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 CWP49/20

UKRI Economic and Social Research Council

Posterior Average Effects*

Stéphane Bonhomme† Martin Weidner‡

October 6, 2020

Abstract

Economists are often interested in estimating averages with respect to distributions of unobservables. Examples are moments of individual fixed-effects, average partial effects in discrete choice models, and counterfactual simulations in structural models. For such quantities, we propose and study posterior average effects (PAE), 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 show that PAE have minimum worst-case bias under local misspecification of the parametric distribution of unobservables. This provides a rationale for reporting these estimators in applications. We introduce a measure of informativeness of the posterior conditioning, which quantifies the bias of PAE relative to parametric model-based estimators, and we study other robustness properties of PAE for estimation and prediction. As illustrations, we report PAE estimates of 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. This includes average partial effects in discrete choice models, moments of individual fixed-effects in panel data, and average welfare effects in structural models, all of which are expectations with respect to some distribution of shocks and heterogeneity. The standard approach in applied work is to assume a parametric form for the distribution of unobservables and to compute the average effect under that assumption. For example, in binary choice, researchers often assume normality of the error term, and compute average partial effects under normality. This "model-based" estimation of average effects is justified under the assumption that the parametric model is *correctly specified*.

In this paper, we consider a different approach, where the average effect is computed *conditional on the observation sample*. We refer to such estimators as "posterior average effects" (PAE). Posterior averaging is appealing for prediction purposes, and it plays a central role in Bayesian and empirical Bayes approaches (e.g., Berger, 1980, Morris, 1983). Here we focus instead 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 when and why posterior conditioning may be useful in estimation. Our main result will show that PAE have robustness properties when the parametric model is *misspecified*.

PAE are closely related to empirical Bayes (EB) estimators, which are increasingly popular in applied economics. Consider a fixed-effects model of teacher quality, which is 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 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, holding fixed the values of the mean and variance parameters. 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 them 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 function of teacher quality evaluated at a point. Since this quantity is an average of indicator functions, the PAE

is simply an average of posterior means — that is, of EB estimates — *of the indicator functions*. This estimator is available in closed form. However, the PAE differs from the empirical distribution of the EB estimates of teacher effects. In particular, while the variance of EB estimates is too small relative to that of latent teacher quality, the PAE has the correct variance. Related applications of PAE include settings involving neighborhood/place effects (Chetty and Hendren, 2017, Finkelstein *et al.*, 2017) or hospital quality (Hull, 2018).

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 using PAE can be beneficial.

To study 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. Throughout the paper, we often simply use *bias* to denote worst-case asymptotic bias in such a neighborhood. In our main theorem, we show that PAE have *minimum local bias* — calculated in an asymptotic where the size of the neighborhood tends to zero — within a large class of estimators. The theorem implies that PAE are least sensitive to small departures from correct specification, and that other estimators will generally have larger bias under local misspecification.

In our examples and illustrations, we find that the information contained in the posterior conditioning is setting-specific. This is intuitive, since although PAE have minimum bias locally, the bias is not zero in general and varies across applications. PAE tend to 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). Consistently with this intuition, our bias analysis suggests quantifying the "informativeness" of the posterior conditioning using an easily computable R^2 coefficient.

While PAE have minimum local bias, they do not have minimum mean squared error in

general. Indeed, in small samples where variance dominates bias, model-based estimators can have smaller mean squared error than PAE. Hence, PAE are best suited for large samples — e.g., when the number of teachers is large. Although one can compute estimators that minimize mean squared error locally, those depend on neighborhood size. An important practical advantage of PAE is that they do *not* require taking a stand on the degree of misspecification through the size of the neighborhood, and they are simple to implement.

To illustrate the scope of PAE for applications, we consider two empirical settings. In the first one, we study the estimation of neighborhood/place effects in the US. Chetty and Hendren (2017) report estimates of the variance of neighborhood effects, as well as EB estimates of those effects. Our goal 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, PAE inform our knowledge of the density of neighborhood effects, and motivate future analyses using more flexible model specifications and individual-level data.

In the 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), where log-income is the sum of a random-walk component and a component that is independent over time. Researchers often estimate the covariance structure of the latent components in a first step. Then, in order to document distributions or to use the income process in a consumption-saving model, they often assume Gaussianity. However, there is increasing evidence that income components are not Gaussian (e.g., Geweke and Keane, 2000, Hirano, 2002, Bonhomme and Robin, 2010, Guvenen *et al.*, 2016). We estimate posterior distribution functions and quantiles of permanent and transitory income components using recent waves from the Panel Study of Income Dynamics (PSID). PAE reveal that both components are non-normal, especially the transitory one.

We analyze several extensions. First, we describe the form of PAE in several models, including binary choice and censored regression, and we illustrate in simulations that PAE can perform substantially better than model-based estimators under misspecification. Second, we discuss how to construct confidence intervals and specification tests based on PAE. Third,

we study the bias properties of PAE under *non-local* misspecification. This complements our main result, which is based on a local asymptotic approach. Specifically, in neighborhoods of fixed size we show that the worst-case bias of PAE is at most twice the minimum bias achievable. Lastly, we revisit the question of optimality of EB estimates for *predicting* individual parameters. By extending our misspecification analysis from worst-case bias of sample averages to worst-case mean squared prediction error, we show that EB estimators remain optimal, up to small-order terms, under local deviations from normality.

Related literature and outline. PAE are closely related to parametric EB estimators (Efron and Morris, 1973, Morris, 1983). For recent econometric applications of shrinkage methods (James and Stein, 1961, Efron, 2012), see Hansen (2016), Fessler and Kasy (2018), and Abadie and Kasy (2018). Recent contributions to nonparametric EB methods are Koenker and Mizera (2014) and Ignatiadis and Wager (2019). Unlike nonparametric EB, 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.[1]

In panel data settings, Arellano and Bonhomme (2009) study the bias of random-effects estimators of averages of functions of covariates and individual effects. They show that, when the distribution of individual effects is misspecified whereas the other features of the model are correctly specified, PAE are consistent as n and T tend to infinity. By contrast, in our setup, only n tends to infinity, and misspecification may affect the entire joint distribution of unobservables. Our analysis also connects to the literature on robustness to model misspecification (e.g., Huber and Ronchetti, 2009, Andrews *et al.*, 2017, 2018, Armstrong and Kolesár, 2018, Bonhomme and Weidner, 2018, Christensen and Connault, 2019). Here our aim is to propose and justify a class of simple, practical 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 present our framework and derive our main theoretical result. In Section 4 we illustrate the use of PAE in two empirical settings. In Section 5 we describe several extensions. Finally, we conclude in Section 6.

[1] Berger (1979) provides a gamma-minimax characterization of Bayes estimators in ϵ-contaminated neighborhoods.

2 Motivating example: a fixed-effects model

To motivate the analysis, 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. There are n teachers and J observations per teacher. 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 with n, $\overline{Y}_i$ 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.[2]

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

[2]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).

covariates, for example. Under the assumption that all components are uncorrelated, μ_α, σ^2_α and σ^2_ε can be consistently estimated using quasi-maximum likelihood or minimum distance based on mean and covariance restrictions.[3]

Given estimates $\widehat{\mu}_\alpha$, $\widehat{\sigma}^2_\alpha$, $\widehat{\sigma}^2_\varepsilon$, we can compute empirical Bayes (EB) 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 $\widehat{\mu}_\alpha$, $\widehat{\sigma}^2_\alpha$, $\widehat{\sigma}^2_\varepsilon$ fixed, and $\widehat{\rho} = \frac{\widehat{\sigma}^2_\alpha}{\widehat{\sigma}^2_\alpha + \widehat{\sigma}^2_\varepsilon/J}$ is a shrinkage factor. Here, Y_i are vectors containing all Y_{ij}, $j = 1, ..., J$. The EB estimates in (2) are well-justified as predictors of the α_i, since (when treating $\widehat{\mu}_\alpha$, $\widehat{\sigma}^2_\alpha$, $\widehat{\sigma}^2_\varepsilon$ 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 EB estimates in order to estimate our target quantity $F_\alpha(a)$. A possible estimator is

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

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

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

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

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

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

where Φ denotes the distribution function of the standard normal. $\widehat{F}^{\mathrm{P}}_\alpha(a)$ is an example of a *posterior average effect* (PAE). One can check that it is consistent for any fixed J when the

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

distribution of $\alpha, \varepsilon_1, ..., \varepsilon_J$ is normal. Under non-normality, $\widehat{F}^{\mathrm{P}}_{\alpha}(a)$ is consistent as J tends to infinity with n, although it is generally biased for small J.[4] Moreover, the mean and variance of $\widehat{F}^{\mathrm{P}}_{\alpha}$ are $(1-\widehat{\rho})\widehat{\mu}_{\alpha} + \widehat{\rho}\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i$ and $(1-\widehat{\rho})\widehat{\sigma}^2_{\alpha} + \widehat{\rho}^2\left[\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i^2 - (\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i)^2\right]$, respectively, which are consistent for μ_{α} and σ^2_{α} for any J.

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

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

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, $\widehat{F}^{\mathrm{M}}_{\alpha}(a)$ is consistent for any fixed J, and it is efficient when $\widehat{\mu}_{\alpha}$ and $\widehat{\sigma}^2_{\alpha}$ are maximum likelihood estimates. Moreover, the mean and variance of $\widehat{F}^{\mathrm{M}}_{\alpha}$ are $\widehat{\mu}_{\alpha}$ and $\widehat{\sigma}^2_{\alpha}$, which are consistent irrespective of normality. However, in contrast to the other estimators above, when $\alpha, \varepsilon_1, ..., \varepsilon_J$ is *not* normally distributed $\widehat{F}^{\mathrm{M}}_{\alpha}(a)$ is generally inconsistent for $F_{\alpha}(a)$ as J tends to infinity. The inconsistency arises from the fact that $\widehat{F}^{\mathrm{M}}_{\alpha}(a)$ only depends on the data through the mean $\widehat{\mu}_{\alpha}$ and the variance $\widehat{\sigma}^2_{\alpha}$. In particular, $\widehat{F}^{\mathrm{M}}_{\alpha}$ 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. Our framework allows for misspecification of the normal distribution of $\alpha, \varepsilon_1, ..., \varepsilon_J$. We show that the PAE $\widehat{F}^{\mathrm{P}}_{\alpha}(a)$ has minimum worst-case bias under local misspecification — i.e., in a small neighborhood around the normal reference distribution. To our knowledge, unlike the other three estimators above, posterior estimators of distributions are novel to practitioners. They are also straightforward to implement. Our characterization provides a justification for reporting them in applications.

