Statistica Sinica Preprint No: SS-2016-0416.R1	
Title	Semi-parametric prediction intervals in small areas when auxiliary data are measured with error
Manuscript ID	SS-2016-0416.R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0416
Complete List of Authors	Gauri Datta Aurore Delaigle Peter Hall and Li Wang
Corresponding Author	Aurore Delaigle
E-mail	aurored@unimelb.edu.au

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

Semi-parametric prediction intervals in small areas when auxiliary data are measured with error

Gauri Datta[1], Aurore Delaigle[2], Peter Hall[2] and Li Wang[3]

[1]Department of Statistics, University of Georgia, Athens, GA 30602, USA.

[2]Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and School of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia.

[3]Department of Statistics and the Statistical Laboratory, Iowa State University, Ames, IA 50011, USA.

Foreword: *Our friend and colleague Peter Hall died in Melbourne, Australia on January 9, 2016, before this manuscript was completed. Peter worked hard on this paper before he fell ill, deriving all the theoretical results of the manuscript, whence our decision to submit this manuscript for the special issue in his honour. His theory is particularly striking since it reveals that the properties of the prediction interval depend on whether or not the contaminated covariate takes the value zero. Peter was very interested by this phenomenon but could not find an intuitive explanation to it. We checked his proofs thoroughly but could not find an intuitive explanation either, except that a similar behaviour is sometimes encountered in other problems.*

Abstract: In recent years, demand for reliable small area statistics has considerably increased, but the size of samples obtained in small areas is too often small to produce accurate predictors of quantities of interest. To overcome this difficulty, a common approach is to use auxiliary data from other areas or other sources, and produce estimators that combine them with direct data. A popular model for combining direct and indirect data sources is the Fay-Herriot model, which assumes that the auxiliary variables are observed accurately. However, these variables are often subject to measurement errors, and not taking this into account can lead to estimators that are even worse than those based exclusively on the direct data. We consider structural measurement error models and a semi-parametric approach based on the Fay-Herriot model to produce reliable prediction intervals for small area characteristics of interest. Our theoretical study reveals the surprising fact that the properties of the prediction interval are not the same for all values of the noisy covariate. Indeed, the convergence rates are slower when the contaminated covariate takes the value zero than in other cases. Our procedure is illustrated with an application and simulation studies.

Keywords: Deconvolution; density estimation; Fay-Herriot model; Fourier transform; Laplace distribution.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

1 Introduction

Small area estimation methods are indispensable statistical tools to the administrators and policy makers in National Statistical Offices and many world organizations. Economic planning and welfare activities of the governments and non-government organizations rely heavily on accurate data measuring income, employment, living and health conditions for various geographic and demographic segments. While nationwide surveys such as the American Community Survey, the National Health and Nutrition Examination Survey (NHANES), and the National Health Interview Survey (NHIS), collect large samples at the national level, the subset of the data collected in local geographic and demographic domains, also known as small areas, is usually of too small size to compute accurate small area statistics.

To produce more reliable estimators at the small area level, a common approach is to use model-based methods which combine data from multiple sources, surveys, administrative records, registers, and social media; see for example Rao (2003), Pfeffermann (2013), Ybarra and Lohr (2008), and Rao and Molina (2015). Suppose we are interested in predicting a random quantity T_j in small area j, where $j = 1, \ldots, n$, and that we have at our disposal a sample of independent pairs $(A_1, Y_1), \ldots, (A_n, Y_n)$, where the component A_j is a vector of auxiliary variables and Y_j is a direct estimator of T_j. The estimator Y_j is computed based on a sample only from area j, and it is referred to in small area estimation literature as a direct estimator (see Rao and Molina (2015)). A popular small area estimation model assumes the decomposition $Y_j = T_j + \epsilon_j$, where $\epsilon_j \sim N(0, \tau_j)$ and the T_j's and the ϵ_j's are completely independent. We shall follow the small area literature and assume that the τ_j's known; see González-Manteiga et al. (2010) for how these can be estimated in practice. See also Otto and Bell (1995). The widely used Fay-Herriot model (Fay and Herriot (1979)) decomposes T_j as $T_j = \beta_0 + \beta_1 A_j + V_j$, where $V_j \sim N(0, \sigma_V^2)$, with various inde-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

pendence models in use. Here we assume that the A_j's, the V_j's and the ϵ_j's are completely independent.

