INFERENCE ON A DISTRIBUTION FROM NOISY DRAWS

Koen Jochmans*
University of Cambridge

Martin Weidner†
University College London

September 10, 2019

Abstract

We consider a situation where the distribution of a random variable is being estimated by the empirical distribution of noisy measurements of that variable. This is common practice in, for example, teacher value-added models and other fixed-effect models for panel data. We use an asymptotic embedding where the noise shrinks with the sample size to calculate the leading bias in the empirical distribution arising from the presence of noise. The leading bias in the empirical quantile function is equally obtained. These calculations are new in the literature, where only results on smooth functionals such as the mean and variance have been derived. Given a closed-form expression for the bias, bias-corrected estimator of the distribution function and quantile function can be constructed. We provide both analytical and jackknife corrections that recenter the limit distribution and yield confidence intervals with correct coverage in large samples. These corrections are non-parametric and easy to implement. Our approach can be connected to corrections for selection bias and shrinkage estimation and is to be contrasted with deconvolution. Simulation results confirm the much-improved sampling behavior of the corrected estimators.

JEL Classification: C14, C23

Keywords: bias correction, estimation noise, nonparametric inference, measurement error, panel data, regression to the mean, shrinkage.

*Address: University of Cambridge, Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, United Kingdom. E-mail: kj345@cam.ac.uk.

†Address: University College London, Department of Economics, Drayton House, 30 Gordon Street, London WC1 H0AX, United Kingdom; and CeMMAP. E-mail: m.weidner@ucl.ac.uk.

We are grateful to Isaiah Andrews, Stéphane Bonhomme, Bo Honoré, Ryo Okui, and Peter Schmidt for comments and discussion.

Jochmans gratefully acknowledges financial support from the European Research Council through Starting Grant nº 715787. Weidner gratefully acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001and from the European Research Council grants ERC-2014-CoG-646917-ROMIA and ERC-2018-CoG-819086-PANEDA.

The first version of this paper dates from March 13, 2018. The current and previous versions can be accessed from https://arxiv.org/abs/1803.04991.

1 Introduction

Let $\theta_1, \ldots, \theta_n$ be a random sample from a distribution F that is of interest. Suppose that we only observe noisy measurements of these variables, say $\vartheta_1, \ldots, \vartheta_n$. A popular approach is to do inference on F and its functionals using the empirical distribution of $\vartheta_1, \ldots, \vartheta_n$. This is common practice when analyzing panel data with heterogenous coefficients. In the literature on student achievement, for example, θ_i is a teacher effect, ϑ_i is an estimator of it obtained from data on student test scores, and we care about the distribution of teacher value-added (see, e.g., Jackson, Rockoff and Staiger 2014 for an overview). In the same vein, Guvenen (2009), Browning, Ejrnæs and Alvarez (2010), and Magnac and Roux (2019) estimate heterogenous earning profiles, while Ahn, Choi, Gale and Kariv (2014) find substantial heterogeneity in ambiguity aversion In a nonlinear fixed-effect model, the marginal effect is heterogenous across units and interest lies in the distribution of these effects as well as its functionals (Chamberlain 1984, Hahn and Newey 2004). Although the plug-in approach is popular, using $\vartheta_1, \ldots, \vartheta_n$ rather than $\theta_1, \ldots, \theta_n$ introduces bias that is almost entirely ignored in practice. Barras, Gagliardini and Scaillet (2018), who are interested in the distribution of the skill of fund managers, find that not accounting for bias leads to substantial overestimation of tail mass and misses to pick up the substantial asymmetry in the skill distribution.

We analyze the properties of the plug-in estimator of F in a location-scale setting where

$$\vartheta_i = \theta_i + \frac{\sigma_i}{\sqrt{m}}\,\varepsilon_i, \qquad \varepsilon_i \,|\, (\theta_i, \sigma_i^2) \sim \text{i.i.d. } (0,1),$$

where m is a parameter that grows with n. As the variance of the (heteroskedastic) noise is σ_i^2/m, this device shrinks the noise as the sample size grows. This is a very natural asymptotic embedding in settings where ϑ_i is an estimator of θ_i obtained from a sample of size m, as in a panel data setting or meta-analysis (Vivalt, 2015). It is related to, yet different from, an approach based on small measurement-error approximations as in Chesher (1991, 2017),[1] and has precedent in the analysis of fixed-effect models for panel

[1] Chesher (1991) provides expansions for densities, while we focus on distribution and quantile functions.

data, although for different purposes, as discussed in more detail below (see, e.g., Hahn and Kuersteiner 2002 and Alvarez and Arellano 2003).

Efron (2011) essentially entertains the homoskedastic setting with normal noise, where

$$\vartheta_i | \, \theta_i \sim N(\theta_i, \sigma^2/m),$$

and defines selection bias as the tendency of the ϑ_i's associated with the (in magnitude) largest θ_i's to be larger than their corresponding θ_i. He proposes to deal with selection bias by using the well-known Empirical Bayes estimator of Robbins (1956), which here is equal to

$$\vartheta_i + \frac{\sigma^2}{m} \nabla^1 \log p(\vartheta_i),$$

where p is the marginal density of the ϑ_i and ∇^1 denotes the first-derivative operator. For example, when $\theta_i \sim N(0, \psi^2)$ this expression then yields the (infeasible) shrinkage estimator

$$\left(1 - \frac{\sigma^2/m}{\sigma^2/m + \psi^2}\right) \vartheta_i,$$

a parametric plug-in estimator of which would be the James and Stein (1961) estimator. More generally, non-parametric implementation would also require estimation of p and its first derivative. Shrinkage to the overall mean (in this case zero) is intuitive, as selection bias essentially manifests itself through the tails of the empirical distribution of the ϑ_i being too thick.[2] Shrinkage is commonly-applied in empirical work (see, e.g., Rockoff 2004; Chetty, Friedman and Rockoff 2014). It should be stressed, though, that, while shrinkage improves on $\vartheta_1, \ldots, \vartheta_n$ in terms of estimation risk, it does not lead to preferable estimators of the distribution F or its moments.

The approach taken here is different from Efron (2011). Without making parametric assumptions on F, we calculate the (leading) bias of the naive plug-in estimator of the

Chesher (2017) discusses the impact of noise in the explanatory variables in a quantile-regression model; this is a different setup than the one considered here.

[2]The same shrinkage factor is applied to each ϑ_i, a consequence of the noise being homoskedastic. How to deal with heteroskedastic noise in an Empirical Bayes framework is not obvious. Discussion and a recent contribution can be found in and Weinstein, Ma, Brown and Zhang (2018).

distribution,

$$\hat{F}(\theta) := n^{-1} \sum_{i=1}^{n} 1\{\vartheta_i \leq \theta\}.$$

This calculation allows to construct estimators that correct for the bias directly. In the James-Stein problem, where $\theta_i \sim N(\eta, \psi^2)$, for example, the bias under homoskedastic noise equals

$$-\frac{\theta - \eta}{2} \frac{\sigma^2/\psi^2}{m} \phi\left(\frac{\theta - \eta}{\psi}\right) + O(m^{-2}).$$

Thus, the empirical distribution is indeed upward biased in the left tail and downward biased in the right tail. A bias order of m^{-1} implies incorrect coverage of confidence intervals unless $n/m^2 \to 0$. We present non-parametric plug-in and jackknife estimators of the leading bias and show that the bias-corrected estimators are asymptotically normal with zero mean and variance $F(\theta)\,(1 - F(\theta))$ as long as $n/m^4 \to 0$. So, bias correction is preferable to the naive plug-in approach for typical data sizes encountered in practice, where m tends to be quite small relative to n. We also provide corresponding bias-corrected estimators of the quantile function of F.

If the distribution of $\sigma_i\,\varepsilon_i$ is fully known, recovering F is a (generalized) deconvolution problem that can be solved for fixed m. Deconvolution-based estimators are well studied (see, e.g., Carroll and Hall 1988 and Delaigle and Meister 2008). However, they have a very slow rate of convergence and it is well documented that they can behave quite poorly in small samples. In response to this, Efron (2016) has recently argued for a return to a more parametric approach. Our approach delivers intuitive non-parametric estimators that enjoy the usual parametric convergence rate and are numerically well behaved. Although it does not deliver a fixed-m consistent estimator, bias correction further ensures that size-correct inference can be performed, provided that n/m^4 is small. It is not clear how to conduct inference based on deconvolution estimators.

Working out the statistical properties of $\hat{F}$ (and of its quantile function) is non-trivial because $\hat{F}$ is a non-smooth function of the data $\vartheta_1, \ldots, \vartheta_n$. As such, the approach taken here is different from, and complementary to, recent work on estimating average marginal effects in panel data models, which only looks at smooth functionals such as the mean

and variance (see, e.g., Fernández-Val and Lee 2013; Okui and Yanagi 2017). The impact of noise on smooth transformations of the ϑ_i can be handled using conventional methods based on Taylor-series expansions. We contrast such an approach with our derivations below. How to perform inference on the quantiles of marginal effects in nonlinear panel models is a long-standing open question (Dhaene and Jochmans, 2015), and the current work can be seen as a first step in that direction.

In work contemporaneous to our own, Okui and Yanagi (2018) derive the bias of a kernel-smoothed estimator of F and its derivative. Such smoothing greatly facilitates the calculation of the bias, making it amenable to conventional analysis. However, it also introduces additional bias terms that require much stronger moment conditions as well as further restrictions on the relative growth rates of n, m, and the bandwidth that governs the smoothing. Nevertheless, the (leading) bias term obtained in Okui and Yanagi (2018, Theorem 3) coincides with ours in Proposition 1 below.

2 Large-sample properties of plug-in estimators

Let F be a univariate distribution on the real line. We are interested in estimation of and inference on F and its quantile function $q(\tau) := \inf_\theta\{\theta : F(\theta) \geq \tau\}$. If a random sample $\theta_1, \ldots, \theta_n$ from F would be available this would be a standard problem. We instead consider the situation where $\theta_1, \ldots, \theta_n$ themselves are unobserved and we observe noisy measurements $\vartheta_1, \ldots, \vartheta_n$, with variances $\sigma_1^2/m, \ldots, \sigma_n^2/m$ for a positive real number m which, in our asymptotic analysis below, will be required to grow with n. We assume the following.

Assumption 1. *The variables $(\theta_i, \sigma_i^2, \vartheta_i)$ are i.i.d. across i, with*

$$E(\vartheta_i \,|\, \theta_i, \sigma_i^2) = \theta_i\,, \qquad\qquad E((\vartheta_i - \theta_i)^2 \,|\, \theta_i, \sigma_i^2) = \frac{\sigma_i^2}{m}\,,$$

and $\sigma_i^2 \in [\underline{\sigma}^2, \overline{\sigma}^2] \subset (0, \infty)$ for all i.

Our setup reflects a situation where the noisy measurements $\vartheta_1, \ldots, \vartheta_n$ converge in squared mean to $\theta_1, \ldots, \theta_n$ at the rate m^{-1}. A leading case is the situation where ϑ_i is an estimator

of θ_i obtained from a sample of size m that converges at the parametric rate.[3] We allow θ_i and σ_i^2 to be correlated, implying that the noise $\vartheta_i - \theta_i$ is not independent of θ_i. Hence, we allow for measurement error to be non-classical. Recovering the distribution of θ_i from a sample of $(\vartheta_i, \sigma_i^2)$ is, therefore, not a standard deconvolution problem.

It is common to estimate $F(\theta)$ by

$$\hat{F}(\theta) := n^{-1} \sum_{i=1}^{n} 1\{\vartheta_i \leq \theta\},$$

the empirical distribution of the ϑ_i at θ. As we will show below, under suitable regularity conditions, such plug-in estimators are consistent and asymptotically normal as $n \to \infty$ provided that m grows with n so that n/m^2 converges to a finite constant. The use of $\vartheta_1, \ldots, \vartheta_n$ rather than $\theta_1, \ldots, \theta_n$ introduces bias of the order m^{-1}, in general. This bias implies that test statistics are size distorted and the coverage of confidence sets is incorrect unless n/m^2 converges to zero.

The bias problem is easy to see (and fix) when interest lies in smooth functionals of F,

$$\mu := E(\varphi(\theta_i)),$$

for a (multiple-times) differentiable function φ. An (infeasible) plug-in estimator based on $\theta_1, \ldots, \theta_n$ would be

$$\tilde{\mu} := n^{-1} \sum_{i=1}^{n} \varphi(\theta_i).$$

Clearly, this estimator is unbiased and satisfies $\tilde{\mu} \overset{a}{\sim} N(\mu, \sigma_\mu^2/n)$ as soon as $\sigma_\mu^2 := \mathrm{var}(\varphi(\theta_i))$

[3] Everything to follow can be readily modified to different convergence rates as well as to the case where

$$\mathrm{var}(\vartheta_i | \, \theta_i, \sigma_i^2) = \sigma_i^2/m_i,$$

with $m_i := p_i m$ for a random variable $p_i \in (0, 1]$. It suffices to redefine σ_i^2 as σ_i^2/p_i. When the ϑ_i represent estimators this device allows for the sample size to vary with i. For example, in a panel data setting, it would cover unbalanced panels under a missing-at-random assumption. Further, the requirement that ϑ_i is unbiased can be relaxed to allow for standard non-linearity bias of order m^{-1}. We do not do this here as it is possible quite generally to reduce the bias down to $O(m^{-2})$, for example via a jackknife or bootstrap correction, making it negligible in our analysis below.

exists. For the feasible plug-in estimator of μ,

$$\hat{\mu} := n^{-1} \sum_{i=1}^{n} \varphi(\vartheta_i),$$

under standard regularity conditions, a Taylor-series expansion of $\varphi_i(\vartheta_i)$ around θ_i yields

$$E(\hat{\mu} - \mu) = \frac{b_\mu}{m} + O(m^{-2}), \qquad b_\mu := \frac{E(\nabla^2 \varphi(\theta_i)\, \sigma_i^2)}{2},$$

and

$$\operatorname{var}(\hat{\mu}) = \frac{\sigma_\mu^2}{n} + O\left(n^{-1} m^{-1}\right).$$

Hence, letting $z \sim N(0,1)$, we have

$$\frac{\hat{\mu} - \mu}{\sigma_\mu / \sqrt{n}} \overset{a}{\sim} z + \sqrt{\frac{n}{m^2}} \frac{b_\mu}{\sigma_\mu} \sim N(c\, b_\mu / \sigma_\mu, \sigma_\mu^2),$$

as $n/m^2 \to c^2 < \infty$ when $n, m \to \infty$. The noise in $\vartheta_1, \ldots, \vartheta_n$ introduces bias unless φ is linear. It can be corrected for by subtracting a plug-in estimator of b_μ / m from $\hat{\mu}$. Doing so, again under regularity conditions, delivers and estimator that is asymptotically unbiased as long as $n/m^4 \to 0$.

2.1 Distribution function

The machinery from above cannot be applied to deduce the bias of $\hat{F}$ as it is a step function and, hence, non-differentiable. We will derive its leading bias under the following conditions. To state them, we let

$$\varepsilon_i := \frac{\vartheta_i - \theta_i}{\sigma_i / \sqrt{m}}$$

and write f for the density function of F.