Note that one may wish to relax normality 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). While these estimators may be consistent even when α is not normal, consistency relies on additional restrictions on the model. For example, the assumptions in Kotlarski (1967) require that α, ε_1, ..., ε_J be mutually *independent*. By contrast, we do *not* impose any such additional conditions in our framework. In Section 3, we will show that

[4] Consistency of $\widehat{F}^{\mathrm{P}}_{\alpha}(a)$ 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 $\mathbf{1}\{\alpha_i \leq a\}$.

asymptotically linear estimators have larger local asymptotic bias than PAE under the form of misspecification that we consider.[5]

To illustrate that an independence assumption among α, ε_1, ..., ε_J can be restrictive, consider a situation where the researcher is concerned that the variance of ε_j depends on α. For instance, the variance of classroom-level shocks may depend on teacher quality. The presence of such conditional heteroskedasticity would invalidate conventional nonparametric deconvolution estimators. By contrast, we will show that $\widehat{F}_{\alpha}^{\mathrm{P}}(a)$ has minimum bias in local neighborhoods of distributions that allow for conditional heteroskedasticity, and more generally for any joint distributions of $(\alpha, \varepsilon_1, ..., \varepsilon_J)$ with given means and variances.

In model (1), the researcher may be interested in estimating other quantities. 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 EB 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 PAE 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. In Section 4 we apply PAE to model (1) and estimate the distribution of neighborhood/place effects in the US (Chetty and Hendren, 2017). In addition, we show that the permanent-transitory model of income dynamics (e.g., Hall and Mishkin, 1982) has

[5] In our framework, we will focus on nonparametric neighborhoods around a parametric reference model (e.g., a normal density). It would be interesting to consider nonparametric reference models, and analyze the properties of posterior estimators in such settings, although this exceeds the scope of this paper.

a structure similar to model (1), and we report PAE estimates in this context. In Appendix S5 we report simulation results for PAE estimators of the skewness and Gini coefficient of teacher effects in model (1). Lastly, in other models — such as static or dynamic discrete choice models and models with censored outcomes — our results motivate the use of PAE as complements to other estimators that researchers commonly report, and we provide examples in Section 5 and analyze them in Appendix S5.

3 Framework and main result

In this section we describe our framework to study PAE, and we present and discuss our local bias characterization.

3.1 Model-based estimators and PAE

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 density of $U \,|\, X$. The expectation is taken with respect to the product $f_0 f_X$, where f_X is the marginal density of X. For conciseness we leave the dependence on f_X implicit.[6]

While the researcher does not know the true f_0, she has a reference parametric density f_σ for $U \,|\, X$, 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

[6] We focus on a scalar δ_β, but our results continue to hold in the vector-valued case, as we show in Appendix S4. This extension is useful to show that our results apply to distribution *functions* and, by inversion, to quantile functions. Moreover, although our focus is on average effects that depend linearly on f_0, in Appendix S4 we also discuss how to estimate quantities that depend on f_0 nonlinearly.

true densities 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 discrete choice and censored regression models, as we discuss in Section 5 and Appendix S5. 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 data.

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

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

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

To construct a posterior estimator, consider the posterior density $p_{\beta,\sigma}$ of $U \,|\, Y, X$. This posterior density is computed using Bayes rule, based on the prior f_σ on $U \,|\, X$ and the likelihood of $Y \,|\, U, X$ 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,

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

We will compute $p_{\beta,\sigma}$ analytically in all our examples. In Appendix S4 we describe a simulation-based computational approach when an analytical expression is not available. We define the *posterior average effect* (PAE) as the posterior estimator

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

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

From a Bayesian perspective, $\widehat{\overline{\delta}}^{\mathrm{P}}$ is a natural estimator to consider when β and σ are known. Indeed, $\widehat{\overline{\delta}}^{\mathrm{P}}$ 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.[7] 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. In the remainder of this section we formalize this intuition in a framework that accounts for misspecification bias.

Model (1) in the notation of this section. To map the fixed-effects model (1) to the general notation, note that in this case 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 $\beta = (\mu_\alpha, \sigma^2_\alpha, \sigma^2_\varepsilon)'$. 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.

3.2 Neighborhoods, estimators, and worst-case bias

Let $P(\beta, f_0)$ denote the true density of (Y, U, X), where as before we omit the reference to the marginal density 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 $\sum_{i=1}^n \psi_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i) = 0$.[8] Given a distance measure

[7] $\widehat{\delta}^{\mathrm{P}}$ 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 nonparametric prior on f_0, and computing the posterior mean of $\overline{\delta}$ under this prior. We discuss this interpretation formally in Appendix S4, in the case where U has finite support.

[8] Throughout we take the estimators $\widehat{\beta}$ and $\widehat{\sigma}$, and the moment function ψ, as given. In particular, we do not address the question of optimal estimation of β under misspecification.

d and a scalar $\epsilon \geq 0$, we define the following *neighborhood* of the reference density 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 densities 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 density f_σ, whereas $\epsilon > 0$ corresponds to misspecification.

For ease of notation we omit the dependence of Γ_ϵ on β, σ_*, and ψ, all of which we consider fixed and given in this section. Indeed, we assume that the researcher has chosen the estimators $\widehat{\beta}$ and $\widehat{\sigma}$ — our theory is silent about where this choice comes from — and that she has already observed their realized values in a large sample. The moment function ψ is determined by this choice of estimators. Moreover, in large samples, the true parameter value β and the pseudo-true value σ_* are arbitrarily close to the observed values $\widehat{\beta}$ and $\widehat{\sigma}$. In our setup, we only consider densities of unobservables f_0 that are consistent with those values, in the sense that the moment restriction $\mathbb{E}_{P(\beta,f_0)}[\psi_{\beta,\sigma_*}(Y,X)] = 0$ holds. This large-sample logic is consistent with our focus on asymptotic bias as a measure of performance. In the fixed-effects model (1) of teacher quality, this logic is best suited to settings where the number n of teachers is large.[9]

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 non-stochastic

[9]Note that the same logic might suggest imposing that other features of the joint population distribution of the data (Y, X), such as means, covariances, higher-order moments, or even the entire distribution, be kept constant for all $f_0 \in \Gamma_\epsilon$. Restricting neighborhoods in this way does not affect the results in the next subsection because those are valid for all possible ψ, and one could thus impose additional moment restrictions on f_0.

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 $\widehat{\delta}_\gamma$ and focus on worst-case bias. Alternatively, one could minimize the worst-case mean squared error of $\widehat{\delta}_\gamma$ as in Bonhomme and Weidner (2018), 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 PAE they would generally depend on ϵ, as we will discuss in Remark 1 below. Relative to such estimators, PAE have the practical advantage that they do not require the researcher to take a stand on the degree of misspecification ϵ. On the other hand, since PAE minimize bias they are best suited for settings with a large number of cross-sectional units.

3.3 Local bias characterization

Before stating our main result, we first characterize the worst-case bias $b_\epsilon(\gamma)$ of estimators $\widehat{\delta}_\gamma$ for small ϵ. The following lemma is instrumental in proving that PAE minimize local bias. For conciseness, in this subsection 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 $P(\beta, f_{\sigma_*})$. The 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

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

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

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, in Lemma 1 we only explicitly present the conditions 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. To accommodate situations where supports are unbounded such as the example of Section 2, 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 that the small-ϵ bias of the PAE is, up to smaller-order terms, proportional to the within-(Y,X) standard deviation of $\delta(U,X)$ under the reference model:[10]

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

For example, in the fixed-effects model (1) of teacher quality the bias of the PAE $\widehat{F}^{\mathrm{P}}_\alpha(a)$ is

$$b_\epsilon(\gamma^{\mathrm{P}}) = \epsilon^{\frac{1}{2}} \left\{ \frac{4}{\phi''(1)} T\left(\frac{a - \mu_\alpha}{\sigma_\alpha}, \sqrt{\frac{1-\rho}{1+\rho}} \right) \right\}^{\frac{1}{2}} + \mathcal{O}(\epsilon),$$

where $T(a,b) = \varphi(a) \int_0^b \frac{\varphi(az)}{1+z^2} dz$ is Owen's T function (Owen, 1956), and φ is the standard normal density. The bias decreases as the number J of observations per teacher increases, and tends to zero as J tends to infinity and the shrinkage factor ρ tends to one.

The next theorem, which is the main result of this section, shows that the PAE has minimum worst-case bias locally.

[10] Similar expressions appear in Bayesian statistics when computing derivatives of posterior quantities with respect to prior densities; see, e.g., Gustafson (2000).

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

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

Theorem 1 provides a rationale for using PAE in applications.[11] For example, in the fixed-effects model (1), it motivates using the posterior distribution estimator $\widehat{F}^{\mathrm{P}}_\alpha(a)$ given by (4). We report PAE estimators of distributions and illustrate their usefulness in two empirical settings in Section 4. In addition, in Appendix S5 we show the results of Monte Carlo simulations for two estimators in model (1). In the simulations, the normal reference model is misspecified, and we find that PAE provide substantial bias reduction relative to parametric model-based estimators (see Appendix Figure S1).