A difficulty in applications is that it is not always possible to measure all the components of the auxiliary vectors A_j accurately, and the techniques developed for covariates without measurement error may perform rather poorly in this case; see for example, Ghosh et al. (2006), Ybarra and Lohr (2008), and Torabi et al. (2009). In particular, if the measurement error is not taken into account, using the auxiliary A_j's may result in estimators that are even less accurate than those based on the direct data from the small areas; see Ybarra and Lohr (2008), who propose a corrected small area predictor based on the empirical best linear unbiased prediction approach. While Ghosh and Sinha (2007), Ybarra and Lohr (2008), and Datta et al. (2010) used the frequentist approach to the problem, Ghosh et al. (2006), Torabi et al. (2009), Arima et al. (2012, 2015), and Datta et al. (2010), proposed a Bayesian approach.

While the aforementioned methods are useful, their focus is on point predictors, whereas in the small area estimation literature where the covariates are measured without error, there is substantial interest in the construction of prediction intervals for the T_j's; see for example Datta et al. (2002), Hall and Maiti (2006), Chatterjee et al. (2008), and Diao et al. (2014). In this work, our goal is to construct prediction intervals for the small area population means $T_1, \ldots, T_n$ by taking the measurement errors of covariates into account and by relaxing some of the distributional assumptions for the random effects and sampling errors often employed in the literature. Although we focus on the construction of those intervals, we note that since our method is based on estimating the conditional distribution of T_j given the observed data, it can also be used to construct predictors under these relaxed assumptions.

In Section 2.1, we introduce a Fay-Herriot model with a covariate subject to measurement error. To construct prediction intervals, we derive the conditional dis-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

tribution of the small area characteristic T_j given the values of its direct estimator Y_j and other observed data, and suggest estimators of this conditional distribution in Section 2.2. Section 3 gives the asymptotic properties of the estimators of the model parameters and the prediction intervals. Simulation studies and an illustrative example are presented in Section 4. In Section 5, we extend our approach to the case where one of the error distributions is unknown. Estimation of the parameters is relatively standard and is deferred to Appendix B. Proofs of the theoretical results are given in Section 6 and in Appendices E and F.

2 Model and estimators

2.1 Model and data

We are interested in predicting T_j, for $j = 1, \ldots, n$. We have at our disposal a sample of independent $(p+2)$-vectors $(W_j, Q_j^{\mathrm{T}}, Y_j)$, for $1 \leq j \leq n$, where the connection between (W_j, Q_j^{T}) and A_j is described in the next pargraph. We assume a measurement error version of the Fay-Herriot model:

$$Y_j = T_j + \epsilon_j\,, \quad T_j = \beta_0 + \beta_1\, X_j + \beta_2^{\mathrm{T}}\, Q_j + V_j\,, \quad W_j = X_j + U_j\,, \tag{2.1}$$

where β_0, β_1, U_j, V_j, W_j, X_j, and Y_j are scalars, β_2 and Q_j are p-vectors, β_0, β_1, and β_2 are unknown, the variables Q_j, U_j, V_j, X_j, and ϵ_j, for $j \geq 1$, are completely independent, the U_j's have a common, known distribution symmetric around zero, the X_j's have a common, unknown distribution, the V_j's have zero mean and unknown variance σ_V^2, the ϵ_j's have a known distribution symmetric around zero and known variance τ_j, and the τ_j's are uniformly bounded. We consider two cases: the distribution of the V_j's is known except for the variance σ_V^2, and the distribution of V is totally unknown.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

The data come from two sources: direct data $Y_1, \ldots, Y_n$, and indirect auxiliary observations or covariates (W_j, Q_j^{T}), which are a partially noisy version of $A_j = (X_j, Q_j^{\mathrm{T}})$; W_j is a noisy version of X_j, and the measurement error U_j reflects the inaccuracy in the measurement process, for example due to sampling variability. The model at (2.1) resembles a classical measurement error linear model, and estimating the unknown parameters can be done using standard methods (Appendix B). However, our prediction problem differs from that one because, since we are in a small area context, we have at our disposal two measurements, Y_j and (W_j, Q_j^T), of the variable T_j to be predicted. In a standard linear prediction problem with errors, we would observed only (W_j, Q_j^T).

Throughout we use $(Q^{\mathrm{T}}, T, W, Y, \epsilon, \tau)$ to denote a generic $(Q_j^{\mathrm{T}}, T_j, W_j, Y_j, \epsilon_j, \tau_j)$, where $\tau = \mathrm{var}(\epsilon)$ is known. (Here, τ denotes the variance of a generic ϵ and so we have dropped the index j in τ_j.) Our aim is to develop methodology for constructing a prediction interval for T, given the value of (Q^{T}, W, Y) using the data $(Q_k^{\mathrm{T}}, W_k, Y_k)$, $1 \le k \le n$. To summarise, we observe $n+1$ triplets: (Q^{T}, W, Y), which corresponds to the individual whose value of T we wish to predict, and $(Q_j^{\mathrm{T}}, W_j, Y_j)$, for $j = 1, \ldots, n$, which we use to construct estimators of all the unknowns in this prediction problem. Of course the procedure can be applied for all individuals in the study, using in each case the other n observations to estimate the unknown quantities.

