Jochmans, Koen; Weidner, Martin

**Working Paper**

# Inference on a distribution from noisy draws

cemmap working paper, No. CWP14/18

**Provided in Cooperation with:**
Institute for Fiscal Studies (IFS), London

*Suggested Citation:* Jochmans, Koen; Weidner, Martin (2018) : Inference on a distribution from noisy draws, cemmap working paper, No. CWP14/18, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2018.1418

This Version is available at:
https://hdl.handle.net/10419/189709

cemmap
centre for microdata methods and practice

# Inference on a distribution from noisy draws

Koen Jochmans
Martin Weidner

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP14/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

# INFERENCE ON A DISTRIBUTION FROM NOISY DRAWS

Koen Jochmans*
University of Cambridge

Martin Weidner†
University College London

February 13, 2018

**Abstract**

We consider a situation where a distribution is being estimated by the empirical distribution of noisy measurements. The measurements errors are allowed to be heteroskedastic and their variance may depend on the realization of the underlying random variable. We use an asymptotic embedding where the noise shrinks with the sample size to calculate the leading bias arising from the presence of noise. Conditions are obtained under which this bias is asymptotically non-negligible. Analytical and jackknife corrections for the empirical distribution are derived that recenter the limit distribution and yield confidence intervals with correct coverage in large samples. Similar adjustments are presented for nonparametric estimators of the density and quantile function. Our approach can be connected to corrections for selection bias and shrinkage estimation. Simulation results confirm the much improved sampling behavior of the corrected estimators. An empirical application to the estimation of a stochastic-frontier model is also provided.

**JEL Classification:** C14, C23

**Keywords:** bias correction, nonparametric inference, regression to the mean.

*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. E-mail: m.weidner@ucl.ac.uk.

We are grateful to Isaiah Andrews and Bo Honoré for comments and discussion and to Peter Schmidt for suggesting to us the empirical application.

Jochmans gratefully acknowledges financial support from the European Research Council through Starting Grant n° 715787. Weidner gratefully acknowledges support from the Economic and Social Research Council through the ESRC Centre for Microdata Methods and Practice grant RES-589-28-0001 and from the European Research Council grant ERC-2014-CoG-646917-ROMIA.

# 1 Introduction

Let $\theta_1, \ldots, \theta_n$ be a random sample from a distribution function $F$ that is of interest. In many empirical problems we only observe noisy measurements of these variables, say $\vartheta_1, \ldots, \vartheta_n$. A common approach in practice, then, is to do inference on $F$ and its functionals using the empirical distribution of $\vartheta_1, \ldots, \vartheta_n$.

In the literature on student achievement (see, e.g, Rivkin, Hanushek and Kain 2005 and references therein), for example, a standard approach is to fit a two-way error-component model to test-score data and quantify the teacher's added value as the sample variance of the estimated teacher fixed effect. A similar argument has been used extensively to estimate worker and firm heterogeneity from wage data (following Abowd, Kramarz and Margolis 1999). While the literature has become aware of the potential bias induced by the presence of sampling noise, the issue has been mostly ignored in practice (see, e.g., Card, Heining and Kline 2013).[1]

In this paper we analyze the effect of using noisy measurements on the accuracy of statistical inference on $F$ in an asymptotic embedding where the noise shrinks with the sample size. More precisely, if we write the variances of $\vartheta_1, \ldots, \vartheta_n$ as $\sigma_1^2/m, \ldots, \sigma_n^2/m$ for some real number $m$, then we consider double asymptotics where $n, m \to \infty$ jointly. This asymptotic embedding is intuitive when $\vartheta_i$ is an estimator of $\theta_i$ obtained from a sample of size $m$. It has been used in models with stratum-specific nuisance parameters and many strata (following Li, Lindsay and Waterman 2003 and Sartori 2003) although the focus there has mostly been on the sampling behavior of estimators of common parameters and not on the (distribution of the) stratum-specific parameters. Fernández-Val and Lee (2013) and Dhaene and Jochmans (2015) do consider inference on average marginal effects in such models, but these only involve smooth functions of the nuisance parameters. Okui and Yanagi (2016) further provide a consistency result for the empirical distribution function

[1] An exception is Rockoff (2004), who notes that the sample variance of his estimated teacher effects will tend to overestimate the true variance, and attempts to deal with the issue through an Empirical Bayes approach.

but subsequently sidestep the bias issue in inference.

We will focus on the case where

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

although our results hold more generally in situations where the

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

are random draws from some well-behaved distribution. The setting here is different from deconvolution, as the measurements are allowed to be heteroskedastic and $\theta_i$ and $\sigma_i^2$ may be dependent. The normality assumption is motivated by the fact that, in large samples, asymptotically-linear estimators behave like normal variables. It also helps to connect our work to the literature on selection bias as recently dealt with by Efron (2011). There, selection bias is defined as the tendency of the $\vartheta_i$'s associated with the (in magnitude) largest $\theta_i$'s to be larger than their corresponding $\theta_i$. Efron (2011) essentially entertains the homoskedastic setting where

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

He proposes to deal with selection bias by using the Empirical Bayes estimators of Robbins (1956), which here would be

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

where $p$ is the marginal density of the $\vartheta_i$ and $\nabla^1$ denotes the first-derivative operator. For example, when $\theta_i \sim N(\eta, \psi^2)$ 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. Non-parametric implementation would require estimation of $p$ and its derivative. The shrinkage to zero is intuitive, as selection bias essentially manifests itself through the tails of the empirical distribution of the $\vartheta_i$ being too thick.

The approach taken here is different. We calculate the bias of the naive plug-in estimator

$$n^{-1} \sum_{i=1}^{n} 1\{\vartheta_i \leq \theta\}$$

of $F(\theta)$ and correct for it directly and non-parametrically. In the James-Stein problem where $\theta_i \sim N(\eta, \psi^2)$, for example, the bias equals

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

Thus, the empirical distribution is indeed upward biased in the left tail and downward biased in the right tail. The 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 $n/m^4 \to 0$. We also provide corresponding bias-corrected estimators for the quantiles of $F$, and discuss similar corrections applied directly to a kernel density estimator.

We provide simulation evidence on the improvement of our approach over the naive plug-in estimator. We also discuss an empirical illustration in which we use our corrections to non-parametrically estimate the distribution of firm inefficiencies in a stochastic-frontier model.

# 2 Estimation and inference

Let $F$ be a univariate distribution on the real line. Here, we are interested in estimation of and inference on $F$ and its 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 $n$. Moreover, we assume the following.