Remark 1. ***(mean squared error)*** *While we have shown that PAE minimize worst-case bias locally, they generally do not have minimum mean squared error (MSE). To see this, let us assume that β and σ_* are known. In a local asymptotic framework where $n\epsilon$ tends to a constant and under suitable regularity conditions, we show in Appendix S4 that the estimator with minimum worst-case MSE is given by*

$$\widehat{\delta}^{\mathrm{MMSE}} = [1 - w_{n\epsilon}]\ \widehat{\delta}^{\mathrm{M}} + w_{n\epsilon}\,\widehat{\delta}^{\mathrm{P}}, \qquad\qquad w_{n\epsilon} := \left(1 + \frac{\phi''(1)}{2n\epsilon}\right)^{-1}, \tag{17}$$

which is a linear combination between the model-based estimator and the PAE. The model-based estimator $\widehat{\delta}^{\mathrm{M}}$, which has the smallest asymptotic variance, will be preferred when ϵ is small relative to $1/n$, while the PAE, which has smallest asymptotic bias, will be preferred when ϵ is large relative to $1/n$. However, in order to implement such estimators $\widehat{\delta}^{\mathrm{MMSE}}$ that minimize worst-case MSE, knowledge of ϵ is required. See Bonhomme and Weidner (2018) for an approach to minimum-MSE estimation.

Remark 2. ***(uniqueness)*** *In the absence of covariates and for known parameters β, σ_*, the proof of Theorem 1 shows that γ^{P} is the unique minimizer of the first order worst-case*

[11] To provide an intuition for the theorem, 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 PAE $\widehat{\delta}^{\mathrm{P}} = \frac{1}{n}\sum_{i=1}^n \gamma^{\mathrm{P}}_{\widehat{\beta},\widehat{\sigma}}(Y_i, X_i)$ minimizes the first-order contribution to the worst-case bias.

bias. More generally, if covariates are present and the parameters β, σ_* are estimated, then the leading order contribution of $b_\epsilon(\gamma)$ is minimized if and only if $\gamma(Y,X) = \gamma^{\mathrm{P}}(Y,X) + \omega(X) + \lambda'\psi(Y,X) + o_{P_*}(1)$, for some λ and ω such that $\mathbb{E}_{f_X}[\omega(X)] = 0$ — see part (ii) of Theorem A1 in Appendix A for a formal statement. Hence, while the PAE is not the unique minimizer of worst-case bias in this case, any bias-minimizing estimator differs from the PAE by a zero-mean function of X and a linear combination of the moment function ψ. In addition, since $\gamma^{\mathrm{P}}(Y,X)$ is orthogonal to $\omega(X) + \lambda'\psi(Y,X)$, $\widehat{\delta}^{\mathrm{P}}$ has smallest variance within the class of minimum bias estimators.[12]

Remark 3. ***(form of misspecification)*** *Theorem 1 is based on nonparametric neighborhoods that consist of unrestricted distributions of $U \mid X$, except for the moment conditions that pin down β and σ_*. However, if one is willing to make additional assumptions on f_0 that further restrict 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, 1998). However, the PAE in (4) is biased for small J. As a consequence, the PAE does not minimize local bias 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 $\overline{\delta}$.[13] *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 PAE $\widehat{\delta}^{\mathrm{P}}$ 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*

[12] This is closely related to Remark 1 and the corresponding derivation of equation (17) in Appendix S4, which show that the PAE estimator $\widehat{\delta}^{\mathrm{P}}$ is obtained from a worst-case MSE problem in the limit where $n \to \infty$ and $n\epsilon \to 0$.

[13] 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.

independent of ε_j, then it is natural to report the consistent estimator $\widehat{\delta}^{\mathrm{FE}}$.

Remark 4. ***(posterior informativeness)*** *Our bias calculations can be used to compare the bias of the PAE $\widehat{\delta}^{\mathrm{P}}$ 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, the ratio of the two worst-case biases satisfies*

$$\lim_{\epsilon \to 0} \frac{b_\epsilon(\gamma^{\mathrm{P}})}{b_\epsilon(\gamma^{\mathrm{M}})} = \frac{\left\{\mathrm{Var}_* \left(v(U,X) - \mathbb{E}_*[v(U,X) \,|\, Y,X]\right)\right\}^{\frac{1}{2}}}{\left\{\mathrm{Var}_* \left(v(U,X)\right)\right\}^{\frac{1}{2}}}, \tag{18}$$

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.[14] *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,

$$R^2 = \frac{\mathrm{Var}_* \left(\mathbb{E}_*[v(U,X) \,|\, Y,X]\right)}{\mathrm{Var}_* \left(v(U,X)\right)}, \tag{19}$$

where with some 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 refer to R^2 in (19) as a measure of the "informativeness" of the posterior conditioning, and we will report it in our illustrations. As an example, for $\widehat{F}^{\mathrm{P}}_\alpha(a)$ in model (1), the informativeness of the posterior conditioning is

$$R^2 = 1 - \frac{2T\left(\frac{a-\widehat{\mu}_\alpha}{\widehat{\sigma}_\alpha}, \sqrt{\frac{1-\widehat{\rho}}{1+\widehat{\rho}}}\right)}{\Phi\left(\frac{a-\widehat{\mu}_\alpha}{\widehat{\sigma}_\alpha}\right)\left[1 - \Phi\left(\frac{a-\widehat{\mu}_\alpha}{\widehat{\sigma}_\alpha}\right)\right]}. \tag{20}$$

In this case the R^2 increases with the number J of observations per teacher, and it tends to one as J tends to infinity.

4 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

[14] 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 $\gamma = \gamma^{\mathrm{M}}$.

study a permanent-transitory model of income dynamics (Hall and Mishkin, 1982, Blundell *et al.*, 2008) using the PSID. 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.

4.1 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 $\widehat{\sigma}^2_c$ of the variances of $\overline{\varepsilon}_c$ for every c. When weighted by population, the fixed-effects estimates $\widehat{\mu}_c$ have mean zero. We treat neighborhoods as independent observations.[15]

We first estimate the variance of place effects μ_c, following CH. We trim the top 1% percentile of $\widehat{\sigma}^2_c$, and weigh all results by population weights.[16] 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 $\widehat{\sigma}^2_c$ weighted by population is $\widehat{\sigma}^2_{\overline{\varepsilon}} = .047$. Given those, we estimate the variance of place effects as $\widehat{\sigma}^2_\mu = \text{Var}(\widehat{\mu}_c) - \widehat{\sigma}^2_{\overline{\varepsilon}} = .030$. In this setting, the shrinkage factor $\widehat{\rho}_c = \widehat{\sigma}^2_\mu / (\widehat{\sigma}^2_\mu + \widehat{\sigma}^2_c)$ 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.[17]

We use a normal with zero mean and variance $\widehat{\sigma}^2_\mu$ as a prior for μ_c. Then, we estimate

[15] The statistics we use for calculations are available on the Equality of Opportunity website; see https://opportunityinsights.org/paper/neighborhoodsii/ Given the aggregate data at hand, we necessarily need to assume that estimates $\widehat{\mu}_c$ are independent across neighborhoods c, although this might be restrictive in this setting.

[16] This differs slightly from CH's approach, which is based on $1/\widehat{\sigma}^2_c$ precision weights and no trimming. We replicated the analysis using precision weights in the un-trimmed sample and found similar results.

[17] It is quantitatively important to account for this heterogeneity. In our initial work on the data we found that imposing a constant shrinkage factor reduced the informativeness of the posterior conditioning.

the density of neighborhood effects μ_c, using the derivative of the posterior estimator of the distribution function (4); that is,

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

where π_c are population weights. In this subsection, in order to ease the visualization of the results, we will show estimates of densities, which are the derivatives of the PAE of distribution functions. With some abuse of terminology we will refer to those as "posterior densities".[18]

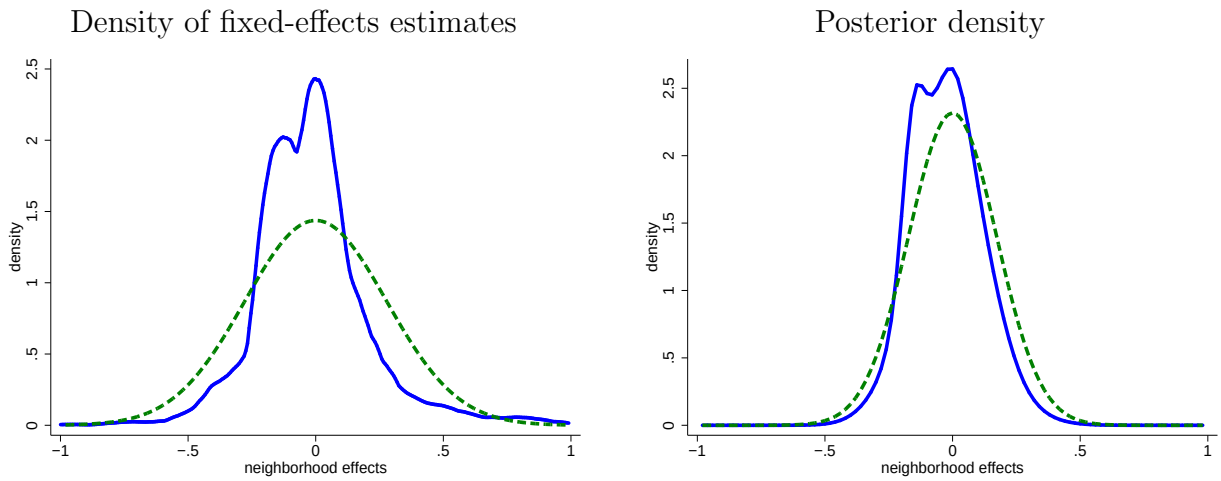
In Figure 1 we report several density estimates. In the left graph, we show a nonparametric kernel density estimate of the fixed-effects $\widehat{\mu}_c$, weighted by population (in solid), together with its best-fitting normal (in dashed). The graph shows substantial non-normality of the fixed-effects estimates. In particular, the large variance appears to be driven by some large positive and negative estimates $\widehat{\mu}_c$. In the right graph we report the posterior density $\widehat{f}_{\mu}^{\mathrm{P}}$ 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). The posterior density of neighborhood effects differs from the normal prior, although the two estimators have the same variance by construction.[19] In addition, a specification test that compares model-based estimator and PAE, which we describe in Appendix S4, suggests that these differences are statistically significant. Indeed, assuming independence across 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 $c = 1, ..., C_{\mathrm{sim}}$, 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

[18] Our theory extends to the multivariate case and it applies in particular to distribution functions (see Appendix S4). In addition, note that the density of μ at a can be approximated for arbitrarily small $h > 0$ by the expectation of $\mathbf{1}\{|\mu - a|/h\}/2h$. Taking the limit of the corresponding PAE as h tends to zero gives $\widehat{f}_{\mu}^{\mathrm{P}}(a)$. For this reason we expect derivatives — such as $\widehat{f}_{\mu}^{\mathrm{P}}(a)$ — of PAE of distribution functions to enjoy the same minimum-bias property.

