators of distributions of permanent and transitory income components. In other models, such as static or dynamic discrete choice models, or in models with censored outcomes, our results motivate the use of posterior average effects as complements to other estimators that researchers commonly report. We provide examples in Section 8.

Finally, it is of interest to analyze the robustness under misspecification of the empirical Bayes estimates (2) as predictors of the individual α_i. Using the same framework, but now focusing on worst-case asymptotic mean squared error (MSE), we will show in Section 6 that, in model (1), the empirical Bayes estimate of α_i is locally MSE-optimal and globally near MSE-optimal under misspecification of the normal model.

3 Model-based and posterior estimators

We consider the following class of models,

$$Y_i = g_\beta(U_i, X_i), \tag{9}$$

where outcomes Y_i and covariates X_i are observed by the researcher, and U_i are unobserved. The function g_β is known up to the finite-dimensional parameter β.

Our aim is to estimate an average effect of the form

$$\overline{\delta} = \mathbb{E}_{f_0}\left[\delta_\beta(U, X)\right], \tag{10}$$

where δ_β is known given β. Here f_0 denotes the true distribution of $U \mid X$. The expectation is taken with respect to $f_0 \times f_X$, where f_X is the marginal distribution of X. For conciseness, we will leave the dependence on f_X implicit throughout the paper. We focus on a scalar δ_β for simplicity, but our optimality results continue to hold in the vector-valued case, as we show in Appendix B. Moreover, although our focus is on average effects that depend linearly on f_0, in Appendix B we also discuss how to estimate quantities that depend on f_0 nonlinearly.

While the researcher does not know the true f_0, she has a reference parametric distribution f_σ for $U \mid X$, which depends on a finite-dimensional parameter σ. We will allow f_σ to be misspecified, in the sense that f_0 may not belong to $\{f_\sigma\}$. However, we will always assume that g_β is correctly specified. In other words, misspecification will only affect the distribution of U and its dependence on X, not the structural link between (U, X) and outcomes.

To estimate $\overline{\delta}$ in (10), we assume that the researcher has an estimator $\widehat{\beta}$ that remains consistent for β under misspecification of f_σ. More precisely, we will only consider potential true distributions f_0 such that $\widehat{\beta}$ tends in probability to the true value β under f_0. In many economic models, the assumptions needed to consistently estimate β are not sufficient to consistently estimate $\overline{\delta}$. This is the case in the fixed-effects model (1), where consistent estimates of means and variances can be obtained in the absence of normality. This is also the case in the models we outline in Section 8. In addition, we assume that the researcher has an estimator $\widehat{\sigma}$ that tends in probability to some σ_* under f_0. Unlike β, the parameter σ_* is a model-specific "pseudo-true value" that is not assumed to have generated the observed data.

Given $\widehat{\beta}$, $\widehat{\sigma}$, a sample $\{Y_i, X_i,\ i = 1, ..., n\}$ from (Y, X), and the parametric distribution f_σ, a *model-based* estimator of $\overline{\delta}$ is

$$\widehat{\delta}^{\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_{\widehat{\sigma}}}\left[\delta_{\widehat{\beta}}(U, X) \,\middle|\, X = X_i\right]. \tag{11}$$

When not available in closed form, this estimator can be computed by numerical integration or simulation under the parametric distribution $f_{\widehat{\sigma}}$. It is easy to see that, under standard conditions, $\widehat{\delta}^{\mathrm{M}}$ is consistent for $\overline{\delta}$ under correct specification; that is, when f_{σ_*} is the true distribution of $U \mid X$.

To construct a posterior estimator, consider the posterior distribution $p_{\beta,\sigma}$ of $U \mid Y, X$. This posterior distribution is computed using Bayes rule, based on the prior f_σ on $U \mid X$ and the likelihood of $Y \mid U, X$ implied by g_β. Formally, let $\mathcal{U}(y, x, \beta) = \{u \, : \, y = g_\beta(u, x)\}$. We define, whenever the denominator is non-zero,

$$p_{\beta,\sigma}(u \mid y, x) = \frac{f_\sigma(u \mid x)\mathbf{1}\{u \in \mathcal{U}(y, x, \beta)\}}{\int f_\sigma(v \mid x)\mathbf{1}\{v \in \mathcal{U}(y, x, \beta)\}dv}. \tag{12}$$

We will compute $p_{\beta,\sigma}$ analytically in all our examples. In Appendix B we describe a simulation-based approach for computing posterior average effects when an analytical expression is not available.

We then define a *posterior estimator* (or *posterior average effect*) of $\overline{\delta}$ as

$$\widehat{\delta}^{\mathrm{P}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{p_{\widehat{\beta},\widehat{\sigma}}}\left[\delta_{\widehat{\beta}}(U, X) \,\middle|\, Y = Y_i, X = X_i\right]. \tag{13}$$

Under standard regularity conditions, it is easy to see that, like $\widehat{\delta}^{\mathrm{M}}$, the posterior average effect $\widehat{\delta}^{\mathrm{P}}$ is consistent for $\overline{\delta}$ under correct specification.

From a Bayesian perspective, $\widehat{\delta}^{\mathrm{P}}$ is a natural estimator to consider. To see this, suppose that β and σ are known. $\widehat{\delta}^{\mathrm{P}}$ is then the posterior mean of $\frac{1}{n}\sum_{i=1}^{n} \delta_\beta(U_i, X_i)$, where the prior on U_i is f_σ, independent across i. $\widehat{\delta}^{\mathrm{P}}$ is also the average of the posterior means of $\delta_\beta(U_i, X_i)$ across individuals. An alternative Bayesian interpretation is obtained by specifying a non-parametric prior on f_0, and computing the posterior mean of $\overline{\delta}$ under this prior. We discuss this interpretation formally in Appendix C, in the case where U has finite support.

However, a frequentist justification for $\widehat{\delta}^{\mathrm{P}}$, and in particular a rationale for preferring $\widehat{\delta}^{\mathrm{P}}$ over $\widehat{\delta}^{\mathrm{M}}$, appear to be lacking in the literature. Indeed, under correct specification of f_σ, both estimators $\widehat{\delta}^{\mathrm{P}}$ and $\widehat{\delta}^{\mathrm{M}}$ are consistent, and, as we pointed out in the previous section, $\widehat{\delta}^{\mathrm{P}}$ may have a higher variance than $\widehat{\delta}^{\mathrm{M}}$.

The key difference between model-based and posterior estimators is that $\widehat{\delta}^{\mathrm{P}}$ is conditional on the observation sample. An intuitive reason for the conditioning is the recognition that realizations Y_i may be informative about the values of the unknown U_i's. As we will see in the next section, this intuition can be formalized in a framework that accounts for misspecification bias.

We end this section by showing how the fixed-effects model (1) can be mapped to the general notation. In Section 8 we present additional examples. In model (1) there are no covariates X, and the vector of unobservables U is

$$U = \left(\frac{\alpha - \mu_\alpha}{\sigma_\alpha}, \frac{\varepsilon_1}{\sigma_\varepsilon}, ..., \frac{\varepsilon_J}{\sigma_\varepsilon} \right)'.$$

The vector β is $\beta = (\mu_\alpha, \sigma_\alpha^2, \sigma_\varepsilon^2)'$. The reference distribution for U is a standard multivariate normal, so there is no other unknown parameter. We assume that the researcher has computed an estimator $\widehat{\beta}$, for example by quasi-maximum likelihood or minimum distance, which remains consistent for β when U is not normally distributed. When focusing on the distribution function of α at a point a, the target parameter is given by (10) with $\delta_\beta(U, X) = \mathbf{1}\{\alpha \leq a\}$, which in this case does not depend on β, X. Lastly, the model-based and posterior estimators $\widehat{\delta}^{\mathrm{M}}$ and $\widehat{\delta}^{\mathrm{P}}$ are given by (5) and (4), respectively.

4 Main results

In this section we describe our main theoretical results on posterior average effects, which are given in Theorems 1 and 2 below.

4.1 Estimators and worst-case bias

Let $P(\beta, f_0)$ denote the true distribution of (Y, U, X), where as before we omit the reference to the marginal distribution of X for conciseness. We assume that, under $P(\beta, f_0)$, $\widehat{\beta}$ is consistent for the true β, and $\widehat{\sigma}$ is consistent for a model-specific "pseudo-true" value σ_*, where $\mathbb{E}_{P(\beta,f_0)}[\psi_{\beta,\sigma_*}(Y, X)] = 0$ for some moment function ψ. For example, $\widehat{\beta}$ and $\widehat{\sigma}$ may be the method-of-moments estimators that solve $\sum_{i=1}^n \psi_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i) = 0$.

Given a distance measure d and a scalar $\epsilon \geq 0$, we define the following *neighborhood* of the reference distribution f_{σ_*}:

$$\Gamma_\epsilon = \left\{ f_0 \, : \, d(f_0, f_{\sigma_*}) \leq \epsilon, \;\; \mathbb{E}_{P(\beta,f_0)}[\psi_{\beta,\sigma_*}(Y, X)] = 0 \right\}.$$

This neighborhood consists of distributions of $U \,|\, X$ that are at most ϵ away from f_{σ_*}, and under which $\widehat{\beta}$ and $\widehat{\sigma}$ converge asymptotically to β and σ_*, respectively. The case $\epsilon = 0$ corresponds to correct specification of the reference distribution f_σ, whereas $\epsilon > 0$ corresponds to misspecification. Note that Γ_ϵ depends on β and σ_*, which are fixed in this section. Also, Γ_ϵ depends on the estimators $\widehat{\beta}$ and $\widehat{\sigma}$ through the moment function ψ. We take these estimators as given, and do not address the question of optimal estimation of β under misspecification.

Let us denote the supports of X and U as $\mathcal{X}$ and $\mathcal{U}$, respectively. We assume that d is a *ϕ-divergence* of the form

$$d(f_0, f_{\sigma_*}) = \int_{\mathcal{X}} \int_{\mathcal{U}} \phi\left(\frac{f_0(u\,|\,x)}{f_{\sigma_*}(u\,|\,x)}\right) f_{\sigma_*}(u\,|\,x)\, f_X(x)\, du\, dx,$$

where ϕ is a convex function that satisfies $\phi(1) = 0$ and $\phi''(1) > 0$. This family contains as special cases the Kullback-Leibler divergence (averaged over X), the Hellinger distance, the χ^2 divergence, and more generally the members of the Cressie-Read family of divergences (Cressie and Read, 1984). It is commonly used to measure misspecification, see Andrews *et al.* (2018) and Christensen and Connault (2019) for recent examples.

We focus on asymptotically linear estimators of $\overline{\delta}$ that satisfy, for a scalar function γ and as n tends to infinity,

$$\widehat{\delta}_\gamma = \frac{1}{n}\sum_{i=1}^{n} \gamma_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i) + o_{P(\beta,f_0)}(1). \tag{14}$$

Note that $\widehat{\delta}_\gamma$ depends on $\widehat{\beta}, \widehat{\sigma}$, but for conciseness we leave the dependence implicit in the notation. Many estimators can be written in this form (see, e.g., Bickel *et al.*, 1993).

Given an estimator $\widehat{\delta}_\gamma$, we define its ϵ-worst-case *bias* as

$$b_\epsilon(\gamma) = \sup_{f_0 \in \Gamma_\epsilon} \left| \mathbb{E}_{P(\beta,f_0)}[\gamma_{\beta,\sigma_*}(Y,X)] - \mathbb{E}_{f_0}[\delta_\beta(U,X)] \right|. \tag{15}$$

The worst-case bias $b_\epsilon(\gamma)$ is our measure of how well an estimator $\widehat{\delta}_\gamma$ performs under misspecification. The results below are specific to this particular objective. The bias here is asymptotic, since $\mathbb{E}_{P(\beta,f_0)}[\gamma_{\beta,\sigma_*}(Y,X)]$ is the probability limit of $\widehat{\delta}_\gamma$ as n tends to infinity under $P(\beta, f_0)$.

In our framework, we do not account for the variance of $\widehat{\delta}_\gamma$. In the fixed-effects model (1) of teacher quality, the variance is inversely proportional to the number of teachers. Using bias as a measure of performance of an estimator will thus be better suited when the cross-section of teachers is large. Alternatively, one could minimize the worst-case mean squared error of $\widehat{\delta}_\gamma$ as in Bonhomme and Weidner (2018),[4] or a weighted bias with respect to some prior on Γ_ϵ. In such cases the optimal estimators would take different forms. In particular, unlike posterior average effects they would generally depend on ϵ.

[4] In Section 6 we will use the MSE as a criterion for predicting individual-specific parameters. Our point here is that the MSE could also be employed as a criterion for the performance of $\widehat{\delta}_\gamma$, although in this paper we focus on bias.

4.2 Local optimality

We are now in a position to state our first main result. For this, we first characterize the bias $b_\epsilon(\gamma)$ of estimators $\widehat{\delta}_\gamma$ for small ϵ. The following lemma is instrumental in proving the local optimality of posterior average effects. For conciseness, from now on we suppress the reference to β, σ_* from the notation, and we denote as $\mathbb{E}_*$ and Var_* expectations and variance that are taken under the reference model $P(\beta, f_{\sigma_*})$. All proofs are in Appendix A.

Lemma 1. *Let* $\widetilde{\psi}(y,x) = \psi(y,x) - \mathbb{E}_*\left[\psi(Y,X)\middle|X=x\right]$. *Suppose that one of the following conditions holds:*

(i) $\phi(1)=0$, $\phi(r)$ *is four times continuously differentiable with* $\phi''(r)>0$ *for all* $r>0$, $\mathbb{E}_*[\psi(Y,X)]=0$, $\mathbb{E}_*\left[\widetilde{\psi}(Y,X)\,\widetilde{\psi}(Y,X)'\right]>0$, *and* $|\gamma(y,x)|$, $|\delta(u,x)|$, $|\psi(y,x)|$ *are bounded over the domain of* Y, U, X.

(ii) Condition (ii) of Lemma A1 in Appendix A holds (this alternative condition allows for unbounded γ, δ, ψ, *but at the cost of stronger assumptions on* $\phi(r)$*).*

Then, as ϵ *tends to zero we have*

$$
\begin{aligned}
b_\epsilon(\gamma) &= |\mathbb{E}_*[\gamma(Y,X)-\delta(U,X)]| \\
&\quad + \epsilon^{\frac{1}{2}}\left\{\frac{2}{\phi''(1)}\mathrm{Var}_*\Big(\gamma(Y,X)-\delta(U,X)-\mathbb{E}_*\left[\gamma(Y,X)-\delta(U,X)\,|\,X\right]-\lambda'\widetilde{\psi}(Y,X)\Big)\right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon),
\end{aligned}
$$

where $\lambda = \left\{\mathbb{E}_*\left[\widetilde{\psi}(Y,X)\,\widetilde{\psi}(Y,X)'\right]\right\}^{-1}\mathbb{E}_*\left[(\gamma(Y,X)-\delta(U,X))\,\widetilde{\psi}(Y,X)\right]$.

To derive the formula for the worst-case bias in Lemma 1, we maximize the bias with respect to f_0 subject to three contraints: f_0 belongs to an ϵ-neighborhood of f_*, it is such that the moment condition is satisfied at (β, σ_*), and it is a density. For ease of exposition, we only explicitly present the conditions in Lemma 1 for the case where γ, δ and ψ are bounded. This is satisfied, for example, if those functions and $g(u,x)$ are all continuous, and the domain of U and X is bounded. In Appendix A we detail the case of unbounded functions γ, δ and ψ, which only requires existence of third moments under the reference distribution. To guarantee that $b_\epsilon(\gamma)$ is well-defined in the unbounded case, we require a regularization of the function $\phi(r)$ for large values of r.

Lemma 1 implies local bias-optimality of posterior average effects, which is our first main result, stated in the following theorem.

Theorem 1. *Suppose that the conditions of Lemma 1 hold, and let*

$$\gamma^{\mathrm{P}}(y,x) = \mathbb{E}_*[\delta(U,X) \,|\, Y=y, X=x]. \tag{16}$$

Then, as ϵ tends to zero we have

$$b_\epsilon(\gamma^{\mathrm{P}}) \leq b_\epsilon(\gamma) + \mathcal{O}(\epsilon).$$

To provide an intuition for Theorem 1, note that, by Lemma 1, γ^{P} sets the first term in $b_\epsilon(\gamma)$ to zero. Moreover, γ^{P} minimizes the second term as well, since $\lambda = 0$ when $\gamma = \gamma^{\mathrm{P}}$. It follows that the posterior average effect $\widehat{\delta}^{\mathrm{P}} = \frac{1}{n}\sum_{i=1}^n \gamma^{\mathrm{P}}_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i)$ minimizes the first-order contribution to the worst-case bias.

In the presence of covariates, γ^{P} is in general not the unique minimizer of the first order bias. Indeed, for any function $\omega(x)$ with $\mathbb{E}_{f_X}[\omega(X)] = 0$, we have $b_\epsilon(\gamma^{\mathrm{P}} + \omega) = b_\epsilon(\gamma^{\mathrm{P}})$. Hence $\gamma^{\mathrm{P}} + \omega$ has also minimum bias locally. Nonetheless, since they have the same large-n limit, we expect $\widehat{\delta}_{\gamma^{\mathrm{P}}}$ and $\widehat{\delta}_{\gamma^{\mathrm{P}}+\omega}$ to behave similarly when the number of observations (e.g., teachers) is large.

Finally, note that, by Lemma 1, we have

$$b_\epsilon(\gamma^{\mathrm{P}}) = \epsilon^{\frac{1}{2}} \left\{ \frac{2}{\phi''(1)} \mathrm{Var}_* \left(\delta(U,X) - \mathbb{E}_*[\delta(U,X) \,|\, Y, X]\right) \right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon),$$

which is, up to smaller-order terms, proportional to the within-(Y,X) standard deviation of $\delta(U,X)$ under the reference model. Similar expressions appear in Bayesian statistics when computing derivatives of posterior quantities with respect to prior densities; see, e.g., Gustafson (2000).

4.3 Global bound

Our second main result is that, for any fixed $\epsilon \geq 0$, the worst-case bias of $\widehat{\delta}^{\mathrm{P}}$ is never larger than twice that of the best possible estimator.

Theorem 2. *Let γ^{P} be as in (16), and assume that $\phi(r)$ is convex with $\phi(1) = 0$. Then, for all $\epsilon > 0$,*

$$b_\epsilon(\gamma^{\mathrm{P}}) \leq 2 \inf_\gamma \, b_\epsilon(\gamma).$$

In Theorem 2 we establish a global bound on the bias of posterior average effects. The infimum in the theorem is taken over all possible functions $\gamma(y,x)$, subject to measurability conditions, which we implicitly assume throughout the paper. Besides this, we only rely on asymptotic linearity of the estimators. The proof of Theorem 2, which we provide in Appendix A, relies on very different arguments than in the small-ϵ case.

While Theorem 2 shows that the bias of posterior average effects is never larger than twice the minimum possible bias, $\widehat{\delta}^{\mathrm{P}}$ is not necessarily bias-optimal under global misspecification. Indeed, one can characterize the estimator that minimizes the worst-case bias for fixed ϵ, but this estimator is not a posterior average effect, and it depends on ϵ in general.[5] Moreover, the factor two in Theorem 2 cannot be improved upon in general, as we show in Appendix B in the context of a simple binary choice model.