To do this we need to construct an estimator of the density of T conditional on (Q^{T}, W, Y). In Appendix A, we prove that it is given by

$$f_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \frac{f_\epsilon(t-y) \int f_V(t-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q)\, f_X(x)\, f_U(w-x)\, dx}{\int f_{V+\epsilon}(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q)\, f_U(w-x)\, f_X(x)\, dx}\,, \quad (2.2)$$

where β_0, β_1, β_2, the variance σ_V^2 of V, and the density f_X are unknown. Estimating the unknown parameters is relatively standard; see Appendix B. In Section 2.2, we show how to estimate the other unknown quantities, and deduce our prediction intervals. In practice it is commonly assumed that f_V and f_ϵ are known (usually normal),

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

and for the main part of this work we shall focus on that setting. However, we shall also see that it is possible to relax this assumption; see Section 5.

2.2 Prediction intervals

To construct a prediction interval for T, we need to estimate the conditional distribution $F_{T\,|\,Q,W,Y}$ corresponding to the density $f_{T\,|\,Q,W,Y}$ at (2.2). The latter depends on f_X, f_V and f_ϵ. In the small area literature, it is often assumed that f_ϵ is known and f_V is known up to its variance σ_V^2, which is the setting we use in this section to derive an estimator of $F_{T\,|\,Q,W,Y}$. In particular, if we let g denote the density of V/σ_V, then $f_V(\cdot) = \sigma_V^{-1} g(\cdot/\sigma_V)$, where g is assumed to be known. In this case, we can use relatively standard deconvolution methods, and the most interesting aspect of this problem is the theory, which reveals unusual and intriguing properties; see Section 3. Our simulation results in Section 4.2 suggest that our procedure seems relatively robust against misspecification. In Section 5 we derive an estimation procedure in the case where one of those two densities is unknown and estimated from the data.

When f_ϵ and g are known, the only unknowns in (2.2) are β_0, β_1, β_2, and σ_V^2, which can be replaced by standard estimators from the measurement error literature, using techniques discussed in Fuller (2009), Buonaccorsi (2010), and Delaigle and Hall (2011) (see Appendix B), and $f_X(x)$, which can be estimated by the kernel deconvolution estimator of Carroll and Hall (1988) and Stefanski and Carroll (1990):

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K_U\Big(\frac{x - W_j}{h}; h\Big)\,,\ K_U(x; h) = \frac{1}{2\pi} \int \exp(-itx)\ \frac{\phi_K(t)}{\phi_U(t/h)}\ dt\,, \quad (2.3)$$

where ϕ_K is the Fourier transform of a kernel function K, $h > 0$ is a smoothing parameter called bandwidth, and, for any random variable R, $\phi_R(t) = E(e^{itR})$ denotes the characteristic function of R.

Moreover, $f_{V+\epsilon}(s) = \int f_\epsilon(v)\, f_V(s - v)\, dv = \sigma_V^{-1} \int f_\epsilon(v)\, g\{(s - v)/\sigma_V\}\, dv$, which

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

can be estimated by $\hat{f}_{V+\epsilon}(s) = \hat{\sigma}_V^{-1} \int f_\epsilon(v)\, g\{(s-v)/\hat{\sigma}_V\}\, dv$. Using (2.2), we can estimate $F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \int_{-\infty}^{t} f_{T\,|\,Q,W,Y}(s\,|\,q,w,y)\, ds$ by

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \frac{\int_{-\infty}^{t} f_\epsilon(s-y) \int f_V(s-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q)\, \hat{f}_X(x)\, f_U(w-x)\, dx\, ds}{\int \hat{f}_{V+\epsilon}(y-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q)\, f_U(w-x)\, \hat{f}_X(x)\, dx} . \quad (2.4)$$

Next, let $\alpha \in (0,1)$ and define $\hat{t}_\alpha = \hat{t}_\alpha(q,w,y)$ to be the solution, in t, of $\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \alpha$. Approximate one-sided prediction intervals of nominal coverage $1-\alpha$ for T given (Q^{T}, W, Y) can be defined by $\hat{\mathcal{I}}_\alpha^L = (-\infty, \hat{t}_{1-\alpha}]$ or $\hat{\mathcal{I}}_\alpha^R = [\hat{t}_\alpha, \infty)$. Approximate two-sided intervals of nominal coverage $1-\alpha$ can be defined by $\hat{\mathcal{I}}_\alpha = [\hat{t}_{\alpha_1}, \hat{t}_{1-\alpha_2}]$, where $\alpha_1 + \alpha_2 = \alpha$; a typical choice is to take $\alpha_1 = \alpha_2 = \alpha/2$.