[19] In comparison, neighborhood-specific empirical Bayes estimates have a substantially lower dispersion. In Appendix Figure S3 we report an estimate of their density $\widehat{f}_{\mu}^{\mathrm{PM}}$. 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 $\widehat{f}_{\mu}^{\mathrm{P}}$ is .030.

Figure 1: Density of neighborhood effects



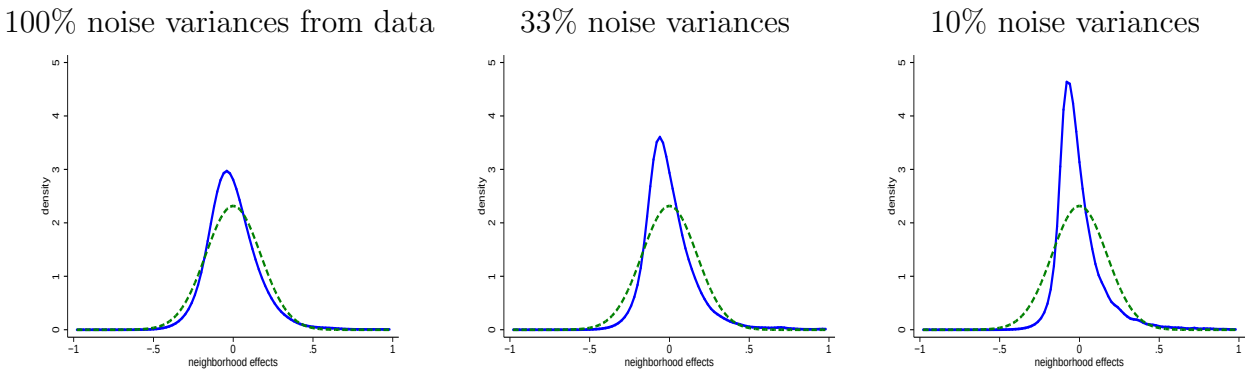
Notes: In the left graph we show the density of fixed-effects estimates $\widehat{\mu}_c$ (solid) and its normal fit (dashed). In the right graph we show the posterior density of μ_c (solid) and the prior density (dashed). Calculations are based on statistics available on the Equality of Opportunity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

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 density 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 (20). Note the R^2 coefficient varies along the distribution. 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.[20] 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

[20]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.

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



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.

model the neighborhood effects and the noise, individual-level data would be needed.

Lastly, we perform two additional exercises as robustness checks. Firstly, we incorporate the mean income $\overline{y}_c$ of permanent residents in county c at the 25% percentile 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 then a correlated random-effects specification, where the mean depends on $\overline{y}_c$ linearly. Appendix Figure S4 shows small differences with 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 Appendix Figure S5 shows that the normal prior and the posterior density are closer to each other than in the case of commuting zones.

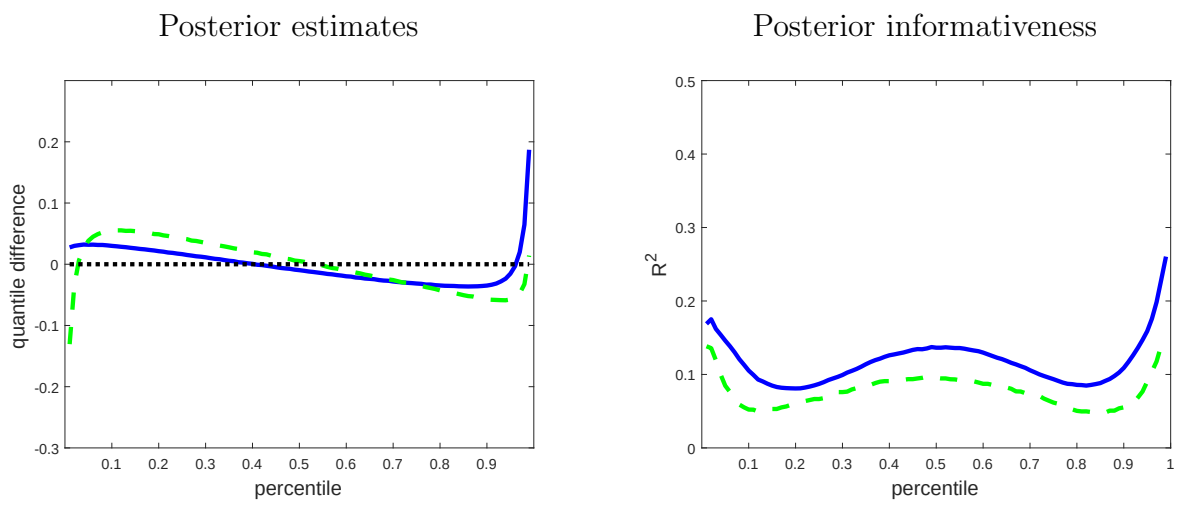
4.2 Income dynamics

In this subsection we consider the permanent-transitory model of household log-income

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

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

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 (19). η_{it} is shown in solid, and ε_{it} is shown in dashed. Sample from the PSID.

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. Note that the fact that we report quantile functions as opposed to distribution functions is not essential, but this helps visualizing the results.

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. 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 we report our posterior informativeness measure 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 Figure S6 we compare our posterior estimates with estimates based on a flexible non-normal and non-linear model from Arellano *et al.* (2017). 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, PAE 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. PAE 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. PAE are also applicable to other — nonlinear — econometric models, as we describe in the next section.

5 Complements and extensions

In this section we outline several complements and extensions that we analyze in detail in the appendix.

5.1 PAE in various settings

PAE are applicable to a wide variety of settings. In many econometric models, semi-parametric estimators — i.e., robust to distributional assumptions on unobservables — of β parameters are available; see Powell (1994) for example. In such models, PAE provide estimators of average effects that enjoy robustness properties when parametric assumptions are violated. As examples, in Appendix S5 we study static binary and ordered choice models, censored regression models, and panel data binary choice models. We also show how the White (1980) formula for robust standard errors in linear regression can be interpreted as a PAE. Lastly, we report illustrative simulations for static binary and ordered choice models, along with simulations for the fixed-effects model (1).

5.2 Confidence intervals and specification test

Under correct specification of the reference model, it is easy to derive the asymptotic distributions of $\widehat{\delta}^{\mathrm{M}}$ and $\widehat{\delta}^{\mathrm{P}}$ using standard arguments. Moreover, under local misspecification, confidence intervals that account for both model uncertainty and sampling uncertainty can be constructed following Armstrong and Kolesár (2018) and Bonhomme and Weidner (2018). However, such confidence intervals require the researcher to set a value for the degree of misspecification ϵ. In Appendix S4 we provide details on confidence intervals calculations. In addition, we explain how to construct a specification test of the reference model based on the difference $\widehat{\delta}^{\mathrm{P}} - \widehat{\delta}^{\mathrm{M}}$.

5.3 Fixed-ϵ bias bound

As a complement to the local analysis of Section 3, we show the following *non-local* bias bound in Appendix S1.

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 fixed-ϵ bound on the bias of PAE.[21] Although $\widehat{\delta}^{\mathrm{P}}$ does not necessarily minimize bias for finite ϵ, Theorem 2 shows that its bias is never larger than twice the minimum worst-case bias in the neighborhood within the class of asymptotically linear estimators. This minimum bias is generally non-zero, whenever the quantity of interest $\overline{\delta}$ is not point-identified.[22] In addition, the factor two in Theorem 2 cannot be improved upon in general, as we show in Appendix S4 in the context of a simple binary choice model.

5.4 Robustness in prediction

In applications such as the fixed-effects model (1) of teacher quality, researchers are often interested in *predicting* the quality α_i of teacher i. Although our 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. It is well-known that EB estimators minimize mean squared prediction 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, conditioning on nonlinear functions of the data may improve prediction accuracy.

In Appendix S3, we use our framework — applied to worst-case mean squared prediction error instead of worst-case bias of a sample average — to provide results on the robustness of EB estimators in the presence of misspecification.[23] We show that EB estimators have minimum worst-case mean squared prediction error, up to smaller-order terms, under local deviations from normality. In addition, we derive a fixed-ϵ, non-local risk bound in the spirit of Theorem 2.

6 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

[21] 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.

[22] Imposing that $f_0 \in \Gamma_\epsilon$ for some $\epsilon > 0$ implies that $\overline{\delta}$ has finite lower and upper bounds.

[23] Note that this differs from the analysis of the MSE properties of *average estimators* in Remark 1.

have shown that posterior average effects (PAE) have minimum worst-case bias under local misspecification of parametric assumptions. PAE are simple to implement, and our analysis provides a rationale for reporting them in applications. As an example, Arnold *et al.* (2020) recently reported PAE to document judge heterogeneity in the context of bail decisions. While we have used a linear fixed-effects model as a running example due to its popularity, there are many other possible applications, some of which we discuss in the appendix.

Lastly, our examples highlight that the information contained in the conditioning is setting-specific. Hence, PAE 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] 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.

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

[4] 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.

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

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

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

[8] Arnold, D., W. S. Dobbie, and P. Hull (2020): "Measuring Racial Discrimination in Bail Decisions," (No. w26999). National Bureau of Economic Research.

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

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

[11] 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.

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

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

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

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

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

[17] 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.

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

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

[20] 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.

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

[22] 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.

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

[24] 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.

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

[26] 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.

[27] 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.

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

[29] 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.

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

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

[32] Hirano, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70(2), 781–799.

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

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

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

[36] 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.

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

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

[39] 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.

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

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

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

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

[44] Owen, D. B. (1956): "Tables for Computing Bivariate Normal Probabilities," *The Annals of Mathematical Statistics*, 27(4), 1075–1090.