Assumption 2. *The distribution of ε_i is absolutely continuous, $E(\varepsilon_i^3) = 0$, $E(\varepsilon_i^4) < \infty$, f is three times differentiable with uniformly bounded derivatives, and one of the following two sets of conditions hold:*

A. (i) The function $E(\sigma_i^{p+1} | \theta_i = \theta)$ is p-times differentiable for $p = 1, 2, 3$; (ii) the joint density of (θ_i, σ_i) exists, and the conditional density function of θ_i given σ_i is three times

differentiable with respect to θ_i and the third derivative is bounded in absolute value by a function $e(\sigma_i)$ such that $E(e(\sigma_i)) < \infty$.

B. (i) There exists a deterministic function σ so that $\sigma_i = \sigma(\theta_i)$ for all i; and (ii) σ is four times differentiable and has uniformly-bounded derivatives.

The first part of Assumption 2 imposes conventional moment and smoothness conditions. The no-skewness assumption is conventional in this setting but can be dispensed with, in which case the higher-order bias would be of order $m^{-3/2}$ as opposed to the m^{-2} reported below. The remainder of Assumption 2 distinguishes between the cases where the relation between θ_i and σ_i^2 is stochastic (Assumption 2.A) and deterministic (Assumption 2.B). It requires smoothess of certain densities and conditional expectations. The bias in the deterministic setting is somewhat more difficult to handle. It is nonetheless useful to include as it covers situations where (conditional) heteroskedasticity is a function of θ_i, which is not uncommon.

Define the function

$$\beta(\theta) := \frac{E(\sigma_i^2|\theta_i = \theta)\, f(\theta)}{2},$$

which is well-behaved under Assumption 2, and let

$$b_F(\theta) := \beta'(\theta)$$

be its derivative. We also introduce the covariance function

$$\sigma_F(\theta, \theta') := F(\theta \wedge \theta') - F(\theta)\, F(\theta'),$$

where we use $\theta \wedge \theta'$ to denote $\min\{\theta, \theta'\}$. Proposition 1 summarizes the large-sample properties of $\hat{F}$.

Proposition 1. *Let Assumptions 1 and 2 hold. Then, as $n, m \to \infty$,*

$$E(\hat{F}(\theta)) - F(\theta) = \frac{b_F(\theta)}{m} + O(m^{-2}), \qquad \operatorname{cov}\big(\hat{F}(\theta), \hat{F}(\theta')\big) = \frac{\sigma_F(\theta, \theta')}{n} + O(n^{-1}m^{-1}),$$

where the order of the remainder terms is uniform in θ. If furthermore $n/m^4 \to 0$, then we have

$$\sqrt{n}\left(\hat{F}(\theta) - F(\theta) - \frac{b_F(\theta)}{m}\right) \rightsquigarrow \mathbb{G}_F(\theta),$$

where $\mathbb{G}_F(\theta)$ is a mean zero Gaussian process with covariance function $\sigma_F(\theta_1, \theta_2)$.

Proof. The proof is in Appendix A. □

To illustrate the result suppose that σ_i^2 is independent of θ_i and that θ_i has density function

$$f(\theta) = \frac{1}{\psi}\phi\left(\frac{\theta - \eta}{\psi}\right),$$

as in the James and Stein (1961) problem. Letting σ^2 denote the mean of the σ_i^2 an application of Proposition 1 yields

$$b_F(\theta) = -\frac{\theta - \eta}{2}\frac{\sigma^2}{\psi^2}\phi\left(\frac{\theta - \eta}{\psi}\right).$$

Thus, $\hat{F}(\theta)$ is upward biased when $\theta < \eta$ and is downward biased when $\theta > \eta$. This finding is a manifestation of the phenomenon of regression to the mean (or selection bias, or the winner's curse; see Efron 2011). It implies that the empirical distribution tends to be too disperse.

2.2 Quantile function

The bias in $\hat{F}$ translates to bias in estimators of the quantile function. A natural estimator of the quantile function is the left-inverse of $\hat{F}$. With this definition, the plug-in estimator of the τth-quantile is

$$\hat{q}(\tau) := \vartheta_{(\lceil \tau n \rceil)},$$

where $\vartheta_{(\lceil \tau n \rceil)}$ is the $\lceil \tau n \rceil$th order statistic of our sample, where $\lceil a \rceil$ delivers the smallest integer at least as large as a.

To calculate the leading bias in $\hat{q}(\tau)$ observe that it is an (approximate) solution to the empirical moment condition

$$\hat{F}(q) - \tau = 0$$

(with respect to q). From Proposition 1 we know that

$$E(\hat{F}(q(\tau))) - \tau = \frac{b_F(q(\tau))}{m} + O(m^{-2}),$$

uniformly in τ, so the moment condition that defines the estimator $\hat{q}(\tau)$ is biased. Letting

$$b_q(\tau) := -\frac{b_F(q(\tau))}{f(q(\tau))}, \qquad \sigma_q^2(\tau) := \frac{\tau(1-\tau)}{f(q(\tau))^2},$$

we obtain the following result.

Proposition 2. *Let the Assumptions 1 and 2 hold. For $\tau \in (0,1)$, assume that $f > 0$ in a neighborhood of $q(\tau)$. Then,*

$$\sqrt{n}\left(\hat{q}(\tau) - q(\tau) - \frac{b_q(\tau)}{m}\right) \xrightarrow{d} N(0, \sigma_q^2(\tau)),$$

as $n, m \to \infty$ with $n/m^2 \to c \in [0, +\infty)$.

Proof. The proof is in Appendix A. □

As an example, when $\theta_i \sim N(\eta, \psi^2)$, independent of σ_i^2, we have

$$b_q(\tau) = \frac{\sigma^2/\psi^2}{2}\,(q(\tau) - \eta),$$

which, in line with our discussion on regression to the mean above, is positive for all quantiles below the median and negative for all quantiles above the median. The median itself is, in this particular case, estimated without plug-in bias of order m^{-1}. It will, of course, still be subject to the usual n^{-1} bias arising from the nonlinear nature of the estimating equation.

3 Estimation and inference

Propositions 1 and 2 complement the existing results on the bias in smooth functionals (Fernández-Val and Lee 2013; Okui and Yanagi 2017) of the distribution of heterogenous parameters in panel data models. Our calculations confirm that the order of the bias in the empirical distribution and in the quantile function is of the same order as in the smooth case, m^{-1}.

3.1 Split-panel jackknife estimation

Our results validate a traditional jackknife approach to bias correction as in Hahn and Newey (2004) and Dhaene and Jochmans (2015). Such an approach exploits the fact that the bias is proportional to m^{-1} and is based on re-estimating $\theta_1, \ldots, \theta_n$ from subsamples. The simplicity of such a method makes it very useful in panel data applications.

To illustrate how the jackknife would work here, consider a stationary (balanced) $n \times m$ panel. Let ϑ_{i,m_1} be an estimator of θ_i constructed from the $n \times m_1$ subpanel consisting of the first m_1 cross sections only. Then

$$\hat{F}_{m_1}(\theta) := n^{-1} \sum_{i=1}^{n} 1\{\vartheta_{i,m_1} \leq \theta\}$$

is the plug-in estimator of $F(\theta)$ based on this subpanel alone. From Proposition 1 it follows that

$$E(\hat{F}_{m_1}(\theta)) = F(\theta) + \frac{b_F(\theta)}{m_1} + O(m^{-2}).$$

Using the remaining $m_2 := m - m_1$ cross section from the full panel we can equally calculate estimators ϑ_{i,m_2} and subsequently construct

$$\hat{F}_{m_2}(\theta) := n^{-1} \sum_{i=1}^{n} 1\{\vartheta_{i,m_2} \leq \theta\},$$

for which

$$E(\hat{F}_{m_2}(\theta)) = F(\theta) + \frac{b_F(\theta)}{m_2} + O(m^{-2})$$

follows in the same way. Consequently,

$$\tilde{b}_F(\theta) := m_1 \hat{F}_{m_1}(\theta) + m_2 \hat{F}_{m_2}(\theta) - m\hat{F}(\theta)$$

is a split-panel jackknife estimator of the leading bias term $b_F(\theta)$. Hence,

$$\tilde{F}(\theta) := \hat{F}(\theta) - \frac{\tilde{b}_F(\theta)}{m}.$$

is a nonparametric bias-corrected estimator.

A jackknife estimator of the quantile function can be defined in the same way. Moreover, let $\vartheta_{(\lceil \tau n \rceil),m_1}$ and $\vartheta_{(\lceil \tau n \rceil),m_2}$ be the $\lceil \tau n \rceil$ order statistic of the re-estimated quantities in the

first and second subsample, respectively. Recall that $\vartheta_{(\lceil \tau n \rceil),m_1}$ is the (approximate) solution to $\hat{F}_{m_1}(q) - \tau = 0$, and so is our estimator of $q(\tau)$ as obtained from the information in the $n \times m_1$ subpanel only. As before,

$$\tilde{b}_q(\tau) := m_1 \vartheta_{(\lceil \tau n \rceil),m_1} + m_2 \vartheta_{(\lceil \tau n \rceil),m_2} - m \vartheta_{(\lceil \tau n \rceil)}$$

is a nonparametric estimator of $b_q(\tau)$ that gives rise to a jackknife bias-corrected estimator of the quantile function.

The large-sample behavior of these jackknife estimators is the same as for the analytic corrections in Propositions 3 and 4 below. The split-sample jackknife is simple to implement but require access to the original data from which $\vartheta_1, \ldots, \vartheta_n$ were computed. This can be infeasible in meta-analysis problems, where each of the ϑ_i is an estimator constructed from a different data set that need not all be accessible. It can also be complicated in structural econometric models, where ϑ_i often will be the solution to a cumbersome optimization programme that can be time-consuming to solve. We, therefore, discuss two alternative bias-correction estimators— one based on a plug-in estimator of the bias and one based on a jackknife estimator—that do not require re-estimation of $\theta_1, \ldots, \theta_n$ in the next two subsections.

3.2 Analytic bias correction

We will formulate regularity conditions for a plug-in estimator of the bias to be consistent under the maintained assumption that

$$\vartheta_i | (\theta_i, \sigma_i^2) \sim N(\theta_i, \sigma_i^2/m) \tag{3.1}$$

where $\sigma_1^2, \ldots, \sigma_m^2$ are known. The normality assumption could be replaced by tail conditions on the noise distribution. Under suitable conditions, the results below will continue to go through when the σ_i^2 are replaced by estimators. We abstract away from these complications here as we feel that they would cloud the exposition.

A bias-corrected estimator based on Proposition 1 takes the form

$$\check{F}(\theta) := \hat{F}(\theta) - \frac{\hat{b}_F(\theta)}{m}, \qquad \hat{b}_F(\theta) := -\frac{(nh^2)^{-1} \sum_{i=1}^n \sigma_i^2 \, \kappa'\left(\frac{\vartheta_i - \theta}{h}\right)}{2},$$

where κ' is the derivative of kernel function κ and h is a non-negative bandwidth parameter. Thus, we estimate the bias using standard kernel methods. For simplicity, we will use a Gaussian kernel throughout, so $\kappa'(\eta) := -\eta\,\phi(\eta)$.

We establish the asymptotic behavior of $\check{F}$ under the following conditions.

Assumption 3.

(i) Equation (3.1) holds.

(ii) The conditional density of θ_i given σ_i is five times differentiable with respect to θ_i and the derivatives are bounded in absolute value by a function $e(\sigma_i)$ such that $E(e(\sigma_i)) < \infty$.

(iii) There exists an integer $\omega > 2$, and real numbers $\kappa > 1 + (1 - \omega^{-1})^{-1}$ and $\eta > 0$ so that $\sup_\theta (1 + |\theta|^\kappa)\, f(\theta) = O(1)$ and $\sup_\theta (1 + |\theta|^{1+\eta})\, |\nabla^1 b_F(\theta)| = O(1)$, and $\sup_\theta |b_F(\theta)| = O(1)$.

Assumption 3 contains simple smoothness and boundedness requirements on the conditional density of θ_i given σ_i^2, as well as tail conditions on the marginal density of the θ_i and on the bias function $b_F(\theta)$.

We have the following result.

Proposition 3. *Let Assumptions 1, 2, and 3 hold and let $\varepsilon := (3 - \omega^{-1})\,\omega^{-1} > 0$. If $h = O(m^{-1/2})$, $h^{-1} = O(m^{2/3 - 4/9\,\varepsilon})$, and $h^{-1} = O(n)$, as $n \to \infty$ and $m \to \infty$ with $n/m^4 \to 0$, then*

$$\sqrt{n}(\check{F}(\theta) - F(\theta)) \rightsquigarrow \mathbb{G}_F(\theta)$$

as a stochastic process indexed by θ, where $\mathbb{G}_F(\theta)$ is a mean zero Gaussian process with covariance function $\sigma_F(\theta_1, \theta_2)$.

Proof. The proof is in Appendix B. □

The implications of Proposition 3 are qualitatively similar to those for smooth functionals discussed above. Indeed, for any fixed θ, it implies that

$$\check{F}(\theta) \overset{a}{\sim} N(F(\theta), F(\theta)(1 - F(\theta))/n)$$

as $n \to \infty$ and $m \to \infty$ with $n/m^4 \to 0$. Thus, the leading bias is removed from $\hat{F}$ without incurring any cost in terms of (asymptotic) precision. Given the correction term, the sample variance of

$$1\{\vartheta_i \leq \theta\} + \frac{1}{2}\frac{1}{mh^2}\sigma_i^2\kappa'\left(\frac{\vartheta_i - \theta}{h}\right)$$

is a more natural basis for inference in small samples than is that of $1\{\vartheta_i \leq \theta\}$.

A data-driven way of choosing h is by cross validation. A plug-in estimator of the integrated squared error $\int_{-\infty}^{+\infty}(\check{F}(\theta) - F(\theta))^2\, d\theta$ (up to multiplicative and additive constants) is

$$v(h) := \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\sigma_i^2\sigma_j^2}{h^2}\underline{\phi}'(\vartheta_i, \vartheta_j; h) + \sum_{i=1}^{n}\sum_{j\neq i}\frac{\sigma_i^2}{h}\left(m\phi'\left(\frac{\vartheta_i - \vartheta_j}{h}\right) - \frac{nm}{n-1}\phi\left(\frac{\vartheta_i - \vartheta_j}{h}\right)\right),$$

where we use the shorthand

$$\underline{\phi}'(\vartheta_i, \vartheta_j; h) := \frac{1}{4}\frac{1}{\sqrt{2}h}\phi\left(\frac{\vartheta_i - \vartheta_j}{\sqrt{2}h}\right)\left(\frac{1}{2} - \frac{(\vartheta_i + \vartheta_j)^2}{4h^2} + \frac{\vartheta_i\vartheta_j}{h^2}\right).$$

See the appendix for details on the derivation. The cross-validated bandwidth then is $\check{h} := \arg\min_h v(h)$ on the interval $(0, +\infty)$.