5 Discussion of main results

Theorems 1 and 2 provide a rationale for using posterior average effects in applications. In the fixed-effects model (1), these results motivate using the posterior distribution estimator $\widehat{F}_{\alpha}^{\mathrm{P}}(a)$ in (4). In this section, we discuss several aspects and implications of our theoretical framework and results. In Appendix B, we also provide formulas for confidence intervals, as well as a test of correct specification of the parametric reference distribution based on comparing model-based and posterior estimators. We start by discussing the form of misspecification that we allow for in the theory.

5.1 Form of misspecification

Our theoretical characterizations are based on nonparametric neighborhoods of the reference parametric model that consist of unrestricted distributions of $U \mid X$, except for the moment conditions that pin down β and σ_*.[6] However, if one is willing to make additional assumptions on f_0 – thus further restricting the neighborhood – then one can construct estimators that are more robust than $\widehat{\delta}^{\mathrm{P}}$ within a particular class.

As an example, consider the fixed-effects model (1). Suppose that, in addition to assuming that α, ε_1, ..., ε_J are mutually uncorrelated, the researcher is willing to assume that they are fully independent. In that case, the distribution of α can be consistently estimated under suitable regularity conditions, provided $J \geq 2$ (Kotlarski, 1967, Li and Vuong,

[5] Alternatively, one may characterize the estimator that minimizes a second-order expansion of the bias for small ϵ. Such an estimator is a function of ϵ too.

[6] Note that taking f_0 in Γ_ϵ may impose restrictions on the data generating process through the structural function g_β.

1998). However, the posterior estimator in (4) is biased for small J. As a consequence, the posterior estimator is not bias-optimal in a semi-parametric neighborhood that consists of distributions with independent marginals.

To elaborate further on this point, consider the coefficient $\overline{\delta}$ in the population regression of α on a covariates vector W, see (6). A possible estimator is the coefficient $\widehat{\delta}^{\mathrm{FE}}$ in the regression of the fixed-effects estimates $\overline{Y}_i$ on W_i, see (7). Under correct specification of the reference model, $\widehat{\delta}^{\mathrm{FE}}$ is consistent for $\overline{\delta}$.[7]

However, $\widehat{\delta}^{\mathrm{FE}}$ may be inconsistent under the type of misspecification that we allow for, since ε_j and W may be correlated under f_0. In other words, in our framework, we allow for the possibility that W may have a direct effect on the outcomes Y_j, in which case $\widehat{\delta}^{\mathrm{FE}}$ is no longer consistent. Theorem 1 shows that, under such misspecification, the posterior estimator $\widehat{\delta}^{\mathrm{P}}$ in (8) has minimum worst-case bias locally. Nevertheless, if the researcher is confident that W should not enter the outcome equation, and that it is independent of ε_j, then it is natural to report the consistent estimator $\widehat{\delta}^{\mathrm{FE}}$.

5.2 Posterior informativeness

Our bias calculations can be used to compare the bias of the posterior estimator $\widehat{\delta}^{\mathrm{P}}$ to that of the model-based estimator $\widehat{\delta}^{\mathrm{M}}$. To see this, let $\gamma^{\mathrm{M}}_{\beta,\sigma}(x) = \mathbb{E}_{f_\sigma}[\delta_\beta(U,X)\,|\,X=x]$. Using Lemma 1, for small ϵ, the ratio of the two worst-case biases satisfies

$$\lim_{\epsilon\to 0}\frac{b_\epsilon(\gamma^{\mathrm{P}})}{b_\epsilon(\gamma^{\mathrm{M}})} = \frac{\left\{\mathrm{Var}_*\left(v(U,X)-\mathbb{E}_*[v(U,X)\,|\,Y,X]\right)\right\}^{\frac{1}{2}}}{\left\{\mathrm{Var}_*\left(v(U,X)\right)\right\}^{\frac{1}{2}}}, \tag{17}$$

where $v(U,X)$ is the population residual of $(\delta(U,X)-\gamma^{\mathrm{M}}(X))$ on $\widetilde{\psi}(Y,X)$, under the parametric reference model.[8] Intuitively, the robustness of $\widehat{\delta}^{\mathrm{P}}$ relative to $\widehat{\delta}^{\mathrm{M}}$ depends on how informative the outcome values Y_i are for the latent individual parameters $\delta(U_i,X_i)$.

In practice, we will report an empirical counterpart to the small-ϵ limit of $1-\frac{b^2_\epsilon(\gamma^{\mathrm{P}})}{b^2_\epsilon(\gamma^{\mathrm{M}})}$. This quantity can be simply expressed as the R^2 in the population nonparametric regression of $v(U,X)$ on Y,X under the reference model; that is,

$$R^2 = \frac{\mathrm{Var}_*\left(\mathbb{E}_*[v(U,X)\,|\,Y,X]\right)}{\mathrm{Var}_*\left(v(U,X)\right)}, \tag{18}$$

[7] In fact, in the illustration in Section 2 we have abstracted from covariates, so if U is independent of W under f then $\widehat{\delta}^{\mathrm{FE}}$ tends to $\overline{\delta}=0$. In the more general case where the normal reference distribution of α depends on some covariates, $\overline{\delta}$ would not be zero in general.

[8] That is, $v(u,x)=\delta(u,x)-\gamma^{\mathrm{M}}(x)+\lambda'\widetilde{\psi}(g(u,x),x)$, where all functions are evaluated at β,σ_*, and λ is as defined in Lemma 1 for the case $\gamma=\gamma^{\mathrm{M}}$.

where – with a slight abuse of notation – here $v(U,X)$ denotes the sample residual of $(\delta_{\widehat{\beta}}(U,X) - \gamma^{\mathrm{M}}_{\widehat{\beta},\widehat{\sigma}}(X))$ on $\widetilde{\psi}_{\widehat{\beta},\widehat{\sigma}}(Y,X)$, and expectations and variances are taken with respect to $P(\widehat{\beta}, f_{\widehat{\sigma}})$.

In the spirit of Andrews *et al.* (2018), we will refer to R^2 in (18) as a measure of the "informativeness" of the posterior conditioning, and we will report it in our illustrations.

6 Robustness in prediction

In applications, the researcher may be interested in predicting the value of $\delta_\beta(U_i, X_i)$ for an individual i. For example, in model (1) she may wish to predict the quality α_i of teacher i. Although our main focus in this paper is on the estimation of population averages, it is interesting to see how different predictors perform under misspecification of the reference distribution. In this section, we use our framework – now applied to mean squared error instead of bias – to provide local and global results on the robustness of empirical Bayes posterior predictors.

Formally, under squared loss, we now wish to find a predictor $\gamma_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i)$, for some function γ, such that the worst-case mean squared error is minimum. That is, our goal is to minimize

$$e_\epsilon(\gamma) = \sup_{f_0 \in \Gamma_\epsilon} \mathbb{E}_{P(\beta,f_0)} \left[(\delta_\beta(U,X) - \gamma_{\beta,\sigma_*}(Y,X))^2 \right]$$

with respect to γ.

Similarly as for our measure of worst-case bias, here the mean squared error is asymptotic, since $\mathbb{E}_{P(\beta,f_0)} \left[(\delta_\beta(U,X) - \gamma_{\beta,\sigma_*}(Y,X))^2 \right]$ is the probability limit of $\frac{1}{n}\sum_{i=1}^n (\delta_\beta(U_i, X_i) - \gamma_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i))^2$ as n tends to infinity under $P(\beta, f_0)$. This measure of performance is thus well-suited for settings with a large cross-section (e.g., settings with many teachers).

We first state the following local result, which is a direct generalization of Lemma 1. In the lemma we use the same shorthand notation as in Section 4.

Lemma 2. *Suppose that one of the following conditions holds:*

(i) Condition (i) of Lemma 1 holds.

(ii) Condition (ii) of Lemma A5 in Appendix A holds.

Let $\widetilde{\psi}$ be defined as in Lemma 1. Then, as ϵ tends to zero we have

$$\begin{aligned} e_\epsilon(\gamma) &= \mathbb{E}_* \left[(\gamma(Y,X) - \delta(U,X))^2 \right] \\ &+ \epsilon^{\frac{1}{2}} \left\{ \frac{2}{\phi''(1)} \mathrm{Var}_* \left((\gamma(Y,X) - \delta(U,X))^2 - \mathbb{E}_* \left[(\gamma(Y,X) - \delta(U,X))^2 \,|X \right] - \lambda' \widetilde{\psi}(Y,X) \right) \right\}^{\frac{1}{2}} \\ &+ \mathcal{O}(\epsilon), \end{aligned}$$

where

$$\lambda = \left\{ \mathbb{E}_* \left[\widetilde{\psi}(Y,X) \widetilde{\psi}(Y,X)' \right] \right\}^{-1} \mathbb{E}_* \left[(\gamma(Y,X) - \delta(U,X))^2 \, \widetilde{\psi}(Y,X) \right].$$

Let γ^{P} as in (16), so $\gamma^{\mathrm{P}}_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i)$ is the empirical Bayes estimate of $\delta_\beta(U_i, X_i)$. Under correct specification of the reference distribution f_σ, the posterior mean $\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y_i, X_i)$ is the minimum mean squared error predictor of $\delta_\beta(U_i, X_i)$ under squared loss. Under misspecification of f_σ, Lemma 2 implies that the leading term of the worst-case mean squared error is minimized at $\gamma = \gamma^{\mathrm{P}}$. Moreover, the lemma also implies the stronger result that the first-order term in the expansion of the worst-case mean squared error (which is a multiple of $\epsilon^{\frac{1}{2}}$) is also minimized at γ^{P}, provided the following condition holds almost surely:

$$\mathbb{E}_* \left[(\delta(U,X) - \gamma^{\mathrm{P}}(Y,X))^3 \,|\, Y, X \right] = 0. \tag{19}$$

While (19) is restrictive in general, it is satisfied in the fixed-effects model (1), when the researcher wishes to predict the quality α_i of teacher i. Indeed, in that case (19) is equivalent to the posterior skewness of α_i being zero, when using the normal reference model as the prior. Since the normal distribution is symmetric, (19) is satisfied, and the empirical Bayes estimator $\gamma^{\mathrm{P}}_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i) = \widehat{\mu}_\alpha + \widehat{\rho}(\overline{Y}_i - \widehat{\mu}_\alpha)$ has minimum worst-case mean squared error, up to second-order terms in $\epsilon^{\frac{1}{2}}$.

In addition to local optimality, we also have the following global bound, in the spirit of Theorem 2.

Theorem 3. *Let γ^{P} as in (16). Then, for all $\epsilon > 0$,*

$$e_\epsilon(\gamma^{\mathrm{P}}) \leq 4 \inf_\gamma e_\epsilon(\gamma).$$

Here, as in Theorem 2, the infimum in the theorem is taken over all possible function $\gamma(y, x)$, subject only to measurability conditions. Theorem 3 shows that empirical Bayes estimators are optimal, up to a factor of at most four, in terms of worst-case mean squared error. In model (1), when $\varepsilon_1, ..., \varepsilon_J$ are normally distributed and $\alpha_1, ..., \alpha_N$ are parameters belonging to an L^2 ball, empirical Bayes James-Stein estimators are known to be optimal in terms of asymptotic minimax mean squared error since they achieve the Pinsker bound (see for example Wasserman, 2006, Chapter 7). Here, by contrast, we show local optimality and global near-optimality for a worst case computed in a set of unrestricted, possibly non-normal joint distributions of $\alpha, \varepsilon_1, ..., \varepsilon_J$.

7 Two empirical illustrations

In this section, we revisit two applications of models with latent variables. In our first illustration, we focus on a model of neighborhood effects following Chetty and Hendren (2017), using data for the US that these authors made public. In our second illustration, we study a permanent-transitory model of income dynamics (Hall and Mishkin, 1982, Blundell *et al.*, 2008), using recent PSID data. In both cases, we rely on a normal reference specification and assess how and by how much the posterior conditioning informs the estimates of the parameters of interest.

7.1 Neighborhood effects

In this subsection, we start with estimates of neighborhood (or "place") effects reported in Chetty and Hendren (2017, CH hereafter). Those were obtained using individuals who moved between different commuting zones at different ages. The outcome variable that we focus on is the causal estimate of the income rank at age 26 of a child whose parents are at the 25 percentile of the income distribution. This is CH's preferred measure of place effect.

CH report an estimate of the variance of neighborhood effects, corrected for noise. In addition, they report individual predictors. Here we are interested in documenting the entire distribution of place effects. To do so, we consider the model $\widehat{\mu}_c = \mu_c + \overline{\varepsilon}_c$, for each commuting zone c, where $\widehat{\mu}_c$ is a neighborhood-specific fixed-effects reported by CH, μ_c is the true effect of neighborhood c, and $\overline{\varepsilon}_c$ is additive estimation noise. CH also report estimates $\widehat{\sigma}_c^2$ of the variances of $\overline{\varepsilon}_c$ for every c. When weighted by population, the fixed-effects estimates $\widehat{\mu}_c$ have mean zero. For simplicity, we will treat neighborhoods as independent observations.[9]

We first estimate the variance of place effects μ_c, following CH. We trim the top 1%

[9]The statistics we use for calculations are available on the Equality of Opportunity website; see https://opportunityinsights.org/paper/neighborhoodsii/

percentile of $\widehat{\sigma}_c^2$, and weigh all results by population weights. This differs slightly from CH's approach, which is based on $1/\widehat{\sigma}_c^2$ precision weights and no trimming.[10] We have information about place effects in $C = 590$ commuting zones c in our sample, compared to 595 in the sample without trimming. We estimate a sizable variance of neighborhood fixed-effects: $\text{Var}(\widehat{\mu}_c) = .077$. In turn, the mean of $\widehat{\sigma}_c^2$ weighted by population is $\widehat{\sigma}_{\overline{\varepsilon}}^2 = .047$. Given those, we estimate the variance of place effects as $\widehat{\sigma}_\mu^2 = \text{Var}(\widehat{\mu}_c) - \widehat{\sigma}_{\overline{\varepsilon}}^2 = .030$. In this setting, the shrinkage factor $\widehat{\rho}_c = \widehat{\sigma}_\mu^2/(\widehat{\sigma}_\mu^2 + \widehat{\sigma}_c^2)$ exhibits substantial heterogeneity across commuting zones. Indeed, the mean of $\widehat{\rho}_c$ is .62, and its 10% and 90% percentiles are .21 and .93, respectively. It is quantitatively important to account for this heterogeneity: in our initial work on this data we found that imposing a constant shrinkage factor reduced the informativeness of the posterior conditioning quite substantially.

We use the normal density with zero mean and variance $\widehat{\sigma}_\mu^2$ as a prior for μ_c. Then, we estimate the density of neighborhood effects μ_c, using the derivative of the posterior estimator of the distribution function (4); that is,

$$\widehat{f}_\mu^{\text{P}}(a) = \frac{1}{\sum_{c=1}^C \pi_c} \sum_{c=1}^C \pi_c \frac{1}{\widehat{\sigma}_\mu \sqrt{1-\widehat{\rho}_c}} \varphi\left(\frac{a - \widehat{\rho}_c \widehat{\mu}_c}{\widehat{\sigma}_\mu \sqrt{1-\widehat{\rho}_c}} \right),$$

where φ denotes the standard normal density, and π_c are population weights.

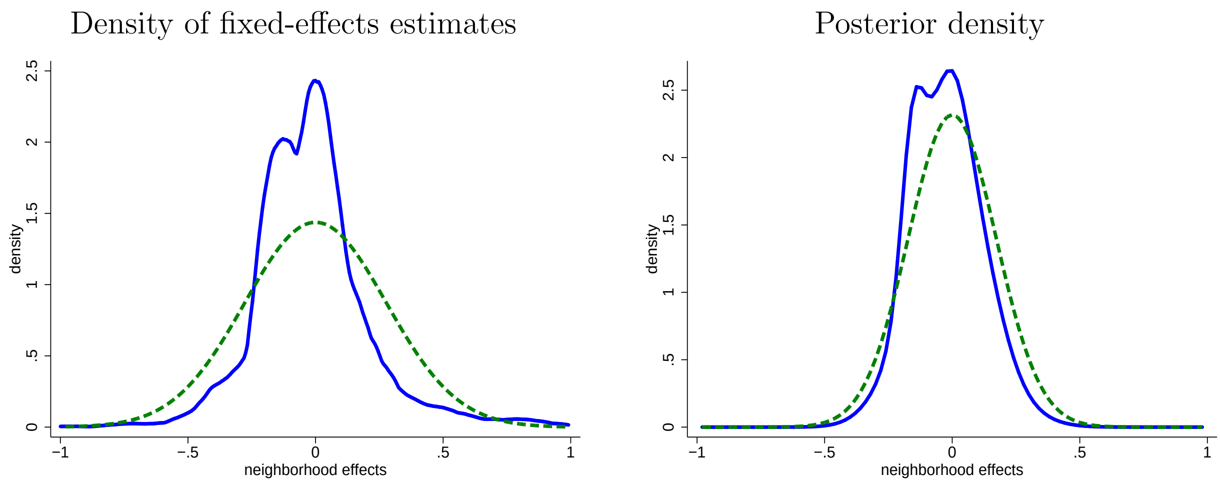
In Figure 1 we report several density estimates. In the left graph, we show a nonparametric kernel density estimate of the fixed-effects $\widehat{\mu}_c$, weighted by population (in solid), together with its best-fitting normal (in dashed). The graph shows substantial non-normality of the fixed-effects estimates. In particular, the large variance appears to be driven by some large positive and negative estimates $\widehat{\mu}_c$.

In the right graph of Figure 1 we report the posterior estimate $\widehat{f}_\mu^{\text{P}}$ of the density of true place effects μ_c (in solid). In addition, we show the normal prior – with zero mean and variance $\widehat{\sigma}_\mu^2$ – that we use to produce the posterior estimate (in dashed). We find that the posterior density of neighborhood effects differs from the normal prior, although the two estimators have the same variance by construction.[11] In addition, the specification test that compares model-based and posterior estimators, which we describe in Appendix B, suggests that these differences are statistically significant. Indeed, assuming independence across

[10] We replicated the analysis using precision weights in the un-trimmed sample and found similar results.

[11] In comparison, neighborhood-specific empirical Bayes estimates have a substantially lower dispersion. In Appendix E we report an estimate of their density $\widehat{f}_\mu^{\text{PM}}$. While $\widehat{\sigma}_\mu^2 = .030$, the variance of the empirical Bayes estimates is .010. By contrast, the variance associated with the posterior density estimator $\widehat{f}_\mu^{\text{P}}$ is .030.