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

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

*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 $\vartheta_i$ is an estimator of $\theta_i$ obtained from a sample of size $m$ that converges at the parametric rate.[2] We allow $\theta_i$ and $\sigma_i^2$ to be correlated, implying that the noise $\vartheta_i - \theta_i$ is not independent of $\theta_i$. Recovering the distribution of $\theta_i$ 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 $\vartheta_i$ at $\theta$. 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 $F$,

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

for a differentiable function $\varphi$. 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 $\sigma_\mu^2 := \mathrm{var}(\varphi(\theta_i))$

[2] 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 $\sigma_i^2$ as $\sigma_i^2/p_i$. When the $\vartheta_i$ 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.

exists. For the feasible plug-in estimator of $\mu$

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

under regularity conditions provided in the Appendix, by a second-order Taylor expansion we have

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

and

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

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

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

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

## 2.1 Estimation of the distribution function

Now consider estimation of the distribution function $F$ using the plug-in estimator $\hat{F}$. Again, the use of noisy measurements introduces bias. The machinery from above cannot be applied to deduce the bias of $\hat{F}$, however, as it is a step function and so non-differentiable.

To derive the bias we impose the following conditions.

**Assumption 2** (Regularity for distribution function)**.** *The density function $f$ 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 $(\theta_i, \sigma_i)$ exists, and the conditional density of $\theta_i$ given $\sigma_i$ is three times differentiable with respect to $\theta_i$ 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 $\sigma$ so that $\sigma_i = \sigma(\theta_i)$ for all $i$; and (ii) $\sigma$ is four times differentiable and has uniformly-bounded derivatives.*

Assumption 2 distinguishes between the cases where the relation between $\theta_i$ and $\sigma_i^2$ is stochastic (Assumption 2.A) and deterministic (Assumption 2.B). It requires smoothess of certain densities and conditional expectations.

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

$$\sigma_F(\theta_1, \theta_2) := F(\theta_1 \wedge \theta_2) - F(\theta_1)\, F(\theta_2),$$

where we use $\theta_1 \wedge \theta_2$ to denote $\min\{\theta_1, \theta_2\}$. Our first theorem gives the leading bias and variance of $\hat{F}$. All proofs are collected in the supplementary appendix.

**Theorem 1** (First-order bias and variance for distribution function)**.** *Let Assumptions 1 and 2 hold. Then, as $n, m \to \infty$,*

$$E(\hat{F}(\theta)) - F(\theta) = \frac{b_F(\theta)}{m} + O(m^{-2}), \qquad \operatorname{cov}(\hat{F}(\theta_1), \hat{F}(\theta_2)) = \frac{\sigma_F(\theta_1, \theta_2)}{n} + O(n^{-1}m^{-1}),$$

*where the order of the remainder terms is uniform in $\theta$.*

To illustrate the result suppose that $\sigma_i^2$ is independent of $\theta_i$ and that $\theta_i$ has density function

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

as in the James and Stein (1961) problem. Letting $\sigma^2$ denote the mean of the $\sigma_i^2$ an application of Theorem 1 yields

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

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, and gives an alternative explanation of why the James and Stein (1961) estimator shrinks toward the overall mean $\eta$. Note, however, that the empirical distribution function of the estimators adjusted through either the Robbins (1956) or James and Stein (1961) formulea discussed above would not constitute an estimator of $F$ that improves on $\hat{F}$ in terms of bias, in general.

A bias-corrected estimator based on Theorem 1 is

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

where $\phi'(\eta) := -\eta\,\phi(\eta)$ and $h$ is a non-negative bandwidth parameter. Thus, we estimate the leading bias using standard kernel methods. The choice of a Gaussian kernel is in no way fundamental and is made only for simplicity of presentation.

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

**Assumption 3** (Regularity for distribution function, cont'd)**.** *(i) The conditional density of $\theta_i$ given $\sigma_i$ is five times differentiable with respect to $\theta_i$ and the derivatives are bounded in absolute value by a function $e(\sigma_i)$ such that $E(e(\sigma_i)) < \infty$. (ii)* $\sup_\theta |b_F(\theta)| = O(1)$*. There exists an integer $\omega > 2$, and real numbers $\kappa > 1 + (1-\omega^{-1})^{-1}$ and $\eta > 0$ so that (iii)* $\sup_\theta (1+|\theta|^\kappa)\, f(\theta) = O(1)$*; and (iv)* $\sup_\theta (1+|\theta|^{1+\eta})\,|\nabla^1 b_F(\theta)| = O(1)$*.*

Parts (i) and (ii) of Assumption 3 are simple smoothness and boundedness requirements. Parts (iii) and (iv) are tail conditions on the marginal density of the $\theta_i$ and on the bias function $b_F(\theta)$.

We have the following result.

**Proposition 1** (Bias correction for distribution function)**.** *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 $n/m^4 \to 0$,*

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

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

The implications of Proposition 1 are qualitatively similar to those for smooth functionals discussed above. Indeed, for any fixed $\theta$, 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

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

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

Theorem 1 equally validates 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 of $\hat{F}$ is proportional to $m^{-1}$ and is based on re-estimating $\theta_1, \ldots, \theta_n$ from subsamples. A somewhat different jackknife procedure can be constructed from the observation that, if $\vartheta_1, \ldots, \vartheta_n$ would have variance $\lambda^2\sigma_1^2, \ldots, \lambda^2\sigma_n^2$, then the bias in $\hat{F}$ would equally be multiplied by $\lambda^2$. This is apparent from the definition of $\beta$ and suggests the estimator

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

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 $\dot{F}$ can be computed without re-estimating $\theta_1, \ldots, \theta_n$. Such an approach bears similarities with the jackknife estimator of a density function introduced in Schucany and Sommers (1977). The reason this estimator is bias-reducing is as follows. By Assumption 1,

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

Further, by a 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 nonparametric density estimation, our estimator attacks bias introduced through estimation noise. Note, finally, that the sample variance of

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

can be used for inference.

It is possible to construct a similar jackknife procedure to correct for the presence of noise in the nonparametric density estimator. Consider the simple kernel-based estimator

$$\hat{f}(\theta) := n^{-1}\sum_{i=1}^{n} \frac{1}{h}\phi\left(\frac{\theta - \vartheta_i}{h}\right),$$

where, as before, $h$ is a bandwidth. Under regularity conditions, this estimator equally suffers from $O(m^{-1})$ bias arising from the noise in the $\vartheta_i$. It also has the usual kernel smoothing bias that is $O(h^2)$. More precisely,

$$E(\hat{f}(\theta)) - f(\theta) = \frac{h^2 f''(\theta)}{2} + \frac{\beta''(\theta)}{m} + O(h^4) + O(m^{-2}).$$

On the other hand, the variance of the kernel density estimator is

$$\mathrm{var}(\hat{f}(\theta)) = \frac{f(\theta)}{2\sqrt{\pi}}\,\frac{1}{nh} + O(n^{-1}h^{-1}m^{-1}) + O(n^{-1}).$$

Thus, if as usual we balance the bias term in $\hat{f}(\theta)$ of order $h^2$ with the variance of order $(nh)^{-1}$ by choosing $h \sim n^{-1/5}$, then the bias from the noisy measurement of order $m^{-1}$ becomes asymptotically dominant whenever $m \ll n^{2/5}$, in which case bias correction is required for correct large sample inference. A jackknife estimator that reduces bias from both sources is

$$\dot{f}(\theta) := \hat{f}(\theta) - \frac{\hat{f}_\lambda(\theta) - \hat{f}(\theta)}{\lambda^2}, \qquad \hat{f}_\lambda(\theta) := n^{-1} \sum_{i=1}^{n} \frac{1}{h_i} \phi\left(\frac{\theta - \vartheta_i}{h_i}\right),$$

where the observation-specific bandwidth parameter is

$$h_i := \sqrt{(1+\lambda^2)\, h^2 + \lambda^2 \sigma_i^2 / m}.$$

The intuition for the form of this jackknife is similar to the one for the distribution function. Also, an analytical correction may equally be constructed based on the bias formula stated above, but we do not go into detail here for the sake of brevity.