Now turn the bias-corrected estimation of the quantile function. Proposition 2 readily suggests a bias-corrected estimator of the form

$$\hat{q}(\tau) - \frac{\hat{b}_q(\tau)}{m}, \qquad \hat{b}_q(\tau) := -\frac{\hat{b}_F(\hat{q}(\tau))}{\hat{f}(\hat{q}(\tau))},$$

using obvious notation. While (under suitable regularity conditions) such an estimator successfully reduces bias it has the unattractive property that it requires a non-parametric estimator of the density f, which further shows up in the denominator.

An alternative estimator that avoids this issue is

$$\check{q}(\tau) := \vartheta_{(\lceil \hat{\tau}^* n \rceil)}, \qquad \hat{\tau}^* := \tau + \frac{\hat{b}_F(\hat{q}(\tau))}{m},$$

The justification for this estimator comes from the fact that $E(\hat{F}(q(\tau))) - \tau^* = O(m^{-2})$, where $\tau^* = \tau + b_F(q(\tau))/m$, and its interpretation is intuitive. Given the noise in the ϑ_i relative to the θ_i, the empirical distribution of the former is too heavy-tailed relative to

the latter, and so $\hat{q}(\tau)$ estimates a quantile that is too extreme, on average. Changing the quantile of interest from τ to τ^* adjusts the naive estimator and corrects for regression to the mean.

Proposition 4. *Let the assumptions stated in Proposition 3 hold. For $\tau \in (0,1)$, assume that $f > 0$ in a neighborhood of $q(\tau)$. Then,*

$$\sqrt{n}\,(\check{q}(\tau) - q(\tau)) \xrightarrow{d} N(0, \sigma_q^2(\tau)),$$

as $n, m \to \infty$ with $n/m^4 \to 0$.

Proof. The proof is in Appendix B. □

The corrected estimator has the same asymptotic variance as the uncorrected estimator. It is well-known that plug-in estimators of σ_q^2 can perform quite poorly in small samples (Maritz and Jarrett 1978). Typically, researchers rely on the bootstrap, and we suggest doing so here. Moreover, draw (many) random samples of size n from the original sample $\vartheta_1, \ldots, \vartheta_n$ and re-estimate $q(\tau)$ by the bias-corrected estimator for each such sample. Then construct confidence intervals for $q(\tau)$ using the percentiles of the empirical distribution of these estimates. Note that, again, this bootstrap procedure does not involve re-estimation of the individual θ_i.

3.3 Jackknife bias correction

A jackknife procedure can be constructed from the observation that, if $\vartheta_1, \ldots, \vartheta_n$ would have variance $\lambda^2\sigma_1^2, \ldots, \lambda^2\sigma_n^2$ for some $\lambda > 0$, then the bias in $\hat{F}$ would equally be multiplied by λ^2. This is apparent from the definition of β and Proposition 1. This observation suggests the jackknife estimator

$$\dot{F}(\theta) := \hat{F}(\theta) - \frac{\dot{b}_F(\theta)}{m} = \frac{1+\lambda^2}{\lambda^2}\,\hat{F}(\theta) - \frac{1}{\lambda^2}\,\hat{F}_\lambda(\theta),$$

where

$$\dot{b}_F(\theta) := m\,\frac{\hat{F}_\lambda(\theta) - \hat{F}(\theta)}{\lambda^2}, \qquad \hat{F}_\lambda(\theta) := n^{-1}\sum_{i=1}^{n} \Phi\left(\frac{1}{\lambda}\frac{\theta - \vartheta_i}{\sigma_i/\sqrt{m}}\right).$$

Note that, contrary to the split-sample jackknife $\tilde{F}$, the estimator $\dot{F}$ can be computed without re-estimating $\theta_1, \ldots, \theta_n$ but, in turn, requires knowledge (or estimates) of the σ_i^2. The current jackknife bears similarities to the jackknife estimator of a density function introduced in Schucany and Sommers (1977).

The reason this estimator is bias-reducing is as follows. By iterated expectations,

$$E(\hat{F}(\theta)) = E\left(\Phi\left(\frac{\theta - \theta_i}{\sigma_i/\sqrt{m}}\right)\right) = F(\theta) + \frac{b_F(\theta)}{m} + O(m^{-2}).$$

Further, by a standard convolution argument,

$$E(\hat{F}_\lambda(\theta)) = E\left(\Phi\left(\frac{1}{\sqrt{1+\lambda^2}}\frac{\theta - \theta_i}{\sigma_i/\sqrt{m}}\right)\right) = F(\theta) + (1+\lambda^2)\frac{b_F(\theta)}{m} + O(m^{-2}).$$

Thus, our $\dot{b}_F(\theta)$ is a sample version of $b_F(\theta)$. Like in Schucany and Sommers (1977), the approach exploits variation in a bandwidth parameter. However, while they address smoothing bias in non-parametric density estimation (in a similar way as would the use of a higher-order kernel), our estimator attacks bias introduced through noise. Note, finally, that the sample variance of

$$1\{\vartheta_i \leq \theta\} - \frac{1}{\lambda^2}\left(\Phi\left(\frac{1}{\lambda}\frac{\theta - \vartheta_i}{\sigma_i/\sqrt{m}}\right) - 1\{\vartheta_i \leq \theta\}\right)$$

can be used for inference instead of that of only $1\{\vartheta_i \leq \theta\}$ although, again, both will be valid asymptotically.

The view of correcting the moment condition that defines $\hat{q}(\tau)$ also suggests the jackknife estimator

$$\dot{q}(\tau) := \frac{1+\lambda^2}{\lambda^2}\hat{q}(\tau) - \frac{1}{\lambda^2}\,\hat{q}_\lambda(\tau),$$

where $\hat{q}_\lambda(\tau) := \min_q\{q : \hat{F}_\lambda(q) \geq \tau\}$, again for some chosen λ. The intuition behind this jackknife correction follows from the discussion on the bias-reducing nature of $\dot{F}$ and the definition of $\hat{q}$.

4 Numerical illustrations

4.1 Normal noise

To support our theory we provide simulation results for a James and Stein (1961) problem where $\theta_i \sim N(0, \psi^2)$ and we have access to an $n \times m$ panel on independent realizations of the random variable

$$x_{it} | \, \theta_i \sim N(\theta_i, \sigma^2).$$

This setup is a simple random-coefficient model. It is similar to the classic many normal means problem of Neyman and Scott (1948). While their focus was on consistent estimation of the within-group variance, σ^2, for fixed m, our focus is on between-group characteristics and the distribution of the θ_i as a whole. We estimate θ_i by the fixed-effect estimator, i.e.,

$$\vartheta_i = m^{-1} \sum_{t=1}^{m} x_{it}.$$

The sampling variance of $\vartheta_i | \theta_i$ is σ^2/m. Rather than assuming this variance to be known we implement our procedure using the estimator

$$s_i^2 := (m-1)^{-1} \sum_{t=1}^{m} (x_{it} - \vartheta_i)^2.$$

We do not make use of the fact that the ϑ_i are homoskedastic in estimating the noise or in constructing the bias correction. Moreover, the implementation of our procedure is non-parametric in the noise distribution.

A deconvolution argument implies that

$$\vartheta_i \sim N(0, \psi^2 + \sigma^2/m).$$

Thus, indeed, the empirical distribution of the fixed-effect estimator is too fat-tailed. In particular, the sample variance of $\vartheta_1, \ldots, \vartheta_n$,

$$\hat{\psi}^2 := \frac{1}{n-1} \sum_{i=1}^{n} (\vartheta_i - \overline{\vartheta})^2, \qquad \overline{\vartheta} := n^{-1} \sum_{i=1}^{n} \vartheta_i,$$

is a biased estimator of ψ^2. To illustrate how this invalidates inference in typically-sized data sets we simulated data for $\psi^2 = 1$ (so F is standard normal) and $\sigma^2 = 5$. The panel dimensions (n, m) reported on are $(50, 3)$, $(100, 4)$, and $(200, 5)$. Table 1 shows the bias and standard deviation of $\hat{\psi}^2$ as well as the empirical rejection frequency of the usual two-sided t-test for the null that $\psi = 1$. The nominal size is set to 5%. In practice, however, the test rejects in virtually each of the $10,000$ replications. The table provides the same summary statistics for the bias-corrected estimator

$$\check{\psi}^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left((\vartheta_i - \overline{\vartheta})^2 - \frac{s_i^2}{m} \right).$$

The adjustment reduces the estimator's bias relative to its standard error and brings down the empirical rejection frequencies to just over their nominal value for the sample sizes considered.

Table 1: Variance estimation under normal noise

		bias		std		se/std		size (5%)	
n	m	$\hat{\psi}^2$	$\check{\psi}^2$	$\hat{\psi}^2$	$\check{\psi}^2$	$\hat{\psi}^2$	$\check{\psi}^2$	$\hat{\psi}^2$	$\check{\psi}^2$
50	3	1.616	-0.054	0.525	0.577	0.964	0.971	0.973	0.082
100	4	1.224	-0.028	0.321	0.337	0.966	0.969	0.997	0.073
200	5	0.989	-0.010	0.199	0.205	0.985	0.985	1.000	0.062

A popular approach in empirical work to deal with noise in $\vartheta_1, \ldots, \vartheta_n$ is shrinkage estimation (see, e.g., Chetty, Friedman and Rockoff 2014). This procedure is not designed to improve estimation and inference of F or its moments, however. In the current setting, the (infeasible, parametric) shrinkage estimator is simply

$$\left(1 - \frac{\sigma^2/m}{\sigma^2/m + \psi^2} \right) \vartheta_i.$$

Its exact sampling variance is

$$\left(\frac{\psi^2}{\sigma^2/m + \psi^2} \right) \psi^2 = \psi^2 - \frac{\sigma^2/\psi^2}{m} + o(m^{-1}).$$

It follows that the sample variance of the shrunken $\vartheta_1, \ldots, \vartheta_n$ has a bias that is of the same order as that in the sample variance of $\vartheta_1, \ldots, \vartheta_n$. Interestingly, note that, here, this

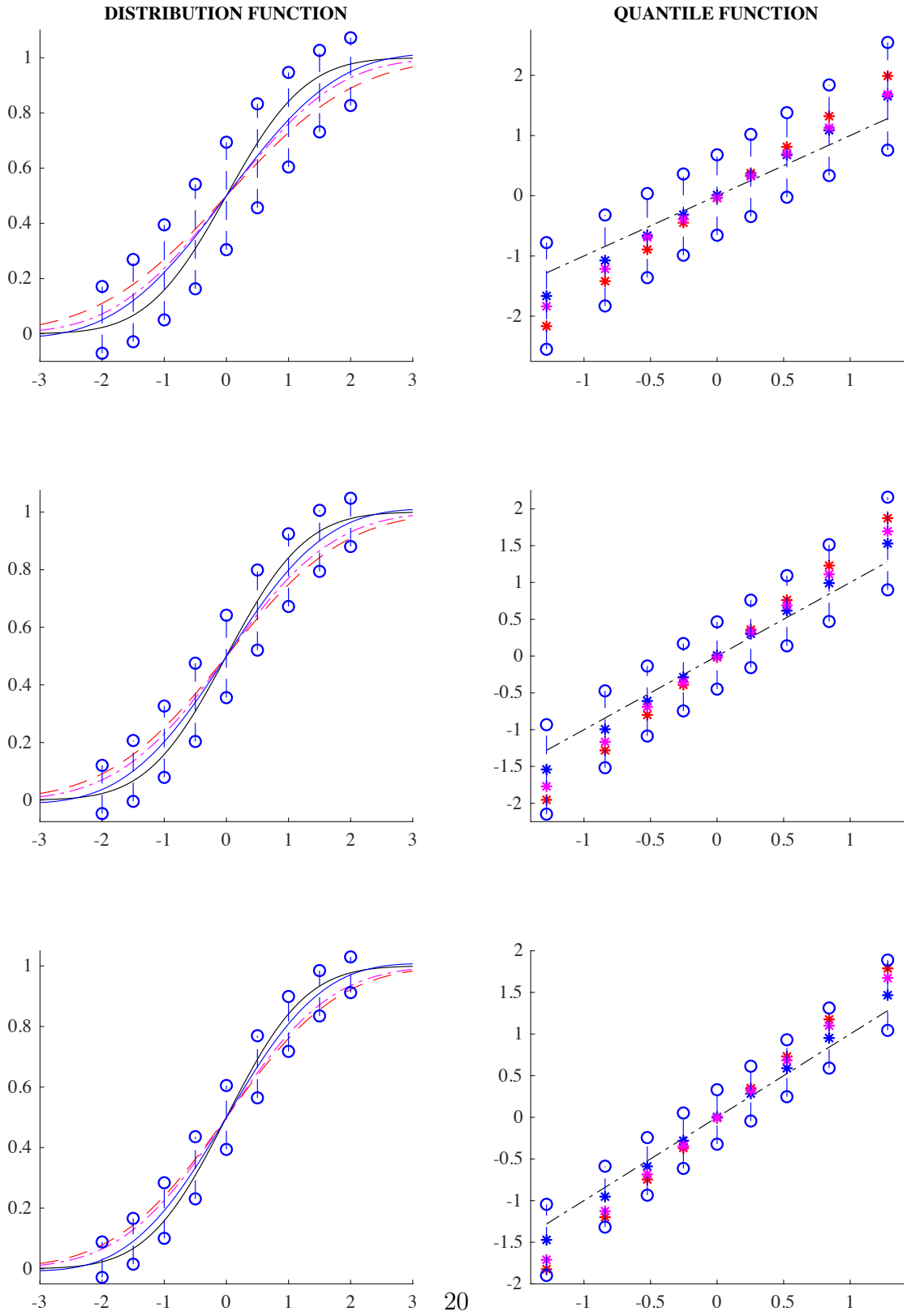
estimator overcorrects for the presence of noise, and so will be underestimating the true variance, ψ^2, on average.

The left plots in Figure 1 provide simulation results for the distribution function F for the same Monte Carlo designs. The upper-, middle-, and lower plots are for the sample size $(50, 3)$, $(100, 4)$, $(200, 5)$, respectively. Each plot contains the true curve F (black; solid) together with (the average over the Monte Carlo replications of) the naive plug-in estimator (red; dashed), the empirical distribution of the Empirical-Bayes point estimates (purple; dashed-dotted), and the analytically bias-corrected estimator (blue; solid). 95% confidence bands are placed around the latter estimator. The bandwidth in the correction term in $\check{F}$ was chosen via the cross-validation procedure discussed above. Empirical Bayes was implemented non-parametrically (and correctly assuming homoskedasticity) based on the formula stated in the introduction using a kernel estimator and the optimal bandwidth that assumes knowledge of the normality of the target distribution. Simulations results for a jackknife correction yielded very similar corrections and are omitted here for brevity (results for the jackknife can be found in previous versions of this paper).