In the small area literature it is often assumed that V and ϵ are normally distributed, with zero means and respective variances σ_V^2 and τ. In this case we have

$$F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \Psi_1(t,y,q,w)\big/\Psi_2(t,y,q,w)\,, \quad (2.5)$$

where, for $k = 1, 2$, $\Psi_k(t,y,q,w) = \int \psi_k(t,y,q,w,x)\, f_X(x)\, dx$, and

$$\psi_k(t,y,q,w,x) = \left[\Phi\left\{t - \frac{\sigma_V^2 y + \tau\left(\beta_0+\beta_1 x+\beta_2^T q\right)}{\sigma_V^2+\tau}; \frac{\tau\sigma_V^2}{\sigma_V^2+\tau}\right\}\right]^{2-k} \times \phi\big(y-\beta_0-\beta_1 x-\beta_2^{\mathrm{T}} q\,;\, \sigma_V^2+\tau\big)\, f_U(w-x)\,, \quad (2.6)$$

with $\phi(x;\sigma^2)$ and $\Phi(x;\sigma^2)$ corresponding to the univariate normal density and distribution functions when the distribution has zero mean and variance σ^2.

In that case, the estimator at (2.4) can be simplified in

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \widehat{\Psi}_1(t,y,q,w)\big/\widehat{\Psi}_2(t,y,q,w)\,, \quad (2.7)$$

where $\widehat{\Psi}_k(t,y,q,w) = \int \hat{\psi}_k(t,y,q,w,x)\, \hat{f}_X(x)\, dx$ and

$$\hat{\psi}_k(t,y,q,w,x) = \left[\Phi\left\{t - \frac{\hat{\sigma}_V^2 y + \tau\left(\hat{\beta}_0+\hat{\beta}_1 x+\hat{\beta}_2^T q\right)}{\hat{\sigma}_V^2+\tau}; \frac{\tau\hat{\sigma}_V^2}{\hat{\sigma}_V^2+\tau}\right\}\right]^{2-k} \times \phi\left(y-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^T q; \hat{\sigma}_V^2+\tau\right)\, f_U(w-x)\,. \quad (2.8)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

3 Theoretical properties

The procedure we have derived for the case where g, the density of V/σ_V, and f_ϵ are known, uses relatively standard arguments from the deconvolution literature. However, establishing theoretical properties of the prediction intervals derived in Section 2.2 is rather difficult and requires a number of steps. An interesting aspect of this problem is that the theory reveals some intriguing properties. Indeed, the accuracy of the prediction interval depends on the value taken by W. Specifically, prediction is more difficult when $W = 0$, where parametric rates are not possible, than for other values where we can reach the parametric rate.

In Appendix B we establish consistency of the parameter estimators derived there. In Section 3.2, we derive theoretical properties of the estimator of the conditional distribution $F_{T\,|\,Q,W,Y}$ from Section 2.2. In Section 3.3 we investigate theoretical properties of the quantile estimators $\hat{t}_\alpha$ defined in Section 2.2. Finally, we deduce the theoretical properties of our prediction intervals in Section 3.4.

Throughout this section we assume that the characteristic function ϕ_U is real-valued and does not vanish on the real line. Moreover, we establish our results under the assumption that V and ϵ are normally distributed. In particular, we consider prediction intervals based on the version of the estimator $\widehat{F}_{T\,|\,Q,W,Y}$ of $F_{T\,|\,Q,W,Y}$ defined at (2.7). It would be possible to extend our results to the prediction intervals under other distributional assumptions, as well as for the fully nonparametric version we shall suggest in Section 5, but the arguments would be even more tedious than those for the estimator from Section 2.2, which are already quite long and technical.

3.1 Conditions

We start by describing conditions that will be needed to prove our results. Recall the notation ψ_k from Section 2.2. Define $\chi(t) = \chi(t\,|\,s, y, q, w) = \int e^{itx}\,\psi_k(s, y, q, w, x)\,dx$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

for either $k = 1$ or $k = 2$, let $\chi_1 = \Re\,\chi$, $\chi_2 = \Im\,\chi$ denote the real and imaginary parts of χ, put $\rho_j = \chi_j/\phi_U$, let $r \geq 1$ be an integer, and set $\beta(t) = t^{2r}\,\phi_U(t)$,

$$\Psi_{kr}(x) = \left(\frac{\partial}{\partial x}\right)^{2r-1} \psi_k(s, q, w, x)\,.$$