## 2.2 Estimation of the quantile function

The bias in $\hat{F}$ translates to bias in estimators of the quantile function. A natural estimator for $\tau$th-quantile $q(\tau)$ is given by $\hat{q}(\tau) := \hat{F}^{\leftarrow}(\hat{\tau})$, where we use $\hat{F}^{\leftarrow}$ to denote the left-inverse of $\hat{F}$. Moreover,

$$\hat{q}(\tau) = \hat{F}^{\leftarrow}(\hat{\tau}) = \vartheta_{(\lceil \tau n \rceil)},$$

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

The quantile estimator is an approximate solution to the empirical moment condition $\hat{F}(q) - \tau = 0$ (with respect to $q$). From Theorem 1 we know that

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

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

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

we obtain the following asymptotic bias result.

**Corollary 1** (First-order bias and variance for quantiles)**.** *Let the Assumptions 1 and 2 hold. For $\tau \in (0,1)$, assume that $f > 0$ in a neighborhood of $q(\tau)$. Then, as $n, m \to \infty$ with $n/m^2 \to 0$ we have*

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

As an example, when $\theta_i \sim N(\eta, \psi^2)$, independent of $\sigma_i^2$, 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 case, estimated without plug-in bias of order $m^{-1}$.

Corollary 1 readily suggests a bias-corrected estimator of the form

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

using obvious notation. While such an estimator successfully reduces bias it has the unattractive property that it requires a nonparametric estimator of the density $f$. An alternative estimator that avoids this issue is

$$\check{q}(\tau) := \hat{F}^{\leftarrow}(\hat{\tau}^*), \qquad \hat{\tau}^* := \tau + \frac{\hat{b}_F(\hat{q}(\tau))}{m}.$$

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 $\vartheta_i$ relative to the $\theta_i$, 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 $\tau$ to $\tau^*$ adjusts the naive estimator and corrects for regression to the mean.

**Proposition 2.** *Let the assumptions stated in Proposition 1 hold. For $\tau \in (0,1)$, assume that $f > 0$ in a neighborhood of $q(\tau)$. Then, as $n, m \to \infty$ with $n/m^4 \to 0$ we have*

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

The corrected estimator has the same asymptotic variance as the uncorrected estimator. It is well-known that plug-in estimators of $\sigma_q^2$ can perform quite poorly in small samples (Maritz and Jarrett 1978). Typically, researchers rely on the bootstap. For the simulations reported on below we used the simple nonparametric bootstrap of Efron (1981), where we draw random samples of size $n$ from the original sample $\vartheta_1, \ldots, \vartheta_n$. Note that we do not re-estimate the individual $\theta_i$ during the bootstrap procedure.

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

$$\dot{q}(\tau) := \hat{q}(\tau) - \frac{\hat{q}_\lambda(\tau) - \hat{q}(\tau)}{\lambda^2}, \qquad \hat{q}_\lambda(\tau) := \min_q\{q : \hat{F}_\lambda(q) \geq \tau\}.$$

The intuition behind this jackknife correction is the same as the one underlying $\dot{F}$ discussed above.

# 3 Numerical illustrations

## 3.1 Simulated data

To illustrate how much of the bias is eliminated in practice we provide simulation results for a James and Stein (1961) problem where the data are an $n \times m$ panel on $x_{it} \sim N(\theta_i, \sigma^2)$ and $\theta_i \sim N(0, \psi^2)$. This setup is a special case of the random-coefficient model from above and 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 $\sigma^2$ for fixed $m$, our focus is on between-group characteristics and the distribution of the $\theta_i$ as a whole, and this as both $n$ and $m$ grow. In this example,

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

and, rather than assuming their individual variances $\sigma_i^2$ to be known we use the estimators

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

Note that we do not make use of the fact that the $\vartheta_i$ are homoskedastic in estimating the noise or in constructing the bias correction.

The left-hand side plots in Figure 1 shows (the average over the Monte Carlo replications of) the bias-corrected estimator $\check{F}$ (solid line) together with the plug-in estimator $\hat{F}$ (dotted line) and the true distribution $F$ (dashed line) for $n = 50$ (top plot) and $n = 100$ (bottom plot), with $m = \lceil \sqrt{n} \rceil$. The bandwidth was selected using cross validation. The right-hand side plots show the corresponding results for the jackknife estimator $\dot{F}$ with $\lambda = 1$. The simulations confirm that $\hat{F}$ is too fat-tailed relative to $F$ and show that bias correction alleviates most of this bias. The plots also provide pointwise 95% confidence bounds (at the quantiles of $F$, marked by *) centered around the corrected estimator.

The winner's curse also translates into the sample variance of $\vartheta_1, \ldots, \vartheta_n$, i.e.,

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

overestimating $\psi$, on average. Indeed, here, $\vartheta_i \sim N(0, \psi^2 + \sigma^2/m)$. This overestimation results in the standard $t$-statistic overrejecting under the null. Table 1 shows this for several configurations for $n$ and $m$ that satisfy $n/m^2 \to 1$ as $n \to \infty$. The table also show that the bias-corrected version of $\psi^2$,

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

corrects the plug-in estimator for most of the bias without notably affecting the estimtors variance. Consequently, our bias-correction procedure yields a $t$-statistic that controls size well.

In Tables 2 and 3 we provide simulation results for the naive quantile estimator and our bias-corrected version based on shifting the quantile of interest. The sample sizes are as in Figure 1 (that is, $n = 50, m = 8$ and $n = 100, m = 11$, respectively). The simulations again confirm our theoretical results. The order statistics are downward biased below the median and upward biased above the median. The bias is more severe as we move into the tails of the distribution. At the median there is no bias of order $m^{-1}$, but there is the usual nonlinearity bias of order $n^{-1}$. At all percentiles our correction substantially reduces bias.