The simulations clearly show the substantial bias in the naive estimator. This bias becomes more pronounced relative to its standard error as the sample size grows and, indeed, $\hat{F}$ starts falling outside of the confidence bands of $\check{F}$ in the middle and bottom plots. The Empirical-Bayes estimator is less biased than $\hat{F}$. However, its bias is of the same order and so, as the sample size grows it does not move toward F but, rather, towards $\hat{F}$.[4] Only $\check{F}$ is sufficiently bias-reducing. Indeed, its confidence band settles around F as the sample grows. We note that, while $\check{F}$ tends to be slightly more volatile than $\hat{F}$ in small samples, the bias-reduction outweighs this in terms of root mean squared error (RMSE). Indeed, the RMSE of $(\hat{F}, \check{F})$ across the designs are $(.0969, .0816)$, $(.0756, .0578)$, and $(.0620, .0424)$, respectively.

[4]Recall that the Empirical-Bayes estimator is not designed for inference on F but, in stead, aims to minimize risk in estimating $\theta_1, \ldots, \theta_n$. In terms of RMSE it dominates $\vartheta_1, \ldots, \vartheta_n$. For the three sample sizes considered here, the RMSEs are 1.667, 1.246, and 1.000 for the plug-in estimators and 1.233, 1.018, .874 for Empirical Bayes.

Figure 1: Estimation of F and q under normal noise



The reduction in bias is again sufficient to bring the empirical size of tests in line with their nominal size. To see this Table 2 provides empirical rejection frequencies of two-sided tests at the 5% level for F at each of its deciles using both $\hat{F}$ and $\check{F}$. The rejection frequencies based on the naive estimator are much too high for all sample sizes and deciles and get worse as the sample gets larger. Empirical size is much closer to nominal size after adjusting for noise, and this is observed at all deciles.

Table 2: Inference on F under normal noise: empirical size

τ	.1	.2	.3	.4	.5	.6	.7	.8	.9
					$(n,m)=(50,3)$				
$\hat{F}$	0.4814	0.5518	0.3695	0.1530	0.0681	0.1598	0.3801	0.5610	0.4828
$\check{F}$	0.0600	0.0928	0.1039	0.0785	0.0563	0.0745	0.1029	0.0891	0.0628
					$(n,m)=(100,4)$				
$\hat{F}$	0.6962	0.7304	0.5564	0.2280	0.0566	0.2312	0.5586	0.7352	0.7034
$\check{F}$	0.0608	0.0848	0.0920	0.0664	0.0494	0.0734	0.0932	0.0782	0.0532
					$(n,m)=(200,5)$				
$\hat{F}$	0.926	0.902	0.7634	0.3288	0.0576	0.3212	0.7646	0.903	0.9146
$\check{F}$	0.0536	0.0828	0.0996	0.0770	0.0496	0.0792	0.0978	0.0780	0.0554

The right plots in Figure 1 provide simulation results for estimators of the deciles of F. The presentation is constructed around a QQ plot of the standard normal, pictured as the black dashed-dotted line in each plot. Along the QQ plot, the average (over the Monte Carlo replications) of the naive estimator (red), Empirical Bayes (purple), and the (analytically) bias-corrected quantiles (blue) are shown by $*$ symbols. Confidence intervals around the latter (in blue,-o) are again equally provided. Like the naive estimator, the Empirical Bayes estimators are the appropriate order statistics of $\vartheta_1, \ldots, \vartheta_n$, after shrinkage has been applied to each. Visual inspection reveals that the results are in line with those obtained for the distribution function. As the sample size grows, only $\check{q}$ successfully adjusts for bias arising from estimation noise in $\vartheta_1, \ldots, \vartheta_n$. More detailed results on inference are available in a previous version of this paper.

4.2 Skew-normal noise

As stated at the beginning of the paper our approach does not hinge on the normality of the noise distribution. To illustrate this numerically we re-did the simulation exercise with errors drawn from a (shifted) skew-normal distribution (Azzalini, 1985) with mean zero, variance five, and skewness parameter equal to unity. This configuration yields a distribution that is strongly right-skewed. This departure from normality does not affect the leading bias term nor the implementation of our estimator. The skewness does imply that the remaining (higher-order) bias is not of order m^{-2} but, rather, of order $m^{-3/2}$, so that the rate requirements on the sample size in our theorems involve n/m^3 rather than n/m^4. A glance at the output in Figure 2 and Table 3 allows to verify that our corrections indeed are equally effective in this case.

Table 3: Inference on F under skew-normal noise: empirical size

τ	.1	.2	.3	.4	.5	.6	.7	.8	.9
					$(n,m) = (50,3)$				
$\hat{F}$	0.4360	0.5678	0.4298	0.2192	0.0778	0.1086	0.3168	0.5366	0.5078
$\check{F}$	0.0606	0.0834	0.0840	0.0658	0.0552	0.0858	0.1024	0.0906	0.0650
					$(n,m) = (100,4)$				
$\hat{F}$	0.6416	0.7480	0.6164	0.3032	0.0564	0.1604	0.5080	0.7404	0.7412
$\check{F}$	0.0548	0.0948	0.0876	0.0592	0.0560	0.0764	0.1080	0.0728	0.0488
					$(n,m) = (200,5)$				
$\hat{F}$	0.8810	0.8944	0.7958	0.4026	0.0626	0.2480	0.7234	0.8990	0.9368
$\check{F}$	0.0590	0.0754	0.0836	0.0590	0.0526	0.0876	0.1042	0.0806	0.0456

4.3 Estimating proportions

A nonlinear example is the estimation of proportions. Let $\theta_i \sim \text{uniform}[0,1]$ represent success probabilities. Given a series of m Bernoulli experiments the maximum-likelihood estimator of θ_i is the success probability in the sample, ϑ_i. Here, $m\,\vartheta_i \sim \text{Binomial}(m,\theta_i)$, and so ϑ_i is unbiased and has variance $\sigma_i^2/m = \theta_i(1-\theta_i)/m$, which is a deterministic

Figure 2: Estimation of F and q under skew-normal noise

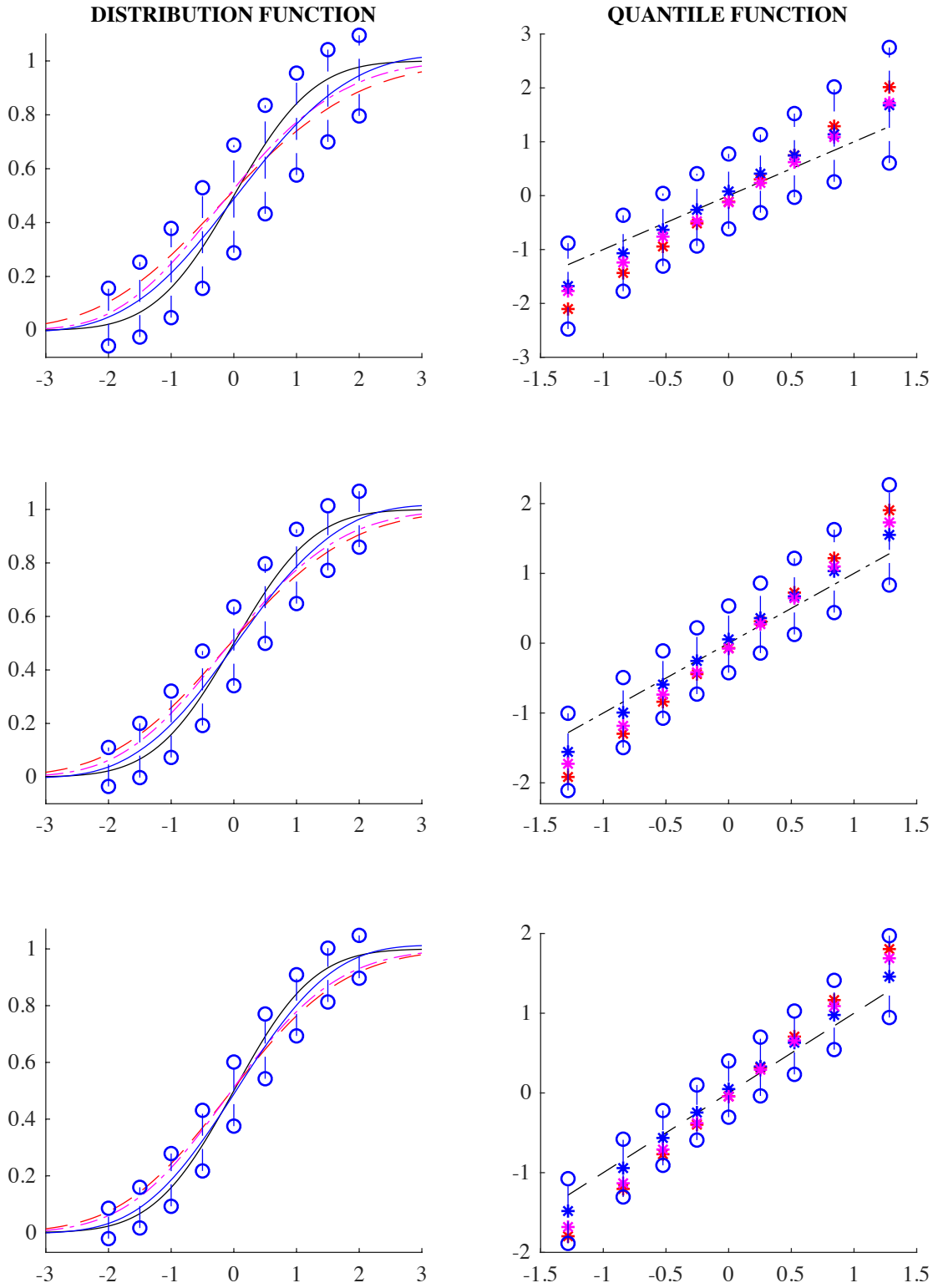
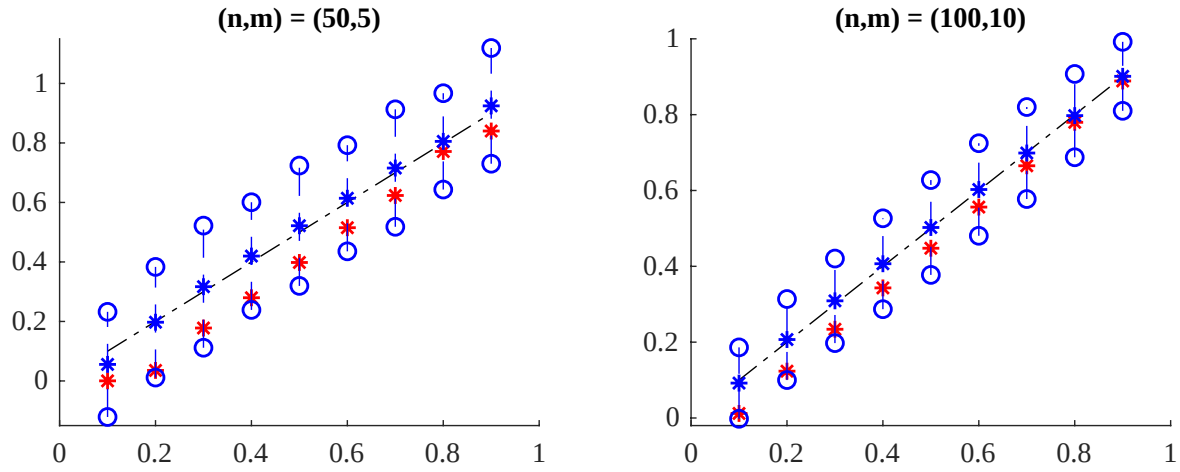


Figure 3: Estimation of q from empirical frequencies



function of θ_i. To evaluate the performance of our approach in this nonlinear problem we provide descriptive statistics for the estimator of the quantile function of the success probabilities in Figure 3. The plots, for $(n, m) = (50, 5)$ (left) and $(n, m) = (100, 10)$ (right) have the same layout as before (although we do not provide results for an Empirical Bayes estimator here). The results reveal that the order statistics are all downward biased and that our correction is near unbiased at all deciles.

5 Conclusions

In this paper, we have considered inference on the distribution of latent variables from noisy measurements. In an asymptotic embedding where the variance of the noise shrinks with the sample size, we have derived the leading bias in the empirical distribution function of the noisy measurements and suggested both an analytical and a jackknife correction. These estimators are straightforward to implement. Moreover, they provide a simple and numerically stable (approximate) solution to a generalized deconvolution problem that, in addition, yields valid inference procedures.

A Appendix

A.1 Proof of Proposition 1

The following known result is useful to prove Proposition 1.

Lemma A.1 (Komlós, Major and Tusnády 1975)**.** *Let $\mathbb{G}_n$ denote the empirical cumulative distribution of an i.i.d. sample of size n from a uniform distribution on [0,1]. Let $\mathbb{B}_n$ denote a sequence of Brownian bridges. Then*

$$\sup_{u\in[0,1]} \left|\sqrt{n}\left(\mathbb{G}_n(u) - u\right) - \mathbb{B}_n(u)\right| = O_p(\log(n)/\sqrt{n}).$$

Proof of Proposition 1. Suppose, first, that Assumption 2.A holds. Then (θ_i, σ_i) have a joint density, $h(\theta_i, \sigma_i)$. We will denote the marginal density of σ_i by $h(\sigma_i)$ and the conditional density of θ_i given σ_i by $h(\theta_i|\sigma_i)$. For any real number δ let

$$G(\theta, \delta) := E(1\{\theta_i + \delta\sigma_i \leq \theta\}) = \int_{\underline{\sigma}}^{\overline{\sigma}} \int_{-\infty}^{\theta-\delta\sigma} h(\theta, \sigma)\, d\theta\, d\sigma.$$

Note that $G(\theta, 0) = F(\theta)$ and that

$$E(\hat{F}(\theta)) = E(1\{\vartheta_i \leq \theta\}) = E\left(1\left\{\theta_i + \frac{\varepsilon_i}{\sqrt{m}}\sigma_i \leq \theta\right\}\right) = E\left(G(\theta, \varepsilon_i/\sqrt{m})\right). \tag{A.1}$$

Assumption 2.A implies that G is smooth and differentiable in its second argument. By definition of the function $e(\sigma_i)$,

$$\sup_\theta \sup_\delta |\nabla_2^4 G(\theta, \delta)| = \sup_\theta \sup_\delta \left|\int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^4\, \nabla_1^3 h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma\right| \leq \int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^4\, e(\sigma)\, h(\sigma)\, d\sigma, \tag{A.2}$$

which equals $E(\sigma_i^4 e(\sigma_i))$ and is finite by assumption. Therefore, by (A.1) and a fourth-order expansion of $G(\theta, \varepsilon_i/\sqrt{m})$ in its second argument around zero we find that

$$E(\hat{F}(\theta)) = F(\theta) + \frac{1}{2}\frac{\nabla_2^2 G(\theta, 0)}{m} + \frac{1}{24}\frac{E(\varepsilon_i^4\, \nabla_2^4 G(\theta, \varepsilon_i^*/\sqrt{m}))}{m^2},$$

where ε_i^* is some value between zero and ε_i, and where, in addition to (A.2), we have used that $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = 1$ by construction, and that $E(\varepsilon_i^3) = 0$ and $E(\varepsilon_i^4) < \infty$ by assumption. By direct calculation,

$$\nabla_2^2 G(\theta, 0) = 2\, b_F(\theta).$$

Therefore,

$$E(\hat{F}(\theta)) = F(\theta) + \frac{b_F(\theta)}{m} + O(m^{-2})$$

under Assumption 2.A, as claimed.