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

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

APPENDIX

A Proofs of Lemma 1 and Theorem 1

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* $\widetilde{\psi}(y,x) = \psi(y,x) - \mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\psi(Y,X)\middle|X=x\right]$, *let* $\widetilde{\gamma}(y,x) = \gamma(y,x) - \mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\gamma(Y,X)\middle|X=x\right]$ *and* $\widetilde{\delta}(u,x) = \delta(u,x) - \mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\delta(U,X)\middle|X=x\right]$. *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* $\left|\gamma_{\beta,\sigma_*}(y,x)\right|$, $\left|\delta_\beta(u,x)\right|$ *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 the following results hold as ϵ *tends to zero.*

(i) We have

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

(ii) If we have $b_\epsilon(\gamma) = b_\epsilon(\gamma^{\mathrm{P}}_{\beta,\sigma_*}) + o(\epsilon^{1/2})$, *then there exist* $\lambda \in \mathbb{R}^{\dim \psi}$ *and a function* $\omega : \mathcal{X} \to \mathbb{R}$ *with* $\mathbb{E}_{f_X}[\omega(X)] = 0$ *such that*

$$\gamma_{\beta,\sigma_*}(Y,X) = \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) + \omega(X) + \lambda' \, \psi_{\beta,\sigma_*}(Y,X) + o_{P(\beta,f_{\sigma_*})}(1).$$

Notice that Theorem A1 in the main text is a special case of part (i) of Theorem A1. Part (ii) of Theorem A1 is discussed in Remark 2 of the main text. The proof of Theorem A1 provides explicit expressions for λ and $\omega(X)$ that appear in part (ii), namely λ is the same as in the last line of Lemma A1, and $\omega(x) = \mathbb{E}_{P(\beta,f_{\sigma_*})}\big[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \lambda' \, \psi_{\beta,\sigma_*}(Y,X)\big|X = x\big] - \mathbb{E}_{P(\beta,f_{\sigma_*})}\big[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \lambda' \, \psi_{\beta,\sigma_*}(Y,X)\big]$.

A.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

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

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*

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

satisfies

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

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 S2. 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 densities $f_0^{(s)}(u|x)$ in Lemma A2 are exponentially tilted versions of the reference density $f_{\sigma_*}(u|x)$. 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

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

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

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 S2.

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 density $P(\beta, f_{\sigma_*})$. 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

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

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 $\lambda^{(1)}_{\text{rem}}(s) \in \mathbb{R}$, $\lambda^{(2)}_{\text{rem}}(s) \in \mathbb{R}$, and $\lambda^{(3)}_{\text{rem}}(s) \in \mathbb{R}^{\dim\psi}$, which will only be specified in (A8) below. We can rewrite

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

with

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

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

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

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 $\lambda^{(1)}_{\text{rem}}(s,x)$, $\lambda^{(2)}_{\text{rem}}(s)$, $\lambda^{(3)}_{\text{rem}}(s)$ such that (A7) holds exactly. Plugging the expansions (A4) into (A7), using the definition of $t(u,x|s)$, as well as $\mathbb{E}_*\left[t_{(0)}(U,X|s)\middle|X=x\right] = 0$, $\mathbb{E}_*\left\{[t_{(0)}(U,X|s)]^2\right\} = 2\,\phi''(1)$, and $\mathbb{E}_*\psi(Y,X)\,t_{(0)}(U,X|s) = 0$, we obtain

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

Those conditions can be rewritten as follows

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

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

$$\begin{aligned}
\lambda^{(1)}_{\text{rem}}(s,x) &= -2[\phi''(1)]^2\,R_1(0)\,\frac{\text{Var}_*\left[\widetilde{\gamma}(Y,X)-\widetilde{\delta}(U,X)-\lambda'\,\widetilde{\psi}(Y,X)\,\middle|\,X=x\right]}{\text{Var}_*\left[\widetilde{\gamma}(Y,X)-\widetilde{\delta}(U,X)-\lambda'\,\widetilde{\psi}(Y,X)\right]}+\mathcal{O}(\epsilon^{1/2}),\\
\lambda^{(2)}_{\text{rem}}(s) &= -\frac{1}{2\,\kappa}\,\mathbb{E}_*\left[t_{(0)}(U,X|s)\right]^3\,R_2(0)+\mathcal{O}(\epsilon^{1/2}),\\
\lambda^{(3)}_{\text{rem}}(s) &= -\phi''(1)\,\left\{\mathbb{E}_*\left[\psi(Y,X)\,\psi(Y,X)'\right]\right\}^{-1}\mathbb{E}_*\left\{\psi(Y,X)\,\left[t_{(0)}(U,X|s)\right]^2\right\}\,R_1(0)+\mathcal{O}(\epsilon^{1/2}).
\end{aligned}\tag{A9}$$

Notice that $\lambda^{(1)}_{\text{rem}}(s,x)$, $\lambda^{(2)}_{\text{rem}}(s)$, $\lambda^{(3)}_{\text{rem}}(s)$ also appear implicitly on the right-hand sides of the equations (A8), because $t_{\text{rem}}(u,x|s)$ depends on those parameters, and (A8) is therefore a system of equations for $\lambda^{(1)}_{\text{rem}}(s,x)$, $\lambda^{(2)}_{\text{rem}}(s)$, $\lambda^{(3)}_{\text{rem}}(s)$. 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 $|R_1(t)|\le c_1$ and $|R_2(t)|\le c_2$. 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

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

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 $\mathbb{E}_*\left|\gamma(Y,X)-\delta(U,X)\right|^3<\infty$, and $\mathbb{E}_*\left|\psi(Y,X)\right|^3<\infty$, 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.2 Proof of Theorem A1

Part (i): We first want to show that $b_\epsilon(\gamma^{\mathrm{P}}_{\beta,\sigma_*})\le b_\epsilon(\gamma)+\mathcal{O}(\epsilon)$. By applying Lemma A1 to both $\gamma_{\beta,\sigma_*}(y,x)$ and $\gamma^{\mathrm{P}}_{\beta,\sigma_*}(y,x)=\mathbb{E}_{p_{\beta,\sigma_*}}[\delta_\beta(U,X)\,|\,Y=y,X=x]$ we obtain, as $\epsilon\to 0$,

$$\begin{aligned} b_\epsilon(\gamma) &= \left|\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma_{\beta,\sigma_*}(Y,X)]-\mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X)]\right| \\ &\qquad + \epsilon^{\frac{1}{2}}\left\{\frac{2}{\phi''(1)}\mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\left(\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_\beta(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right)^2\right]\right\}^{\frac{1}{2}}+\mathcal{O}(\epsilon), \\ b_\epsilon(\gamma^{\mathrm{P}}) &= \epsilon^{\frac{1}{2}}\left\{\frac{2}{\phi''(1)}\mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\left(\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)^2\right]\right\}^{\frac{1}{2}}+\mathcal{O}(\epsilon), \end{aligned} \tag{A12}$$

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 $b_\epsilon(\gamma^{\mathrm{P}})$ we used that by the law of iterated expectations we have that $\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)]-\mathbb{E}_{f_{\sigma_*}}[\delta_\beta(U,X)]=0$ (that is, the first term in $b_\epsilon(\gamma)$ is not present in $b_\epsilon(\gamma^{\mathrm{P}})$) and also $\mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\left(\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)-\delta_\beta(U,X)\right)\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right]=0$ (that is, the vector λ is equal to zero for γ^{P}). We also use that under the reference model $\widetilde{\gamma}_{\beta,\sigma_*}(Y,X)-\widetilde{\delta}_\beta(U,X)-\lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)$ has zero mean, implying that its variance equals its second moment.

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 $b_\epsilon(\gamma^{\mathrm{P}}) \leq b_\epsilon(\gamma)$ 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 to

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

Again applying the law of iterated expectations we find that

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

Using this we obtain

$$\begin{aligned}
&\mathbb{E}_{P(\beta,f_{\sigma_*})} \left\{ \left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right] - \left[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right] \right\}^2 \\
&= \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right]^2 + \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right]^2 \\
&\quad - 2\, \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right] \left[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right] \\
&= \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right]^2 - \mathbb{E}_{P(\beta,f_{\sigma_*})} \left[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right]^2 .
\end{aligned} \tag{A13}$$

Since $\mathbb{E}_{P(\beta,f_{\sigma_*})} \left\{ \left[\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda' \widetilde{\psi}_{\beta,\sigma_*}(Y,X) \right] - \left[\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) \right] \right\}^2 \geq 0$ we thus conclude that

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

and therefore we obtain that

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

This is the first statement of the theorem. This concludes the proof of part (i) of Theorem A1, of which Theorem 1 in the main text is a special case.

Part (ii): Next, let $\gamma_{\beta,\sigma_*}(y,x)$ be such that

$$b_\epsilon(\gamma) = b_\epsilon(\gamma^{\mathrm{P}}_{\beta,\sigma_*}) + o(\epsilon^{1/2}). \tag{A14}$$

Then, the bias expansions in (A12) are still valid, and using those we conclude that we must have

$$\mathbb{E}_{P(\beta,f_{\sigma_*})}[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)] = o(1), \tag{A15}$$

because otherwise that term dominates all other terms in (A14). We also conclude that we must have

$$\begin{aligned}\mathbb{E}_{P(\beta,f_{\sigma_*})}&\left[\left(\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X)\right)^2\right] \\ &\leq \mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\left(\gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)\right)^2\right] + o(1)\end{aligned}$$

for (A14) to hold. Furthermore, the calculation in (A13) is still valid here, and the inequality in the last display can therefore equivalently be rewritten as

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

where we write $=$ instead of $\leq$, because the left hand side expression is non-negative. Applying Markov's inequality we thus find that

$$\widetilde{\gamma}_{\beta,\sigma_*}(Y,X) - \widetilde{\delta}_\beta(U,X) - \lambda'\,\widetilde{\psi}_{\beta,\sigma_*}(Y,X) = \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) + o_{P(\beta,f_{\sigma_*})}(1).$$