Figure 1: Estimation of $F$ in the James-Stein problem

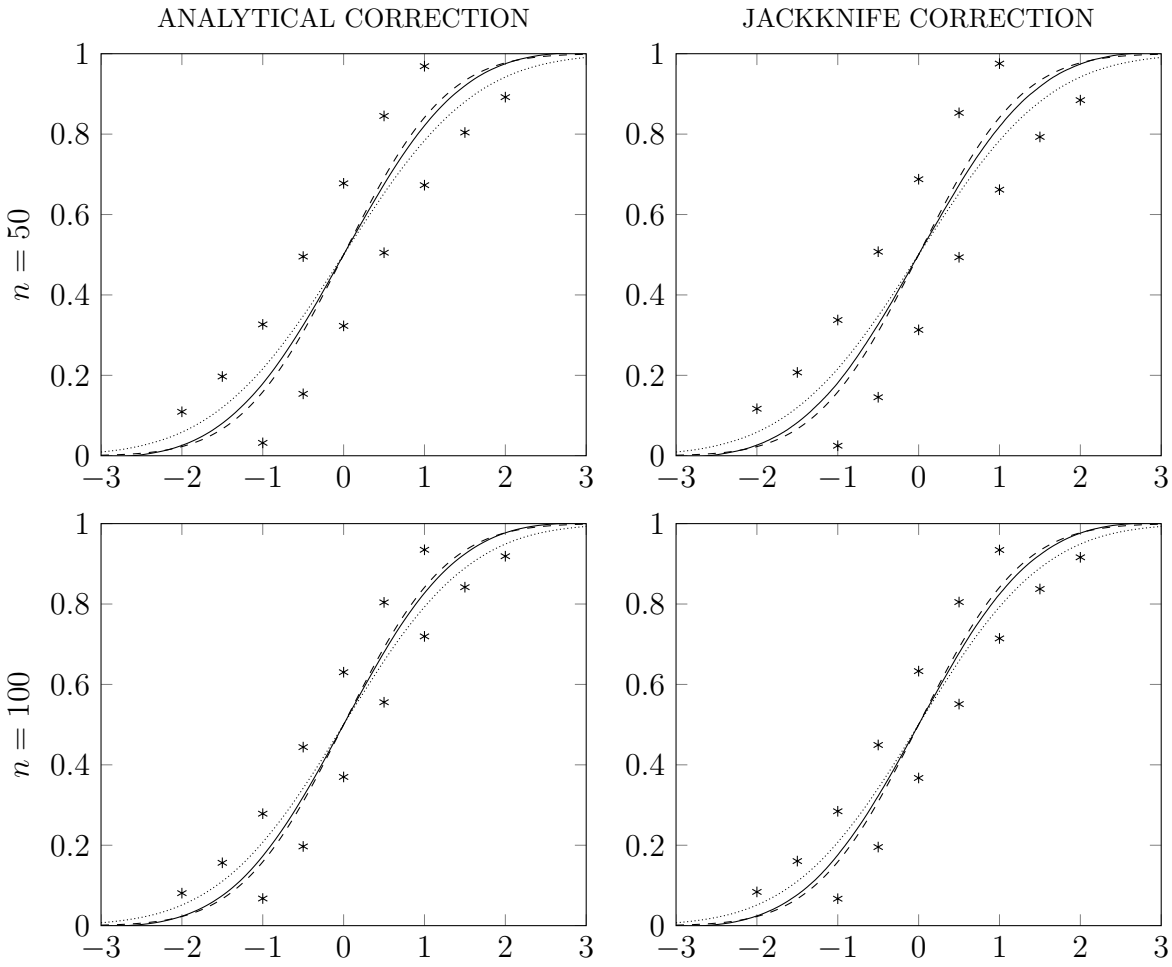


Figure notes: Mean (solid) and 95% confidence bands (*) over 10,000 Monte Carlo replications of $\check{F}$ (left plots) and $\dot{F}$ (right plots) for $n = 50$ (top plots) and $n = 100$ (bottom plots). Each plot also provides $F$ (dashed) and the mean of $\hat{F}$ (dotted). Data generation: $\vartheta_i \sim N(0,1)$ and $\vartheta_i \sim N(\theta_i, 5/\lceil\sqrt{n}\rceil)$. Bandwidth selection via cross validation. Jackknife correction implemented with $\lambda = 1$.

Adjusting for $m^{-1}$ bias is equally effective in correcting the rejection rates of the $t$-statistics downward toward their nominal level of 5%.

Table 1: Estimation of $\psi^2$ in the James-Stein problem

| | | bias | | std | | se/std | | size (5%) | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $m$ | $\hat{\psi}$ | $\check{\psi}^2$ | $\hat{\psi}^2$ | $\check{\psi}^2$ | $\hat{\psi}^2$ | $\check{\psi}^2$ | $\hat{\psi}^2$ | $\check{\psi}^2$ |
| 50 | 8 | 0.597 | -0.028 | 0.326 | 0.330 | 0.945 | 0.945 | 0.463 | 0.097 |
| 100 | 11 | 0.488 | -0.012 | 0.211 | 0.213 | 0.982 | 0.979 | 0.681 | 0.073 |
| 200 | 15 | 0.327 | -0.006 | 0.133 | 0.134 | 0.988 | 0.987 | 0.728 | 0.062 |
| 500 | 23 | 0.215 | -0.002 | 0.077 | 0.077 | 0.998 | 0.998 | 0.831 | 0.057 |

Data generation: $\theta_i \sim N(0,1)$ and $\vartheta_i \sim N(\theta_i, 5/m)$. $10,000$ Monte Carlo replications. Standard errors computed via the Delta method.

Table 2: Estimation of quantiles in the James-Stein problem ($n = 50, m = 8$)

| | $\tau$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| bias | $\hat{q}$ | -0.409 | -0.267 | -0.174 | -0.099 | -0.031 | 0.032 | 0.101 | 0.181 | 0.270 |
| | $\check{q}$ | -0.126 | -0.074 | -0.045 | -0.018 | 0.004 | 0.021 | 0.039 | 0.068 | 0.112 |
| std | $\hat{q}$ | 0.312 | 0.262 | 0.242 | 0.232 | 0.229 | 0.233 | 0.238 | 0.253 | 0.294 |
| | $\check{q}$ | 0.355 | 0.302 | 0.277 | 0.267 | 0.265 | 0.270 | 0.280 | 0.297 | 0.358 |
| se/std | $\hat{q}$ | 1.053 | 1.029 | 1.001 | 1.018 | 1.012 | 1.003 | 1.023 | 1.043 | 1.066 |
| | $\check{q}$ | 1.022 | 1.011 | 1.021 | 1.016 | 1.010 | 0.999 | 1.012 | 1.042 | 1.041 |
| size (5%) | $\hat{q}$ | 0.247 | 0.186 | 0.127 | 0.087 | 0.076 | 0.077 | 0.085 | 0.117 | 0.141 |
| | $\check{q}$ | 0.088 | 0.081 | 0.073 | 0.076 | 0.081 | 0.082 | 0.079 | 0.070 | 0.075 |

Data generation: $\theta_i \sim N(0,1)$ and $\vartheta_i \sim N(\theta_i, 5/m)$. $10,000$ Monte Carlo replications. Standard errors computed via the nonparametric bootstrap.