Figure 1: Density of neighborhood effects



Notes: In the left graph we show the density of fixed-effects estimates $\widehat{\mu}_c$ *(solid) and its normal fit (dashed). In the right graph we show the posterior density of* μ_c *(solid) and the prior density (dashed). Calculations are based on statistics available on the Equality of Opportunity website.*

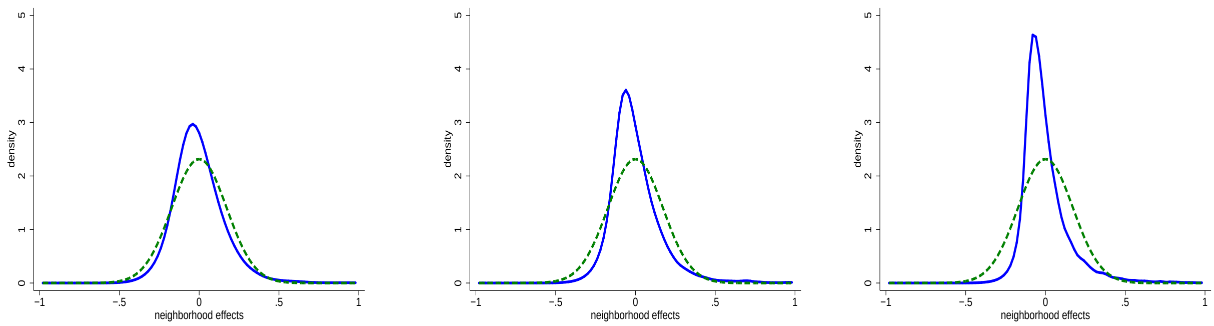
commuting zones, we obtain pvalues below .01 at all deciles except the bottom two. To assess how likely it is that the posterior estimator approximates the shape of the density of true neighborhood effects, we now perform two different exercises, based on a simulation and on numerical calculations motivated by our theory.

We start with a simulation, where μ_c, for $c = 1, ..., C_{\text{sim}}$, are *log-normally* distributed with zero mean and variance $\widehat{\sigma}_{\mu}^2$, and $\overline{\varepsilon}_c$ are normally distributed independent of μ_c with zero mean. We consider three scenarios for the noise variances $\widehat{\sigma}_c^2$: the estimates from CH, one-third of those values, and one-tenth of those values. In this exercise we again weight by population. We show the results for $C_{\text{sim}} = 100,000$ simulated neighborhoods. In the left graph of Figure 2 we see that, when the noise variances are the ones from the data, the posterior density is more skewed than the normal, yet the posterior shape is quite different from the true log-normal density of μ_c. When reducing the noise variances in the middle and right graphs, the posterior density estimate gets closer to the log-normal. In the right graph, where the shrinkage factor is .90 on average (as opposed to .62 in the data), the posterior average effect approximates the highly non-normal shape of the true distribution of neighborhood effects very well.

We next turn to our posterior informativeness measure, which is given by equation (18)

Figure 2: Density of neighborhood effects in simulated data with log-normal μ_c

100% noise variances from data | 33% noise variances | 10% noise variances



Notes: Simulation with μ_c *log-normal and* $\overline{\varepsilon}_c$ *normal. The posterior density is shown in solid, the prior density is shown in dashed. The left graph corresponds to the noise variances* $\widehat{\sigma}_c^2$ *of the data, the middle one corresponds to the noise variances divided by* 3*, and the right graph corresponds to the noise variances divided by* 10*.*

applied to $\delta_\beta(U, X) = \mathbf{1}\{\alpha \leq a\}$ for different values a. In this case, the R^2 coefficient varies along the distribution. We compute it by simulation, using 100,000 draws. We find that the weighted average R^2 across values of a is 28%, where we weigh across cutoff values a by the reference distribution for α.[12] This value is consistent with the message of the simulation exercise as it suggests that, while the posterior conditioning informs the shape of the distribution of neighborhood effects, the signal-to-noise ratio is not high enough to be confident about the exact shape of the density. To provide additional insights, it would be interesting to refine the reference model using a non-normal parametric or semi-parametric specification. However, to flexibly model the neighborhood effects and the noise, one would need to use the individual-level data.

Lastly, we perform two additional exercises as robustness checks and report the corresponding results in Appendix E. Firstly, we incorporate the mean income of permanent residents in country c at the 25% percentile – say, $\overline{y}_c$ – as a covariate. CH rely on information on permanent residents' income to improve the accuracy of individual predictions. Here we use it to refine the reference distribution and to improve the estimation of the distribution of neighborhood effects. Specifically, our reference model for μ_c is now a correlated random-effects specification, where the mean depends on $\overline{y}_c$ linearly. We find small differences with

[12]In addition, we compute the value of the R^2 when the noise variances are one-third or one-tenth of their values in the data. We find that the R^2 is 36% on average in the former case, and 47% in the latter case.

our baseline estimates. Secondly, we re-do our main analysis at the county level, instead of the commuting zone level. In that case the signal-to-noise ratio is lower, our posterior informativeness R^2 measure is 17% on average, and we find that the normal prior and the posterior density are closer to each other than in the case of commuting zones.

7.2 Income dynamics

In this subsection we consider a permanent-transitory model of household log-income given by

$$Y_{it} = \eta_{it} + \varepsilon_{it}, \qquad \eta_{it} = \eta_{i,t-1} + V_{it}, \qquad i = 1, ..., n, \quad t = 1, ..., T,$$

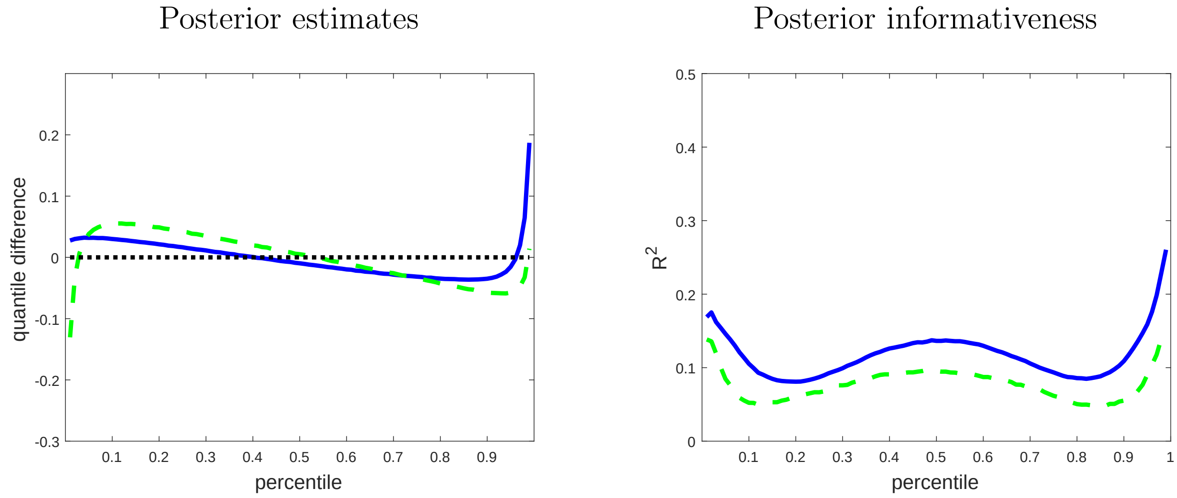
where ε_{it} and V_{it} are independent at all lags and leads, and independent of η_{i0}. This process is commonly used as an input for life-cycle consumption/savings models. Researchers often estimate covariances in a first step using minimum distance, and then impose a normality assumption for further analysis. However, there is increasing evidence that income components are *not* normally distributed. Instead of using a more flexible model – as has been done by Carlton and Hall (1978) and a large subsequent literature – here we compute posterior estimates. The advantages of this approach are that no additional assumptions are needed, and that implementation is straightforward.

We focus on six recent waves of the PSID 1999-2009 (every other year), see Blundell *et al.* (2016) for a description of the data. We use the same sample selection as in Arellano *et al.* (2017), and work with a balanced panel of $n = 792$ households over $T = 6$ periods. Y_{it} are residuals of log total pre-tax household labor earnings on a set of demographics, which include cohort interacted with education categories for both household members, race, state, and large-city dummies, a family size indicator, number of kids, a dummy for income recipient other than husband and wife, and a dummy for kids out of the household.

Our aim is to estimate the quantiles of η_{it} and ε_{it}. To do so, we compare normal model-based estimates with posterior estimates, by plotting differences of quantile functions averaged over time periods. We compute the quantiles by inverting the posterior estimates of the distribution functions. The model's structure is similar to that of the fixed-effects model (1), and analytical expressions for posterior estimators are easy to derive.

In the left graph of Figure 3, we show the quantile differences for η_{it} in solid, and the ones for ε_{it} in dashed. In both cases, quantiles in the lower (respectively, upper) part of the distribution are higher (resp., lower) under posterior estimates than under normal estimates.

Figure 3: Quantiles of income components



Notes: The left graph shows quantile differences between posterior and model-based estimators. The right graph shows the posterior informativeness R^2 measure, see equation (18). η_{it} is shown in solid, and ε_{it} is shown in dashed. Sample from the PSID.

This suggests that the distributions of both latent components show excess kurtosis (i.e., "peakedness") relative to the normal. Moreover, our posterior estimates suggest stronger violation of normality for ε_{it} than for η_{it}.

In the right graph of Figure 3 we report our posterior informativeness measure, as given by (18), at different quantiles. The estimates suggest that there is information in the posterior conditioning, especially for the permanent income component η_{it}. At the same time, the R^2 never exceeds 25%, which suggests that posterior estimates may still be biased when the reference distribution is misspecified.

Several papers have already documented the presence of excess kurtosis in income components using parametric or semi-parametric methods. In Appendix E we compare our posterior estimates with estimates based on a flexible non-normal and non-linear model from Arellano *et al.* (2017). We find that, although both sets of estimates show qualitatively similar evidence of excess kurtosis, the non-normality of the posterior estimates is less pronounced than the non-normality of the estimates from Arellano *et al.* (2017), especially in the case of the transitory component ε_{it}.

Overall, these illustrations give two examples where, starting from a normal prior, the posterior conditioning is informative about the true unknown distributions. In both settings, posterior average effects are not normal. Yet, as indicated by the R^2 values we report, the signal-to-noise ratios are not high enough to be certain about the exact shapes of the densities of interest, thus motivating further analyses using non-normal specifications. Posterior estimators should be useful in other environments where model (1) and its extensions are widely used, for example in teacher value-added applications where the signal-to-noise ratio is driven by the number of observations per teacher.

8 Posterior average effects in various settings

Posterior average effects are applicable in a wide variety of settings. In this section, we provide additional examples of models where they may be of interest. In Appendix D we show illustrative simulations for two models.

Linear regression. Consider the linear regression

$$Y_i = X_i'\beta + U_i.$$

Suppose that $\mathbb{E}[XU] = 0$, and that the OLS estimator $\widehat{\beta}$ is consistent for β. Suppose also that the researcher is interested in the average effect $\overline{\delta} = \mathbb{E}_{f_0}[U^2 XX']$.[13] In this context, a model-based approach consists in modeling $U \mid X$, say, as a normal with zero mean and variance σ^2, and computing

$$\widehat{\delta}^{\mathrm{M}} = \widehat{\sigma}^2 \frac{1}{n}\sum_{i=1}^n X_i X_i',$$

where $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2$ is the maximum likelihood estimator of σ^2 under normality.

By contrast, a posterior average effect is

$$\begin{aligned}\widehat{\delta}^{\mathrm{P}} &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p_{\widehat{\beta},\widehat{\sigma}}}\left[U^2 XX' \,\middle|\, Y = Y_i, X = X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2 X_i X_i'.\end{aligned}$$

This is the central piece in the White (1980) variance formula. $\widehat{\delta}^{\mathrm{P}}$ remains consistent for $\overline{\delta}$ absent normality or homoskedasticity of U. In this very special case, $\widehat{\delta}^{\mathrm{P}}$ is thus fully robust to misspecification of f_σ, since U_i is a deterministic function of Y_i, X_i and β.

[13]In this example $\overline{\delta}$ is multi-dimensional; see Appendix B.

Censored regression. Consider next the censored regression model

$$Y_i = \max(Y_i^*, 0), \text{ where } Y_i^* = X_i'\beta + U_i. \tag{20}$$

In this model, β can be consistently estimated under weak conditions. For example, Powell's (1986) symmetrically trimmed least-squares estimator is consistent for β when $U \mid X$ is symmetric around zero, under suitable regularity conditions. In this setting, suppose that we are interested in a moment of the potential outcomes Y_i^*, such as $\overline{\delta} = \mathbb{E}_{f_0}[h(Y^*)]$ for some function h. As an example, the researcher may wish to estimate a feature of the distribution of wages using a sample affected by top- or bottom-coding.

Following a model-based approach, let us assume that $U \mid X \sim \mathcal{N}(0, \sigma^2)$, and estimate σ^2 using maximum likelihood. A model-based estimator is then $\widehat{\delta}^{\text{M}} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{f_{\widehat{\sigma}}}[h(X_i'\widehat{\beta} + U)]$. By contrast, a posterior average effect is

$$\widehat{\delta}^{\text{P}} = \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbf{1}\{Y_i > 0\}h(Y_i)}_{\text{uncensored}} + \frac{1}{n}\sum_{i=1}^n \underbrace{\mathbf{1}\{Y_i = 0\}\mathbb{E}_{p_{\widehat{\beta},\widehat{\sigma}}}\left[h(X_i'\widehat{\beta} + U) \,\middle|\, X_i'\widehat{\beta} + U \leq 0\right]}_{\text{censored}}.$$

This estimator relies on actual Y's for uncensored observations, and on imputed Y's for censored ones.

The censored regression model illustrates an aspect related to the class of neighborhoods that our theoretical characterizations rely on. In model (20), the researcher might want to impose that $U \mid X$ be symmetric around zero, which is the main assumption for consistency of the Powell (1986) estimator. It is possible to construct estimators that minimize local worst-case bias in an ϵ-neighborhood that only consists of symmetric distributions f_0. However, posterior average effects are no longer bias-optimal in this class.

More generally, the assumptions that justify the use of a particular estimator $\widehat{\beta}$ may suggest further restrictions on the neighborhood. Our optimality results are based on a class where such restrictions are not imposed. Indeed, the only additional restriction on f_0, beyond belonging to an ϵ-neighborhood around f_{σ_*}, is that the population moment condition $\mathbb{E}_{P(\beta,f_0)}[\psi_{\beta,\sigma_*}(Y, X)] = 0$ is assumed to hold.

Binary choice. Consider now the binary choice model

$$Y_i = \mathbf{1}\{X_i'\beta + U_i > 0\}. \tag{21}$$

In this model, Manski (1975, 1985) shows that β is identified up to scale as soon as the median of $U \mid X$ is zero, under sufficiently large support of X. In addition, he provides conditions for consistency of the maximum score estimator $\widehat{\beta}$, again up to scale. Manski's conditions, however, are not sufficient to consistently estimate the average structural function (ASF, Blundell and Powell, 2004)

$$\overline{\delta}(x) = \mathbb{E}_{f_0}[\mathbf{1}\{x'\beta + U > 0\}].$$

Let us take as reference parametric distribution for $U \mid X$ a normal with zero mean and variance σ^2, and let $\widehat{\sigma}^2$ denote the maximum likelihood estimator of σ^2 given $\widehat{\beta}$, based on normality.[14] A model-based estimator of the ASF is $\widehat{\delta}^{\text{M}}(x) = \Phi\left(\frac{x'\widehat{\beta}}{\widehat{\sigma}}\right)$, and a posterior estimator is

$$\widehat{\delta}^{\text{P}}(x) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i \frac{\min\left(\Phi\left(\frac{x'\widehat{\beta}}{\widehat{\sigma}}\right), \Phi\left(\frac{X_i'\widehat{\beta}}{\widehat{\sigma}}\right)\right)}{\Phi\left(\frac{X_i'\widehat{\beta}}{\widehat{\sigma}}\right)} + (1 - Y_i)\frac{\max\left(\Phi\left(\frac{x'\widehat{\beta}}{\widehat{\sigma}}\right) - \Phi\left(\frac{X_i'\widehat{\beta}}{\widehat{\sigma}}\right), 0\right)}{1 - \Phi\left(\frac{X_i'\widehat{\beta}}{\widehat{\sigma}}\right)}\right].$$

Unlike $\widehat{\delta}^{\text{M}}(x)$, the posterior ASF estimator $\widehat{\delta}^{\text{P}}(x)$ depends directly on the observations of the binary Y_i's, in addition to the indirect data dependence through $\widehat{\beta}$ and $\widehat{\sigma}^2$. In Appendix D we present simulations from an ordered choice model, which suggest that the informativeness of the posterior conditioning – and the robustness properties of posterior estimators compared to model-based estimators – depend crucially on the support of the dependent variable. In particular, our simulations suggest that robustness gains might be modest in binary choice settings.

Panel data discrete choice. Our last example is the panel data model

$$Y_{it} = \mathbf{1}\{X_{it}'\beta + \alpha_i + \varepsilon_{it} > 0\}, \quad i = 1, ..., n, \quad t = 1, ..., T.$$

When ε_{it} are i.i.d. standard logistic, β can be consistently estimated using the conditional logit estimator (Andersen, 1970, Chamberlain, 1984). However, additional assumptions are needed to consistently estimate average partial effects such as the effect of a discrete shift of Δ along the k-th component of X,

$$\overline{\delta} = (\mathbb{E}_{f_0}[\mathbf{1}\{(X_t + \Delta \cdot e_k)'\beta + \alpha + \varepsilon_t > 0\}] - \mathbb{E}_{f_0}[\mathbf{1}\{X_t'\beta + \alpha + \varepsilon_t > 0\}])/\Delta,$$

where e_k is a vector of zeros with a one in the k-th position.

The standard approach is to postulate a parametric random-effects specification for the conditional distribution of α given $X_1, ..., X_T$, and to compute an average effect $\widehat{\delta}^{\text{M}}$ with

[14] Specifically, $\widehat{\sigma}$ maximizes the probit log-likelihood $\sum_{i=1}^{n} Y_i \log \Phi\left(\frac{X_i'\widehat{\beta}}{\sigma}\right) + (1 - Y_i) \log \Phi\left(-\frac{X_i'\widehat{\beta}}{\sigma}\right)$.

respect to that distribution. By contrast, a posterior estimator is computed conditional on the observations $Y_{i1}, ..., Y_{iT}$, for every individual i. As T tends to infinity, such estimators are robust to misspecification of α, provided ε_t is correctly specified (Arellano and Bonhomme, 2009). Our analysis shows that they also have robustness properties when ε_t is not logistic.

Aguirregabiria *et al.* (2018) show that conditional logit-like estimators can also be used to consistently estimate parameters in structural dynamic discrete choice settings. As an example, they study the Rust (1987) model of bus engine replacement in the presence of unobserved heterogeneity in maintenance and replacement costs. In such structural models, estimating average welfare effects of policies requires averaging with respect to the distribution of unobservables. Posterior estimators provide an alternative to the standard parametric model-based approach in this context.

9 Conclusion

Posterior averages are commonly used to predict individual parameters such as teacher quality or neighborhood effects, and they play a central role in Bayesian and empirical Bayes approaches. In this paper, we have provided a frequentist justification for posterior conditioning when the goal of the researcher is to estimate a population average quantity.