Next, suppose that Assumption 2.B holds. Then we have a deterministic relationship between θ_i and σ_i. We may define $G(\theta, \delta)$ as above but have to take care when Taylor expanding in δ, as the function may be non-continuous. A non-continuity occurs whenever the number of solutions t (on the real line) to the equation $t + \delta\sigma(t) = \theta$ changes. However, at $\delta = 0$ the only solution to this equation is $t = \theta$, and because we assume that the function $\sigma(\theta)$ has uniformly bounded derivative σ', there always exists $\eta > 0$ such that for all $\delta \in (-\eta, \eta)$ and all real θ the equation $t + \delta\sigma(t) = \theta$ has a unique solution in t on the real line. We denote this solution by $t^*(\theta, \delta)$, that is, we have $t^*(\theta, \delta) + \delta\sigma(t^*(\theta, \delta)) = \theta$. Using this we find that for $\delta \in (-\eta, \eta)$ we have

$$G(\theta, \delta) = F(t^*(\theta, \delta)), \qquad \nabla_2^1 t^*(\theta, \delta) = -\frac{\sigma(t^*(\theta, \delta))}{1 + \delta\, \sigma'(t^*(\theta, \delta))},$$

where the last equation is obtained by taking derivatives of $t^*(\theta, \delta) + \delta\sigma(t^*(\theta, \delta)) = \theta$ with respect to δ and then solving for the derivative. Because we have that $t^*(\theta, 0) = \theta$ we then find

$$G(\theta, 0) = F(\theta), \qquad \nabla_2^1 G(\theta, 0) = -\sigma(\theta) f(\theta), \qquad \nabla_2^2 G(\theta, 0) = 2b_F(\theta).$$

Differentiating further we see that $\nabla_2^3 G(\theta, 0)$, and $\nabla_2^4 G(\theta, 0)$ are functions of the derivatives of f and σ up to third and fourth order, resepctively, our assumption that these derivatives are uniformly bounded implies that

$$\sup_{\theta} \sup_{\delta \in (-\eta, \eta)} \left| \nabla_2^4 G(\theta, \delta) \right| < \infty \tag{A.3}$$

for some $\eta > 0$. The only obstacle that now prevents us from proceeding with an expansion as we did under Assumption 2.A is that the bound (A.3) is restricted to a neighborhood around zero.

To complete the proof we argue that the restriction that $\delta \in (-\eta, \eta)$ relaxes sufficiently fast as m grows. We do so as follows. First, note that we still have

$$E\hat{F}(\theta) = E\left(G(\theta, \varepsilon_i/\sqrt{m})\right).$$

Because $E(\varepsilon_i^4) < \infty$, an application of Markov's inequality yields

$$P\left(|\varepsilon_i| > m^\alpha\right) = O(m^{-4/\alpha}),$$

as $m \to \infty$, for any $\alpha > 0$. In the following let $\alpha \in (0, 1/2)$. We have

$$\begin{aligned} E(\hat{F}(\theta)) &= E\left(1\left\{|\varepsilon_i| \leq m^\alpha\right\} G\left(\theta, \frac{\varepsilon_i}{\sqrt{m}}\right)\right) + E\left(1\left\{|\varepsilon_i| > m^\alpha\right\} G\left(\theta, \frac{\varepsilon_i}{\sqrt{m}}\right)\right) \\ &= E\left(1\left\{|\varepsilon_i| \leq m^\alpha\right\} G\left(\theta, \frac{\varepsilon_i}{\sqrt{m}}\right)\right) + o(m^{-2}), \end{aligned}$$

uniformly in θ. This follows from the observation that

$$\sup_\theta E\left(1\left\{|\varepsilon_i| > m^\alpha\right\} G\left(\theta, \frac{\varepsilon_i}{\sqrt{m}}\right)\right) \leq P\left(|\varepsilon_i| > m^\alpha\right) = O(m^\alpha) = o(m^{-2}),$$

where we have used the fact that $G(\theta, \delta)$ is restricted to the unit interval. A Taylor expansion gives

$$E(\hat{F}(\theta)) = G(\theta, 0) + E(\varepsilon_i)\frac{\nabla_2^1 G(\theta, 0)}{m^{1/2}} + \frac{E(\varepsilon_i^2)}{2}\frac{\nabla_2^2 G(\theta, 0)}{m} + \frac{E(\varepsilon_i^3)}{6}\frac{\nabla_2^3 G(\theta, 0)}{m^{3/2}} + r(\theta) + o(m^{-2}),$$

where we let

$$r(\theta) := r_2(\theta) - r_1(\theta)$$

for

$$\begin{aligned} r_1(\theta) := P(|\varepsilon_i| > m^\alpha)\, G(\theta, 0) \\ + E(1\{|\varepsilon_i| > m^\alpha\}\, \varepsilon_i)\, \frac{\nabla_2^1 G(\theta, 0)}{m^{1/2}} \\ + \frac{E(1\{|\varepsilon_i| > m^\alpha\}\, \varepsilon_i^2)}{2}\, \frac{\nabla_2^2 G(\theta, 0)}{m} \\ + \frac{E(1\{|\varepsilon_i| > m^\alpha\}\, \varepsilon_i^3)}{6}\, \frac{\nabla_2^3 G(\theta, 0)}{m^{3/2}} \end{aligned}$$

and

$$r_2(\theta) := m^{-2}\, \frac{E\left(1\{|\varepsilon_i| \leq m^\alpha\}\, \varepsilon_i^4\, \nabla_2^4 G(\theta, \varepsilon_i^*/\sqrt{m})\right)}{24}$$

for random variables ε_i between zero and ε_i. As all relevant derivatives are bounded we have

$$\sup_\theta |r_1(\theta)| = o(m^{-2}) \sup_\theta \left(1 + \left|\nabla_2^1 G(\theta,0)\right| + \left|\nabla_2^2 G(\theta,0)\right| + \left|\nabla_2^3 G(\theta,0)\right|\right) = o(m^{-2}).$$

Also, using (A.3) we obtain, with $\rho := 1/2 - \alpha > 0$,

$$\sup_\theta |r_2(\theta)| \leq m^{-2} \frac{E(\varepsilon_i^4)}{24} \sup_{\delta \in (-m^{-\rho}, m^\rho)} |\nabla_2^4 G(\theta,\delta)| = O(m^{-2}).$$

Hence, $\sup_\theta |r(\theta)| = O(m^{-2})$. We then immediately obtain that

$$E(\hat{F}(\theta)) = F(\theta) + \frac{b_F(\theta)}{m} + O(m^{-2})$$

uniformly in θ. This completes the proof of the bias expression under Assumption 2.B.

For the result on the covariance, note that

$$\operatorname{cov}(\hat{F}(\theta_1), \hat{F}(\theta_2)) = \frac{E(\hat{F}(\theta_1 \wedge \theta_2)) - E(\hat{F}(\theta_1))\, E(\hat{F}(\theta_2))}{n}$$

depends only on $E(\hat{F}(\theta))$ which, up to $O(m^{-2})$ and uniformly in θ, has been calculated above. Moreover,

$$\begin{aligned}
\operatorname{cov}(\hat{F}(\theta_1), \hat{F}(\theta_2)) &= \frac{(F(\theta_1 \wedge \theta_2) + O(m^{-1})) - (F(\theta_1) + O(m^{-1}))\,(F(\theta_2) + O(m^{-1}))}{n} \\
&= \frac{F(\theta_1 \wedge \theta_2) - F(\theta_1)F(\theta_2)}{n} + O(n^{-1}m^{-1}) \\
&= \frac{\sigma_F(\theta_1, \theta_2)}{n} + O(n^{-1}m^{-1}),
\end{aligned}$$

as stated in the proposition.

To complete the proof it remains only to verify the limit distribution of the scaled empirical distribution function. Let $F_m(\theta) := E(1\{\vartheta_i \leq \theta\})$, the distribution function of ϑ_i. Our assumptions imply that F_m is continuous and that it has no mass points. With $u_i := F_m(\vartheta_i)$, we therefore have that u_i is i.i.d. uniformly distributed on $[0,1]$ by the probability integral transform. An application of Lemma A.1 with $u = F_m(\theta)$ and exploiting monotonicity of distribution functions then gives

$$\sup_\theta \left| \sqrt{n}(\hat{F}(\theta) - F_m(\theta)) - \mathbb{B}_n(F_m(\theta)) \right| = O_p(\log(n)/\sqrt{n}).$$

We have already shown that, uniformly in θ,

$$F_m(\theta) = F(\theta) + \frac{b_F(\theta)}{m} + O(m^{-2}).$$

Therefore, using that $n/m^4 \to 0$,

$$\sqrt{n}(\hat{F}(\theta) - F_m(\theta)) = \sqrt{n}\left(\hat{F}(\theta) - F(\theta) - \frac{b_F(\theta)}{m}\right) + o(1),$$

holds uniformly in θ. Furthermore, our bias calculation implies that $F_m(\theta) - F(\theta)$ converges to zero uniformly in θ as $m \to 0$, so that applying Lévy's modulus-of-continuity theorem, that is,

$$\lim_{\epsilon \to 0} \sup_{t \in [0,1-\epsilon]} \frac{|\mathbb{B}_n(t) - \mathbb{B}_n(t+\epsilon)|}{\sqrt{\epsilon \log(1/\epsilon)}} = O(1), \qquad \epsilon > 0,$$

to our problem yields $\sup_\theta |\mathbb{B}_n(F_m(\theta)) - \mathbb{B}_n(F(\theta))| \xrightarrow{p} 0$ as $m \to \infty$. We thus have that $\mathbb{B}_n(F_m(\theta)) \rightsquigarrow \mathbb{B}_n(F(\theta))$. Putting everything together and noting that, by definition, $\mathbb{B}_n(F(\theta)) = \mathbb{G}_F(\theta)$, we obtain

$$\sup_\theta \left| \sqrt{n}\left(\hat{F}(\theta) - F(\theta) - \frac{b_F(\theta)}{m}\right) - \mathbb{G}_F(\theta) \right| = o_p(1),$$

which completes the proof of the proposition. □

Proof of Proposition 2

Lemma A.2. *Let Assumptions 1 and 2 hold. Let f_m denote the density function of ϑ_i. Then,*

(i) $\sup_\theta |f_m(\theta) - f(\theta)| = O(m^{-1})$,

(ii) $\sup_\theta |\nabla^1 f_m(\theta) - \nabla^1 f(\theta)| = O(m^{-1})$,

(iii) $\sup_\theta |\nabla^2 f_m(\theta) - \nabla^2 f(\theta)| = O(1)$,

(iv) $\sup_\theta |\nabla^3 f_m(\theta) - \nabla^3 f(\theta)| = O(1)$.

Proof. For brevity, we only show the result on Assumption 2.A. From the argument in the proof of Proposition 1 we have

$$F_m(\theta) - F(\theta) = \frac{1}{2}\frac{E(\varepsilon_i^2\, H(\theta, \varepsilon_i^*/\sqrt{m}))}{m}$$

by a second-order expansion, where ε_i^* is a value between zero and ε_i and we introduce the function

$$H(\theta,\delta) := \int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^2 \nabla_1^1 h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma,$$

where $h(\theta_i|\sigma_i)$ and $h(\sigma_i)$ are the density functions of θ_i given σ_i and of σ_i, respectively. Differentiating with respect to θ yields the first conclusion of the lemma as

$$\sup_\theta |f_m(\theta) - f(\theta)| = \sup_\theta \left| \frac{1}{2} \frac{E(\varepsilon_i^2\, \nabla_1^1 H(\theta, \varepsilon_i^*/\sqrt{m}))}{m} \right| \leq \frac{E(\sigma_i^2)}{m}\, \frac{\sup_\theta \sup_\delta |\nabla_1^1 H(\theta,\delta)|}{2} = O(m^{-1}),$$

which follows from the inequality

$$\sup_\theta \sup_\delta |\nabla_1^1 H(\theta,\delta)| = \sup_\theta \sup_\delta \left| \int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^3\, \nabla_1^2\, h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma \right| \leq \int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^3\, e(\sigma)\, h(\sigma)\, d\sigma < \infty$$

and the definition of the function $e(\sigma)$ in Assumption 2.A. The second conclusion of the lemma follows in the same manner, differentiating once more. Finally, the third and fourth conclusion are obtained similarly. The point of departure is now the following identity, which is derived in the proof of Proposition 1,

$$F_m(\theta) = E\left(G(\theta, \varepsilon_i^*/\sqrt{m})\right)$$

where

$$G(\theta,\delta) := \int_{\underline{\sigma}}^{\overline{\sigma}} \int_{-\infty}^{\theta - \delta\sigma} h(\vartheta|\sigma)\, h(\sigma)\, d\vartheta\, d\sigma.$$

Repeated differentiation shows that

$$\sup_\theta \sup_\delta |\nabla_1^3 G(\theta,\delta)| = \sup_\theta \sup_\delta \left| \int_{\underline{\sigma}}^{\overline{\sigma}} \nabla_1^2 h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma \right| \leq \left| \int_{\underline{\sigma}}^{\overline{\sigma}} e(\sigma)\, h(\sigma)\, d\sigma \right| < \infty,$$

$$\sup_\theta \sup_\delta |\nabla_1^4 G(\theta,\delta)| = \sup_\theta \sup_\delta \left| \int_{\underline{\sigma}}^{\overline{\sigma}} \nabla_1^3 h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma \right| \leq \left| \int_{\underline{\sigma}}^{\overline{\sigma}} e(\sigma)\, h(\sigma)\, d\sigma \right| < \infty,$$

and so $\sup_\theta |\nabla^3 F_m(\theta)| = O(1)$ and $\sup_\theta |\nabla^4 F_m(\theta)| = O(1)$ follow. Furthermore,

$$\sup_\theta |\nabla^2 f_m(\theta) - \nabla^2 f(\theta)| \leq \sup_\theta |\nabla^2 f_m(\theta)| + \sup_\theta |\nabla^2 f(\theta)| = O(1),$$

$$\sup_\theta |\nabla^3 f_m(\theta) - \nabla^3 f(\theta)| \leq \sup_\theta |\nabla^3 f_m(\theta)| + \sup_\theta |\nabla^3 f(\theta)| = O(1),$$

follows because f has uniformly bounded derivatives up to third order by assumption. This completes the proof. □