Table 3: Estimation of quantiles in the James-Stein problem ($n = 100, m = 11$)

| | $\tau$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|
| bias | $\hat{q}$ | -0.310 | -0.201 | -0.127 | -0.067 | -0.013 | 0.043 | 0.101 | 0.168 | 0.253 |
| | $\check{q}$ | -0.083 | -0.045 | -0.026 | -0.012 | 0.003 | 0.018 | 0.033 | 0.052 | 0.084 |
| std | $\hat{q}$ | 0.210 | 0.173 | 0.159 | 0.153 | 0.151 | 0.153 | 0.159 | 0.171 | 0.202 |
| | $\check{q}$ | 0.241 | 0.201 | 0.185 | 0.180 | 0.178 | 0.180 | 0.188 | 0.204 | 0.245 |
| se/std | $\hat{q}$ | 1.051 | 1.046 | 1.027 | 1.036 | 1.040 | 1.035 | 1.036 | 1.049 | 1.055 |
| | $\check{q}$ | 1.056 | 1.042 | 1.037 | 1.025 | 1.030 | 1.030 | 1.029 | 1.037 | 1.039 |
| size | $\hat{q}$ | 0.294 | 0.197 | 0.125 | 0.083 | 0.065 | 0.072 | 0.100 | 0.154 | 0.219 |
| | $\check{q}$ | 0.076 | 0.066 | 0.066 | 0.066 | 0.068 | 0.070 | 0.073 | 0.074 | 0.077 |

Data generation: $\theta_i \sim N(0,1)$ and $\vartheta_i \sim N(\theta_i, 5/m)$. $10,000$ Monte Carlo replications. Standard errors computed via the nonparametric bootstrap.

Figure 2: Estimation of $F$ in the dairy-farm data

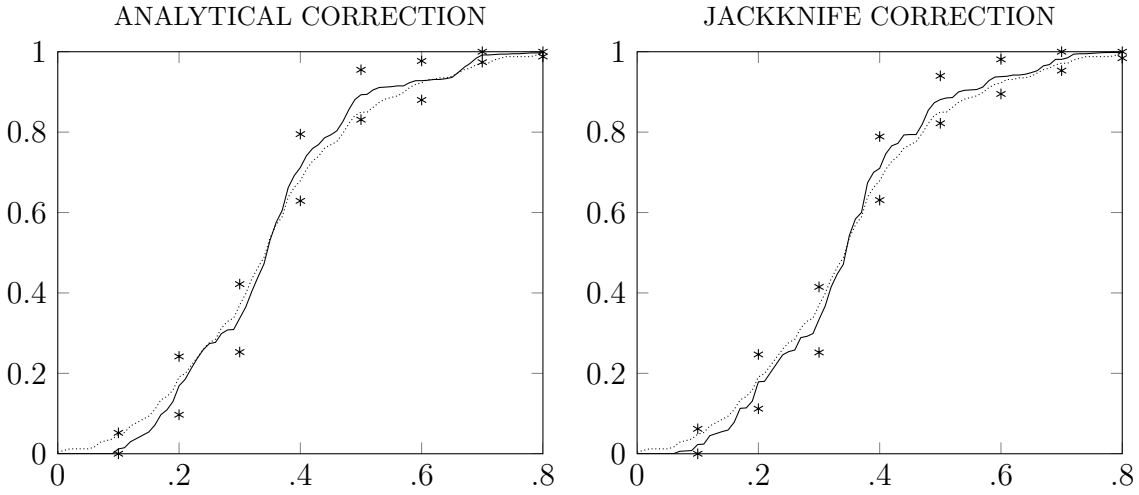


Figure notes: Uncorrected (dotted) and bias-corrected (solid) estimator of the distribution of technical inefficiency with 95% confidence bands (*). Bandwidth selection via cross validation. Jackknife correction implemented with $\lambda = 1$.

## 3.2 Empirical example

As an illustration of our approach we estimate a stochastic-frontier model as in Schmidt and Sickles (1984). We follow Belotti, Daidone, Ilardi and Atella (2013) and estimate a translog production function for milk (in liters per year) from a panel data set of 247 Spanish dairy farms over the period 1993-1995. The regressors are (the logs of) the number of milking cows, the number of man-equivalent units, the number of hectares devoted to pasture and crops, and the kilograms of feedstuffs fed to the dairy cows, as well as the interactions between all these inputs. Time dummies are equally included. Letting $y_{it}$ denote log output and $x_{it}$ the vector of all regressors the fixed-effect version of the stochastic-frontier model is

$$y_{it} = \alpha + x_{it}'\beta - \theta_i + \varepsilon_{it},$$

where $\varepsilon_{it}$ is a zero mean normal error and $\theta_i \geq 0$ represents technical inefficiency. The distribution of this (in)efficiency measure is of interest. If we rewrite the above model as $y_{it} = x_{it}'\beta + \alpha_i + \varepsilon_i$, for $\alpha_i := \alpha - \theta_i$, it takes the form of a standard linear fixed-effect model. We follow Schmidt and Sickles (1984) and first estimate the $\alpha_i$ by standard linear regression for each farm $i$, say $\hat{\alpha}_i$, and next construct the estimator $\vartheta_i = \max_i(\hat{\alpha}_i) - \hat{\alpha}_i$ for the (in)efficiency parameter $\theta_i$ (we use heteroskedasticity-robust standard errors). By doing so we are normalizing the most efficient firm in the sample as being 100% efficient.

Standard statistical packages report estimates of the mean and standard deviation of the technical inefficiency measure. In our data, these conventional plug-in estimates equal .3490 (with a standard error of .0103) and .1611 (with a standard error of .0078), respectively. The bias-corrected estimator of the standard deviation is .1361 (with a standard error of .0092).

In Figure 2 we present analytically- and jackknife-corrected nonparametric estimates of the full distribution function of the inefficiency parameter (solid blue), along with 95% confidence bands (*). Each plot also contains the (uncorrected) empirical distribution function of the $\vartheta_i$ (solid black). Both corrections yield similar estimators of $F$, and both have smaller tails than the plug-in estimator.

# References


Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*, 251–333.

Belotti, F., S. Daidone, G. Ilardi, and V. Atella (2013). Stochastic frontier analysis using Stata. *Stata Journal 13*, 719–758.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of West German wage inequality. *Quarterly Journal of Economics 128*, 967–1015.

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

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika 68*, 589–599.

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

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

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

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.

Li, H., B. Lindsay, and R. Waterman (2003). Efficiency of projected score methods in rectangular array asymptotics. *Journal of the Royal Statistical Society, Series B 65*, 191–208.

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.

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

Okui, R. and T. Yanagi (2016). Panel data analysis with heterogenous dynamics. Mimeo.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*, 417–458.