We assume that, for constants $C_1, C_2, \ldots$,

(i) Ψ_{kr} has a jump discontinuity of size s_k at w, and $(1 + |x|)\,|\Psi_{kr}^{(\ell)}(x)|$ is bounded and integrable on $(-\infty, w) \cup (w, \infty)$ for $\ell = 1, 2, 3, 4$; (ii) $\phi_U(t)$ is real valued and does not vanish for any real t, ϕ_K is real valued and compactly supported, $|\phi_K'|$ is bounded, and $|\rho_j'|$ is bounded; (iii) $|\phi_W(t)| \leq C_1\,(1+|t|)^{-C_2}$, where $C_1 > 0$, $C_2 > 2$; (iv) $\beta(t) = b_1 + O(|t|^{-b_2})$ as $|t| \to \infty$, where $b_1, b_2 > 0$, and $|\beta'(t)| \leq C_3\,(1+|t|)^{-C_4}$ for all t, where $C_3 > 0$, $C_4 > 1$; (3.1)

and that for each $w \in \mathbb{R}$,

$$\max_k \sup_{s,y,q} \sup_{-\infty<t<\infty} |\chi(t\,|\,s, y, q, w)/\phi_U(t)| < \infty\,, \tag{3.2}$$

where $\max_k$ denotes the maximum over $k = 1, 2$ and the second supremum is over s, y, and q in compact sets $\mathcal{S} \subset \mathbb{R}$, $\mathcal{Y} \subset \mathbb{R}$, and $\mathcal{Q} \subset \mathbb{R}^p$, respectively. See Appendix C for a discussion of these conditions. We assume too that, for an integer $\ell \geq 1$,

the functions $\lambda_k(u\,|\,s, y, q, w) \equiv \int \psi_k(s, y, q, w, x + u)\,f_X(x)\,dx$, for $k = 1, 2$ have ℓ partial derivatives with respect to u, for u in a neighbourhood of the origin, and those derivatives are bounded uniformly in $s \in \mathcal{S}$, $y \in \mathcal{Y}$, and $q \in \mathcal{Q}$. (3.3)

Defining the function $\lambda_k^{[\ell]}(u\,|\,s, y, q, w) = (\partial/\partial u)^\ell\,\lambda_k(u\,|\,s, y, q, w)$, we assume that:

for $k = 1, 2$ $\lambda_k^{[\ell]}(u\,|\,s, y, q, w)$ is continuous in a neighbourhood of $u = 0$. (3.4)

Of the kernel K and bandwidth h we assume additionally that:

$\int (1 + |u|)^\ell\,|K(u)|\,du < \infty$, where ℓ is as in (3.3), $\kappa_j \equiv \int u^j\,K(u)\,du = 0$ for $1 \leq j \leq \ell - 1$, and $\int K = 1$; and $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$. (3.5)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

Another assumption will be useful to prove some of our results:

(a) the bandwidth h is chosen such that $\int E(\hat{f}_X^2) = O(n^a)$, where $a \geq 0$, and $n^{1-\eta}\, h \to \infty$ for some $\eta > 0$; (b) if a is as in part (a), then $n^{a+\varepsilon}\, h = O(1)$ as $n \to \infty$, for some $\varepsilon > 0$; and (c) the random quantities Q, U, V and X all have at least ν finite moments, where the value of $\nu \geq 4$ depends on a and ε in parts (a) and (b). (3.6)

Part (a) of (3.6) is milder than the condition usually assumed, where h would be chosen so that $\int E(\hat{f}_X - f_X)^2$ converges to zero as $n \to \infty$, and hence $\int E(\hat{f}_X^2) = O(1)$.

3.2 Consistency of conditional distribution estimator

The next theorem establishes consistency of the conditional distribution estimator $\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y)$ defined at (2.7). Its proof is provided in Section 6.1.

Theorem 1. *Assume the conditions imposed in Theorem 4, and that* (3.1)–(3.3) *and* (3.5) *hold. Then: (i) For each real t and y, and each $q \in \mathbb{R}^p$,*

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) - F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \begin{cases} O_p\Big\{(nh)^{-1/2} + h^\ell\Big\} & \text{if } w = 0 \\ O_p\big(n^{-1/2} + h^\ell\big) & \text{if } w \neq 0\,, \end{cases} \tag{3.7}$$

where, if (3.4) *holds, when $w = 0$, the term $O_p(h^\ell)$ can be written more explicitly as $\big(c_{01}/\tau_{02} - \tau_{01}\; c_{02}/\tau_{02}^2\big)\, h^\ell + o_P\big(h^\ell\big)$, where τ_{0k} and c_{0k}, $k = 1, 2$, are defined in* (6.1). *(ii) For each $\eta > 0$,*

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) - F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \begin{cases} O_p\Big\{\big(n^{1-\eta}h\big)^{-1/2} + h^\ell\Big\} & \text{if } w = 0 \\ O_p\big(n^{-(1-\eta)/2} + h^\ell\big) & \text{if } w \neq 0\,, \end{cases}$$

uniformly in t, q and y in any compact subsets of their respective domains, where in the case $w = 0$ we ask in addition that $n^{1-\eta}h \to \infty$.