Defining

$$\begin{aligned}\omega(x) := \mathbb{E}_{P(\beta,f_{\sigma_*})}&\left[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \lambda'\,\psi_{\beta,\sigma_*}(Y,X)\middle|X = x\right] \\ &- \mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \lambda'\,\psi_{\beta,\sigma_*}(Y,X)\right],\end{aligned}$$

we therefore obtain

$$\begin{aligned}\gamma_{\beta,\sigma_*}(Y,X) &= \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) + \omega(X) + \lambda'\,\psi_{\beta,\sigma_*}(Y,X) \\ &\quad + \mathbb{E}_{P(\beta,f_{\sigma_*})}\left[\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X) - \lambda'\,\psi_{\beta,\sigma_*}(Y,X)\right] + o_{P(\beta,f_{\sigma_*})}(1) \\ &= \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X) + \omega(X) + \lambda'\,\psi_{\beta,\sigma_*}(Y,X) + o_{P(\beta,f_{\sigma_*})}(1),\end{aligned}$$

where in the last step we have used (A15) and $\mathbb{E}_{P(\beta,f_{\sigma_*})}[\psi_{\beta,\sigma_*}(Y,X)] = 0$. Finally, notice that by construction we have

$$\mathbb{E}_{f_X}[\omega(X)] = 0.$$

ONLINE APPENDIX — NOT FOR PUBLICATION

S1 Proof of Theorem 2

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

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

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

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

Lemma S1. *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 S2. Notice that both Theorem 2 and Lemma S1 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

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

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 S1 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

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

∎

S2 Proofs 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,

$$\begin{aligned} \widetilde{f}_0(u|x; s) = \operatorname*{argmax}_{f_0 \in [0,\infty)} \Bigg\{ & s\, [\gamma(g(u, x), x) - \delta(u, x)]\, f_X(x)\, f_0 - \mu_1(s, x)\, f_X(x)\, f_0 \\ & - \mu_2(s)\, \phi\left(\frac{f_0}{f_*(u|x)} \right) f_*(u|x)\, f_X(x) - \mu_3'(s)\, \psi(g(u, x), x)\; f_X(x)\, f_0 \Bigg\}, \end{aligned} \tag{S1}$$

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

$$\begin{aligned} \int_{\mathcal{U}\times\mathcal{X}} \widetilde{f}_0(u|x; s)\, f_X(x)\, du\, dx &= 1, \\ \int_{\mathcal{U}\times\mathcal{X}} \phi\left(\frac{\widetilde{f}_0(u|x; s)}{f_*(u|x)} \right) f_*(u|x)\, f_X(x)\, du\, dx &= \epsilon, \\ \int_{\mathcal{U}\times\mathcal{X}} \psi(g(u, x), x)\, \widetilde{f}_0(u|x; s)\, f_X(x)\, du\, dx &= 0 \end{aligned} \tag{S2}$$

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 (S1) 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 (S2) are exactly the conditions (A3) imposed in the lemma. Under the conditions of the lemma we therefore have

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

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

$$\phi_*(t) = \max_{r\geq 0} \left[r\, t - \phi(r)\right] = \rho(t)\, t - \phi(\rho(t)). \tag{S3}$$

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

$$|R_1(t)| \,\leq\, \widetilde{c}_1\,t^2, \quad \text{and} \quad |R_2(t)| \,\leq\, \widetilde{c}_2\,t^3. \tag{S4}$$

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

$$|\rho(t) - 1| = \frac{\left|t - \overline{\phi}'(\rho(t))\right|}{\nu} \leq \frac{|t|}{\nu}. \tag{S5}$$

Using (S4) and (S5), 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 (S3) 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 (S5) therefore give

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

for all $t \in \mathbb{R}$. Using this and (S4), 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 S1. Let $f_0 \in \Gamma_\epsilon$. Remember the definition of the posterior density $p_{\beta,\sigma_*}(u \,|\, y, x)$ in (12). Define

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

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

$$\mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\zeta(U,X) \,|\, Y, X\right]\right\} = \mathbb{E}_{P(\beta,\widetilde{f}_0)}\left[\zeta(U,X)\right]. \tag{S6}$$

We also find that

$$\mathbb{E}_{P(\beta,\widetilde{f}_0)}[\psi_{\beta,\sigma_*}(Y,X)] = \mathbb{E}_{P(\beta,f_0)}\left\{\mathbb{E}_{p_{\beta,\sigma_*}}\left[\psi_{\beta,\sigma_*}(Y,X) \,|\, Y, X\right]\right\} = \mathbb{E}_{P(\beta,f_0)}[\psi_{\beta,\sigma_*}(Y,X)] = 0. \tag{S7}$$

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

$$\begin{aligned} d(\widetilde{f}_0, f_{\sigma_*}) &\leq \int_{\mathcal{X}}\int_{\mathcal{U}}\int_{\mathcal{U}} \phi\left(\frac{f_0(\tilde{u}|x)}{f_{\sigma_*}(\tilde{u}\,|\,x)}\right) K_{\beta,\sigma_*}(\tilde{u}|u,x)\,d\tilde{u}\, f_{\sigma_*}(u\,|\,x) f_X(x)\,du\,dx \\ &= \int_{\mathcal{X}}\int_{\mathcal{U}} \phi\left(\frac{f_0(\tilde{u}|x)}{f_{\sigma_*}(\tilde{u}\,|\,x)}\right) \underbrace{\left[\int_{\mathcal{U}} f_{\sigma_*}(u\,|\,x)\, K_{\beta,\sigma_*}(\tilde{u}|u,x)\,du\right]}_{=f_{\sigma_*}(\tilde{u}\,|\,x)} f_X(x)\,d\tilde{u}\,dx \\ &= d(f_0, f_{\sigma_*}) \leq \epsilon. \end{aligned} \tag{S8}$$

Because $\widetilde{f}_0$ satisfies (S7) and (S8) 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 (S6) 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

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

∎

S3 Robustness in prediction

Under squared loss, we 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 prediction 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 to our measure of worst-case bias, here the mean squared prediction error is asymptotic, hence 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.

Lemma S2. *In addition to defining* $\widetilde{\psi}(y,x) = \psi(y,x) - \mathbb{E}_*\left[\psi(Y,X)\big|X=x\right]$, *let* $\widetilde{\gamma}(y,x) = \gamma(y,x) - \mathbb{E}_*\left[\gamma(Y,X)\big|X=x\right]$ *and* $\widetilde{\delta}(u,x) = \delta(u,x) - \mathbb{E}_*\left[\delta(U,X)\big|X=x\right]$. *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* $\left|\gamma_{\beta,\sigma_*}(y,x)\right|$, $\left|\delta_\beta(u,x)\right|$ *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*

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

where

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

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 density 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 S2 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 prediction error (which is a multiple of $\epsilon^{\frac{1}{2}}$) is also minimized at γ^{P}, provided the following condition holds almost surely:

$$
\mathbb{E}_{p_{\beta,\sigma_*}}\left[\left(\delta_\beta(U,X) - \gamma^{\mathrm{P}}_{\beta,\sigma_*}(Y,X)\right)^3 \,|\, Y,X\right] = 0. \tag{S9}
$$

While (S9) 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 (S9) 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, (S9) 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 prediction error, up to second-order terms in $\epsilon^{\frac{1}{2}}$.

We also have a fixed-ϵ bound in the spirit of Theorem 2.

Theorem S1. *Let* $\gamma^{\mathrm{P}}_{\beta,\sigma_*}$ *as in (A1). Then, for all* $\epsilon > 0$,

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

Theorem S1 shows that EB estimators are optimal, up to a factor of at most four, in terms of worst-case mean squared prediction 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 Wasserman, 2006, Chapter 7). Here, by contrast, we consider a worst case computed in a set of unrestricted, possibly non-normal joint distributions of $\alpha, \varepsilon_1, ..., \varepsilon_J$.

Proof of Lemma S2. This statement of the lemma is obtained from Lemma A1 by replacing $(\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X))$ by $\big(\gamma_{\beta,\sigma_*}(Y,X) - \delta_\beta(U,X)\big)^2$. The proof is obtained by the same replacement from the proof of Lemma A1. ∎

Proof of Theorem S1. By definition we have

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

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

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

We furthermore have

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

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 S1 and the definition of $e_\epsilon(\gamma)$. Combining the results of the last two displays we obtain that

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

∎

S4 Extensions

In this section of the appendix we consider eight issues in turn: how to compute PAE 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, how to construct confidence intervals, how to perform specification tests, how to derive the form of minimum-MSE estimators, and how to interpret PAE as Bayesian estimators in models where U has finite support.

S4.1 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 density 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 $f_{\widehat{\sigma}}(\cdot \,|\, X_i)$, 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.

S4.2 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

$$\varphi_\beta(f_0) = \varphi_\beta(f_{\sigma_*}) + \nabla\varphi_\beta(f_{\sigma_*})[f_0 - f_{\sigma^*}] + o(\epsilon^{\frac{1}{2}}), \tag{S10}$$

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 S5 we report model-based and posterior estimates of Gini coefficients based on simulated data.

S4.3 The constant in Theorem 2

The binary choice model that we describe in Section S5 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 the binary choice model (S13) of Section S5 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,

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

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

S4.4 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 minimum-bias results for PAE for scalar $\overline{\delta}$ then apply to $b_\epsilon(\gamma, v)$ for every given vector v, and the minimum-bias property is maintained after taking the supremum over the set of vectors v with $\|v\|_* = 1$. Thus, for the multi-dimensional case, we expect PAE to minimize local bias in the sense of Theorem 1, and to satisfy a bias bound with a factor of two as in Theorem 2, although a formal proof of local bias minimization 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 $a \in \mathcal{A} \subset \mathbb{R}$. Let us take the supremum norm $\|\cdot\|_\infty$ over functions $v(a)$ of a. This amounts to taking an ℓ^1 norm on distribution functions. Letting $\delta^{(a)}(U,X) = \mathbf{1}\{\alpha \leq a\}$, the local bias of the PAE is then

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

The ℓ^1-bias properties of distribution functions will translate into similar properties for quantile functions, subject to suitable (i.e., Lipschitz) conditions.