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.

Sartori, N. (2003). Modified profile likelihood in models with stratum nuisance parameters. *Biometrika 90*, 533–549.

Schmidt, P. and R. C. Sickles (1984). Production frontiers and panel data. *Journal of Business & Economic Statistics 2*, 367–374.

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

# INFERENCE ON A DISTRIBUTION FROM NOISY DRAWS

## SUPPLEMENTARY MATERIAL

Koen Jochmans* University of Cambridge

Martin Weidner† University College London

January 18, 2018

**Notational convention:** we let $\nabla_p^q \varphi$ denote the $q$th derivative of $\varphi$ with respect to its $p$th argument. We omit the subscript for univariate $\varphi$.

# Appendix A: Auxiliary results

**Lemma A.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 $n \to \infty$,*

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

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

**Lemma A.2** (Komlós, Major and Tusnády 1975)**.** *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]. Let $\mathbb{B}_n$ denote a sequence of Brownian bridges. Then*

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

*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. E-mail: m.weidner@ucl.ac.uk.

**Lemma A.3** (Weak convergence for distribution function)**.** *Let Assumptions 1 and 2 hold. Then,*

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

*as $n \to \infty$ and $m \to \infty$ so that $n/m^4 \to 0$, where $\mathbb{G}_F(\theta)$ is a mean zero Gaussian process with covariance function $\sigma_F(\theta_1, \theta_2)$.*

*Proof.* Let $F_m(\theta) := E(1\{\vartheta_i \leq \theta\})$, the distribution function of $\vartheta_i$. 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.2 with $u = F_m(\theta)$ and exploiting monotonicity of distribution functions then gives

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

Now, Theorem 1 states that, uniformly in $\theta$,

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

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 $\theta$. Furthermore, Theorem 1 implies that $F_m(\theta) - F(\theta)$ converges to zero uniformly in $\theta$ 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 lemma. □

**Lemma A.4.** *Let Assumptions 1 and 2 hold. Let $f_m$ denote the density function of $\vartheta_i$. 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 Theorem 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 $\varepsilon_i^*$ is a value between zero and $\varepsilon_i$ 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 $\theta_i$ given $\sigma_i$ and of $\sigma_i$, respectively. Differentiating with respect to $\theta$ 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 Theorem 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

$$\sup_\theta \sup_\delta |\nabla_1^3 G(\theta,\delta)| = \sup_\theta \sup_\delta |\int_{\underline{\sigma}}^{\overline{\sigma}} \nabla_1^2 h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma| \leq |\int_{\underline{\sigma}}^{\overline{\sigma}} e(\sigma)\, h(\sigma)\, d\sigma| < \infty,$$
$$\sup_\theta \sup_\delta |\nabla_1^4 G(\theta,\delta)| = \sup_\theta \sup_\delta |\int_{\underline{\sigma}}^{\overline{\sigma}} \nabla_1^3 h(\theta - \delta\sigma|\sigma)\, h(\sigma)\, d\sigma| \leq |\int_{\underline{\sigma}}^{\overline{\sigma}} e(\sigma)\, h(\sigma)\, d\sigma| < \infty,$$

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. □

**Lemma A.5.** *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 $\theta$ and that there exists a function $e$ so that $|\nabla_1^{q+2} h(\theta|\sigma)| \leq e(\sigma)$ and $E(e(\sigma_i)) < \infty$.*

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

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

A change of variable from $\theta$ 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 $\theta$,

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

as claimed. This completes the proof. □

**Lemma A.6.** *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 $\theta_i$ evaluated in $\varepsilon$ 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

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

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

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

Consider the second term on the right-hand side in (A.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 $\vartheta$. To see that this is so, note that

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

is bounded by

$$\int_{-\infty}^{\infty} \max_{\sigma\in[\underline{\sigma},\overline{\sigma}]} \left\{ \tfrac{1}{\sigma/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\sigma/\sqrt{m}}\right)\right\} d\varepsilon = 2\int_{0}^{\infty} \max_{\sigma\in[\underline{\sigma},\overline{\sigma}]} \left\{ \tfrac{1}{\sigma/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\sigma/\sqrt{m}}\right)\right\} d\varepsilon.$$

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

$$\int_{0}^{\infty} \max_{\sigma\in[\underline{\sigma},\overline{\sigma}]} \left\{ \tfrac{1}{\sigma/\sqrt{m}}\phi\left(\tfrac{\varepsilon}{\sigma/\sqrt{m}}\right)\right\} d\varepsilon = \tfrac{e^{-1/2}}{\sqrt{2\pi}} \log(\overline{\sigma}/\underline{\sigma}) + \tfrac{1}{2} < \infty,$$

as claimed. For the first right-hand side term in (A.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

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

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 $\vartheta$, as claimed. This completes the proof of the lemma. □

**Lemma A.7.** *Let the assumptions of Proposition 1 hold. Then,*

$$\sup_\theta E(\hat{b}_F(\theta) - b_F(\theta)) = O(m^{-1}) + O(h^2), \qquad \sup_\theta \operatorname{var}(\hat{b}_F(\theta)) = O(n^{-1}h^3).$$

*Proof.* Consider the bias result first. 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

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

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 $\varepsilon^*$ lies between $\varepsilon$ and zero. From Lemma A.5 we have

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

uniformly in $\theta$, 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.

Next, to establish the variance result note that

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

Now, with

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

we have

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

which is $O(n^{-1}h^3)$ uniformly in $\theta$ as $\sup_\theta |\beta_m^2(\theta)| < \infty$ because $\sigma_i$ 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 $\theta$. This completes the proof. □

**Lemma A.8.** *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 $\epsilon > 0$,*

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

*Proof.* 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,

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

where we have used the definition of $b_i(\theta)$ in the first step, boundedness of the $\sigma_i$ 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 $\theta$, in the final step. This completes the proof. □

**Lemma A.9.** *Let the assumptions of Proposition 1 hold. Then,*

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

*Proof.* 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 $\sigma_i^2$ 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 $\delta_a$ is Dirac's delta at $a$. Then, writing $\nabla_u$ for the derivative with respect to $u$, we get

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

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,

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

Clearly, this term is bounded uniformly on any finite interval. To evaluate it for large

values of $\theta$, observe that

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

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

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

and $f_m(\theta) = O(|\theta|^{-\kappa})$ as $|\theta|\to\infty$ by Lemma A.6, 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 (A.3) is $O(h|\theta|^{-\kappa}\,\log(1+|\theta|))$ uniformly in $\theta$. Next, for the second term in (A.2) we use Lemma A.1 to establish that, for any $\epsilon\in(0,1/2]$, we have

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

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

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

uniformly in $\theta$. 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. □

# Appendix B: Proofs of results in the main text

**Proof of Theorem 1.** We first derive the bias expression under Assumption 2.A and Assumption 2.B, in turn, and then establish the variance result.