We have established two properties of posterior estimators of average effects under misspecification of parametric assumptions: a local bias-optimality result, and a bound on global bias. Posterior average effects are simple to implement. Our results provide a rationale for reporting them in many applications. We have used a linear fixed-effects model as a running example, and we have mentioned other possible applications in Section 8. In addition, we have used our framework to establish novel results on the robustness of empirical Bayes estimators of individual effects.

Lastly, our examples highlight that the information contained in the conditioning is setting-specific. Hence, posterior average effects are complements to – but not substitutes for – other approaches that rely on additional assumptions, such as semi-parametric approaches under point or partial identification (e.g., Powell, 1994, Tamer, 2010), or recent approaches that aim for robustness within a specific class of models (e.g., Bonhomme and Weidner, 2018, Armstrong and Kolesár, 2018, Christensen and Connault, 2019).

References

[1] Abadie, A., and M. Kasy (2018): "The Risk of Machine Learning," to appear in the *Review of Economics and Statistics.*

[2] Aguirregabiria, V., J. Gu, and Y. Luo (2018): "Sufficient Statistics for Unobserved Heterogeneity in Structural Dynamic Logit Models," arXiv preprint arXiv:1805.04048.

[3] Andersen, E. B. (1970): "Asymptotic Properties of Conditional Maximum-Likelihood Estimators," *Journal of the Royal Statistical Society. Series B (Methodological)*, 283–301.

[4] Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132(4), 1553–1592.

[5] Andrews, I., M. Gentzkow, and J. M. Shapiro (2018): "On the Informativeness of Descriptive Statistics for Structural Estimates," unpublished manuscript.

[6] Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017): "Leveraging Lotteries for School Value-Added: Testing and Estimation," *Quarterly Journal of Economics*, 132(2), 871–919.

[7] Arellano, M., Blundell, R., and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85(3), 693–734.

[8] Arellano, M., and S. Bonhomme, S. (2009): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77(2), 489–536.

[9] Arellano, M., and S. Bonhomme (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," *Review of Economic Studies*, 79(3), 987–1020.

[10] Armstrong, T. B., and M. Kolesár (2018): "Sensitivity Analysis Using Approximate Moment Condition Models," arXiv preprint arXiv:1808.07387.

[11] Azzalini, A. (2013): *The Skew-Normal and Related Families.* Vol. 3. Cambridge University Press.

[12] Berger, J. (1980): *Statistical Decision Theory: Foundations, Concepts, and Methods.* Springer.

[13] Berger, R. L. (1979): "Gamma Minimax Robustness of Bayes Rules: Gamma Minimax Robustness," *Communications in Statistics – Theory and Methods*, 8(6), 543–560.

[14] Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993): *Efficient and Adaptive Inference in Semiparametric Models.* Johns Hopkins University Press.

[15] Blundell, R., L. Pistaferri, and I. Preston (2008): "Consumption Inequality and Partial Insurance," *American Economic Review*, 98(5): 1887–1921.

[16] Blundell, R., L. Pistaferri, and I. Saporta-Eksten (2016): "Consumption Smoothing and Family Labor Supply," *American Economic Review*, 106(2), 387–435.

[17] Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71(3), 655–679.

[18] Bonhomme, S., and J. M. Robin (2010): "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics," *Review of Economic Studies*, 77(2), 491–533.

[19] Bonhomme, S., and Weidner, M. (2018): "Minimizing sensitivity to model misspecification," arXiv preprint arXiv:1807.02161.

[20] Carlton, D. W., and R. E. Hall (1978): "The Distribution of Permanent Income," in *Income Distribution and Economic Inequality.* New York: Halsted.

[21] Chamberlain, G. (1984): "Panel Data", in Griliches, Z. and M. D. Intriligator (eds.), *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

[22] Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014): "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 104(9), 2593-2632.

[23] Chetty, R., and N. Hendren (2018): "The Impacts of Neighborhoods on Intergenerational Mobility: County-Level Estimates," *Quarterly Journal of Economics*, 133(2), 1163-1228.

[24] Christensen, T., and B. Connault (2019): "Counterfactual Sensitivity and Robustness," unpublished manuscript.

[25] Cressie, N., and T. R. C. Read (1984): "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society Series B*, 46(3), 440–464.

[26] Delaigle, A., P. Hall, and A. Meister (2008): "On Deconvolution with Repeated Measurements," *Annals of Statistics*, 36(2), 665–685.

[27] Dobbie, W., and R. G. Fryer Jr (2013): "Getting Beneath the Veil of Effective Schools: Evidence from New York City," *American Economic Journal: Applied Economics*, 5(4), 28–60.

[28] Efron, B. (2012): *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Vol. 1. Cambridge University Press.

[29] Efron, B., and C. Morris (1973): "Stein's Estimation Rule and its Competitors – An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68(341), 117-130.

[30] Fessler, P., and M. Kasy (2018): "How to Use Economic Theory to Improve Estimators," to appear in the *Review of Economics and Statistics.*

[31] Finkelstein, A., M. Gentzkow, P. Hull, and H. Williams (2017): "Adjusting Risk Adjustment – Accounting for Variation in Diagnostic Intensity," *New England Journal of Medicine*, 376, 608–610.

[32] Florios, K., and S. Skouras (2008): "Exact Computation of Max Weighted Score Estimators," *Journal of Econometrics*, 146(1), 86–91.

[33] Geweke, J., and M. Keane (2000): "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96(2), 293–356.

[34] Gustafson, P. (2000): "Local Robustness in Bayesian Analysis," in *Robust Bayesian Analysis* (pp. 71-88). Springer, New York, NY.

[35] Guvenen, F., F. Karahan, S. Ozcan, and J. Song (2016): "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Risk?" unpublished manuscript.

[36] Guvenen, F., S. Ozcan, and J. Song (2014): "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 122(3), 621–660.

[37] Hall, R., and F. Mishkin (1982): "The sensitivity of Consumption to Transitory Income: Estimates from Panel Data of Households," *Econometrica*, 50(2): 261–81.

[38] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986): *Robust Statistics: The Approach Based on Influence Functions.* Wiley Series in Probability and Statistics.

[39] Hansen, B. E. (2016): "Efficient Shrinkage in Parametric Models," *Journal of Econometrics*, 190(1), 115–132.

[40] Hansen, L. P., and T. J. Sargent (2008): *Robustness.* Princeton University Press.

[41] Huber, P. J., and E. M. Ronchetti (2009): *Robust Statistics.* Second Edition. Wiley.

[42] Hull, P. (2018): "Estimating Hospital Quality with Quasi-Experimental Data," unpublished manuscript.

[43] Ignatiadis, N., and S. Wager (2019): "Bias-Aware Confidence Intervals for Empirical Bayes Analysis," arXiv preprint arXiv:1902.02774.

[44] James, W., and C. Stein (1961): "Estimation with Quadratic Loss," in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, 1, 361–379. Univ. of California Press.

[45] Jochmans, K., and Weidner, M. (2018): "Inference on a distribution from noisy draws," arXiv preprint arXiv:1803.04991.

[46] Kane, T. J., and Staiger, D. O. (2008): "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation", National Bureau of Economic Research (No. w14607).

[47] Koenker, R., and I. Mizera (2014): "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules," *Journal of the American Statistical Association*, 109(506), 674–685.

[48] Kotlarski, I. (1967): "On Characterizing the Gamma and the Normal Distribution," *Pacific Journal of Mathematics*, 20(1), 69–76.

[49] Li, T., and Q. Vuong (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65(2), 139–165.

[50] Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), 205–228.

[51] Manski, C. F. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27(3), 313–333.

[52] Moon, H., and F. Schorfheide (2012): "Bayesian and Frequentist Inference in Partially Identified Models," *Econometrica*, 80(2), 755–782.

[53] Morris, C. N. (1983): "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78(381), 47–55.

[54] Mueller, U. K. (2013): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," *Econometrica*, 81(5), 1805–1849.

[55] Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics* 4, 2111–2245.

[56] Okui, R., and Yanagi, T. (2018): "Kernel Estimation for Panel Data with Heterogeneous Dynamics," arXiv preprint arXiv:1802.08825.

[57] Powell, J. L. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 1435–1460.

[58] Powell, J. L. (1994): "Estimation of Semiparametric Models," *Handbook of Econometrics*, 4, 2443–2521.

[59] Robbins, H. (1964): *The Empirical Bayes Approach to Statistical Decision Problems*, *Annals Math. Stat.*, 35(1), 1-20.

[60] Rust, J. (1987): "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 999–1033.

[61] Tamer, E. (2010): "Partial Identification in Econometrics," *Annu. Rev. Econ.*, 2(1), 167–195.

[62] Wasserman, L. (2006): *All of Nonparametric Statistics*. Springer Science & Business Media.

[63] White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 817-838.

[64] Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

APPENDIX (for online publication)

A Proofs

A.1 Local optimality

The following is an extended version of Lemma 1 and Theorem 1 in the main text, which also covers the case of unbounded functions $\gamma_{\beta,\sigma_*}(y,x)$, $\delta_\beta(u,x)$ and $\psi_{\beta,\sigma_*}(y,x)$. In addition, we make explicit again the dependence on β and σ_*, which we suppressed in the main text.

Lemma A1. *In addition to defining* $\widetilde{\psi}(y,x) = \psi(y,x) - \mathbb{E}_* \left[\psi(Y,X)\middle|X=x\right]$, *let* $\widetilde{\gamma}(y,x) = \gamma(y,x) - \mathbb{E}_* \left[\gamma(Y,X)\middle|X=x\right]$ *and* $\widetilde{\delta}(u,x) = \delta(u,x) - \mathbb{E}_* \left[\delta(U,X)\middle|X=x\right]$. *Suppose that* $\phi(r) = \overline{\phi}(r) + \nu\,(r-1)^2$, *with* $\nu \geq 0$, *and a function* $\overline{\phi}(r)$ *that is four times continuously differentiable with* $\overline{\phi}(1) = 0$ *and* $\overline{\phi}''(r) > 0$, *for all* $r \in (0,\infty)$. *Assume* $\mathbb{E}_{P(\beta,f_{\sigma*})}\psi_{\beta,\sigma_*}(Y,X) = 0$ *and* $\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)'\right] > 0$. *Furthermore, assume that one of the following holds:*

(i) $\nu = 0$, *and the functions* $\left|\gamma_{\beta,\sigma_*}(y,x)\right|$, $|\delta_\beta(u,x)|$ *and* $\left|\psi_{\beta,\sigma_*}(y,x)\right|$ *are bounded over the domain of* Y, U, X.

(ii) $\nu > 0$, *and* $\mathbb{E}_{P(\beta,f_{\sigma*})}\left|\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)\right|^3 < \infty$, *and* $\mathbb{E}_{P(\beta,f_{\sigma*})}\left|\psi_{\beta,\sigma_*}(Y,X)\right|^3 < \infty$.

Then, as $\epsilon \to 0$ *we have*

$$b_\epsilon(\gamma) = \left|\mathbb{E}_{P(\beta,f_{\sigma*})}[\gamma_{\beta,\sigma_*}(Y,X)] - \mathbb{E}_{f_{\sigma*}}[\delta_\beta(U,X)]\right| + \epsilon^{\frac{1}{2}} \left\{\frac{2}{\phi''(1)}\mathrm{Var}_{P(\beta,f_{\sigma*})}\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]\right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon),$$

where

$$\lambda = \left\{\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)'\right]\right\}^{-1}\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\left(\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)\right)\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right].$$

Theorem A1. *Suppose that the conditions of Lemma A1 hold, and let*

$$\gamma^{\mathrm{P}}_{\beta,\sigma_*}(y,x) = \mathbb{E}_{p_{\beta,\sigma_*}}[\delta_\beta(U,X)\,|\,Y=y, X=x]. \tag{A1}$$

Then, as ϵ *tends to zero we have*

$$b_\epsilon(\gamma^{\mathrm{P}}_{\beta,\sigma_*}) \leq b_\epsilon(\gamma) + \mathcal{O}(\epsilon).$$

A.1.1 Proof of Lemma A1 (containing Lemma 1 as a special case)

We first introduce some additional notation and establish some helpful intermediate results. We write $\mathcal{B}$ and $\mathcal{S}$ for the set of possible values of the parameters β and σ, respectively. Lemma A1 is for given values $\beta \in \mathcal{B}$ and $\sigma_* \in \mathcal{S}$, and given functions $\gamma_{\beta,\sigma_*}(y,x)$, $\delta_\beta(u,x)$, $\psi_{\beta,\sigma_*}(y,x)$, and those values and functions are also taken as given in following two intermediate lemmas. Remember also that Γ_ϵ depends on the function $\phi : [0,\infty) \to \mathbb{R} \cup \{\infty\}$, which is assumed to be strictly convex in Lemma A1. We define the corresponding function $\rho : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ by

$$\rho(t) := \begin{cases} \operatorname{argmax}_{r \geq 0} \left[r\, t - \phi(r) \right] & \text{if this "argmax" exists,} \\ \infty & \text{otherwise.} \end{cases} \tag{A2}$$

For $t = \phi'(r)$ we have $\rho(t) = r$, that is, for those values of t the function $\rho(t)$ is simply the inverse function of the first derivative ϕ'. For $t < \inf_{r>0} \phi'(r)$ we have $\rho(t) = 0$, and for $t > \sup_{r>0} \phi'(r)$ the value of $\rho(t)$ is defined to be ∞. The following lemma provides a characterization of the ϵ-worst-case bias $b_\epsilon(\gamma)$ that was defined in (15).

Lemma A2. *Let $\epsilon > 0$. Assume that $\phi(r)$ is strictly convex with $\phi(1) = 0$. Suppose that for $s \in \{-1, 1\}$ and $x \in \mathcal{X}$ there exists $\lambda^{(1)}_{\beta,\sigma_*}(s,x) \in \mathbb{R}$, $\lambda^{(2)}_{\beta,\sigma_*}(s) > 0$, $\lambda^{(3)}_{\beta,\sigma_*}(s) \in \mathbb{R}^{\dim \psi}$ such that*

$$t_{\beta,\sigma_*}(u,x|s) := \lambda^{(1)}_{\beta,\sigma_*}(s,x) + s\, \lambda^{(2)}_{\beta,\sigma_*}(s) \left[\gamma_{\beta,\sigma_*}(g_\beta(u,x),x) - \delta_\beta(u,x) \right] + \lambda^{(3)\prime}_{\beta,\sigma_*}(s)\, \psi_{\beta,\sigma_*}(g_\beta(u,x),x))$$

satisfies

$$\begin{aligned} \forall x \in \mathcal{X}: \quad \mathbb{E}_{P(\beta,f_{\sigma_*})} \left\{ \rho\left[t_{\beta,\sigma_*}(U,X|s) \right] \,\middle|\, X = x \right\} &= 1, \\ \mathbb{E}_{P(\beta,f_{\sigma_*})}\, \phi \left\{ \rho\left[t_{\beta,\sigma_*}(U,X|s) \right] \right\} &= \epsilon, \\ \mathbb{E}_{P(\beta,f_{\sigma_*})} \left\{ \psi_{\beta,\sigma_*}(Y,X)\ \rho\left[t_{\beta,\sigma_*}(U,X|s) \right] \right\} &= 0. \end{aligned} \tag{A3}$$

Then the maximizer ($s = +1$) and minimizer ($s = -1$) of $\mathbb{E}_{P(\beta,f_0)} \left[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right]$ over $f_0 \in \Gamma_\epsilon$ are given by

$$f_0^{(s)}(u|x) = f_{\sigma_*}(u|x)\, \rho\left[t_{\beta,\sigma_*}(u,x|s) \right],$$

and for the worst-case absolute bias we therefore have

$$b_\epsilon(\gamma) = \max_{s \in \{-1,1\}} \left\{ s\, \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\left[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right]\ \rho\left[t_{\beta,\sigma_*}(U,X|s) \right] \right] \right\}.$$

The proof of Lemma A2 is given in Section A.4. Notice that for $\phi(r) = r[\log(r) - 1]$, when $d(f_0, f_{\sigma_*})$ is the Kullback-Leibler divergence, we have $\rho(t) = \exp(t)$, and the worst case distributions $f_0^{(s)}(u|x)$ in Lemma A2 are exponentially tilted versions of the reference distribution $f_{\sigma_*}(u|x)$. Lemma A2 shows that, more generally, the required "tilting function" is given by $\rho(t)$.

We impose $\phi(1) = 0$ throughout the paper to guarantee that $d(f_0, f_{\sigma_*}) \geq 0$ (by an application of Jensen's inequality). In addition, we now impose the normalization $\phi'(1) = 0$. This is without loss of generality, because we can always redefine $\phi(r) \mapsto \phi(r) - (r-1)\,\phi'(1)$, which has no effect on $d(f_0, f_{\sigma_*})$ and guarantees $\phi'(1) = 0$ for the redefined function.

The goal of the following lemma is to establish Taylor expansions of $\rho(t)$ and $\phi(\rho(t))$ around $t = 0$ of the form

$$\rho(t) = 1 + \frac{t}{\phi''(1)} + t^2\, R_1(t), \qquad\qquad \phi(\rho(t)) = \frac{t^2}{2\,\phi''(1)} + t^3\, R_2(t), \tag{A4}$$

where the remainder terms are defined by

$$R_1(t) := \begin{cases} t^{-2}\,[\rho(t) - 1 - t/\phi''(1)] & \text{if } t \neq 0, \\ -\phi'''(1)/\{2\,[\phi''(1)]^3\} & \text{if } t = 0, \end{cases}$$

$$R_2(t) := \begin{cases} t^{-3}\,[\phi(\rho(t)) - t^2/\{2\phi''(1)\}] & \text{if } t \neq 0, \\ -\phi'''(1)/\{3\,[\phi''(1)]^3\} & \text{if } t = 0. \end{cases}$$

Notice that the expansions (A4) are trivially true by definition of $R_1(t)$ and $R_2(t)$, but the following lemma provides bounds on $R_1(t)$ and $R_2(t)$, which are useful for the proof of Lemma A1 afterwards.

Lemma A3. *For all $r \geq 0$ let $\phi(r) = \overline{\phi}(r) + \nu\,(r-1)^2$, for $\nu \geq 0$, and a function $\overline{\phi} : [0, \infty) \to \mathbb{R} \cup \{\infty\}$ that is four times continuously differentiable with $\overline{\phi}(1) = \overline{\phi}'(1) = 0$ and $\overline{\phi}''(r) > 0$, for all $r \in (0, \infty)$. The lemma has two parts:*

(i) Assume in addition that $\nu = 0$. Then, there exist constants $c_1 > 0$, $c_2 > 0$ and $\eta > 0$ such that for all $t \in [-\eta, \eta]$ we have

$$|R_1(t)| \leq c_1, \qquad \text{and} \qquad |R_2(t)| \leq c_2, \tag{A5}$$

and the functions $R_1(t)$ and $R_2(t)$ are continuous within $[-\eta, \eta]$.

(ii) Assume in addition that $\nu > 0$. Then, there exist constants $c_1 > 0$ and $c_2 > 0$ such that the two inequalities in (A5) hold for all $t \in \mathbb{R}$, and the functions $R_1(t)$ and $R_2(t)$ are everywhere continuous.