S4.5 Confidence intervals

Consider first the correctly specified case. 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

$$n^{\frac{1}{2}} \begin{pmatrix} \widehat{\delta}^{\mathrm{M}} - \overline{\delta} \\ \widehat{\delta}^{\mathrm{P}} - \overline{\delta} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \tag{S11}$$

Here, $\Sigma_{11} = \mathrm{Var}_*\left(G_1' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, X]\right)$, $\Sigma_{12} = \mathrm{Cov}_*\left(G_1' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, X],\right.$ $\left. G_2' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, Y,X]\right)$, $\Sigma_{21} = \Sigma_{12}$, and $\Sigma_{22} = \mathrm{Var}_*\left(G_2' h(Y,X) + \mathbb{E}_*[\delta(U,X) \,|\, Y,X]\right)$, for $G_1 = \partial_{\beta,\sigma} \mathbb{E}_{\beta,\sigma_*}[\delta_\beta(U,X)]$ and $G_2 = \mathbb{E}_{\beta,\sigma_*}\left\{\partial_{\beta,\sigma} \mathbb{E}_{p_{\beta,\sigma_*}}[\delta_\beta(U,X) \,|\, Y,X]\right\}$, where $\partial_\theta g(\theta_1)$ denotes the gradient of $g(\theta)$ at $\theta = \theta_1$. Note that in (S11) we allow δ_β to be non-smooth in β (e.g., an indicator function).

Consider next the locally misspecified case. 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 interpret ϵ by relating it to the local power of a specification test.

S4.6 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 $\widetilde{\Sigma} = \mathrm{Var}_* \left(\mathbb{E}_*[\delta(U,X) \,|\, Y, X] - \mathbb{E}_*[\delta(U,X) \,|\, X] + (G_2 - G_1)' h(Y,X) \right)$. Hence, under correct specification,

$$n \left(\widehat{\delta}^{\mathrm{P}} - \widehat{\delta}^{\mathrm{M}} \right)' \widetilde{\Sigma}^{-1} \left(\widehat{\delta}^{\mathrm{P}} - \widehat{\delta}^{\mathrm{M}} \right) \xrightarrow{d} \chi^2_1.$$

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

S4.7 Minimum local worst-case MSE estimator

Here we explain why $\widehat{\delta}^{\mathrm{MMSE}}$ in (17) gives the estimator with minimum worst-case MSE in a local neighborhood around the reference model (i.e., for small ϵ). We only consider the case where β and σ_* are known and not estimated; that is, we have $\psi(y,x) = 0$. Then, finding $\gamma^{\mathrm{MMSE}}(y,x)$ such that $\widehat{\delta}^{\mathrm{MMSE}}$ minimizes worst-case MSE over $f_0 \in \Gamma_\epsilon$ can, to leading order in ϵ and n^{-1}, be shown to be equivalent to minimizing

$$[b_\epsilon(\gamma)]^2 \, + \, \frac{1}{n} \mathrm{Var}_*[\gamma(Y,X)].$$

See Bonhomme and Weidner (2018) for details.

Next, applying Lemma 1 and noting that $\mathbb{E}_*[\gamma(Y,X) - \delta(U,X)] = 0$ is required for MSE minimization,[1] we find that to leading order in ϵ and n^{-1} the worst-case MSE reads

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

This expression for the approximate worst-case MSE depends on the distribution of X, which is unknown. For the minimum local worst-case bias result in Theorem 1 it does not matter that the distribution of X is unknown, because that distribution is identified from the sample as $n \to \infty$. However, for the MSE result here we have to take a stand on how to deal with the randomness in the observed covariates. In the following we *condition on the observed sample of covariates*, and replace all population expectations over X by sample averages over

[1] Adding a constant to $\gamma(y,x)$ such that $\mathbb{E}_*[\gamma(Y,X) - \delta(U,X)] = 0$ has no effect on the higher order bias terms in Lemma 1, nor on $\mathrm{Var}_*[\gamma(Y,X)]$. It is therefore always optimal to eliminate the leading bias term $\mathbb{E}_*[\gamma(Y,X) - \delta(U,X)]$ in this way.

X_i, $i = 1, \ldots, n$. We write $\widehat{\mathbb{E}}_X$ for those sample averages. The worst-case MSE objective function in the last display then reads

$$\frac{2\,\epsilon}{\phi''(1)}\,\widehat{\mathbb{E}}_X \mathrm{Var}_*\left[\gamma(Y,X) - \delta(U,X)\,\middle|\,X\right] + \frac{1}{n}\,\widehat{\mathbb{E}}_X\,\mathbb{E}_*\left(\left\{\gamma(Y,X) - \widehat{\mathbb{E}}_X\mathbb{E}_*[\delta(U,X)|X]\right\}^2\,\middle|\,X\right).$$

By the law of total variance we have

$$\begin{aligned}
&\mathrm{Var}_*\left[\gamma(Y,X) - \delta(U,X)\,\middle|\,X\right] \\
&= \mathbb{E}_*\left\{\mathrm{Var}_*\left[\gamma(Y,X) - \delta(U,X)\,\middle|\,Y,X\right]\,\middle|\,X\right\} + \mathrm{Var}_*\left\{\mathbb{E}_*\left[\gamma(Y,X) - \delta(U,X)\,\middle|\,Y,X\right]\,\middle|\,X\right\} \\
&= \mathbb{E}_*\left\{\mathrm{Var}_*\left[\delta(U,X)\,\middle|\,Y,X\right]\,\middle|\,X\right\} + \mathrm{Var}_*\left\{\mathbb{E}_*\left[\gamma(Y,X) - \delta(U,X)\,\middle|\,Y,X\right]\,\middle|\,X\right\}.
\end{aligned}$$

In the following we can ignore the term $\mathbb{E}_*\left\{\mathrm{Var}_*\left[\delta(U,X)\,\middle|\,Y,X\right]\,\middle|\,X\right\}$, because it does not depend on $\gamma(y,x)$. Then, the leading approximation to the worst-case MSE is given by the sample average over X of

$$\frac{2\,\epsilon}{\phi''(1)}\,\mathrm{Var}_*\left\{\gamma(Y,X) - \mathbb{E}_*\left[\delta(U,X)\,\middle|\,Y,X\right]\,\middle|\,X\right\} + \frac{1}{n}\,\mathbb{E}_*\left(\left\{\gamma(Y,X) - \widehat{\mathbb{E}}_X\mathbb{E}_*[\delta(U,X)|X]\right\}^2\,\middle|\,X\right).$$

Clearly, if for any given $X = x$ we find $\gamma(y,x)$ that minimizes this objective function, then its expected value over the sample distribution of X is also minimized. The corresponding first-order condition for $\gamma^{\mathrm{MMSE}}(Y,X)$ reads

$$\begin{aligned}
\frac{1}{n}\left\{\gamma^{\mathrm{MMSE}}(y,x) - \widehat{\mathbb{E}}_X\mathbb{E}_*[\delta(U,X)|X]\right\} + \frac{2\,\epsilon}{\phi''(1)}\Bigg\{&\gamma^{\mathrm{MMSE}}(y,x) - \mathbb{E}_*\left[\delta(U,X)\,\middle|\,Y=y,\,X=x\right] \\
&- \mathbb{E}_*\left[\gamma^{\mathrm{MMSE}}(Y,x)\,\middle|\,X=x\right] + \mathbb{E}_*\left[\delta(U,X)\,\middle|\,X=x\right]\Bigg\} = 0.
\end{aligned}$$

The solution to this first-order condition is

$$\begin{aligned}
\gamma^{\mathrm{MMSE}}(y,x) \;=\; &\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_*[\delta(U,X)|X=X_i] \\
&+ \left(1 + \frac{\phi''(1)}{2n\epsilon}\right)^{-1}\Big\{\mathbb{E}_*[\delta(U,X)\,|\,Y=y,X=x] - \mathbb{E}_*[\delta(U,X)\,|\,X=x]\Big\},
\end{aligned}$$

where we have now written $\widehat{\mathbb{E}}_X$ as $\frac{1}{n}\sum_{i=1}^n$.

The corresponding minimum local MSE estimator for $\overline{\delta} = \mathbb{E}_*\left[\delta_\beta(U,X)\right]$ is then given by

$$\begin{aligned}
\widehat{\delta}^{\mathrm{MMSE}} = \frac{1}{n}\sum_{i=1}^{n}\gamma^{\mathrm{MMSE}}(Y_i,X_i) = &\left[1 - \left(1 + \frac{\phi''(1)}{2n\epsilon}\right)^{-1}\right]\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_*[\delta(U,X)|X_i] \\
&+ \left(1 + \frac{\phi''(1)}{2n\epsilon}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_*[\delta(U,X)\,|\,Y=Y_i,X=X_i],
\end{aligned}$$

which is the result stated in equation (17) of the main text.

S4.8 Finite support

Here 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 $\widehat{\overline{\delta}}^{\mathrm{P}} = \frac{1}{n}\sum_{k=1}^K n_k \delta_k$. This estimator does not depend on the assumed f. Moreover, as $\min_{k=1,...,K} n_k$ tends to infinity we have

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

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

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

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

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,

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

Through $\overline{\delta}_\ell^U$, $\widehat{\overline{\delta}}^{\mathrm{P}}$ depends on the prior ω^U in general, even as $\min_{\ell=1,...,L} n_\ell$ 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 $k = 1, ..., K$. We will focus on the posterior mean

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

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 $\overline{\omega}_\ell = \sum_{k=1}^K \mathbf{1}\{\ell(k) = \ell\}\omega_k$. $(\overline{\omega}_1, ..., \overline{\omega}_L)$ follows the Dirichlet distribution $\text{Dir}(L, \overline{\alpha})$. 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

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

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

S5 Posterior average effects in various settings

In this section, we provide additional examples of models where PAE may be of interest, and we show illustrative simulations for two models.

S5.1 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 $\overline{\delta} = \mathbb{E}_{f_0}[U^2 XX']$.[2] In this context, a model-based approach consists in modeling $U \,|\, X$, say, as a normal with zero mean and

[2] In this example $\overline{\delta}$ is multi-dimensional; see Appendix S4.

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 PAE 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.