Under Assumption 2.A $(\theta_i, \sigma_i)$ have a joint density which we will write as $h(\theta_i, \sigma_i)$. We will denote the marginal density of $\sigma_i$ by $h(\sigma_i)$ and the conditional density of $\theta_i$ given $\sigma_i$ by $h(\theta_i|\sigma_i)$. For any real number $\delta$ 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

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

Assumption 2.A implies that $G$ is smooth and differentiable in its second argument. A fourth-order expansion of $G(\theta, \varepsilon_i/\sqrt{m})$ around zero gives

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

where $\varepsilon_i^*$ is some value between zero and $\varepsilon_i$ and where we have exploited that $\varepsilon_i \sim N(0, 1)$, and so its odd moments are zero. Now, by direct calculation,

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

and, by definition of the function $e(\sigma_i)$,

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

which equals $E(\sigma_i^4 e(\sigma_i))$ and is finite by assumption. Therefore,

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

under Assumption 2.A, as claimed.

Under Assumption 2.B we have a deterministic relationship between $\theta_i$ and $\sigma_i$. We may define $G(\theta, \delta)$ as above but have to take care when Taylor expanding it in $\delta$, as the function

may be non-continuous. A non-continuity in 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 $\sigma'$, there always exists $\eta > 0$ such that for all $\delta \in (-\eta, \eta)$ and all real $\theta$ 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 $\delta$ 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 $\sigma$ up to third and fourth order, resepctively, our assumption that these derivatives are uniformly bounded implies that

$$\sup_{\theta} \sup_{\delta \in (-\eta, \eta)} \left| \nabla_2^4 G(\theta, \delta) \right| < \infty \tag{B.1}$$

for some $\eta > 0$. The only obstacle that now prevents us from proceeding with an expansion as we did under Assumption 2.B is that the bound (B.1) 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 $\varepsilon_i$ is normally distributed we also have that

$$P\left(|\varepsilon_i| > m^{\alpha}\right) = O(m^{-\gamma}) \qquad \text{for any } \alpha, \gamma > 0$$

as $m \to \infty$; we set $\alpha \in (0, 1/2)$ in the argument to follow. 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 $\theta$. 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 $\varepsilon_i$ between zero and $\varepsilon_i$. Because $\varepsilon_i$ is normally distributed 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 (B.1) 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 $\theta$. This completes the proof of the bias expression under Assumption 2.B.

For the result on the covariance, finally, 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 $\theta$, 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 theorem. □

**Proof of Proposition 1.** 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 $\nu$, 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 $\tau$ lives in a bounded interval. From Lemma A.9, using the chain rule of differentiation, it follows that

$$\sup_{\tau \in (-\nu^{-1}, \nu^{-1})} \left| \nabla^1_\tau \hat{b}_F(t^{-1}(\tau)) - \nabla^1_\tau b_F(t^{-1}(\tau)) \right| = O_p(h^{-(1+\omega^{-1})}), \tag{B.2}$$

where we use the notation $\nabla_\tau$ to indicate derivatives with respect to $\tau$. We therefore have that $\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))$, as a function $\tau$, 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 (B.2) 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

$$\max_{\tau \in I_h} |\hat{b}_F(t^{-1}(\tau)) - b_F(t^{-1}(\tau))| + O_p(h^2). \tag{B.3}$$

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 the first part of Lemma A.7 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

$$\begin{aligned} E\left(\left(\max_{\tau \in I_h} \left|\hat{\Delta}(t^{-1}(\tau))\right|\right)^\omega\right) &= E\left(\max_{\tau \in I_h} \left|\hat{\Delta}(t^{-1}(\tau))\right|^\omega\right) \\ &\leq E\left(\sum_{\tau \in I_h} \left|\hat{\Delta}(t^{-1}(\tau))\right|^\omega\right) \\ &= \sum_{\tau \in I_h} E\left(\left|\hat{\Delta}(t^{-1}(\tau))\right|^\omega\right) \leq \lceil h^{-3+1/\omega} \rceil \sup_{\theta \in \mathbb{R}} E\left|\hat{\Delta}(\theta)\right|^\omega. \end{aligned} \tag{B.4}$$

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 $\omega$. Using the second part of Lemma A.7 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.8 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 (B.4) 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 (B.3) and (B.4) 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 1 now follows immediately from Lemma A.3. □

**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. □

**Proof of Corollary 1.** The $\vartheta_i$ are i.i.d. draws from the distribution $F_m$ which according to Lemma A.4 has non-degenerate density $f_m$, that is, the $\vartheta_i$ are continuously distributed. Thus,

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

is the $k$th 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

$$\sqrt{n}(u_{(k)} - \tau) \xrightarrow{d} N(0, \tau(1-\tau)). \tag{B.5}$$

Let $q_m(\tau) := F_m^{-1}(\tau)$, the $\tau$th-quantile of $F_m$. By expanding the function $F_m^{-1}$ around $\tau$ 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 (B.5) we have $u_{(k)} - \tau = O_P(n^{-1/2})$. This implies that $\xi_{(k)} \xrightarrow{p} \tau$. Using Lemma A.4 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.4 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

$$\hat{q}(\tau) = q_m(\tau) + \frac{u_{(k)}-\tau}{f(q(\tau))} + O_p(n^{-1} + n^{-1/2}m^{-1}). \tag{B.6}$$

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

$$q_m(\tau) = q(\tau) - \frac{b_F(q(\tau))/f(q(\tau))}{m} + O(m^{-2}). \tag{B.7}$$

Combining (B.5), (B.6), and (B.7) gives the statement of the theorem. □

**Proof of Proposition 2.** 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.2 and Theorem 1 in Doss and Gill (1992) give

$$\sup_{\tau\in[0,1]} \left|\sqrt{n}\left(\mathbb{G}_n^{\leftarrow}(\tau) - \tau\right) + \mathbb{B}_n(\tau)\right| = o_P(1), \tag{B.8}$$

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.2.

Equation (B.8) 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.4 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 $\tau$. The last result therefore also implies that

$$\check{q}(\tau) - \hat{q}(\tau) = O_p(m^{-1}) + o_p(n^{-1/2}). \tag{B.9}$$

Next, The result (B.8) 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 Theorem 1 we know that $F_m(\theta) = E(\hat{F}(\theta)) = F(\theta) + b_F(\theta)/m + O(m^{-2})$, uniformly in $\theta$. 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 1 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 (B.9) 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. □

# Appendix C: Results for smooth functionals

Here we provide formal derivations for the plug-in estimator

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

of $\mu := E(\varphi(\theta_i))$ when $\varphi$ is smooth.

**Assumption C.1** (Regularity for linear functionals)**.** *$E(\varphi(\theta_i)^2) < \infty$ and there exists a $q \geq 3$ such that (i) $\varphi$ is $(q+1)$ times differentiable; (ii) $E(\nabla^p\varphi(\theta_i)^4) < \infty$ for $p = 1, \ldots, q$; and (iii) $\sup_\theta|\nabla^{q+1}\varphi(\theta)| < \infty$.*

Assumption C.1 collects standard conditions that validate a $q$th-order Taylor expansion and allow to control the remainder term.