The proof of Lemma A3 is given in Section A.4.

Comment: Part (i) and part (ii) of Lemma A3 give the same approximations of $\rho(t)$ and $\phi(\rho(t))$, but the difference is that in part (i) the result only holds locally in a neighborhood of $t = 0$, while in part (ii) the inequalities are established globally for all $t \in \mathbb{R}$. Notice that the result of part (ii) cannot hold under the assumptions of part (i) only, because $\rho(t)$ is equal to infinity for all $t > t_{\sup}$, where $t_{\sup} = \sup_{r\in(0,\infty)} \phi'(r)$ can be finite. The regularization $\phi(r) = \overline{\phi}(r) + \nu\,(r-1)^2$, with $\nu > 0$, guarantees that $\rho(t)$ is finite and well-defined for all $t \in \mathbb{R}$. This property of the regularized $\phi(r)$ is key whenever the moment functions γ, δ, ψ are unbounded (i.e., for case (ii) of the assumptions of Lemma A1).

Using the intermediate Lemmas A2 and A3 we can now show Lemma A1, which contains Lemma 1 as a special case.

Proof of Lemma A1. # Additional notation and definitions: In this proof we again drop the arguments β and σ_* everywhere for ease notation, and we write $\mathbb{E}_*$ and Var_* for expectations and variances under the reference distribution $P(\beta, f_{\sigma_*})$. We also continue to use the normalization $\phi'(1) = 0$, which is without loss of generality, as explained above. Let $\lambda \in \mathbb{R}^{\dim\psi}$ be as defined in the statement of the lemma, and furthermore define

$$\kappa = \left\{ \frac{\mathrm{Var}_*\left[\widetilde{\gamma}(Y,X) - \widetilde{\delta}(U,X) - \lambda'\,\widetilde{\psi}(Y,X)\right]}{2\,\phi''(1)} \right\}^{1/2} .$$

For $s \in \{-1,+1\}$ and $\epsilon > 0$, let

$$t(u,x|s) = \lambda^{(1)}(s,x) + s\,\lambda^{(2)}(s)\,[\gamma(g(u,x),x) - \delta(u,x)] + \lambda^{(3)\prime}(s)\,\psi(g(u,x),x),$$

with

$$\begin{aligned}
\lambda^{(1)}(s,x) &= -\epsilon^{1/2}\,s\,\kappa^{-1}\,\mathbb{E}_*\left[\gamma(Y,X) - \delta(U,X) - \lambda'\,\psi(Y,X)\,\middle|\,X=x\right] \\
&\qquad + \epsilon\left\{\lambda^{(1)}_{\mathrm{rem}}(s,x) - s\,\lambda^{(2)}_{\mathrm{rem}}(s)\,\mathbb{E}_*\left[\gamma(Y,X) - \delta(U,X) - \lambda'\,\psi(Y,X)\,\middle|\,X=x\right]\right\}, \\
\lambda^{(2)}(s) &= \epsilon^{1/2}\kappa^{-1} + \epsilon\,\lambda^{(2)}_{\mathrm{rem}}(s), \\
\lambda^{(3)}(s) &= -\epsilon^{1/2}\,s\,\kappa^{-1}\,\lambda + \epsilon\left[\lambda^{(3)}_{\mathrm{rem}}(s) - s\,\lambda^{(2)}_{\mathrm{rem}}(s)\,\lambda\right].
\end{aligned}$$

Here, we are explicit about the leading order terms (of order $\epsilon^{1/2}$), but the higher order terms (of order ϵ) contain the coefficients $\lambda^{(1)}_{\mathrm{rem}}(s) \in \mathbb{R}$, $\lambda^{(2)}_{\mathrm{rem}}(s) \in \mathbb{R}$, and $\lambda^{(3)}_{\mathrm{rem}}(s) \in \mathbb{R}^{\dim\psi}$, which will only be specified in (A8) below. We can rewrite

$$t(u,x|s) = \epsilon^{1/2} t_{(0)}(u,x|s) + \epsilon\, t_{\rm rem}(u,x|s), \tag{A6}$$

with

$$\begin{aligned} t_{(0)}(u,x|s) &= s\,\kappa^{-1}\left[\widetilde{\gamma}(g(u,x),x) - \widetilde{\delta}(u,x) - \lambda'\,\widetilde{\psi}(g(u,x),x)\right], \\ t_{\rm rem}(u,x|s) &= \lambda^{(1)}_{\rm rem}(s,x) + \lambda^{(2)}_{\rm rem}(s)\,\kappa\, t_{(0)}(u,x|s) + \lambda^{(3)\prime}_{\rm rem}(s)\,\psi(g(u,x),x). \end{aligned}$$

Here, $t(u,x|s)$, $\lambda^{(1)}(s,x)$, $\lambda^{(2)}(s)$, etc, also depend on ϵ, but we do not make this dependence explicit in our notation. Our goal is to apply Lemma A2 with $t_{\beta,\sigma_*}(u,x|s)$ in the lemma equal to $t(u,x|s)$ as defined here. However, in order to apply that lemma we need to satisfy the conditions (A3), which in current notation read

$$\mathbb{E}_*\,\rho\left[t(U,X|s)\big|X=x\right] = 1, \quad \mathbb{E}_*\,\phi\left\{\rho\left[t(U,X|s)\right]\right\} = \epsilon, \quad \mathbb{E}_*\left\{\psi(Y,X)\;\rho\left[t(U,X|s)\right]\right\} = 0. \tag{A7}$$

The definition of $t(u,x|s)$ above is already designed to satisfy (A7) to leading order in ϵ, but we still need to find $\lambda^{(1)}_{\rm rem}(s,x)$, $\lambda^{(2)}_{\rm rem}(s)$, $\lambda^{(3)}_{\rm rem}(s)$ such that (A7) holds exactly. Plugging the expansions (A4) into (A7), using the definition of $t(u,x|s)$, as well as $\mathbb{E}_*\left[t_{(0)}(U,X|s)\big|X=x\right] = 0$, $\mathbb{E}_*\left\{[t_{(0)}(U,X|s)]^2\right\} = 2\,\phi''(1)$, and $\mathbb{E}_*\psi(Y,X)\,t_{(0)}(U,X|s) = 0$, we obtain

$$\begin{aligned} \mathbb{E}_*\left\{\frac{\epsilon\, t_{\rm rem}(U,X|s)}{\phi''(1)} + [t(U,X|s)]^2\, R_1\left[t(U,X|s)\right]\,\middle|\, X=x\right\} &= 0, \\ \mathbb{E}_*\left\{\frac{2\,\epsilon^{3/2}\, t_{\rm rem}(U,X|s)\, t_{(0)}(U,X|s) + \epsilon^2\left[t_{\rm rem}(U,X|s)\right]^2}{2\phi''(1)} + \left[t(U,X|s)\right]^3 R_2\left[t(U,X|s)\right]\right\} &= 0, \\ \mathbb{E}_*\left\{\frac{\epsilon\,\psi(Y,X)\, t_{\rm rem}(U,X|s)}{\phi''(1)} + \psi(Y,X)\;\left[t(U,X|s)\right]^2 R_1\left[t(U,X|s)\right]\right\} &= 0. \end{aligned}$$

Those conditions can be rewritten as follows

$$\begin{aligned} \lambda^{(1)}_{\rm rem}(s,x) &= -\phi''(1)\,\mathbb{E}_*\left\{\left[t_{(0)}(U,X|s) + \epsilon^{1/2} t_{\rm rem}(U,X|s)\right]^2\; R_1\left[t(U,X|s)\right]\,\middle|\, X=x\right\}, \\ \lambda^{(2)}_{\rm rem}(s) &= -\frac{1}{2\,\kappa}\,\mathbb{E}_*\left\{\left[t_{(0)}(U,X|s) + \epsilon^{1/2} t_{\rm rem}(U,X|s)\right]^3\; R_2\left[t(U,X|s)\right] + \frac{\epsilon^{1/2}\left[t_{\rm rem}(U,X|s)\right]^2}{2\phi''(1)}\right\}, \\ \lambda^{(3)}_{\rm rem}(s) &= -\phi''(1)\;\left\{\mathbb{E}_*\left[\psi(Y,X)\,\psi(Y,X)'\right]\right\}^{-1} \\ &\quad \times \mathbb{E}_*\left\{\psi(Y,X)\;\left[t_{(0)}(U,X|s) + \epsilon^{1/2} t_{\rm rem}(U,X|s)\right]^2 R_1\left[t(U,X|s)\right]\right\}. \end{aligned} \tag{A8}$$

Thus, as $\epsilon \to 0$ we have

$$\lambda^{(1)}_{\rm rem}(s,x) = -2[\phi''(1)]^2\, R_1(0)\, \frac{\mathrm{Var}_*\left[\widetilde{\gamma}(Y,X) - \widetilde{\delta}(U,X) - \lambda'\,\widetilde{\psi}(Y,X)\,\Big|\, X=x\right]}{\mathrm{Var}_*\left[\widetilde{\gamma}(Y,X) - \widetilde{\delta}(U,X) - \lambda'\,\widetilde{\psi}(Y,X)\right]} + \mathcal{O}(\epsilon^{1/2}),$$

$$\lambda^{(2)}_{\rm rem}(s) = -\frac{1}{2\,\kappa}\,\mathbb{E}_*\left[t_{(0)}(U,X|s)\right]^3\; R_2(0) + \mathcal{O}(\epsilon^{1/2}),$$

$$\lambda^{(3)}_{\rm rem}(s) = -\phi''(1)\; \left\{\mathbb{E}_*\left[\psi(Y,X)\,\psi(Y,X)'\right]\right\}^{-1}\mathbb{E}_*\left\{\psi(Y,X)\;\left[t_{(0)}(U,X|s)\right]^2\right\}\; R_1(0) + \mathcal{O}(\epsilon^{1/2}). \tag{A9}$$

Notice that $\lambda^{(1)}_{\rm rem}(s,x)$, $\lambda^{(2)}_{\rm rem}(s)$, $\lambda^{(3)}_{\rm rem}(s)$ also appear implicitly on the right-hand sides of the equations (A8), because $t_{\rm rem}(u,x|s)$ depends on those parameters, and (A8) is therefore a system of equations for $\lambda^{(1)}_{\rm rem}(s,x)$, $\lambda^{(2)}_{\rm rem}(s)$, $\lambda^{(3)}_{\rm rem}(s)$. Our assumptions guarantee that the system (A8) has a solution for sufficiently small ϵ, as will be explained below for the two different cases distinguished in the lemma.

Proof for case (i): The assumptions for this case guarantee that $t(u,x|s)$ is uniformly bounded over u and x. Part (i) of Lemma A3 guarantees existence of $c_1 > 0$, $c_2 > 0$, $\eta > 0$ such that for all $t \in [-\eta,\eta]$ we have $|R_1(t)| \leq c_1$ and $|R_2(t)| \leq c_2$. For sufficiently small ϵ we have $t(u,x|s) \in [-\eta,\eta]$ for all u and x, implying that as $\epsilon \to 0$ there exists a solution of (A8) that satisfies (A9), which in particular implies

$$\sup_{x\in\mathcal{X}}\left|\lambda^{(1)}(s,x)\right| = \mathcal{O}(1), \qquad \lambda^{(2)}(s) = \mathcal{O}(1), \qquad \lambda^{(3)}(s) = \mathcal{O}(1), \tag{A10}$$

and by construction the conditions (A7) are satisfied for that solution. Thus, for sufficiently small ϵ the $t(u,x|s)$ defined above satisfies the conditions of Lemma A2. Applying that lemma we thus obtain that, for sufficiently small ϵ, we have

$$b_\epsilon(\gamma) = \max_{s\in\{-1,1\}}\left\{s\,\mathbb{E}_*\left[\left[\gamma(Y,X) - \delta(U,X)\right]\;\rho\left[t(U,X|s)\right]\right]\right\}.$$

Again applying the expansion for $\rho(t)$ in (A4), and part (i) of Lemma A3 we thus obtain that

$$\begin{aligned} b_\epsilon(\gamma) &= \max_{s\in\{-1,1\}}\left\{s\,\mathbb{E}_*\left[\gamma(Y,X) - \delta(U,X)\right]\right\} \\ &\qquad + \epsilon^{1/2}\left\{\frac{2}{\phi''(1)}\mathrm{Var}_*\left[\widetilde{\gamma}(Y,X) - \widetilde{\delta}(U,X) - \lambda'\,\widetilde{\psi}(Y,X)\right]\right\}^{1/2} + \mathcal{O}(\epsilon) \\ &= \left|\mathbb{E}_*\left[\gamma(Y,X) - \delta(U,X)\right]\right| + \epsilon^{1/2}\left\{\frac{2}{\phi''(1)}\mathrm{Var}_*\left[\widetilde{\gamma}(Y,X) - \widetilde{\delta}(U,X) - \lambda'\,\widetilde{\psi}(Y,X)\right]\right\}^{1/2} + \mathcal{O}(\epsilon). \end{aligned} \tag{A11}$$

This is what we wanted to show.

Proof for case (ii): In this case, according to part (ii) of Lemma A3 the functions $R_1(t)$ and $R_2(t)$ are continuous and bounded over all $t \in \mathbb{R}$. In addition, we have assumed that $\mathbb{E}_* |\gamma(Y,X) - \delta(U,X)|^3 < \infty$, and $\mathbb{E}_* |\psi(Y,X)|^3 < \infty$, which guarantees that all of the expectations in (A8) are finite. We therefore again conclude that for small ϵ the equations (A8) have a solution such that (A10) holds. The remainder of the proof is equivalent to the proof of part (i), that is, we again apply Lemma A2 and Lemma A3 to obtain (A11). ∎

A.1.2 Proof of Theorem A1

By applying Lemma A1 to both $\gamma_{\beta,\sigma_*}(y,x)$ and $\gamma^{\rm P}_{\beta,\sigma_*}(y,x) = \mathbb{E}_{p_{\beta,\sigma_*}}[\delta_\beta(U,X)\,|\,Y=y, X=x]$ we obtain, as $\epsilon \to 0$,

$$
\begin{aligned}
b_\epsilon(\gamma) &= \left|\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma_{\beta,\sigma_*}(Y,X)] - \mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X)]\right| \\
&\quad + \epsilon^{\frac{1}{2}} \left\{ \frac{2}{\phi''(1)} \mathrm{Var}_{P(\beta,f_{\sigma_*})} \left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \, \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right] \right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon), \\
b_\epsilon(\gamma^{\rm P}) &= \epsilon^{\frac{1}{2}} \left\{ \frac{2}{\phi''(1)} \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\left(\gamma^{\rm P}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right)^2 \right] \right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon),
\end{aligned}
$$

where

$$
\lambda = \left\{ \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\widetilde{\psi}_{\beta,\sigma_*}(Y,X) \, \widetilde{\psi}_{\beta,\sigma_*}(Y,X)' \right] \right\}^{-1} \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\left(\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right) \, \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right].
$$

Here, to simplify $b_\epsilon(\gamma^{\rm P})$ we used that by the law of iterated expectations we have that $\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)] - \mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X)] = 0$ (that is, the first term in $b_\epsilon(\gamma)$ is not present in $b_\epsilon(\gamma^{\rm P})$) and also $\mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\left(\gamma^{\rm P}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right) \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right] = 0$ (that is, the vector λ is equal to zero for $\gamma^{\rm P}$).

For any $\gamma_{\beta,\sigma_*}(y,x)$ with $\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma_{\beta,\sigma_*}(Y,X)] - \mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X)] \neq 0$ we have $b_\epsilon(\gamma^{\rm P}) \leq b_\epsilon(\gamma)$ for sufficiently small ϵ, and the statement of the theorem thus holds in that case. In the following we therefore consider the case that $\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma_{\beta,\sigma_*}(Y,X)] - \mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X)] = 0$. The expression for $b_\epsilon(\gamma)$ then simplifies analogously to the expression for $b_\epsilon(\gamma^{\rm P})$; that is, we have

$$
b_\epsilon(\gamma) = \epsilon^{\frac{1}{2}} \left\{ \frac{2}{\phi''(1)} \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\left(\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \, \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right)^2 \right] \right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon).
$$

Again applying the law of iterated expectations we find that

$$\begin{aligned}
&\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_{\beta}(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]\\
&\quad=\mathbb{E}_{P(\beta,f_{\sigma*})}\left[-\delta_{\beta}(U,X)\right]\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]\\
&\quad=\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]\\
&\quad=\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]^2.
\end{aligned}$$

Using this we obtain

$$\begin{aligned}
&0\leq\mathbb{E}_{P(\beta,f_{\sigma*})}\left\{\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_{\beta}(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]-\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]\right\}^2\\
&=\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_{\beta}(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]^2+\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]^2\\
&\quad-2\,\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_{\beta}(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]\\
&=\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_{\beta}(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]^2-\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]^2,
\end{aligned}$$

that is, we have shown that

$$\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_{\beta}(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]^2\geq\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\gamma^{\rm P}_{\beta,\sigma_*}(Y,X)-\delta_{\beta}(U,X)\right]^2,$$

and therefore we obtain that

$$b_\epsilon(\gamma^{\rm P}_{\beta,\sigma_*})\leq b_\epsilon(\gamma)+\mathcal{O}(\epsilon).$$

A.2 Global bound

We are now going to show Theorem 2, which we restate here.

Theorem. *Let $\gamma^{\rm P}_{\beta,\sigma_*}$ as in (A1). Then, for all $\epsilon>0$,*

$$b_\epsilon(\gamma^{\rm P}_{\beta,\sigma_*})\leq 2\,\inf_{\gamma}\,b_\epsilon(\gamma_{\beta,\sigma_*}).$$

The following lemma is useful for the proof of this theorem (Theorem 2 in the main text).

Lemma A4. *Let $\epsilon\geq 0$, $\beta\in\mathcal{B}$, $\sigma_*\in\mathcal{S}$, and let $\zeta:\mathcal{U}\times\mathcal{X}\to\mathbb{R}$. Then we have*

$$\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\zeta(U,X)\,|\,Y,X\right]\right\}\right|\leq\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left[\zeta(U,X)\right]\right|.$$

The proof of this lemma is given in Section A.4. Notice that both Theorem 2 and Lemma A4 require that $\phi(r)$ is convex with $\phi(1)=0$, but they do not require $\phi''(1)>0$. For example, $\phi(r)=|r-1|/2$ is allowed here, which gives the total variation distance for $d(f_0,f_{\sigma_*})$.