When $w = 0$, we deduce from Theorem 1 that, by choosing $h \asymp n^{-1/(2\ell+1)}$, $\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y)$ converges to $F_{T\,|\,Q,W,Y}(t\,|\,q,w,y)$ at the rate $n^{-\ell/(2\ell+1)}$, in a pointwise sense. When $w \neq 0$ we obtain root-n consistency: $\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

$F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) + O_p\big(n^{-1/2}\big)$, for any choice of h satisfying $h = O(n^{-1/(2\ell)})$. It follows from the proof of the theorem that the rates at (3.7) for $w = 0$ cannot be improved. In particular, the convergence rates for $w = 0$ are slower than those for $w \neq 0$. We do not have an intuitive explanation for this. However, it is encountered in related problems, also without an intuitive justification; see for example Hall and Lahiri (2008), where the authors prove that the rates of their estimator are optimal even in the neighbourhood of the origin.

The methods used to derive Theorem 1 can be employed to show that, under the same conditions, all partial derivatives of $\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y)$ with respect to t converge at the same rate to the respective derivatives of $F_{T\,|\,Q,W,Y}(t\,|\,q,w,y)$. See Theorem 5 in Appendix D. These results are of independent interest, but they are also particularly useful for deriving the properties of our prediction intervals in Section 3.4.

3.3 Theoretical properties of quantiles estimators

Let $\alpha \in (0,1)$, take $t_\alpha = t_\alpha(q,w,y)$ to be the solutions, in t, of $F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \alpha$, and recall the definition of $\hat{t}_\alpha = \hat{t}_\alpha(q,w,y)$ in Section 2.2. Strictly speaking, there is a small probability that $\hat{t}_\alpha$ is not uniquely defined, but since the probability converges to 0 at a rate faster than n^{-1} then the event of non-uniqueness can be neglected in Theorems 2 and 3. Our proofs are valid under the assumption that $\hat{t}_\alpha$ is the solution nearest to t_α, in cases where there is ambiguity.

It can be seen from part (i) of Theorem 2 that asymptotic properties of $\hat{t}_\alpha - t_\alpha$ are readily and directly deducible from Theorem 1. In particular, $\hat{t}_\alpha - t_\alpha$ converges to zero at the same rate as $\widehat{F}_{T\,|\,Q,W,Y}(t_\alpha\,|\,q,w,y) - F_{T\,|\,Q,W,Y}(t_\alpha\,|\,q,w,y)$, with the same distinction between the cases $w = 0$ and $w \neq 0$ as in Thoerem 1. See Appendix G for a proof.

Theorem 2. *Assume the conditions imposed in Theorem 4, and that* (3.1)–(3.3) *and*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

(3.5) *hold. Then: (i) for each* q, w *and* y,

$$\hat{t}_\alpha(q,w,y) - t_\alpha(q,w,y) \;=\; -\,\frac{\widehat{F}_{T\,|\,Q,W,Y}(t_\alpha\,|\,q,w,y) - F_{T\,|\,Q,W,Y}(t_\alpha\,|\,q,w,y)}{(\partial/\partial t)\,F_{T\,|\,Q,W,Y}(t\,|\,q,w,y)\,\big|_{t=t_\alpha}} + \begin{cases} O_p\big\{(nh)^{-1} + h^{2\ell}\big\} & \text{if } w = 0 \\ O_p\big(n^{-1} + h^{2\ell}\big) & \text{if } w \neq 0\,; \end{cases}$$

and (ii) for each $\eta > 0$,

$$\hat{t}_\alpha(q,w,y) - t_\alpha(q,w,y) = \begin{cases} O_p\Big\{\big(n^{1-\eta}h\big)^{-1/2} + h^\ell\Big\} & \text{if } w = 0 \\ O_p\big(n^{-(1-\eta)/2} + h^\ell\big) & \text{if } w \neq 0\,, \end{cases}$$

uniformly in q *and* y *in any compact subsets of their respective domains, where in the case* $w = 0$ *we ask in addition that* $n^{1-\eta}h \to \infty$.

3.4 Theoretical properties of prediction intervals

In Theorem 3, we explore the coverage accuracy of the prediction interval $(-\infty, \hat{t}_\alpha]$; similar results can be established for two-sided intervals. In stating the theorem we assume that the vector $(T, Q^{\mathrm{T}}, W, Y)$ is independent of the data $(Q_j^{\mathrm{T}}, W_j, Y_j)$, for $1 \le j \le n$, from which $\hat{t}_\alpha$ is computed. See Appendix H for a proof of the theorem.