Proof of Proposition 2. The ϑ_i are i.i.d. draws from the distribution F_m which according to Lemma A.2 has non-degenerate density f_m, that is, the ϑ_i are continuously distributed. Thus,

$$u_{(k)} := F_m(\vartheta_{(k)})$$

is the kth order statistic of a uniform sample. We set $k = \lceil \tau n \rceil$ for the rest of the proof. Then $\hat{q}(\tau) = \vartheta_{(k)}$. Since $k/n \to \tau$ by construction, it is well-known that

$$\sqrt{n}(u_{(k)} - \tau) \xrightarrow{d} N(0, \tau(1-\tau)). \tag{A.4}$$

Let $q_m(\tau) := F_m^{-1}(\tau)$, the τth-quantile of F_m. By expanding the function F_m^{-1} around τ we find that

$$\hat{q}(\tau) = F_m^{-1}(u_{(k)}) = q_m(\tau) + \frac{u_{(k)} - \tau}{f_m(q_m(\tau))} + r_{(k)}$$

for remainder term

$$r_{(k)} := -\frac{f_m'(\xi_{(k)})}{f_m(\xi_{(k)})^3} \left(u_{(k)} - \tau\right)^2,$$

where $\xi_{(k)}$ is a value between $F_m^{-1}(\tau)$ and $F_m^{-1}(u_{(k)})$. From (A.4) we have $u_{(k)} - \tau = O_P(n^{-1/2})$. This implies that $\xi_{(k)} \xrightarrow{p} \tau$. Using Lemma A.2 we may conclude that $f_m(\xi_{(k)}) \xrightarrow{p} f_m(\tau) \to f(\tau) > 0$, and, therefore, that $r_{(k)} = O_p(n^{-1})$. We thus have

$$\hat{q}(\tau) = q_m(\tau) + \frac{u_{(k)} - \tau}{f_m(q_m(\tau))} + O_p(n^{-1}).$$

Again using Lemma A.2 and our assumption that $f(\theta) > 0$ in a neighborhood of $q(\tau) = F^{-1}(\tau)$ we have $f_m(q_m(\tau))^{-1} = f(q(\tau))^{-1} + O(m^{-1})$, and therefore

$$\hat{q}(\tau) = q_m(\tau) + \frac{u_{(k)} - \tau}{f(q(\tau))} + O_p(n^{-1} + n^{-1/2}m^{-1}). \tag{A.5}$$

From Proposition 1 we know $F_m(\theta) = E(\hat{F}(\theta)) = F(\theta) + b_F(\theta)/m + O(m^{-2})$, and therefore

$$q_m(\tau) = q(\tau) - \frac{b_F(q(\tau))/f(q(\tau))}{m} + O(m^{-2}). \tag{A.6}$$

Combining (A.4), (A.5), and (A.6) gives the statement of the theorem. □

Proof of Proposition 3

Lemma A.3. *Let the assumptions of Proposition 3 hold. Then,*

(i) $\sup_\theta E(\hat{b}_F(\theta) - b_F(\theta)) = O(m^{-1}) + O(h^2)$,

(ii) $\sup_\theta \operatorname{var}(\hat{b}_F(\theta)) = O(n^{-1}h^3)$,

(iii) $(1+|\theta|^{1+\eta})|\nabla^1 \hat{b}_F(\theta) - \nabla^1 b_F(\theta)| = O_p(h^{-(\omega+1)/\omega})$.

Lemma A.4. *Let Assumptions 1 hold and define*

$$b_i(\theta) := -\frac{\sigma_i^2}{h^2}\frac{\phi'\left(\frac{\vartheta_i-\theta}{h}\right)}{2}.$$

If f is bounded, then, for any $\epsilon > 0$,

$$\sup_\theta E(|b_i(\theta) - E(b_i(\theta))|^\epsilon)^{1/\epsilon} = O(h^{-2+\epsilon^{-1}}).$$

The proof of those two lemmas is provided below, after the proof of the main text results.

Proof of Proposition 3. We first show that

$$\sup_{\theta\in\mathbb{R}} \left|\hat{b}_F(\theta) - b_F(\theta)\right| = O(m^{-1}) + O(h^2) + O(n^{-1/2}\, h^{-3/2-\varepsilon}).$$

The result of the proposition then follows readily. For a finite ν, introduce the function

$$t(\theta) := \operatorname{sgn}(\theta)\,\frac{1-(1+|\theta|)^{-\nu}}{\nu}.$$

Note that t maps to the finite interval $(-\nu^{-1}, \nu^{-1})$ and is monotone increasing; moreover, $\nabla^1 t(\theta) = (1+|\theta|)^{-(1+\nu)}$. Now consider the reparametrization $\tau = t(\theta)$; note that τ lives in a bounded interval. From Lemma A.3(iii), using the chain rule of differentiation, it follows that

$$\sup_{\tau\in(-\nu^{-1},\nu^{-1})} \left|\nabla_\tau^1 \hat{b}_F(t^{-1}(\tau)) - \nabla_\tau^1 b_F(t^{-1}(\tau))\right| = O_p(h^{-(1+\omega^{-1})}), \tag{A.7}$$

where we use the notation ∇_τ to indicate derivatives with respect to τ. We therefore have that $\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))$, as a function τ, has a uniformly-bounded Lipschitz constant.

Now let I_h be a partition of $(-\nu, -\nu^{-1})$ with subintervals that are (approximately) of length $l_h := h^{3-\omega^{-1}}$. Then (A.7) implies that

$$\sup_\theta |\hat{b}_F(\theta) - b_F(\theta)| = \sup_{\tau \in (-\nu,\nu)} |\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))|$$

is equal to

$$\max_{\tau \in I_h} |\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))| + O_p(h^2). \tag{A.8}$$

Here, the order of the remainder terms follows from the choice of l_h. Now introduce the shorthand

$$\hat{\Delta}(\theta) := \hat{b}_F(\theta) - E(\hat{b}_F(\theta)).$$

Then

$$\max_{\tau \in I_h} |\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))| \leq \max_{\tau \in I_h} |\hat{\Delta}(t^{-1}(\tau))| + \sup_\theta |E(\hat{b}_F(\theta)) - b_F(\theta)|$$

and so Lemma A.3(i) implies that

$$\max_{\tau \in I_h} |\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))| \leq \max_{\tau \in I_h} |\hat{\Delta}(t^{-1}(\tau))| + O(m^{-1} + h^2).$$

Moving on, observe that the number of subintervals making up I_h is equal to $\lceil l_h^{-1} \rceil = \lceil h^{-3+\omega^{-1}} \rceil$, where $\lceil a \rceil$ delivers the smallest integer at least as large as a. We therefore have

$$\begin{aligned} E\left(\left(\max_{\tau \in I_h} \left|\hat{\Delta}(t^{-1}(\tau))\right|\right)^\omega\right) &= E\left(\max_{\tau \in I_h} \left|\hat{\Delta}(t^{-1}(\tau))\right|^\omega\right) \\ &\leq E\left(\sum_{\tau \in I_h} \left|\hat{\Delta}(t^{-1}(\tau))\right|^\omega\right) \\ &= \sum_{\tau \in I_h} E\left(\left|\hat{\Delta}(t^{-1}(\tau))\right|^\omega\right) \leq \lceil h^{-3+1/\omega} \rceil \sup_{\theta \in \mathbb{R}} E\left|\hat{\Delta}(\theta)\right|^\omega. \end{aligned} \tag{A.9}$$

Let $b_i(\theta) := -\frac{1}{2} h^{-2} \sigma_i^2 \phi'\left(\frac{\vartheta_i - \theta}{h}\right)$ and $\Delta_i(\theta) := b_i(\theta) - E b_i(\theta)$. We may then write $\hat{\Delta}(\theta) = n^{-1} \sum_{i=1}^n \Delta_i(\theta)$. Notice that $\Delta_i(\theta)$ are independent and mean zero. By Rosenthal (1970, Theorem 3) we therefore have that

$$\left(E\left(\left|n^{-1/2} \sum_{i=1}^n \Delta_i(\theta)\right|^\omega\right)\right)^{1/\omega}$$

is bounded from above by

$$c \max \left\{ \left(n^{-1} \sum_{i=1}^{n} E\left(\Delta_i(\theta)^2\right) \right)^{1/2}, \; n^{-1/2} \left(\sum_{i=1}^{n} E\left(|\Delta_i(\theta)|^{\omega}\right) \right)^{1/\omega} \right\},$$

where the constant c only depends on ω. Using Lemma A.3(ii) we obtain

$$\sup_{\theta\in\mathbb{R}} \left(n^{-1} \sum_{i=1}^{n} E(\Delta_i(\theta)^2) \right)^{1/2} = \sup_{\theta\in\mathbb{R}} \left(n \operatorname{var} \hat{b}_F(\theta) \right)^{1/2} = O(h^{-3/2}).$$

Using Lemma A.4 we obtain

$$\begin{aligned} n^{-1/2} \sup_{\theta\in\mathbb{R}} \left(\sum_{i=1}^{n} E\left(|\Delta_i(\theta)|^{\omega}\right)^{1/\omega} \right) &= n^{-1/2+1/\omega} \sup_{\theta\in\mathbb{R}} \left(E\,|\Delta_i(\theta)|^{\omega}\right)^{1/\omega} \\ &= O(n^{-1/2+1/\omega}\, h^{-2+1/\omega}) = O(h^{-3/2}), \end{aligned}$$

where in the last step we used the condition that $h^{-1} = O(n)$. We can therefore conclude from Rosenthal's inequality above that

$$\left(\sup_{\theta\in\mathbb{R}} E\left(|\hat{\Delta}(\theta)|^{\omega}\right) \right)^{1/\omega} = n^{-1/2} \left(E\left(\left| n^{-1/2} \sum_{i=1}^{n} \Delta_i(\theta) \right|^{\omega} \right) \right)^{1/\omega} = O(n^{-1/2} h^{-3/2}).$$

Using this and (A.9) we obtain

$$\max_{\tau\in I_h} \left| \hat{\Delta}(t^{-1}(\tau)) \right| = O(h^{(-3+1/\omega)/\omega}\, n^{-1/2}\, h^{-3/2}) = O(n^{-1/2}\, h^{-3/2-\varepsilon}),$$

where $\varepsilon = 3/\omega - 1/\omega^2$. Combining this with (A.8) and (A.9) we thus conclude

$$\sup_{\theta\in\mathbb{R}} \left| \hat{b}_F(\theta) - b_F(\theta) \right| = O(m^{-1}) + O(h^2) + O(n^{-1/2}\, h^{-3/2-\varepsilon}),$$

as claimed.

Now, with $h = O(m^{-1/2})$ and $h^{-1} = O(n^{1-2\omega^{-1}})$ we find

$$\begin{aligned} \sup_{\theta\in\mathbb{R}} \frac{\sqrt{n}}{m} \left| \hat{b}_F(\theta) - b_F(\theta) \right| &= O_P(n^{1/2} m^{-1} h^2 + n^{1/2} m^{-2} + m^{-1} h^{-3/2-\varepsilon}) \\ &= O_P(n^{1/2} m^{-2} + m^{-4/9\epsilon^2}) \\ &= o_P(1), \end{aligned}$$

where in the last step we also used that $n/m^4 \to 0$ and that $m \to \infty$. The result of Proposition 3 now follows immediately from Proposition 1. □

Proof of Proposition 4

Let $\mathbb{G}_n(u) := \hat{F}(F_m^{-1}(u))$ be the empirical distribution function of the i.i.d. sample $u_i = F_m(\vartheta_i)$. Lemma A.1 and Theorem 1 in Doss and Gill (1992) give

$$\sup_{\tau\in[0,1]} \left|\sqrt{n}\left(\mathbb{G}_n^{\leftarrow}(\tau) - \tau\right) + \mathbb{B}_n(\tau)\right| = o_P(1), \tag{A.10}$$

where $\mathbb{G}_n^{\leftarrow}$ again denotes the left inverse of $\mathbb{G}_n$ $\mathbb{B}_n(\tau)$ is the sequence of Brownian bridges that previously appeared in Lemma A.1.

Equation (A.10) yields

$$\mathbb{G}_n^{\leftarrow}(\hat{\tau}^*) - \mathbb{G}_n^{\leftarrow}(\tau) = (\hat{\tau}^* - \tau) - n^{-1/2}\left[\mathbb{B}_n(\hat{\tau}^*) - \mathbb{B}_n(\tau)\right] + o_p(n^{-1/2}).$$

Also, $\hat{\tau}^* - \tau = O_p(m^{-1})$ follows from the results above. Lévy's modulus-of-continuity theorem then implies that $\mathbb{B}_n(\hat{\tau}^*) - \mathbb{B}_n(\tau) = o_P(1)$. Therefore,

$$\mathbb{G}_n^{\leftarrow}(\hat{\tau}^*) - \mathbb{G}_n^{\leftarrow}(\tau) = O_p(m^{-1}) + o_p(n^{-1/2}).$$

By definition we have $\check{q}(\tau) = \hat{F}^{\leftarrow}(\hat{\tau}^*)$ and $\hat{q}(\tau) = \hat{F}^{\leftarrow}(\tau)$, and also that $\mathbb{G}_n^{\leftarrow}(\tau) = F_m(\hat{F}^{\leftarrow}(\tau))$. Substituting this into the last displayed equation yields

$$F_m(\check{q}(\tau)) - F_m(\hat{q}(\tau)) = O_p(m^{-1}) + o_p(n^{-1/2}).$$

Lemma A.2 and our assumptions guarantee that $F_m(\tau)$ has a density $f_m(\tau)$ that is bounded from below in a neighborhood of $q(\tau)$ for the quantile of interest τ. The last result therefore also implies that

$$\check{q}(\tau) - \hat{q}(\tau) = O_p(m^{-1}) + o_p(n^{-1/2}). \tag{A.11}$$

Next, The result (A.10) implies $\sqrt{n}(\mathbb{G}_n^{\leftarrow}(\tau) - \tau) \rightsquigarrow \mathbb{B}(\tau)$ for a Brownian bridge $\mathbb{B}$. For $\check{q}(\tau) = \hat{F}^{\leftarrow}(\hat{\tau}^*)$ we have $F_m(\check{q}(\tau)) = \mathbb{G}_n^{\leftarrow}(\hat{\tau}^*)$, and therefore

$$\sqrt{n}(F_m(\check{q}(\tau)) - \hat{\tau}^*) \rightsquigarrow \mathbb{B}(\tau).$$

From Proposition 1 we know that $F_m(\theta) = E(\hat{F}(\theta)) = F(\theta) + b_F(\theta)/m + O(m^{-2})$, uniformly in θ. We then find

$$\sqrt{n}\left(F(\check{q}(\tau)) - \tau + \frac{b_F(\check{q}(\tau)) - \hat{b}_F(\hat{q}(\tau))}{m} + O(m^{-2})\right) \xrightarrow{d} N(0, \tau(1-\tau)),$$

From the proof of Proposition 3 we also know that $\sup_\theta(\sqrt{n}/m)\left|\hat{b}_F(\theta) - b_F(\theta)\right| = o_p(1)$, and therefore

$$\sqrt{n}\left(F(\check{q}(\tau)) - \tau + \frac{b_F(\check{q}(\tau)) - b_F(\hat{q}(\tau))}{m} + O(m^{-2})\right) \xrightarrow{d} N(0, \tau(1-\tau)).$$

Smoothness of the function b_F and (A.11) imply $b_F(\check{q}(\tau)) - b_F(\hat{q}(\tau)) = O(m^{-1}) + o_p(n^{-1/2})$. We thus obtain $\sqrt{n}\,(F(\check{q}(\tau)) - \tau) \xrightarrow{d} N(0, \tau(1-\tau))$ An application of the delta method with transformation F^{-1} then gives the result. This completes the proof. □