Proof of Theorem 2. By definition we have

$$b_\epsilon(\gamma) = \sup_{f_0\in\Gamma_\epsilon} \left|\mathbb{E}_{P(\beta,f_0)}[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)]\right|,$$
$$b_\epsilon(\gamma^{\mathrm{P}}) = \sup_{f_0\in\Gamma_\epsilon} \left|\mathbb{E}_{P(\beta,f_0)}[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)]\right|.$$

By writing $\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) = \gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \left[\gamma_{\beta,\sigma_*}(Y,X) - \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)\right]$ we obtain

$$\begin{aligned}
b_\epsilon(\gamma^{\mathrm{P}}) &= \sup_{f_0\in\Gamma_\epsilon} \left|\mathbb{E}_{P(\beta,f_0)}\left[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)\right] - \mathbb{E}_{P(\beta,f_0)}\left[\gamma_{\beta,\sigma_*}(Y,X) - \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)\right]\right| \\
&\leq b_\epsilon(\gamma) + \sup_{f_0\in\Gamma_\epsilon} \left|\mathbb{E}_{P(\beta,f_0)}\left[\gamma_{\beta,\sigma_*}(Y,X) - \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)\right]\right| \\
&= b_\epsilon(\gamma) + \sup_{f_0\in\Gamma_\epsilon} \left|\mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\gamma_{\beta,\sigma_*}(g_\beta(U,X),X) - \delta_\beta(U,X)\,|\,Y,X\right]\right\}\right| \\
&\leq b_\epsilon(\gamma) + \sup_{f_0\in\Gamma_\epsilon} \left|\mathbb{E}_{P(\beta,f_0)}[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)]\right| = 2\,b_\epsilon(\gamma),
\end{aligned}$$

where in the second-to-last step we have used Lemma A4 with $\zeta(u,x) = \gamma_{\beta,\sigma_*}(g_\beta(u,x),x) - \delta_\beta(u,x)$. We have thus shown that $b_\epsilon(\gamma^{\mathrm{P}}) \leq 2\,b_\epsilon(\gamma)$ holds for any function $\gamma_{\beta,\sigma_*}(y,x)$, which implies that

$$b_\epsilon(\gamma^{\mathrm{P}}) \leq 2 \inf_\gamma b_\epsilon(\gamma).$$

∎

A.3 Robustness in prediction

Remember the definition

$$e_\epsilon(\gamma) = \sup_{f_0\in\Gamma_\epsilon} \mathbb{E}_{P(\beta,f_0)}\left[(\delta_\beta(U,X) - \gamma_{\beta,\sigma_*}(Y,X))^2\right].$$

We first restate Lemma 2 and Theorem 3 in the main text.

Lemma A5. *In addition to defining $\widetilde{\psi}(y,x) = \psi(y,x) - \mathbb{E}_*\left[\psi(Y,X)\middle|X=x\right]$, let $\widetilde{\gamma}(y,x) = \gamma(y,x) - \mathbb{E}_*\left[\gamma(Y,X)\middle|X=x\right]$ and $\widetilde{\delta}(u,x) = \delta(u,x) - \mathbb{E}_*\left[\delta(U,X)\middle|X=x\right]$. Suppose that $\phi(r) = \overline{\phi}(r) + \nu\,(r-1)^2$, with $\nu \geq 0$, and a function $\overline{\phi}(r)$ that is four times continuously differentiable with $\overline{\phi}(1) = 0$ and $\overline{\phi}''(r) > 0$, for all $r \in (0,\infty)$. Assume $\mathbb{E}_{P(\beta,f_{\sigma_*})}\psi_{\beta,\sigma_*}(Y,X) = 0$ and $\mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\widetilde{\psi}_{\beta,\sigma_*}(Y,X)'\right] > 0$. Furthermore, assume that one of the following holds:*

(i) $\nu = 0$, and the functions $\left|\gamma_{\beta,\sigma_*}(y,x)\right|$, $|\delta_\beta(u,x)|$ and $\left|\psi_{\beta,\sigma_*}(y,x)\right|$ are bounded over the domain of Y, U, X.

(ii) $\nu > 0$, *and* $\mathbb{E}_{P(\beta,f_{\sigma*})}\left|\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right|^6<\infty$, *and* $\mathbb{E}_{P(\beta,f_{\sigma*})}\left|\psi_{\beta,\sigma_*}(Y,X)\right|^3<\infty$.

Then, as $\epsilon \to 0$ *we have*

$$
\begin{aligned}
e_\epsilon(\gamma) = {} & \mathbb{E}_{P(\beta,f_{\sigma*})}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2\right] \\
& + \epsilon^{\frac{1}{2}}\Bigg(\frac{2}{\phi''(1)}\mathrm{Var}_{P(\beta,f_{\sigma*})}\Bigg\{\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2 \\
& \qquad - \mathbb{E}_{P(\beta,f_{\sigma*})}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2\Big|X\right]-\lambda'\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\Bigg\}\Bigg)^{\frac{1}{2}} \\
& + \mathcal{O}(\epsilon),
\end{aligned}
$$

where

$$
\lambda = \left\{\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\widetilde{\psi}_{\beta,\sigma_*}(Y,X)'\right]\right\}^{-1}\mathbb{E}_{P(\beta,f_{\sigma*})}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right].
$$

Theorem A2. *Let* $\gamma^{\mathrm{P}}_{\beta,\sigma_*}$ *as in (A1). Then, for all* $\epsilon > 0$,

$$
e_\epsilon(\gamma^{\mathrm{P}}_{\beta,\sigma_*}) \leq 4 \inf_\gamma e_\epsilon(\gamma_{\beta,\sigma_*}).
$$

Proof of Lemma A5. This statement of the lemma is obtained from Lemma A1 by replacing $(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X))$ by $\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2$. The proof is obtained by the same replacement from the proof of Lemma A1. ∎

Proof of Theorem A2. By definition we have

$$
\begin{aligned}
e_\epsilon(\gamma) &= \sup_{f_0\in\Gamma_\epsilon}\mathbb{E}_{P(\beta,f_0)}[(\delta_\beta(U,X)-\gamma_{\beta,\sigma_*}(Y,X))^2], \\
e_\epsilon(\gamma^{\mathrm{P}}) &= \sup_{f_0\in\Gamma_\epsilon}\mathbb{E}_{P(\beta,f_0)}[(\delta_\beta(U,X)-\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X))^2].
\end{aligned}
$$

Using that $(a-b)^2 \leq 2(a^2+b^2)$ with $a=\delta_\beta(U,X)-\gamma_{\beta,\sigma_*}(Y,X)$ and $b=\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)-\gamma_{\beta,\sigma_*}(Y,X)$ we obtain

$$
\begin{aligned}
e_\epsilon(\gamma^{\mathrm{P}}) &\leq 2\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left[\left(\delta_\beta(U,X)-\gamma_{\beta,\sigma_*}(Y,X)\right)^2\right]+\mathbb{E}_{P(\beta,f_0)}\left[\left(\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)-\gamma_{\beta,\sigma_*}(Y,X)\right)^2\right]\right| \\
&\leq 2e_\epsilon(\gamma)+2\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)\right)^2\right]\right|.
\end{aligned}
$$

We furthermore have

$$\begin{aligned}
&\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)\right)^2\right]\right| \\
&\quad=\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left\{\left[\mathbb{E}_{p_{\beta,\sigma_*}}\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\,|\,Y,X\right)\right]^2\right\}\right| \\
&\quad\le\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2\,|\,Y,X\right]\right\}\right| \\
&\quad\le\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left[\left(\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2\right]\right|=e_\epsilon(\gamma),
\end{aligned}$$

where in the first step we used the definition of $\gamma^{\mathrm{P}}_{\beta,\sigma_*}(y,x)$, in the second step we applied the Cauchy-Schwarz inequality, and in the last line we used Lemma A4 and the definition of $e_\epsilon(\gamma)$. Combining the results of the last two displays we obtain that

$$e_\epsilon(\gamma^{\mathrm{P}}_{\beta,\sigma_*})\le 4\inf_{\gamma}\, e_\epsilon(\gamma_{\beta,\sigma_*}).$$

∎

A.4 Proof of Technical Lemmas

Proof of Lemma A2. In the following we assume that $f_{\sigma_*}(u|x)f_X(x)>0$ for all (u,x) in the joint domain of (U,X). This is without loss of generality, because we can define the joint domain of (U,X) such that this is the case. With a slight abuse of notation we continue to write $\mathcal{U}\times\mathcal{X}$ for the joint domain, even though this need not be a product set.

To account for the absolute value in the definition of $b_\epsilon(\gamma)$ in (15) we let

$$b_\epsilon(\gamma,s)=\sup_{f_0\in\Gamma_\epsilon}\left\{s\,\mathbb{E}_{P(\beta,f_0)}\left[\gamma_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right]\right\},$$

for $s\in\{-1,1\}$. We then have $b_\epsilon(\gamma)=\max_{s\in\{-1,1\}}b_\epsilon(\gamma,s)$. In the following we drop the arguments β and σ_* everywhere, that is, we simply write $g(u,x)$, $\gamma(y,x)$, $\delta(u,x)$, $f_*(u|x)$, $\psi(y,x)$, $\lambda^{(1)}(s,x)$, $\lambda^{(2)}(s)$, $\lambda^{(3)}(s)$ instead of $g_\beta(u,x)$, $\gamma_{\beta,\sigma_*}(y,x)$, $\delta_\beta(u,x)$, $f_{\sigma_*}(u|x)$, $\psi_{\beta,\sigma_*}(y,x)$, $\lambda^{(1)}_{\beta,\sigma_*}(s)$, $\lambda^{(2)}_{\beta,\sigma_*}(s)$, $\lambda^{(3)}_{\beta,\sigma_*}(s)$. The optimal $f_0(u|x)$ in the definition of $b_\epsilon(\gamma,s)$ solves, for $u,x\in\mathcal{U}\times\mathcal{X}$ almost surely under the reference distribution,

$$\begin{aligned}
\widetilde{f}_0(u|x;s)=\operatorname*{argmax}_{f_0\in[0,\infty)}\Bigg\{&s\left[\gamma(g(u,x),x)-\delta(u,x)\right]f_X(x)\,f_0-\mu_1(s,x)\,f_X(x)\,f_0 \\
&-\mu_2(s)\,\phi\left(\frac{f_0}{f_*(u|x)}\right)\,f_*(u|x)\,f_X(x)-\mu_3'(s)\,\psi(g(u,x),x)\;f_X(x)\,f_0\Bigg\},
\end{aligned}\tag{A12}$$

where $\mu_1(s,x) \in \mathbb{R}$, $\mu_2(s) > 0$, $\mu_3(s) \in \mathbb{R}^{\dim \psi}$ are Lagrange multipliers, which we choose to reparameterize as follows

$$\mu_1(s,x) = -\frac{\lambda^{(1)}(s,x)}{\lambda^{(2)}(s)}, \qquad \mu_2(s) = \frac{1}{\lambda^{(2)}(s)}, \qquad \mu_3(s) = -\frac{\lambda^{(3)}(s)}{\lambda^{(2)}(s)}.$$

Those (reparameterized) Lagrange multipliers need to be chosen such that the constraints

$$\begin{aligned} \int_{\mathcal{U}\times\mathcal{X}} \widetilde{f}_0(u|x;s)\, f_X(x)\, du\, dx &= 1, \\ \int_{\mathcal{U}\times\mathcal{X}} \phi\left(\frac{\widetilde{f}_0(u|x;s)}{f_*(u|x)}\right) f_*(u|x)\, f_X(x)\, du\, dx &= \epsilon, \\ \int_{\mathcal{U}\times\mathcal{X}} \psi(g(u,x),x)\, \widetilde{f}_0(u|x;s)\, f_X(x)\, du\, dx &= 0 \end{aligned} \tag{A13}$$

are satisfied. We need $\lambda^{(2)}(s) > 0$ because the second constraint here is actually an inequality constraint ($\leq \epsilon$). Our assumptions guarantee that $f_*(u|x) > 0$ and $f_X(x) > 0$. We can therefore rewrite (A12) as follows,

$$\frac{\widetilde{f}_0(u|x;s)}{f_*(u|x)} = \operatorname*{argmax}_{r\geq 0} \left\{ r\, t(u,x|s) - \phi(r) \right\},$$

where $r = f_0\, f_*(u|x)$, the objective function was multiplied with $f_{\sigma_*}(u|x) f_X(x)$ (which does not change the value of the argmax), and $t(u,x|s) = t_{\beta,\sigma_*}(u,x|s)$ is defined in the statement of the lemma. Comparing the last display with the definition of $\rho(t)$ in (A2) we find that if $\rho\left[t(u,x|s)\right] < \infty$, then

$$\widetilde{f}_0(u|x;s) = f_*(u|x)\, \rho\left[t(u,x|s)\right].$$

The condition $\rho\left[t(u,x|s)\right] < \infty$ is implicitly imposed in the statement of the lemma, because otherwise we could not have $\mathbb{E}_{P(\beta,f_{\sigma_*})}\, \rho\left[t_{\beta,\sigma_*}(U,X|s)\right] = 1$. Using the result in the last display we find that the constraints (A13) are exactly the conditions (A3) imposed in the lemma. Under the conditions of the lemma we therefore have

$$\begin{aligned} b_\epsilon(\gamma,s) &= \sup_{f_0\in\Gamma_\epsilon} \left\{ s\, \mathbb{E}_{P(\beta,f_0)} \left[\gamma(Y,X) - \delta(U,X)\right] \right\} \\ &= \int_{\mathcal{U}\times\mathcal{X}} \left[\gamma(g(u,x),x) - \delta(u,x)\right] \widetilde{f}_0(u|x;s)\, f_X(x)\, du\, dx \\ &= s\, \mathbb{E}_{P(\beta,f_{\sigma*})} \Big\{ \left[\gamma(Y,X) - \delta(U,X)\right]\ \rho\left[t(U,X|s)\right] \Big\}, \end{aligned}$$

and from $b_\epsilon(\gamma) = \max_{s\in\{-1,1\}} b_\epsilon(\gamma,s)$ we thus obtain the statement of the lemma. ∎

Proof of Lemma A3. # Part (i): For $\nu = 0$ we have $\phi = \overline{\phi}$. Our assumptions imply that there exists $\tau > 0$ such that $\phi'(r)$, $\phi''(r)$, $\phi'''(r)$ and $\phi''''(r)$ are all uniformly bounded over $r \in [1-\tau, 1+\tau]$. We can choose $\eta > 0$ such that $[\rho(-\eta), \rho(\eta)] \subset [1-\tau, 1+\tau]$. The conjugate of the convex function $\phi : \mathbb{R} \to \mathbb{R}$ is given by

$$\phi_*(t) = \max_{r \geq 0} \, [r\, t - \phi(r)] = \rho(t)\, t - \phi(\rho(t)). \tag{A14}$$

We have $\rho(t) = \phi_*'(t)$, which is the inverse function of $\phi'(r)$; that is, $\phi'(\rho(t)) = t$. We can express all derivatives of ϕ_* in terms of derivatives of ϕ, for example, $\phi_*''(t) = 1/\phi''(\rho(t))$ and $\phi_*'''(t) = -\phi'''(\rho(t))/[\phi''(\rho(t))]^3$. A Taylor expansion of $\rho(t) = \phi_*'(t)$ around $t = 0 = \phi'(1)$ reads

$$\rho(t) = 1 + \frac{t}{\phi''(1)} + t^2\, R_1(t),$$

where by the mean-value formula for the remainder term we have

$$|R_1(t)| \leq \frac{1}{2} \sup_{t' \in [-\eta,\eta]} |\phi_*'''(t')| \leq \underbrace{\frac{1}{2} \sup_{r \in [1-\tau, 1+\tau]} \left| \frac{\phi'''(r)}{[\phi''(r)]^3} \right|}_{=: c_1 < \infty}.$$

Similarly, a Taylor expansion of $\phi(\rho(t)) = t\, \rho(t) - \phi_*(t)$ around $t = 0$ reads

$$\phi(\rho(t)) = \frac{t^2}{2\, \phi''(1)} + t^3\, R_2(t),$$

where again by the mean-value formula for the remainder we have

$$|R_2(t)| \leq \underbrace{\frac{1}{6} \sup_{r \in [1-\tau, 1+\tau]} \left| -\frac{2\phi'''(r)}{[\phi''(r)]^3} + \frac{3\phi'(r)[\phi'''(r)]^2}{[\phi''(r)]^5} - \frac{\phi'(r)\phi''''(r)}{[\phi''(r)]^4} \right|}_{=: c_2 < \infty}.$$

Continuity of $R_1(t)$ and $R_2(t)$ in a neighborhood of $t = 0$ is also guaranteed by $\phi'(r)$ being four times continuously differentiable in neighborhood around $r = 1$. This concludes the proof of part (i).

Part (ii): For $\nu > 0$ the function $\phi(r) = \overline{\phi}(r) + \nu\, (r-1)^2$ still satisfies all the assumptions of part (i) of the lemma, that is, we can apply part (i) to find that there exists $\widetilde{c}_1 > 0$, $\widetilde{c}_2 > 0$ and $\eta > 0$ such that for all $t \in [-\eta, \eta]$ we have

$$|R_1(t)| \leq \widetilde{c}_1\, t^2, \quad \text{and} \quad |R_2(t)| \leq \widetilde{c}_2\, t^3. \tag{A15}$$

What is left to show here is that there exists constant $c_1 > 0$ and $c_2 > 0$ such that (A5) also holds for $t < -\eta$ and for $t > \eta$.

We have $\phi'(r) = \overline{\phi}'(r) + \nu(r-1)$. Plugging in $r = \rho(t)$ we have $\phi'(\rho(t)) = t$, and therefore $t = \overline{\phi}'(\rho(t)) + \nu[\rho(t) - 1]$. Our assumptions imply that $\overline{\phi}'(\rho(t)) > 0$ for $t > 0$ and $\overline{\phi}'(\rho(t)) < 0$ for $t < 0$. We therefore find that

$$|\rho(t) - 1| = \frac{\left|t - \overline{\phi}'(\rho(t))\right|}{\nu} \leq \frac{|t|}{\nu}. \tag{A16}$$

Using (A15) and (A16), and choosing $c_1 = \max\{\widetilde{c}_1, [1/\nu + 1/\phi''(1)]/\eta\}$, we obtain

$$\left|\rho(t) - 1 - \frac{t}{\phi''(1)}\right| \leq c_1 t^2,$$

for all $t \in \mathbb{R}$. This is the first inequality that we wanted to show.