Censored regression. Consider next the censored regression model

$$Y_i = \max(Y_i^*, 0), \text{ where } Y_i^* = X_i'\beta + U_i. \tag{S12}$$

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 $U \,|\, X$ 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 $U \,|\, X \sim \mathcal{N}(0, \sigma^2)$, and estimate σ^2 using maximum likelihood. A model-based estimator is then $\widehat{\delta}^{\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{f_{\widehat{\sigma}}}[h(X_i'\widehat{\beta} + U)]$. By contrast, a PAE is

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

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 (S12), the researcher might want to impose that $U \,|\, X$ 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 densities f_0. However, PAE may no longer have minimum bias 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 bias 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, and we do not impose further restrictions that might be natural in order to justify the validity of this moment condition.

Binary choice. Consider now the binary choice model

$$Y_i = \mathbf{1}\{X_i'\beta + U_i > 0\}. \tag{S13}$$

In this model, Manski (1975, 1985) shows that β is identified up to scale as soon as the median of $U \,|\, X$ 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 $U \,|\, X$ 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.[3] A model-based estimator of the ASF is $\widehat{\delta}^{\mathrm{M}}(x) = \Phi\left(\frac{x'\widehat{\beta}}{\widehat{\sigma}}\right)$, and a posterior estimator is

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

Unlike $\widehat{\delta}^{\mathrm{M}}(x)$, the posterior ASF estimator $\widehat{\delta}^{\mathrm{P}}(x)$ 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 the next subsection 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.

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

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 $\widehat{\delta}^{\mathrm{M}}$ with 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 T is fixed and n tends to infinity.

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. PAE provide an alternative to the standard parametric model-based approach in this context.

S5.2 Simulations

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

S5.2.1 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 PAE 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 S1 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 biased for small J.

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

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

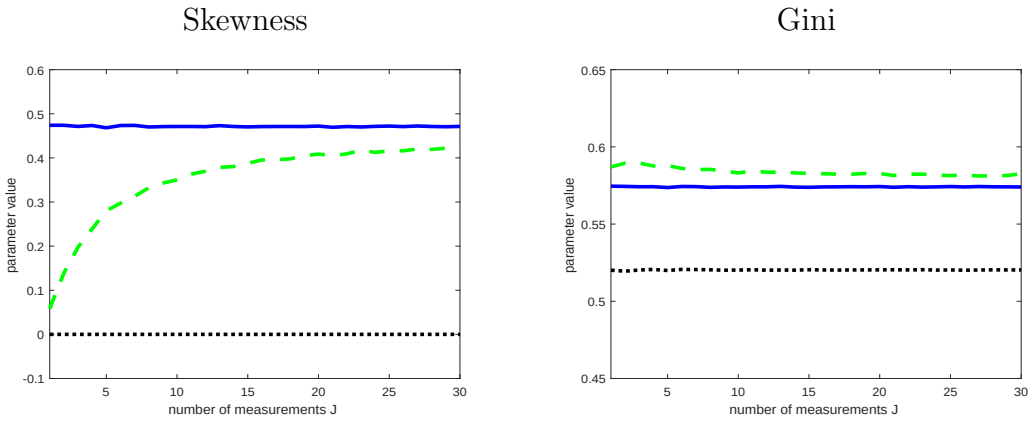
In this case, a model-based estimator is

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

while a PAE is, following (S10),

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

Figure S1: Skewness and Gini estimates in the fixed-effects model



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

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 S1 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.

S5.2.2 Ordered choice model

We next consider the ordered choice model

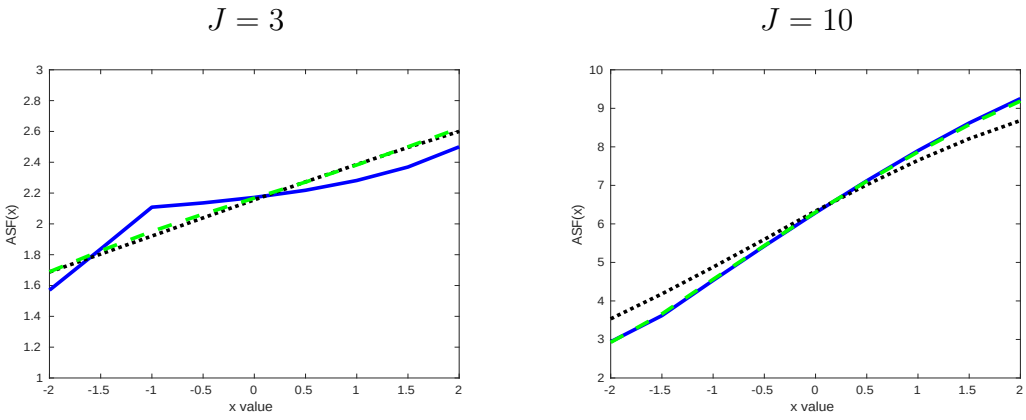
$$Y_i = \sum_{j=1}^{J} j\mathbf{1}\{\mu_{j-1} \leq Y_i^* \leq \mu_j\}, \text{ where } Y_i^* = X_i'\beta + U_i,$$

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

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

We take as reference distribution $U \,|\, X \sim \mathcal{N}(0, \sigma^2)$. 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

Figure S2: Average structural function in the ordered choice model



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

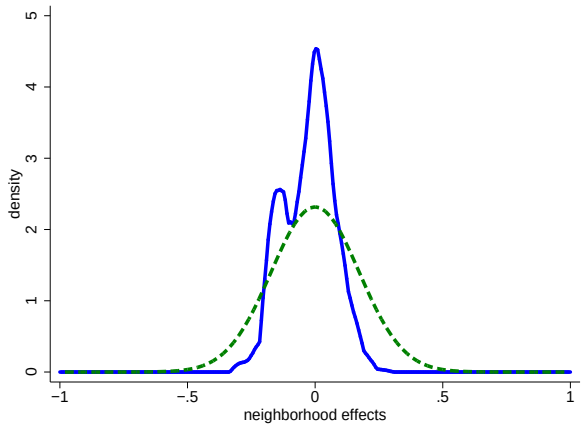
score (Manski, 1985).[4] For computation of maximum score, we use the mixed integer linear programming algorithm of Florios and Skouras (2008).

In Figure S2 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.

[4]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.

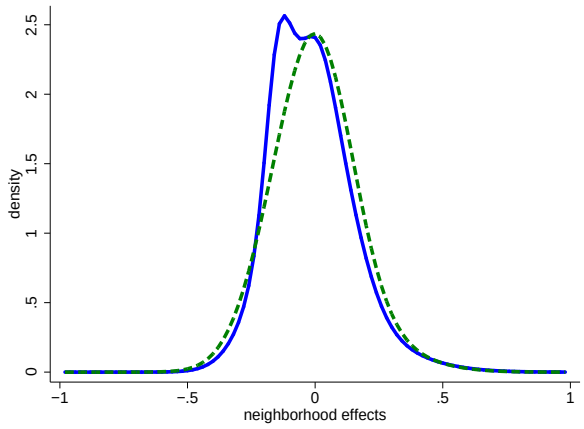
S6 Additional empirical results

Figure S3: Density of posterior means of neighborhood effects



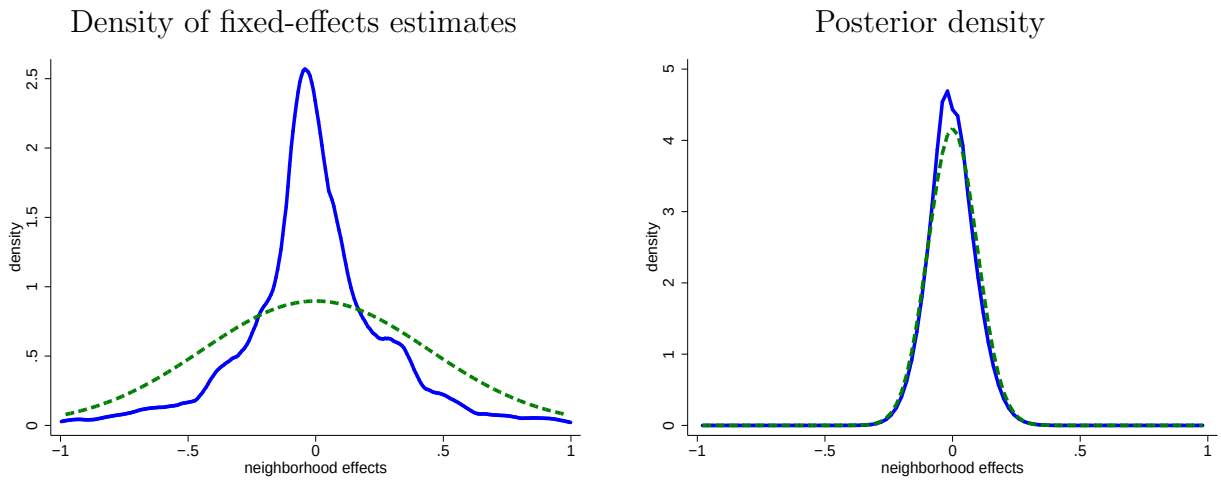
Notes: Density of posterior means of μ_c (solid) and prior density (dashed). Calculations are based on statistics available on the Equality of Opportunity website.

Figure S4: Posterior density of neighborhood effects, correlated random-effects specification



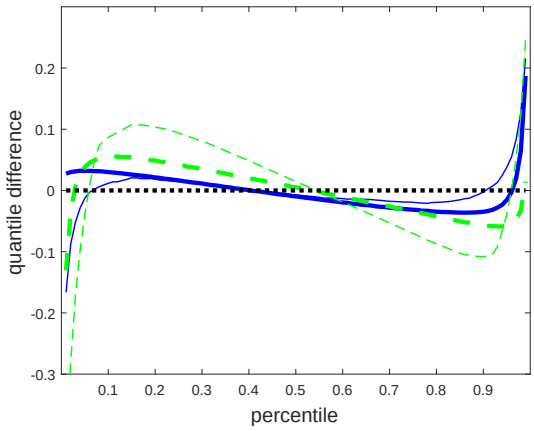
Notes: Posterior density of μ_c (solid) and prior density (dashed), based on a correlated random-effects specification allowing for correlation between the place effects μ_c and the mean income of permanent residents $\overline{y}_c$. Calculations are based on statistics available on the Equality of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

Figure S5: 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 S6: 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.*

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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