We let

$$b_\mu := \frac{E(\nabla^2\varphi(\theta_i)\,\sigma_i^2)}{2}$$

and $\sigma_\mu^2 := \mathrm{var}(\varphi(\theta_i))$ in the following theorem.

**Theorem C.1** (First-order bias and variance for linear functionals)**.** *Let Assumptions 1 and C.1 hold. Then, as* $n, m \to \infty$*,*

$$E(\hat{\mu}) - \mu = \frac{b_\mu}{m} + O(m^{-2}), \quad \text{var}(\hat{\mu}) = \frac{\sigma_\mu^2}{n} + O\left(n^{-1}m^{-1}\right).$$

*Proof.* A Taylor expansion gives

$$\varphi(\vartheta_i) - \varphi(\theta_i) = \sum_{p=1}^{q} \frac{1}{p!} \nabla^p \varphi(\theta_i)(\vartheta_i - \theta_i)^p + \frac{1}{(q+1)!} \nabla^{q+1} \varphi(\vartheta_i^*)(\vartheta_i - \theta_i)^{(q+1)},$$

where $\vartheta_i^*$ lies between $\vartheta_i$ and $\theta_i$.

We first compute the mean of this expression. The normal distribution is completely determined by its first two moments. Moreover,

$$E((\vartheta_i - \theta_i)^p \,|\, \theta_i, \sigma_i) = \begin{cases} 0 & \text{if } p \text{ is odd} \\ (p-1)!!\, \sigma_i^p / m^{p/2} & \text{if } p \text{ is even} \end{cases},$$

where $p!!$ denotes the double factorial of $p$. Therefore, the law of iterated expectations gives

$$E(\varphi(\vartheta_i) - \varphi(\theta_i)) = \sum_{p \text{ is even}}^{q} \frac{(p-1)!!}{p!} \frac{E(\nabla^p \varphi(\theta_i)\, \sigma_i^p)}{m^{p/2}} + O(m^{-(q+1)/2}).$$

The order of the remainder term follows from the fact that $\sup_\theta |\nabla^{q+1} \varphi(\theta)| < b$ for some finite $b$, so that

$$E(|\nabla^{q+1} \varphi(\vartheta_i^*)(\vartheta_i - \theta_i)^{(q+1)}|) \leq b\, E(|(\vartheta_i - \theta_i)^{(q+1)}|) = O(m^{-(q+1)/2}),$$

where the last step follow from

$$E(|(\vartheta_i - \theta_i)^{(q+1)}|) \leq \begin{cases} E(\sigma_i^{q+1})/m^{(q+1)/2} & \text{if } q \text{ is uneven} \\ (E(\sigma_i^{q+2})/m^{(q+2)/2})^{\frac{q+1}{q+2}} & \text{if } q \text{ is even} \end{cases},$$

and we have used Hölder's inequality for the case where $q$ is even. We have established that

$$E(\hat{\mu} - \mu) = m^{-1} \frac{E(\nabla^2 \varphi(\theta_i)\, \sigma_i^2)}{2} + O(m^{-2}),$$

which corresponds to the bias expression in the theorem.

To obtain the variance it suffices to show that all right-hand side terms in the expansion have a variance that is $O(m^{-1})$ as $n \to \infty$. For the $q$ leading terms, proceeding as before gives

$$E(\nabla^p \varphi(\theta_i)^2 (\vartheta_i - \theta_i)^{2p})^2 \leq E(\nabla^p \varphi(\theta_i)^4)\, E((\vartheta_i - \theta_i)^{4p}) = O(m^{-2p}).$$

Similarly,

$$E(|\nabla^{q+1} \varphi(\vartheta_i^*)^2 (\vartheta_i - \theta_i)^{2(q+1)}|) \leq b^2\, E((\vartheta_i - \theta_i)^{2(q+1)}) = O(m^{-(q+1)}).$$

Therefore,

$$\operatorname{var}(\hat{\mu}) = \frac{\operatorname{var}(\varphi(\vartheta_i))}{n} = \frac{\operatorname{var}(\varphi(\theta_i)) + O(m^{-1})}{n}.$$

This is the variance result stated in the theorem. □

Define the plug-in estimator of $b_\mu$

$$\breve{\mu} := \hat{\mu} - \frac{\hat{b}_\mu}{m}, \qquad \hat{b}_\mu := \frac{n^{-1} \sum_{i=1}^n \nabla^2 \varphi(\vartheta_i)\, \sigma_i^2}{2},$$

and impose the following additional requirement.

**Assumption C.2** (Regularity for linear functionals, cont'd)**.** $E(\nabla^p \varphi(\theta_i)^8) < \infty$ *for* $p = 1, \ldots, q$.

Then we have the following result.

**Proposition 1** (Bias correction for linear functionals)**.** *Let Assumptions 1–C.2 hold. Then*

$$\breve{\mu} \overset{a}{\sim} N\left(\mu, \sigma_\mu^2 / n\right)$$

*as* $n \to \infty$ *and* $m \to \infty$ *so that* $n/m^4 \to 0$.

*Proof.* Recall that the leading bias in $\hat{\mu}$ equals

$$b_\mu := \frac{E(\nabla^2 \varphi(\theta_i)\, \sigma_i^2)}{2}.$$

An expansion of its plug-in estimator $\hat{b}_\mu$ around $\theta_i$ allows to write

$$\hat{b}_\mu = \frac{n^{-1} \sum_{i=1}^n \nabla^2 \varphi(\theta_i)\, \sigma_i^2}{2} + n^{-1} \sum_{i=1}^n r_i$$

for

$$r_i := \sum_{p=3}^{q} \frac{\nabla^p \varphi(\theta_i)\, \sigma_i^2\, (\vartheta_i - \theta_i)^{p-2}}{2\,(p-2)!} + \frac{\nabla^{q+1} \varphi(\vartheta_i^*)\, \sigma_i^2\, (\vartheta_i - \theta_i)^{q-1}}{2\,(q-1)!}.$$

Using the same arguments as used in the proof of Theorem C.1 it can be verified that each of the contributions to this remainder term has mean $O(m^{-1})$ and variance $O(m^{-1})$. Also, the dominant term in $\hat{b}_\mu$ is an unbiased estimator of $b_\mu$ whose variance is

$$E\left(\left(\frac{n^{-1}\sum_{i=1}^{n} \nabla^2 \varphi(\theta_i)\, \sigma_i^2}{2} - b_\mu\right)^2\right) = O(n^{-1}).$$

Therefore, $\hat{b}_\mu = b_\mu + O(m^{-1}) + O_p(n^{-1/2})$ which leads to the limit result stated in the proposition. □

# References


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.

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.

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

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.