Theorem 3. *If* (3.1)–(3.3), (3.5), *and* (3.6) *hold, then*

$$P\Big(T \le \hat{t}_\alpha \;\Big|\; Q = q, W = w, Y = y\Big) = \alpha + \begin{cases} O_p\big\{(nh)^{-1/2} + h^\ell\big\} & \text{if } w = 0 \\ O_p\big(n^{-1/2} + h^\ell\big) & \text{if } w \neq 0\,, \end{cases} \tag{3.8}$$

uniformly in q *and* y *in any compact subsets of their respective domains, and in* $\alpha \in [\alpha_1, \alpha_2]$ *for any* $0 < \alpha_1 < \alpha_2 < 1$.

If we assume (3.4) in addition to the conditions in Theorem 3, and slightly strengthen (3.6)(a), and if we use a more intricate argument than that in Section 6 to determine the "remainder" term in (3.8), then the right-hand side of (3.8)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

can be refined to $\alpha + C_1\,(nh)^{-1/2} + C_6\,h^\ell + o\{(nh)^{-1/2} + h^\ell\}$ when $w = 0$, and to $\alpha + C_5\,n^{-1/2} + C_6\,h^\ell + o(n^{-1/2} + h^\ell)$ otherwise, where C_5 and C_6 are constants. (The more intricate argument is sketched in the last paragraph of the proof of Theorem 3.) However, the implications of this property are rather complex. For example, when $w = 0$ and in cases where C_5 and C_6 are both nonzero, the absolute value of the coverage error is minimised, at $O(n^{-\ell/(\ell+1)})$, by taking h to be of size $n^{-1/(\ell+1)}$, although it is easy to give examples where choosing h so as to produce over- or under-coverage might be advantageous. (Under-coverage, in the context of (3.8), is relevant if our interest is in a prediction interval $[\hat{t}_\alpha, \infty)$ rather than $(-\infty, \hat{t}_\alpha]$.)

4 Numerical properties

4.1 Smoothing parameter choice

To compute our prediction interval, we need to choose the bandwidth h used by $\hat{f}_X$ at (2.3), a notoriously difficult task for nonparametric prediction and confidence intervals. In our contaminated data case, we can exploit the error structure to suggest a selection technique inspired by ideas used in nonparametric errors-in-variable regression. There, instead of trying to consistently estimate optimal smoothing parameters, Delaigle and Hall (2008) suggest numerical approximation procedures based on mimicking the contamination process via resamples of data contaminated with additional levels of noise. We propose a method of that type tailored to our problem.

We describe our approach for the interval $\hat{\mathcal{I}}_\alpha$ (it is straightforward to adapt it to $\hat{\mathcal{I}}^L_\alpha$ and $\hat{\mathcal{I}}^R_\alpha$). If we had access to direct data $T_1, \ldots, T_n$, we would choose h so as to minimise an estimator of coverage error, e.g. $\big\{n^{-1}\sum_{i=1}^n I\big(T_i \in \hat{\mathcal{I}}_{\alpha,-i}\big) - (1-\alpha)\big\}^2$, or, to make h less variable, $\int_{\alpha/2}^{3\alpha/2} \big\{n^{-1}\sum_{i=1}^n I\big(T_i \in \hat{\mathcal{I}}_{\alpha,-i}\big) - (1-\alpha)\big\}^2\,d\alpha$. Since we do not observe the T_i's, we cannot compute this error. Instead we mimic it

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

using contaminated versions of the data. Roughly speaking, our idea is to create a new sample of observations $(Q_i, W_i^\circ, Y_i^\circ)$, $1 \le i \le n$, where, compared to the original (W_i, Y_i)'s, the (W_i°, Y_i°)'s are contaminated with an additional level of error, so that the relationship between X_i and W_i in the original sample is mimicked by that between W_i and W_i°. In this new sample, the variable T_i° that plays the role of T_i in the original sample is observed, and thus we can compute the prediction error of a prediction interval for T_i°. Since the new sample is created in a way that mimics the model at (2.1), our h can be well approximated by minimising that version of prediction error.