Derivation of the least-squares cross validation objective function

The integrated squared error of

$$\check{F}(\theta) = \hat{F}(\theta) - \frac{\hat{b}_F(\theta)}{m}$$

is

$$\int(\check{F}(\theta) - F(\theta))^2\,d\theta = \frac{\int \hat{b}_F(\theta)^2\,d\theta}{m^2} - \frac{2\int(\hat{F}(\theta) - F(\theta))\,\hat{b}_F(\theta)\,d\theta}{m} + \text{term independent of } h.$$

Using the definition of $\hat{b}_F$ and expanding the square the first right-hand side term can be written as

$$\frac{\int \hat{b}_F(\theta)^2\,d\theta}{m^2} = \frac{m^{-2}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\sigma_i^2\sigma_j^2}{h^2}\,\frac{1}{4}\int \frac{1}{h}\phi'\left(\frac{\vartheta_i - \theta}{h}\right)\,\frac{1}{h}\phi'\left(\frac{\vartheta_j - \theta}{h}\right)\,d\theta,$$

and using properties of the normal distribution we calculate

$$\int \phi'\left(\frac{\vartheta_i - \theta}{h}\right)\,\phi'\left(\frac{\vartheta_j - \theta}{h}\right)\,d\theta = \frac{1}{\sqrt{2}h}\phi\left(\frac{\vartheta_i - \vartheta_j}{\sqrt{2}h}\right)\left(\frac{h^2}{2} - \frac{(\vartheta_i + \vartheta_j)^2}{4} + \vartheta_i\vartheta_j\right).$$

Next, exploiting that $\phi'(\eta) = -\eta\,\phi(\eta)$ and using well-known results on the truncated normal distribution

$$\begin{aligned}
-\frac{2\int \hat{F}(\theta)\,\hat{b}_F(\theta)\,d\theta}{m} &= \frac{m^{-1}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\sigma_j^2}{h^2}\int_{\vartheta_i}^{+\infty}\phi'\left(\frac{\vartheta_j-\theta}{h}\right)d\theta \\
&= \frac{m^{-1}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\sigma_j^2}{h^2}\int_{\vartheta_i}^{+\infty}\left(\frac{\theta-\vartheta_j}{h}\right)\phi\left(\frac{\theta-\vartheta_j}{h}\right)d\theta \\
&= \frac{m^{-1}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\sigma_j^2}{h}\left(\frac{\vartheta_i-\vartheta_j}{h}\right)\phi\left(\frac{\vartheta_i-\vartheta_j}{h}\right) \\
&= \frac{m^{-1}}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\sigma_i^2}{h}\phi'\left(\frac{\vartheta_i-\vartheta_j}{h}\right).
\end{aligned}$$

Omitting terms for which $j = i$ in the last expression is justified by the fact that $\phi'(0) = 0$. Finally, for the last term, integrating by parts shows that

$$\frac{2\int F(\theta)\,\hat{b}_F(\theta)\,d\theta}{m} = -\frac{m^{-1}}{n}\sum_{i=1}^{n}\frac{\sigma_i^2}{h}\int\phi\left(\frac{\vartheta_i-\theta}{h}\right)f(\theta)\,d\theta.$$

The integral in the right-hand side expression represents an expectation taken with respect to f. A leave-one-out estimator of the entire term is

$$-\frac{m^{-1}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{\sigma_i^2}{h}\phi\left(\frac{\vartheta_i-\vartheta_j}{h}\right).$$

Combining results and multiplying the entire expression through with n^2m^2 yields the cross-validation objective function stated in the main text.

B Proof of Lemmas A.3 and A.4

Before proving Lemmas A.3 and A.4 we first state one known results, and also establish two further intermediate lemmas.

Lemma B.1 (Mason 1981)**.** *Let $\mathbb{G}_n$ be the empirical cumulative distribution of an i.i.d. sample of size n from a uniform distribution on [0,1]. Then, as $n \to \infty$,*

$$\sup_{u\in(0,1)}\left[u(1-u)\right]^{-1+\epsilon}\left|\mathbb{G}_n(u)-u\right| \to 0,$$

almost surely, for any $0 < \epsilon \leq 1/2$.

Lemma B.2. *Let Assumption 1 hold. Then, if* $\sup_\theta(1+|\theta|^\kappa)\,f(\theta)<\infty$,

$$\sup_\theta(1+|\theta|^\kappa)\,f_m(\theta)=O_p(1).$$

holds.

Proof. The conditional density of $\vartheta_i-\theta_i$ given θ_i evaluated in ε is

$$p(\varepsilon|\,\theta):=E\left(\frac{1}{\sigma_i/\sqrt{m}}\phi\left(\frac{\varepsilon}{\sigma_i/\sqrt{m}}\right)\middle|\,\theta_i=\theta\right).$$

We thus have

$$f_m(\vartheta)=\textstyle\int_{-\infty}^{\infty}p(\vartheta-\theta|\,\theta)\,f(\theta)\,d\theta=\int_{-\infty}^{\vartheta/2}p(\vartheta-\theta|\,\theta)\,f(\theta)\,d\theta+\int_{\vartheta/2}^{\infty}p(\vartheta-\theta|\,\theta)\,f(\theta)\,d\theta.$$

Without loss of generality we will take the value ϑ to be positive throughout. We have the bound

$$f_m(\vartheta)\leq\sup_\theta f(\theta)\textstyle\int_{-\infty}^{\vartheta/2}p(\vartheta-\theta|\,\theta)\,d\theta+\sup_{\theta\geq\vartheta/2}f(\theta)\int_{\vartheta/2}^{\infty}p(\vartheta-\theta|\,\theta)\,d\theta.\tag{B.1}$$

Consider the second term on the right-hand side in (B.1). $\sup_{\theta\geq\vartheta/2}f(\theta)=O(1+|\vartheta/2|^{-\kappa})$ by assumption and so it suffices to show that the integral is finite for all ϑ. To see that this is so, note that

$$\textstyle\int_{\vartheta/2}^{\infty}p(\vartheta-\theta|\,\theta)\,d\theta=\int_{-\infty}^{\vartheta/2}p(\varepsilon|\,\vartheta-\varepsilon)\,d\varepsilon=\int_{-\infty}^{\vartheta/2}E\left(\frac{1}{\sigma_i/\sqrt{m}}\phi\left(\frac{\varepsilon}{\sigma_i/\sqrt{m}}\right)\middle|\,\theta_i=\vartheta-\varepsilon\right)\,d\varepsilon$$

is bounded by

$$\textstyle\int_{-\infty}^{\infty}\max_{\sigma\in[\underline{\sigma},\overline{\sigma}]}\left\{\frac{1}{\sigma/\sqrt{m}}\phi\left(\frac{\varepsilon}{\sigma/\sqrt{m}}\right)\right\}\,d\varepsilon=2\int_0^{\infty}\max_{\sigma\in[\underline{\sigma},\overline{\sigma}]}\left\{\frac{1}{\sigma/\sqrt{m}}\phi\left(\frac{\varepsilon}{\sigma/\sqrt{m}}\right)\right\}\,d\varepsilon.$$

The optimizer and optimum of the constrained optimization problem inside the integral are

$$\sigma/\sqrt{m}=\begin{cases}\underline{\sigma}/\sqrt{m} & \text{if } \varepsilon<\underline{\sigma}\\ \varepsilon & \text{if } \varepsilon\in[\underline{\sigma},\overline{\sigma}]\\ \overline{\sigma}/\sqrt{m} & \text{if } \varepsilon>\overline{\sigma}\end{cases},\qquad \max_\sigma=\begin{cases}\frac{1}{\underline{\sigma}/\sqrt{m}}\phi\left(\frac{\varepsilon}{\underline{\sigma}/\sqrt{m}}\right) & \text{if } \varepsilon<\underline{\sigma}\\ \frac{\phi(1)}{\varepsilon} & \text{if } \varepsilon\in[\underline{\sigma},\overline{\sigma}]\\ \frac{1}{\overline{\sigma}/\sqrt{m}}\phi\left(\frac{\varepsilon}{\overline{\sigma}/\sqrt{m}}\right) & \text{if } \varepsilon>\overline{\sigma}\end{cases}.$$

Splitting the integral we find

$$\int_0^\infty \max_{\sigma\in[\underline{\sigma},\bar{\sigma}]} \left\{ \tfrac{1}{\sigma/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\sigma/\sqrt{m}}\right)\right\} d\varepsilon = \tfrac{e^{-1/2}}{\sqrt{2\pi}}\log(\bar{\sigma}/\underline{\sigma}) + \tfrac{1}{2} < \infty,$$

as claimed. For the first right-hand side term in (B.1), recall that $\sup_\theta f(\theta) < \infty$, and so we need to show that the integral vanishes sufficiently fast as $\vartheta \to \infty$. To see that this is the case we proceed as before by observing that

$$\int_{-\infty}^{\vartheta/2} p(\vartheta-\theta|\,\theta)\,d\theta = \int_{\vartheta/2}^\infty E\left(\tfrac{1}{\sigma_i/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\sigma_i/\sqrt{m}}\right)\Big|\,\theta_i = \vartheta-\varepsilon\right)\,d\varepsilon$$

is bounded by

$$\int_{\vartheta/2}^\infty \max_{\sigma\in[\underline{\sigma},\bar{\sigma}]} \tfrac{1}{\sigma/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\sigma/\sqrt{m}}\right)\,d\varepsilon = \int_{\vartheta/2}^\infty \tfrac{1}{\bar{\sigma}/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\bar{\sigma}/\sqrt{m}}\right)\,d\varepsilon = 1-\Phi\left(\tfrac{\vartheta/2}{\bar{\sigma}/\sqrt{m}}\right).$$

Because the tails of the normal distribution decay at an exponential rate this implies that

$$f_m(\vartheta) = O(1+|\vartheta/2|^{-\kappa})$$

uniformly in ϑ, as claimed. This completes the proof of the lemma. □

Lemma B.3. *Let Assumptions 1 and 2 hold and let*

$$\gamma_m^r(\theta) := E(\sigma_i^r|\vartheta_i=\theta)\,f_m(\theta), \qquad \gamma^r(\theta) := E(\sigma_i^r|\theta_i=\theta)\,f(\theta).$$

Then, for any integer r,

$$\sup_\theta |\nabla^q\gamma_m^r(\theta) - \nabla^q\gamma^r(\theta)| = O(m^{-1})$$

provided that the conditional density $h(\theta|\sigma)$ is $(q+2)$ times differentiable with respect to θ and that there exists a function e so that $|\nabla_1^{q+2}h(\theta|\sigma)| \le e(\sigma)$ and $E(e(\sigma_i)) < \infty$.

Proof. Fix r throughout the proof. First note that, by Bayes' rule and Assumption 1, we may write

$$\gamma_m^r(\vartheta) = \int_{\underline{\sigma}}^{\bar{\sigma}}\int_{-\infty}^\infty \sigma^r \tfrac{1}{\sigma/\sqrt{m}}\phi\left(\tfrac{\vartheta-\theta}{\sigma/\sqrt{m}}\right)\,h(\theta,\sigma)\,d\sigma\,d\theta$$

A change of variable from θ to $\varepsilon := (\vartheta - \theta)/(\sigma/\sqrt{m})$ then allows to write

$$\gamma_m^r(\vartheta) = E\left(B_r(\vartheta, \varepsilon_i/\sqrt{m})\right), \qquad B_r(\theta, \delta) := \int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^r \, h(\theta - \delta\sigma, \sigma)\, d\sigma.$$

Observe that $B_r(\vartheta, 0) = \gamma^r(\vartheta)$. Now, by a Taylor expansion,

$$\nabla^q \gamma_m^r(\vartheta) - \nabla^q \gamma^r(\vartheta) = \frac{E\left(\varepsilon_i^2 \, \nabla_1^q \nabla_2^2 B_r(\vartheta, \varepsilon_i^*/\sqrt{m})\right)}{m}.$$

Also, as

$$\nabla_1^p \nabla_2^q B_r(\theta, \delta) = (-1)^q \int_{\underline{\sigma}}^{\overline{\sigma}} \sigma^{r+q} \, \nabla_1^{p+q} h(\theta - \delta\sigma, \sigma)\, d\sigma$$

for any pair of integers (p, q), we have that

$$\sup_\theta \sup_\delta |\nabla_1^q \nabla_2^2 B_r(\theta, \delta)| \leq \overline{\sigma}^{r+q} \sup_\theta \sup_\delta \left|\int_{\underline{\sigma}}^{\overline{\sigma}} \nabla_1^{2+q} h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma\right| \leq \overline{\sigma}^{r+q} \int_{\underline{\sigma}}^{\overline{\sigma}} e(\sigma)\, h(\sigma)\, d\sigma,$$

which is finite. Therefore, uniformly in θ,

$$\nabla^q \gamma_m^r(\theta) - \nabla^q \gamma^r(\theta) = O(m^{-1}),$$

as claimed. This completes the proof. □

Proof of Lemma A.3. /Part (i): With

$$\beta_m(\theta) := \frac{E(\sigma_i^2 | \vartheta_i = \theta)\, f_m(\theta)}{2},$$

a change of variable and integration by parts yield

$$E(\hat{b}_F(\theta)) = -\int_{-\infty}^{\infty} \frac{\beta_m(\vartheta)}{h^2} \, \phi'\left(\tfrac{\vartheta - \theta}{h}\right) d\vartheta = \int_{-\infty}^{\infty} \nabla^1 \beta_m(\theta + h\varepsilon)\, \phi(\varepsilon)\, d\varepsilon.$$

Taylor expanding $\nabla^1 \beta_m$ around $\varepsilon = 0$ and exploiting properties of the normal distribution we obtain

$$E(\hat{b}_F(\theta)) = \nabla^1 \beta_m(\theta) + h^2 \, \frac{\int_{-\infty}^{\infty} \nabla^3 \beta_m(\theta + h\varepsilon^*)\, \varepsilon^2 \, \phi(\varepsilon)\, d\varepsilon}{2},$$

where ε^* lies between ε and zero. From Lemma B.3 we have

$$\nabla^1 \beta_m(\theta) = \nabla^1 \beta(\theta) + O(m^{-1}) = b_F(\theta) + O(m^{-1}),$$

uniformly in θ, and $\sup_\theta|\nabla^3\beta_m(\theta)| < \infty$. Therefore,

$$E(\hat{b}_F(\theta)) = b_F(\theta) + O(m^{-1}) + O(h^2),$$

as claimed.