Using again the convex conjugate defined in (A14) we have

$$\phi(\rho(t)) = t\,\rho(t) - \phi_*(t) = t\,\rho(t) - \max_{r \geq 0}\,[r\,t - \phi(r)] \leq t[\rho(t) - 1] = |t|\;|\rho(t) - 1|\;,$$

where in the second to last step we used that $r = 1$ is one possible choice for $r \geq 0$, and we have $\phi(1) = 0$, and in the last step we used that $\mathrm{sign}[\rho(t) - 1] = \mathrm{sign}(t)$. Our assumptions imply that $\phi(r) \geq 0$, that is, $|\phi(r)| = \phi(r)$. The result in the last display together with (A16) therefore give

$$|\phi(\rho(t))| \leq \frac{t^2}{\nu},$$

for all $t \in \mathbb{R}$. Using this and (A15), and choosing $c_2 = \max\{\widetilde{c}_2, [1/\nu + 1/\{2\phi''(1)\}]/\eta\}$, we thus obtain

$$\left|\phi(\rho(t)) - \frac{t^2}{2\,\phi''(1)}\right| \leq c_2\,t^3,$$

for all $t \in \mathbb{R}$, which is the second inequality that we wanted to show. Continuity of $R_1(t)$ and $R_2(t)$ in $\mathbb{R}$ is also guaranteed by $\phi'(r)$ being four times continuously differentiable in $r \in (0, \infty)$. This concludes the proof of part (ii). ∎

Proof of Lemma A4. Let $f_0 \in \Gamma_\epsilon$. Remember the definition of the posterior distribution $p_{\beta,\sigma_*}(u\,|\,y, x)$ in (12). Define

$$\widetilde{f}_0(u|x) := \mathbb{E}_{P(\beta,f_0)}\,[p_{\beta,\sigma_*}(u\,|\,Y, x)] = \int_{\mathcal{U}} p_{\beta,\sigma_*}(u\,|\,g_\beta(\tilde{u}, x), x)\, f_0(\tilde{u}|x)\, d\tilde{u}.$$

Then, for any $x \in \mathcal{X}$ we have $\widetilde{f}_0(u|x) \geq 0$, for all $u \in \mathcal{U}$, and $\int_{\mathcal{U}} \widetilde{f}_0(u|x) du = 1$; that is, $\widetilde{f}_0(u|x)$ is a probability density over $\mathcal{U}$. Furthermore, by construction we have

$$\mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\,[\zeta(U, X)\,|\,Y, X]\right\} = \mathbb{E}_{P(\beta,\widetilde{f}_0)}\,[\zeta(U, X)]\,. \tag{A17}$$

We also find that

$$\mathbb{E}_{P(\beta,\widetilde{f}_0)}[\psi_{\beta,\sigma_*}(Y,X)] = \mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\psi_{\beta,\sigma_*}(Y,X)\,|\,Y,X\right]\right\} = \mathbb{E}_{P(\beta,f_0)}[\psi_{\beta,\sigma_*}(Y,X)] = 0. \tag{A18}$$

Furthermore, we have

$$\begin{aligned} d(\widetilde{f}_0, f_{\sigma_*}) &= \int_{\mathcal{X}}\int_{\mathcal{U}} \phi\left(\frac{\widetilde{f}_0(u\,|\,x)}{f_{\sigma_*}(u\,|\,x)}\right) f_{\sigma_*}(u\,|\,x) f_X(x)\,du\,dx \\ &= \int_{\mathcal{X}}\int_{\mathcal{U}} \phi\left(\frac{\int_{\mathcal{U}} p_{\beta,\sigma_*}(u\,|\,g_\beta(\tilde{u},x),x)\, f_0(\tilde{u}|x)\,d\tilde{u}}{f_{\sigma_*}(u\,|\,x)}\right) f_{\sigma_*}(u\,|\,x) f_X(x)\,du\,dx \\ &= \int_{\mathcal{X}}\int_{\mathcal{U}} \phi\left(\int_{\mathcal{U}} \frac{f_0(\tilde{u}|x)}{f_{\sigma_*}(\tilde{u}\,|\,x)}\, K_{\beta,\sigma_*}(\tilde{u}|u,x)\,d\tilde{u}\right) f_{\sigma_*}(u\,|\,x) f_X(x)\,du\,dx, \end{aligned}$$

where we defined

$$K_{\beta,\sigma_*}(\tilde{u}|u,x) = \frac{f_{\sigma_*}(\tilde{u}\,|\,x)\, p_{\beta,\sigma_*}(u\,|\,g_\beta(\tilde{u},x),x)}{f_{\sigma_*}(u\,|\,x)}.$$

Using the definition of $p_{\beta,\sigma_*}(u\,|\,y,x)$ one can verify that $K_{\beta,\sigma_*}(\tilde{u}|u,x) \geq 0$, for all $\tilde{u} \in \mathcal{U}$, and $\int_{\mathcal{U}} K_{\beta,\sigma_*}(\tilde{u}|u,x)d\tilde{u} = \frac{\mathbb{E}_{P(\beta,f_{\sigma_*})}\left[p_{\beta,\sigma_*}(u\,|\,Y,x)\right]}{f_{\sigma_*}(u\,|\,x)} = 1$, almost surely (under $P(\beta, f_{\sigma_*})$) for $u \in \mathcal{U}$ and $x \in \mathcal{X}$. Thus, $K_{\beta,\sigma_*}(\tilde{u}|u,x)$ is a probability density over $\tilde{u} \in \mathcal{U}$, for all u, x. Also using that $\phi(r)$ is convex, we can therefore apply Jensen's inequality to obtain

$$\begin{aligned} d(\widetilde{f}_0, f_{\sigma_*}) &\leq \int_{\mathcal{X}}\int_{\mathcal{U}}\int_{\mathcal{U}} \phi\left(\frac{f_0(\tilde{u}|x)}{f_{\sigma_*}(\tilde{u}\,|\,x)}\right) K_{\beta,\sigma_*}(\tilde{u}|u,x)\,d\tilde{u}\, f_{\sigma_*}(u\,|\,x) f_X(x)\,du\,dx \\ &= \int_{\mathcal{X}}\int_{\mathcal{U}} \phi\left(\frac{f_0(\tilde{u}|x)}{f_{\sigma_*}(\tilde{u}\,|\,x)}\right) \underbrace{\left[\int_{\mathcal{U}} f_{\sigma_*}(u\,|\,x)\, K_{\beta,\sigma_*}(\tilde{u}|u,x)\,du\right]}_{=f_{\sigma_*}(\tilde{u}\,|\,x)} f_X(x)\,d\tilde{u}\,dx \\ &= d(f_0, f_{\sigma_*}) \leq \epsilon. \end{aligned} \tag{A19}$$

Because $\widetilde{f}_0$ satisfies (A18) and (A19) we thus have $\widetilde{f}_0 \in \Gamma_\epsilon$. We have thus shown that for every $f_0 \in \Gamma_\epsilon$ there exists $\widetilde{f}_0 \in \Gamma_\epsilon$ such that (A17) holds. Let $\widetilde{\Gamma}_\epsilon$ be the set of all such $\widetilde{f}_0$ obtained for an $f_0 \in \Gamma_\epsilon$. Since $\widetilde{\Gamma}_\epsilon \subset \Gamma_\epsilon$ we find that

$$\sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\zeta(U,X)\,|\,Y,X\right]\right\}\right| = \sup_{\widetilde{f}_0\in\widetilde{\Gamma}_\epsilon}\left|\mathbb{E}_{P(\beta,\widetilde{f}_0)}\left[\zeta(U,X)\right]\right| \leq \sup_{f_0\in\Gamma_\epsilon}\left|\mathbb{E}_{P(\beta,f_0)}\left[\zeta(U,X)\right]\right|.$$

∎

B Extensions

In this section of the appendix we consider five issues in turn: how to compute posterior average effects when they are not available in closed form, how to estimate quantities of interest that are nonlinear in f_0, whether the constant two appearing in Theorem 2 can be improved upon, how our framework can account for multi-dimensional parameters of interest, and how to construct confidence intervals and develop a specification test.

B.1 Computation

$\widehat{\delta}^{\mathrm{P}}$ can be computed in closed form in simple models, such as all the examples in this paper. However, in complex models such as structural models, the likelihood function or posterior distribution may not be available in closed form. A simple approach in such cases is to proceed by simulation.

Specifically, for all $i = 1, ..., n$ we first draw $U_i^{(s)}$, $s = 1, ..., S$ according to $f_{\widehat{\sigma}}(\cdot \mid X_i)$, and compute $Y_i^{(s)} = g_{\widehat{\beta}}(U_i^{(s)}, X_i)$. Then, we regress $\delta_{\widehat{\beta}}(U_i^{(s)}, X_i)$ on $Y_i^{(s)}$, for $s = 1, ..., S$. Any nonparametric/machine learning regression estimator can be used for this purpose. This procedure requires virtually no additional coding given simulation codes for outcomes and counterfactuals.

B.2 Nonlinear effects

The researcher may be interested in a nonlinear function of f_0. Specifically, here we abstract from covariates X and focus on $\overline{\delta} = \varphi_{\beta}(f_0)$, for some functional φ_{β}. As an example, in the fixed-effects model (1), $\overline{\delta}$ may be the Gini coefficient of α. The analysis in the linear case applies verbatim to this case, since under regularity conditions

$$\varphi_{\beta}(f_0) = \varphi_{\beta}(f_{\sigma_*}) + \nabla\varphi_{\beta}(f_{\sigma_*})[f_0 - f_{\sigma^*}] + o(\epsilon^{\frac{1}{2}}), \tag{B20}$$

which is linear in f_0, up to smaller-order terms. Here $\nabla\varphi_{\beta}$ denotes the gradient of $\varphi_{\beta}(f)$ with respect to f. In Appendix D we report model-based and posterior estimates of Gini coefficients based on simulated data.

B.3 The constant in Theorem 2

The binary choice model that we described in Section 8 is helpful to see that the global bound in Theorem 2, which depends on the constant two, cannot be improved upon in general. To see this, consider model (21) with three simplifications: X consists of a single value, β is known, and $\sigma_* = 1$ is fixed. We assume that $x'\beta > X'\beta$.

In this example, for ϵ large enough the worst-case biases of $\widehat{\delta}^{\mathrm{M}}$ and $\widehat{\delta}^{\mathrm{P}}$ are

$$\mathrm{Bias}_{\mathrm{M}} = \max(\Phi(x'\beta), 1 - \Phi(x'\beta)),$$

and

$$\mathrm{Bias}_{\mathrm{P}} = \frac{\max(\Phi(x'\beta) - \Phi(X'\beta), 1 - \Phi(x'\beta))}{1 - \Phi(X'\beta)},$$

respectively.

From this, we first see that the bias of the posterior estimator is smaller than twice that of the model-based estimator. In addition, taking $X'\beta = 0$ and $x'\beta = \eta$, we have, for small η,

$$\frac{\mathrm{Bias}_{\mathrm{P}}}{\mathrm{Bias}_{\mathrm{M}}} = \frac{2(1 - \Phi(\eta))}{\Phi(\eta)} \overset{\eta \to 0}{\longrightarrow} 2.$$

This shows that two is indeed the smallest possible constant in Theorem 2.

B.4 Multi-dimensional average effects

In the main text, we considered the case where the target parameter $\overline{\delta}$ in (10) is scalar. However, our results can be extended to multi-dimensional parameters. The definition of worst-case bias in (15) is then modified to

$$b_{\epsilon}(\gamma) = \sup_{f_0 \in \Gamma_{\epsilon}} \left\| \mathbb{E}_{P(\beta, f_0)}[\gamma(Y, X) - \delta(U, X)] \right\|,$$

where $\| \cdot \|$ is some norm over the vector space in which $\gamma(Y, X)$ and $\delta(U, X)$ take values.

If $\|\cdot\|_*$ denotes the corresponding dual norm, then we can rewrite $b_{\epsilon}(\gamma) = \sup_{\|v\|_* = 1} b_{\epsilon}(\gamma, v)$, where $b_{\epsilon}(\gamma, v) = \sup_{f_0 \in \Gamma_{\epsilon}} \left| \mathbb{E}_{P(\beta, f_0)}[v'\gamma(Y, X) - v'\delta(U, X)] \right|$. Our optimality results for posterior average effects for scalar $\overline{\delta}$ then apply to $b_{\epsilon}(\gamma, v)$ for every given vector v, and optimality is maintained after taking the supremum over the set of vectors v with $\|v\|_* = 1$. Thus, for the multi-dimensional case, we expect posterior average effects to be locally optimal in the sense of Theorem 1, and globally optimal up to a factor of two as in Theorem 2, although a formal proof of local optimality requires making our ϵ-expansion uniform in v.

In the motivating example in Section 2, suppose we are interested in the entire distribution function F_{α} of α, which is an infinite-dimensional parameter. In this case the average effect is a function indexed by $a \in \mathbb{R}$, and we can choose the supremum norm $\| \cdot \|_{\infty}$ over functions of a. Letting $\delta^{(a)}(U, X) = \mathbf{1}\{\alpha \leq a\}$, the local bias of the posterior average effect then reads

$$b_{\epsilon}(\gamma^{\mathrm{P}}) = \epsilon^{\frac{1}{2}} \left\| \left\{ \frac{2}{\phi''(1)} \mathrm{Var}_* \left(\delta^{(a)}(U, X) - \mathbb{E}_*[\delta^{(a)}(U, X) \,|\, Y, X] \right) \right\}^{\frac{1}{2}} \right\|_{\infty} + \mathcal{O}(\epsilon).$$

B.5 Confidence intervals and specification test

In our theory we solely focus on bias, and do not consider variance or confidence intervals. Under correct specification of f_σ, it is easy to derive the asymptotic distributions of $\widehat{\delta}^{\mathrm{M}}$ and $\widehat{\delta}^{\mathrm{P}}$. Specifically, suppose that $\widehat{\beta}$ and $\widehat{\sigma}$ are asymptotically linear in the sense that, for some mean-zero function h, we have

$$\begin{pmatrix} \widehat{\beta} \\ \widehat{\sigma} \end{pmatrix} = \begin{pmatrix} \beta \\ \sigma_* \end{pmatrix} + \frac{1}{n}\sum_{i=1}^{n} h(Y_i, X_i) + o_P(n^{-\frac{1}{2}}).$$

Then, under standard conditions (e.g., Newey and McFadden, 1994), we have

$$n^{\frac{1}{2}} \begin{pmatrix} \widehat{\delta}^{\mathrm{M}} - \overline{\delta} \\ \widehat{\delta}^{\mathrm{P}} - \overline{\delta} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \tag{B21}$$

Here, $\Sigma_{11} = \mathrm{Var}_* \left(G_1' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, X]\right)$, $\Sigma_{12} = \mathrm{Cov}_* \big(G_1' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, X],$ $G_2' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, Y, X]\big)$, $\Sigma_{21} = \Sigma_{12}$, and $\Sigma_{22} = \mathrm{Var}_* \left(G_2' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, Y, X]\right)$, for $G_1 = \partial_{\beta,\sigma} \mathbb{E}_{\beta,\sigma_*} \left[\delta_\beta(U,X)\right]$ and $G_2 = \mathbb{E}_{\beta,\sigma_*} \left\{\partial_{\beta,\sigma} \mathbb{E}_{p_{\beta,\sigma_*}} \left[\delta_\beta(U,X) \,|\, Y, X\right]\right\}$, where $\partial_\theta g(\theta_1)$ denotes the gradient of $g(\theta)$ at $\theta = \theta_1$. Note that in (B21) we allow δ_β to be non-smooth in β (e.g., an indicator function).

Under local, $(1/\sqrt{n})$-misspecification of f_σ, confidence intervals that account for model uncertainty in addition to sampling uncertainty can be constructed as in Armstrong and Kolesár (2018) and Bonhomme and Weidner (2018). A simple possibility to ensure uniform coverage within an ϵ-neighborhood is to add $b_\epsilon(\gamma)$ on both sides of a standard confidence interval of $\overline{\delta}$. For example, one may construct the 95% interval

$$\left[\widehat{\delta}^{\mathrm{P}} \pm \left(\epsilon^{\frac{1}{2}} \left\{ \frac{2}{\phi''(1)} \mathrm{Var}_* \left(\delta(U,X) - \mathbb{E}_*[\delta(U,X) \,|\, Y, X]\right) \right\}^{\frac{1}{2}} + 1.96 n^{-\frac{1}{2}} \widehat{\Sigma}_{22}^{\frac{1}{2}} \right) \right],$$

for $\widehat{\Sigma}_{22} = \mathrm{Var}_* \left(G_2' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, Y, X]\right)$, where expectations and variances are taken with respect to $P(\widehat{\beta}, f_{\widehat{\sigma}})$, and δ, G_2, and h are evaluated at $\widehat{\beta}$ and $\widehat{\sigma}$. Note that this confidence interval requires setting a value for ϵ. Building on Hansen and Sargent (2008), Bonhomme and Weidner (2018) propose to calibrate ϵ by targeting a model detection error probability.

Specification test. Using the asymptotic distribution of $(\widehat{\delta}^{\mathrm{M}}, \widehat{\delta}^{\mathrm{P}})$ under correct specification of f_σ, we obtain

$$n^{\frac{1}{2}} \left(\widehat{\delta}^{\mathrm{P}} - \widehat{\delta}^{\mathrm{M}} \right) \xrightarrow{d} \mathcal{N}\left(0, \widetilde{\Sigma}\right),$$

where $\widetilde{\Sigma} = \text{Var}_*\left(\mathbb{E}_*[\delta(U,X)\,|\,Y,X] - \mathbb{E}_*[\delta(U,X)\,|\,X] + (G_2 - G_1)'h(Y,X)\right)$. Hence, under correct specification,

$$n\left(\widehat{\delta}^{\text{P}} - \widehat{\delta}^{\text{M}}\right)'\widetilde{\Sigma}^{-1}\left(\widehat{\delta}^{\text{P}} - \widehat{\delta}^{\text{M}}\right) \xrightarrow{d} \chi_1^2.$$

Plugging-in a consistent empirical counterpart for $\widetilde{\Sigma}$ in this expression, we obtain a simple test of correct specification of the parametric distribution f_σ.

C Finite support

In this section of the appendix we consider the case where U has finite support and takes the values $u_1, u_2, ..., u_K$ with probability $\omega_1^0, ..., \omega_K^0$. Here we abstract away from β, σ, and covariates X.

Injective and non-injective models. Let $\delta_k = \delta(u_k)$, and denote $g_k = g(u_k)$ where $Y = g(U)$. Let $\overline{g}_1, ..., \overline{g}_L$ denote the $L \leq K$ equivalence classes of $g_1, ..., g_K$. We will denote as $\ell(k) \in \{1, ..., L\}$ the index corresponding to the equivalence class of g_k, for all k. In addition, let $n_\ell = \sum_{i=1}^n \mathbf{1}\{Y_i = \overline{g}_\ell\}$ for all ℓ, and denote $\omega_k^U = f(u_k)$ for all k.

It is useful to distinguish two cases. When g is *injective*, $K = L$ and $\mathbb{E}_{p(f)}[\delta(U)\,|\,g(U) = g_k] = \delta_k$. So we have $\widehat{\delta}^{\text{P}} = \frac{1}{n}\sum_{k=1}^K n_k \delta_k$. This estimator does not depend on the assumed f. Moreover, as $\min_{k=1,...,K} n_k$ tends to infinity we have

$$\widehat{\delta}^{\text{P}} \xrightarrow{p} \sum_{k=1}^K \omega_k^0 \delta_k = \overline{\delta}.$$

Hence $\widehat{\delta}^{\text{P}}$ is consistent for $\overline{\delta}$, irrespective of the choice of the reference distribution f, provided $\omega_k^U > 0$ for all k.