Part (ii): Note that

$$\text{var}(\hat{b}_F(\theta)) = E(\hat{b}_F(\theta)^2) - E(\hat{b}_F(\theta))^2 = \frac{n^{-1}}{4} E\left(\frac{\sigma_i^4}{h^4}\phi'\left(\frac{\vartheta-\theta}{h}\right)^2\right) - b_F(\theta)^2 + o(n^{-1}).$$

Now, with

$$\beta_m^2(\theta) := \frac{E(\sigma_i^4|\vartheta_i=\theta)\, f_m(\theta)}{4},$$

we have

$$\frac{n^{-1}}{4} E\left(\frac{\sigma_i^4}{h^4}\phi'\left(\frac{\vartheta-\theta}{h}\right)^2\right) = \int_{-\infty}^{\infty} \frac{\beta_m^2(\vartheta)}{h^4}\phi'\left(\tfrac{\vartheta-\theta}{h}\right)^2 d\vartheta \leq \frac{\sup_\theta|\beta_m^2(\theta)|}{n} \frac{\int_{-\infty}^{\infty}\phi'\left(\frac{\vartheta-\theta}{h}\right)^2 d\vartheta}{h^4}$$

which is $O(n^{-1}h^3)$ uniformly in θ as $\sup_\theta|\beta_m^2(\theta)| < \infty$ because σ_i is finite and f_m is bounded, and

$$\int_{-\infty}^{\infty}\phi'\left(\tfrac{\vartheta-\theta}{h}\right)^2 d\vartheta = \tfrac{h}{4\sqrt{\pi}},$$

independent of θ. This completes the proof.

Part (iii): First observe that

$$\nabla^1 b_F(\theta) = \nabla^2\beta(\theta)/2,$$

so that $(1+|\theta|^{1+\eta})\,|\nabla^1 b_F(\theta)| < \infty$ follows directly from Assumption 3. What is left to show is that

$$\sup_\theta(1+|\theta|^{1+\eta})\,|\nabla^1\hat{b}_F(\theta)| = O_p(-(1+\omega^{-1})).$$

Note that

$$\nabla^1\hat{b}_F(\theta) = \frac{(nh^2)^{-1}}{2}\sum_{i=1}^n \sigma_i^2\,\phi''\left(\frac{\vartheta_i-\theta}{h}\right).$$

By Hölder's inequality,

$$|\nabla^1\hat{b}_F(\theta)| \leq h^{-2}\left\{\left(n^{-1}\sum_{i=1}^n(\sigma_i^2/2)^\omega\right)^{\omega^{-1}}\right\} \times \left\{\left(n^{-1}\sum_{i=1}^n\left|\phi''\left(\frac{\vartheta_i-\theta}{h}\right)\right|^\psi\right)^{\psi^{-1}}\right\},$$

where $\psi := (1-\omega^{-1})^{-1}$. The first term in braces is bounded in probability because the σ_i^2 are finite. For the second term in braces, write $\mathbb{G}_n$ for the empirical cumulative distribution of an i.i.d. sample of size n from the uniform distribution on $[0,1]$ and let $\mathbb{G}'_n(u) := n^{-1}\sum_{i=1}^n \delta_{u_i-u}$, where δ_a is Dirac's delta at a. Then, writing ∇_u for the derivative with respect to u, we get

$$\begin{aligned} n^{-1}\sum_{i=1}^n \left|\phi''\left(\frac{\vartheta_i-\theta}{h}\right)\right|^\psi &= \int_0^1 \left|\phi''\left(\frac{F_m^{-1}(u)-\theta}{h}\right)\right|^\psi \mathbb{G}'_n(u)\,du \\ &= -\int_0^1 \nabla_u^1 \left|\phi''\left(\frac{F_m^{-1}(u)-\theta}{h}\right)\right|^\psi \mathbb{G}_n(u)\,du \\ &= -\int_0^1 \nabla_u^1 \left|\phi''\left(\frac{F_m^{-1}(u)-\theta}{h}\right)\right|^\psi u\,du \\ &\quad -\int_0^1 \nabla_u^1 \left|\phi''\left(\frac{F_m^{-1}(u)-\theta}{h}\right)\right|^\psi (\mathbb{G}_n(u)-u)\,du \end{aligned} \tag{B.2}$$

where we have used integration by parts in the first step and replaced $\mathbb{G}_n(u)$ by $u+(\mathbb{G}_n(u)-u)$ in the second step. We now consider each of the integrals on the right-hand side in turn. First, integrating by parts,

$$-\int_0^1 \nabla_u^1 \left|\phi''\left(\frac{F_m^{-1}(u)-\theta}{h}\right)\right|^\psi u\,du = E\left(\left|\phi''\left(\frac{\vartheta_i-\theta}{h}\right)\right|^\psi\right). \tag{B.3}$$

Clearly, this term is bounded uniformly on any finite interval. To evaluate it for large values of θ, observe that

$$\begin{aligned} \frac{1}{h}E\left(\left|\phi''\left(\frac{\vartheta_i-\theta}{h}\right)\right|^\psi\right) &= \int_{-\infty}^{+\infty}\frac{1}{h}\left|\phi''\left(\frac{\vartheta-\theta}{h}\right)\right|^\psi f_m(\vartheta)\,d\vartheta \\ &= \int_{\theta-h\log(1+|\theta|)}^{\theta+h\log(1+|\theta|)}\frac{1}{h}\left|\phi''\left(\frac{\vartheta-\theta}{h}\right)\right|^\psi f_m(\vartheta)\,d\vartheta \\ &\quad + \int_{\log(1+|\theta|)}^{\infty}|\phi''(z)|^\psi\, f_m(\theta+zh)\,dz \\ &\quad + \int_{\log(1+|\theta|)}^{\infty}|\phi''(z)|^\psi\, f_m(\theta-zh)\,dz. \end{aligned}$$

Here,

$$\int_{\theta-h\log(1+|\theta|)}^{\theta+h\log(1+|\theta|)}\frac{1}{h}\left|\phi''\left(\frac{\vartheta-\theta}{h}\right)\right|^\psi f_m(\vartheta)\,d\vartheta \le O(\log(1+|\theta|))\sup_\theta|f_m(\theta)| = O(\log(1+|\theta|)),$$

because $\sup_\theta |\phi''(\theta)|^\psi = O(1)$ and f_m is bounded. Further, because

$$\int_x^\infty |\phi''(z)|^\psi \, dz = O(x^{2\psi-1} \, e^{-\psi \, x^2/2}), \qquad \text{as } x \to \infty,$$

and $f_m(\theta) = O(|\theta|^{-\kappa})$ as $|\theta| \to \infty$ by Lemma B.2, we have

$$\begin{aligned}
\textstyle\int_{\log(1+|\theta|)}^\infty |\phi''(z)|^\psi \; f_m(\theta + zh) \, dz &= O\left(\log(1+|\theta|)^{2\psi-1} \, e^{-\psi \log(1+|\theta|)^2/2}\right), \\
\textstyle\int_{\log(1+|\theta|)}^\infty |\phi''(z)|^\psi \; f_m(\theta - zh) \, dz &= O\left(\log(1+|\theta|)^{2\psi-1} \, e^{-\psi \log(1+|\theta|)^2/2}\right).
\end{aligned}$$

Then, as

$$e^{-\psi \log(1+|\theta|)^2/2} = o(|\theta|^a) \text{ for any } a > 0 \text{ as } |\theta| \to \infty$$

we may conclude that the term in (B.3) is $O(h|\theta|^{-\kappa} \log(1+|\theta|))$ uniformly in θ. Next, for the second term in (B.2) we use Lemma B.1 to establish that, for any $\epsilon \in (0, 1/2]$, we have

$$\begin{aligned}
&\left| \textstyle\int_0^1 \nabla_u^1 \left| \phi'' \left(\frac{F_m^{-1}(u) - \theta}{h} \right) \right|^\psi \; (\mathbb{G}_n(u) - u) \; du \right| \\
&\leq o_p(1) \left| \textstyle\int_0^1 \left| \nabla_u^1 \left| \phi'' \left(\frac{F_m^{-1}(u) - \theta}{h} \right) \right|^\psi \right| \; (u^{1-\epsilon} \, (1-u)^{1-\epsilon}) \; du \right| \\
&= o_p(1) \left| \textstyle\int_{-\infty}^{+\infty} \left| \nabla_u^1 \left| \phi'' \left(\frac{F_m^{-1}(u) - \theta}{h} \right) \right|^\psi \right| \; (F_m(\vartheta)^{1-\epsilon} \, (1 - F_m(\vartheta))^{1-\epsilon}) \; d\vartheta \right|,
\end{aligned}$$

where the $o_p(1)$ term is independent of θ. The integral term can be bounded in the same way as (B.3). Hence,

$$\left| \textstyle\int_0^1 \nabla_u^1 \left| \phi'' \left(\frac{F_m^{-1}(u) - \theta}{h} \right) \right|^\psi \; (\mathbb{G}_n(u) - u) \; du \right| = o_p(h|\theta|^{(1-\epsilon)\,(1-\kappa)} \; \log(1+|\theta|))$$

uniformly in θ. We therefore have that

$$\sup_\theta |\hat{b}_F(\theta)| \leq h^{-2} \, O_p(1) \left\{ (O(h|\theta|^{-\kappa} \; \log(1+|\theta|)) + o_p(h|\theta|^{(1-\epsilon)\,(1-\kappa)} \; \log(1+|\theta|))^{\psi^{-1}} \right\}.$$

For any $\eta > (\kappa - 1)(1-\epsilon)(1-1/\omega) - 1 > 0$ it then follows that

$$\sup_\theta \; \left(1 + |\theta|^{1+\eta}\right) \; |\hat{b}_F(\theta)| = O_P\left(h^{-(1+\omega^{-1})}\right).$$

Here, our assumption $\kappa > 1 + (1 - 1/\omega)^{-1}$ guarantees that we can find $\epsilon > 0$ such that $\eta > (\kappa - 1)(1-\epsilon)(1-1/\omega) - 1 > 0$ holds. This concludes the proof. □

Proof of Lemma A.4. First observe that, for any $\epsilon > 0$,

$$\sup_\theta E(|b_i(\theta) - E(b_i(\theta))|^\epsilon) \leq \sup_\theta \sum_{p=0}^{\epsilon} \binom{\epsilon}{p} E(|b_i(\theta)|^p)\, E(|b_i(\theta)|^{\epsilon-p}) \leq 2^\epsilon \sup_\theta E(|b_i(\theta)|^\epsilon).$$

Therefore,

$$\begin{aligned}
\sup_\theta E(|b_i(\theta) - E(b_i(\theta))|^\epsilon)^{\epsilon^{-1}} &\leq 2 \sup_\theta (E(|b_i(\theta)|^\epsilon))^{\epsilon^{-1}} \\
&= \sup_\theta \left(\int_{-\infty}^{\infty} \tfrac{E(\sigma_i^{2\epsilon}|\vartheta_i=\vartheta)\, f_m(\vartheta)}{h^2} \left|\phi'\left(\tfrac{\vartheta-\theta}{h}\right)\right|^\epsilon d\vartheta \right)^{\epsilon^{-1}} \\
&\leq \sup_\vartheta (E(\sigma_i^{2\epsilon}|\vartheta_i = \vartheta)\, f_m(\vartheta))^{\epsilon^{-1}} \frac{\left(\sup_\theta \int_{-\infty}^{\infty} \left|\phi'\left(\tfrac{\vartheta-\theta}{h}\right)\right|^\epsilon d\vartheta\right)^{\epsilon^{-1}}}{h^2} \\
&= O(h^{\epsilon^{-1}-2}),
\end{aligned}$$

where we have used the definition of $b_i(\theta)$ in the first step, boundedness of the σ_i and f_m in the second step, and the fact that

$$\int_{-\infty}^{\infty} \left|\phi'\left(\tfrac{\vartheta-\theta}{h}\right)\right|^\epsilon d\vartheta = O(h),$$

independent of θ, in the final step. This completes the proof. □

References

Ahn, D., S. Choi, D. Gale, and S. Kariv (2014). Estimating ambiguity aversion in a portfolio choice experiment. *Quantitative Economics 5*, 195–223.

Alvarez, J. and M. Arellano (2003). The time series and cross-section asymptotics of dynamic panel data estimators. *Econometrica 71*, 1121–1159.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*, 171–178.

Barras, L., P. Gagliardini, and O. Scaillet (2018). The cross-sectional distribution of fund skill measures. Mimeo.

Browning, M., M. Ejrnæs, and J. Alvarez (2010). Modeling income processes with lots of heterogeneity. *Review of Economic Studies 77*, 1353–1381.

Carroll, R. J. and P. Hall (1988). Optimal rates of convergence for deconvoluting a density. *Journal of the American Statistical Association 83*, 1184–1186.

Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intriligator (Eds.), *Handbook of Econometrics*, Volume 2 of *Handbook of Econometrics*, Chapter 22, pp. 1247–1315. Elsevier.

Chesher, A. (1991). The effect of measurement error. *Biometrika 78*, 451–462.

Chesher, A. (2017). Understanding the effect of measurement error on quantile regressions. *Journal of Econometrics 200*, 223–237.

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

Delaigle, A. and A. Meister (2008). Density estimation with heteroscedastic error. *Bernoulli 14*, 562–579.

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies 82*, 991–1030.

Doss, H. and R. D. Gill (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data. *Journal of the American Statistical Association 87*(419), 869–877.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*, 1602–1614.

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika 103*, 1–20.

Fernández-Val, I. and J. Lee (2013). Panel data models with nonadditive unobserved heterogeneity: Estimation and inference. *Quantitative Economics 4*, 453–481.

Guvenen, F. (2009). An empirical investigation of labor income processes. *Review of Economic Dynamics 12*, 58–79.

Hahn, J. and G. Kuersteiner (2002). Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica 70*, 1639–1657.

Hahn, J. and W. K. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica 72*, 1295–1319.

Jackson, C. K., J. E. Rockoff, and D. O. Staiger (2014). Teacher effects and teacher related policies. *Annual Review of Economics 6*, 801–825.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Volume I, pp. 361–379.

Komlós, J., P. Major, and G. Tusnády (1975). An approximation of partial sums of independent RV'-s, and the sample DF. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 32*, 111–131.

Magnac, T. and S. Roux (2019). Heterogeneity and wage inequalities over the life cycle. Mimeo.

Maritz, J. S. and R. G. Jarrett (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association 73*, 194–196.

Mason, D. M. (1981). Bounds for weighted empirical distribution functions. *The Annals of Probability 9*, 881–884.

Neyman, J. and E. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1–32.

Okui, R. and T. Yanagi (2017). Panel data analysis with heterogeneous dynamics. Available on SSRN at http://dx.doi.org/10.2139/ssrn.2694627.

Okui, R. and T. Yanagi (2018). Kernel estimation for panel data with heterogenous dynamics. Available on arXiv.org as arXiv:1802.08825v2 [econ.EM].

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume I, pp. 157–163.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94*, 247–252.

Rosenthal, H. P. (1970). On the subspaces of L_p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics 8*, 273–303.

Schucany, W. and J. Sommers (1977). Improvement of kernel type density estimators. *Journal of the American Statistical Association 72*, 420–423.

Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation. *American Economic Review: Papers & Proceedings 105*, 467–470.

Weinstein, A., Z. Ma, L. D. Brown, and C.-H. Zhang (2018). Group-linear Empirical Bayes estimates for a heteroscedastic normal mean. Forthcoming in *Journal of the Americal Statistical Association*.