When g is *not injective*, $K \neq L$ and we have

$$\widehat{\delta}^{\text{P}} = \frac{1}{n}\sum_{i=1}^n \sum_{\ell=1}^L \mathbf{1}\{Y_i = \overline{g}_\ell\}\mathbb{E}_{p(f)}[\delta(U)\,|\,g(U) = \overline{g}_\ell] = \frac{1}{n}\sum_{\ell=1}^L n_\ell \mathbb{E}_{p(f)}[\delta(U)\,|\,g(U) = \overline{g}_\ell].$$

Moreover,

$$\begin{aligned}\mathbb{E}_{p(f)}[\delta(U)\,|\,g(U) = \overline{g}_\ell] &= \textstyle\sum_{k=1}^K \Pr_{p(f)}(U = U_k\,|\,g(U) = \overline{g}_\ell)\delta_k \\ &= \textstyle\sum_{k=1}^K \frac{\omega_k^U \mathbf{1}\{\ell(k)=\ell\}}{\sum_{k'=1}^K \omega_{k'}^U \mathbf{1}\{\ell(k')=\ell\}}\delta_k =: \overline{\delta}_\ell^U.\end{aligned}$$

Hence,

$$\widehat{\delta}^{\text{P}} = \frac{1}{n}\sum_{\ell=1}^L n_\ell \overline{\delta}_\ell^U.$$

Through $\overline{\delta}_\ell^U$, $\widehat{\delta}^{\mathrm{P}}$ depends on the prior ω^U in general, even as $\min_{\ell=1,\dots,L} n_\ell$ tends to infinity.

Bayesian interpretation. From a Bayesian perspective, one may view ω^0 as a parameter, and put a prior on it. A simple conjugate prior specification is a Dirichlet distribution $\omega \sim \mathrm{Dir}(K, \alpha)$, where $\alpha_k > 0$ for $k = 1, \dots, K$. We will focus on the posterior mean

$$\widehat{\delta}^{\mathrm{D}} = \mathbb{E}\left[\sum_{k=1}^{K} \delta_k \omega_k \,|\, Y\right] = \sum_{k=1}^{K} \delta_k \mathbb{E}\left[\omega_k \,|\, Y\right],$$

for a Dirichlet prior with $\alpha_k = M\omega_k^U$ for all k, where $M > 0$ is a constant.

For all ℓ, let $\overline{\alpha}_\ell = \sum_{k=1}^{K} \mathbf{1}\{\ell(k) = \ell\}\alpha_k$, and $\overline{\omega}_\ell = \sum_{k=1}^{K} \mathbf{1}\{\ell(k) = \ell\}\omega_k$. $(\overline{\omega}_1, \dots, \overline{\omega}_L)$ follows the Dirichlet distribution $\mathrm{Dir}(L, \overline{\alpha})$. Moreover, for all k, $\omega_k/\overline{\omega}_{\ell(k)}$ is a component of a Dirichlet distribution with mean $\alpha_k/\overline{\alpha}_{\ell(k)}$.

Unlike the $\overline{\omega}_\ell$'s, the $\omega_k/\overline{\omega}_{\ell(k)}$'s are not updated in light of the data since they do not enter the likelihood. Notice the link with the Bayesian analysis of partially identified models in Moon and Schorfheide (2012): here the $\overline{\omega}_\ell$'s are identified but the ω_k's are not, since for identical g_k's the data provides no information to discriminate across ω_k's.

As a result, we have

$$\begin{aligned}\mathbb{E}[\omega_k \,|\, Y] &= \mathbb{E}\left[\frac{\omega_k}{\overline{\omega}_{\ell(k)}}\overline{\omega}_{\ell(k)} \,|\, Y\right] = \mathbb{E}\left[\frac{\omega_k}{\overline{\omega}_{\ell(k)}}\right]\mathbb{E}\left[\overline{\omega}_{\ell(k)} \,|\, Y\right] \\ &= \frac{\alpha_k}{\overline{\alpha}_{\ell(k)}}\frac{n_\ell + \overline{\alpha}_\ell}{n + M} \overset{M\to 0}{\to} \frac{\omega_k^U}{\sum_{k'=1}^{K}\omega_{k'}^U \mathbf{1}\{\ell(k') = \ell(k)\}}\frac{n_{\ell(k)}}{n}.\end{aligned}$$

It thus follows that

$$\widehat{\delta}^{\mathrm{D}} \overset{M\to 0}{\to} \sum_{k=1}^{K} \delta_k \frac{\omega_k^U}{\sum_{k'=1}^{K}\omega_{k'}^U \mathbf{1}\{\ell(k') = \ell(k)\}}\frac{n_{\ell(k)}}{n} = \widehat{\delta}^{\mathrm{P}}.$$

Hence, under a diffuse Dirichlet prior centered around ω^U, the Bayesian posterior mean coincides with the posterior average effect we focus on in this paper.

D Additional simulations

In this section of the appendix we report the results of two simulation exercises, based on the fixed-effects model (1), and on an ordered choice model.

D.1 Fixed-effects model

Skewness. Let us consider the fixed-effects model (1). Suppose the parameter of interest is the skewness of α

$$\overline{\delta} = \mathbb{E}_{f_0}\left[\alpha^3 - 3\frac{\mu_\alpha}{\sigma_\alpha} - \left(\frac{\mu_\alpha}{\sigma_\alpha}\right)^3\right].$$

For example, it is of interest to estimate the skewnesses of income components and how they evolve over time (Guvenen *et al.*, 2014). Since the normal distribution is symmetric, the model-based normal estimator of skewness is simply $\widehat{\delta}^{\mathrm{M}} = 0$, irrespective of the observations Y_{ij}. Hence, $\widehat{\delta}^{\mathrm{M}}$ is not informed by the data, even when the empirical distribution of the fixed-effects $\overline{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$ indicates strong asymmetry.

By contrast, a posterior average effect based on a normal reference distribution is

$$\widehat{\delta}^{\mathrm{P}} = \frac{1}{\widehat{\sigma}_\alpha^3}\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{p(f_{\widehat{\sigma}})}\left[\alpha^3 \,\middle|\, Y = Y_i\right] - 3\frac{\widehat{\mu}_\alpha}{\widehat{\sigma}_\alpha} - \left(\frac{\widehat{\mu}_\alpha}{\widehat{\sigma}_\alpha}\right)^3.$$

It can be verified that

$$\widehat{\delta}^{\mathrm{P}} = \widehat{\rho}^3 \frac{1}{\widehat{\sigma}_\alpha^3}\frac{1}{n}\sum_{i=1}^{n}\left(\overline{Y}_i - \overline{Y}\right)^3,$$

where $\widehat{\rho} = \frac{\widehat{\sigma}_\alpha^2}{\widehat{\sigma}_\alpha^2 + \widehat{\sigma}_\varepsilon^2/J}$. Under mild conditions, and in contrast with $\widehat{\delta}^{\mathrm{M}}$, the posterior estimator $\widehat{\delta}^{\mathrm{P}}$ is consistent for the true skewness of α as J tends to infinity. However, $\widehat{\delta}^{\mathrm{P}}$ is biased for small J in general.

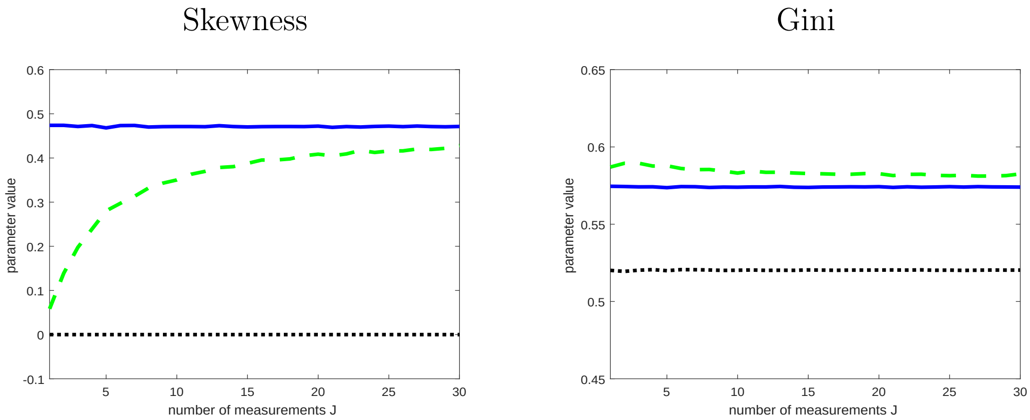
To provide intuition about the magnitude of the bias, we simulate data where all latent components are independent, ε_j are standard normal, and α follows a skew-normal distribution (e.g., Azzalini, 2013) with zero mean, variance 1, and skewness $\approx .47$ corresponding to the skew-normal parameter $\delta = .99$. We take $n = 1000$, and run 100 simulations varying J from 1 to 30. We estimate means and variances using minimum-distance based on first and second moment restrictions.

In the left panel of Figure D1 we show the results. We see that the model-based estimator is equal to zero irrespective of the number J of individual measurements. By contrast, the posterior estimator converges to the true skewness of α as J increases, although it is substantially biased for small J.

Gini coefficient. We next focus on the Gini coefficient of α:

$$G = \frac{1}{2\mathbb{E}_{f_0}[\exp(\alpha)]}\iint |\exp(\alpha') - \exp(\alpha)| f_0(\alpha) f_0(\alpha') d\alpha d\alpha'.$$

Figure D1: Skewness and Gini estimates in the fixed-effects model



Notes: true (solid), posterior (dashed), model-based (dotted). $n = 1000$, 100 *simulations.*

In this case, a model-based estimator is

$$\widehat{G}^{\mathrm{M}} = 2\Phi(\widehat{\sigma}_{\alpha}/\sqrt{2}) - 1,$$

while a posterior average effect is, following (B20),

$$\widehat{G}^{\mathrm{P}} = \widehat{G}^{\mathrm{M}} + \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\nabla\widehat{G}(\alpha)\,|\,Y_i] - \mathbb{E}[\nabla\widehat{G}(\alpha)]\right),$$

where

$$\nabla\widehat{G}(\alpha) = -\exp\left(\alpha - \widehat{\mu}_{\alpha} - \frac{1}{2}\widehat{\sigma}_{\alpha}^{2}\right)\left(\widehat{G}^{\mathrm{M}} + 1 - 2\Phi\left(\frac{\alpha - \widehat{\mu}_{\alpha}}{\widehat{\sigma}_{\alpha}}\right)\right) + \left(1 - 2\Phi\left(\frac{\alpha - \widehat{\mu}_{\alpha}}{\widehat{\sigma}_{\alpha}} - \widehat{\sigma}_{\alpha}\right)\right).$$

In the right panel of Figure D1 we show the simulation results. We see that in this case also the model-based estimator is insensitive to J. The posterior estimator has a lower bias, especially for larger J.

D.2 Ordered choice model

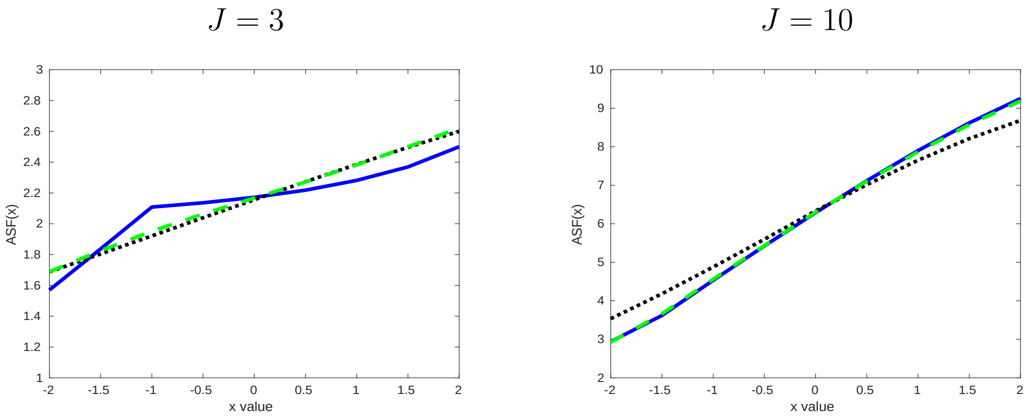
We next consider the ordered choice model

$$Y_i = \sum_{j=1}^{J} j\mathbf{1}\{\mu_{j-1} \leq Y_i^* \leq \mu_j\}, \text{ where } Y_i^* = X_i'\beta + U_i,$$

for a sequence of *known* thresholds $-\infty = \mu_0 < \mu_1 < ... < \mu_{J-1} < \mu_J = +\infty$. This model may be of interest to analyze data on wealth or income, say, where only a bracket containing the true observation is recorded. We focus on the average structural function

$$\overline{\delta}(x) = \mathbb{E}_{f_0}\left[\sum_{j=1}^{J} j\mathbf{1}\{\mu_{j-1} \leq x'\beta + U \leq \mu_j\}\right].$$

Figure D2: Average structural function in the ordered choice model



Notes: true (solid), posterior (dashed), model-based (dotted). $n = 1000$, 100 *simulations.*

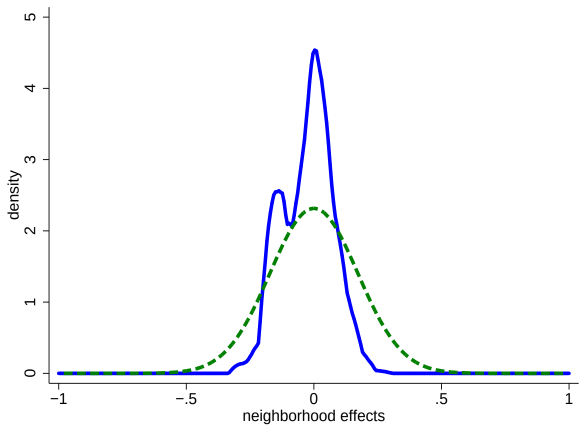
We take as reference distribution $U \mid X \sim \mathcal{N}(0, \sigma^2)$. In the simulated data generating process, U is independent of X, distributed as a re-centered χ^2 with mean zero and variance one. We simulate a scalar standard normal X. We set $n = 1000$, $\beta_1 = .5$, $\beta_0 = 0$, $\sigma = 1$, and μ as uniformly distributed between -2 and 2. We estimate β up to scale using maximum score (Manski, 1985).[15] For computation of maximum score, we use the mixed integer linear programming algorithm of Florios and Skouras (2008).

In Figure D2 we report the results for $J = 3$ (left) and $J = 10$ (right). We see that, when $J = 3$, model-based and posterior estimators are similarly biased. By contrast, when $J = 10$, the posterior estimator aligns well with the true average structural function, even though the model-based estimator is substantially biased.

[15]Specifically, using maximum score we regress $\mathbf{1}\{Y_i \leq j\}$ on X_i and a constant, for all j, imposing that the coefficient of X_i is one. We then regress the J estimates on a common constant and the μ_j, and obtain the implied estimate for β by rescaling.

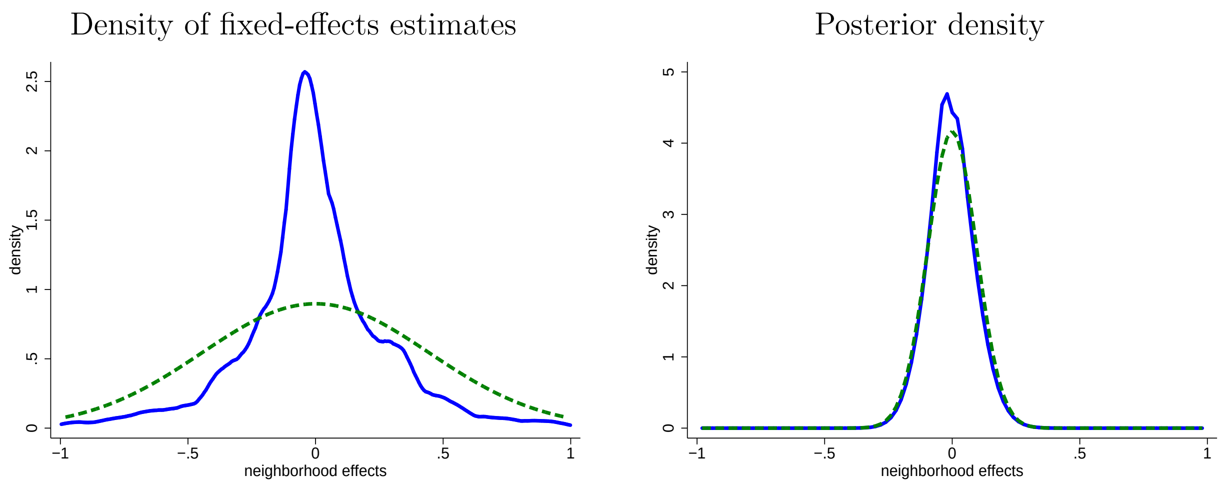
E Additional empirical results

Figure E3: Density of posterior means of neighborhood effects



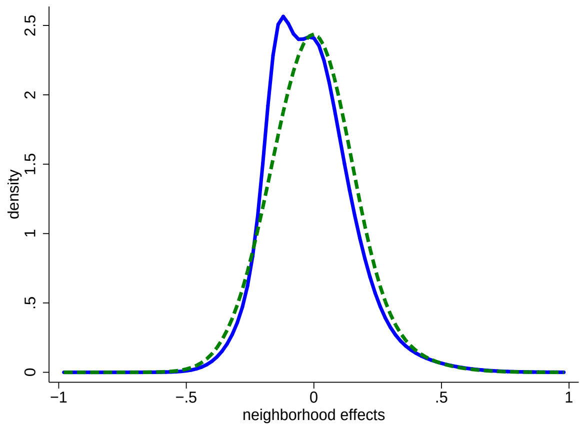
Notes: Density of posterior means of μ_c (solid) and prior density (dashed). Calculations are based on statistics available on the Equality of Opportunity website.

Figure E4: Density of neighborhood effects at the county level



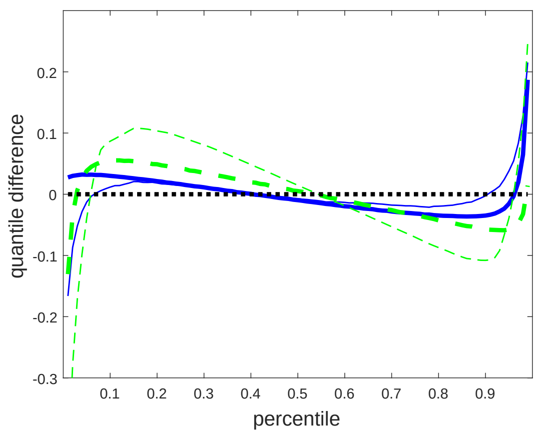
Notes: In the left graph we show the density of fixed-effects estimates $\widehat{\mu}_c^{\text{county}}$ (solid) and normal fit (dashed). In the right graph we show the posterior density of μ_c^{county} (solid) and prior density (dashed). Calculations are based on statistics available on the Equality of Opportunity website.

Figure E5: Posterior density of neighborhood effects, correlated random-effects specification



Notes: Posterior density of μ_c *(solid) and prior density (dashed), based on a correlated random-effects specification allowing for correlation between the place effects* μ_c *and the mean income of permanent residents* $\overline{y}_c$*. Calculations are based on statistics available on the Equality of Opportunity website.*

Figure E6: Quantiles of income components, comparison to Arellano *et al.* (2017)



Notes: The graph shows quantile differences between posterior and model-based estimators in thick font, and estimates from Arellano et al. (2017) in thinner font. η_{it} *is shown in solid and* ε_{it} *is shown in dashed. Sample from the PSID.*