To implement these ideas, first we draw a conventional bootstrap resample, $\mathcal{X}^* = \{(Q_i^{*\mathrm{T}}, W_i^*, Y_i^*) : 1 \le i \le n\}$, with replacement from the dataset $\mathcal{X} = \{(Q_i^\mathrm{T}, W_i, Y_i) : 1 \le i \le n\}$. Each triple $(Q_i^{*\mathrm{T}}, W_i^*, Y_i^*)$ is identical to one of the data triples in $\mathcal{X}$; let that triple be $(Q_{j_i^*}^\mathrm{T}, W_{j_i^*}, Y_{j_i^*})$, where, conditional on $\mathcal{X}$, the j_i^*'s, for $1 \le i \le n$, are independent and identically distributed on $1, \ldots, n$. Write $\tau_{j_i^*}$ for the associated value of the variance of $\epsilon_{j_i^*}$, conditional on j_i^*, and let $(\epsilon_i^\dagger, U_i^\dagger)$ denote a pair of random variables that, conditional on both $\mathcal{X}$ and $\mathcal{X}^*$, have respectively the distribution with density $\tau_{j_i^*}^{-1/2} f_1(\,\cdot\,/\tau_{j_i^*}^{1/2})$, with f_1 the density of $\epsilon_i/\tau_i^{1/2}$, and the distribution of U. (We use the notation $\epsilon_i^\dagger$ and $U_i^\dagger$, rather than ϵ_i^* and U_i^*, to recall that $\epsilon_i^\dagger$ and $U_i^\dagger$ are drawn by sampling from known distributions rather than by resampling from a sample.) We take the pairs $(\epsilon_i^\dagger, U_i^\dagger)$, for $1 \le i \le n$, to be independent, conditional on the data. If $Y_i^\circ = Y_i^* + \hat\beta_1 W_i^* + \epsilon_i^\dagger$ and $W_i^\circ = W_i^* + U_i^\dagger$, then $Y_i^\circ = T_i^\circ + \epsilon_i^\dagger$, where $T_i^\circ = Y_i^* + \hat\beta_1 W_i^* = \beta_0 + \hat\beta_1 W_{j_i^*} + \beta_2^\mathrm{T} Q_{j_i^*} + V_i^\circ$ and $V_i^\circ = \beta_1 X_{j_i^*} + V_{j_i^*} + \epsilon_{j_i^*}$.

In essence, the dataset $\mathcal{X}^\circ = \{(Q_i^{*\mathrm{T}}, W_i^\circ, Y_i^\circ) : 1 \le i \le n\}$ is generated by the model at (2.1), the difference being that V_i in (2.1) is replaced here by V_i°, which is generally more variable than V_i. This motivates us to choose h by minimising the prediction error of a prediction interval for T_i° constructed from $\mathcal{X}^\circ$. Specifically, we

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

omit the ith data triple from $\mathcal{X}^\circ$, obtaining $\mathcal{X}^\circ_{-i}$, say, and we use the methodology from Section 2.3 to construct, from $\mathcal{X}^\circ_{-i}$, a prediction interval $\hat{\mathcal{I}}^\circ_{\alpha,-i}$, say, for T°_i, having nominal coverage $1-\alpha$. To reduce variability, we repeat this procedure B times, generating in this way B datsets $\mathcal{X}^\circ_b = \{({Q^*_{i,b}}^{\mathrm{T}}, W^\circ_{i,b}, Y^\circ_{i,b}) : 1 \le i \le n\}$, $b = 1, \ldots, B$, and obtaining their corresponding $T^\circ_{i,b}$ and $\hat{\mathcal{I}}^\circ_{\alpha,-i,b}$. Then, recalling that for these data, the $T^\circ_{i,b}$'s are known, we choose h by minimising the objective function

$$J(h) = \frac{1}{B}\sum_{b=1}^{B}\int_{\alpha/2}^{3\alpha/2}\left\{\frac{1}{n}\sum_{i=1}^{n} I\big(T^\circ_{i,b} \in \hat{\mathcal{I}}^\circ_{\alpha,-i,b}\big) - (1-\alpha)\right\}^2 d\alpha\,. \tag{4.1}$$

Based on our experience, taking $B = 10$ or 20 often suffices to find appropriate smoothing parameters. In our numerical work, we took $B = 10$. All our codes were written in MATLAB, and to reduce computational burden, h was chosen from the grid $h_{PI} \times (0.25, 0.5, 1, 1.5, 2)$, where h_{PI} is the plug-in bandwidth of Delaigle and Gijbels (2004, 2004), computed using the MATLAB code `PI_deconvUknownth4`, available at `http://www.ms.unimelb.edu.au/~aurored/links.html#Code`.

4.2 Simulations

We applied our methodology for constructing two-sided prediction intervals $\hat{\mathcal{I}}_\alpha = [\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2}]$ on simulated examples. We generated data $(W_j, Q_j^{\mathrm{T}}, Y_j)$, $1 \le j \le n$, for $n = 30$ and 50 small areas, from the model:

$$Y_j = T_j + \epsilon_j\,, \quad T_j = 5 + 3\,X_j + 2\,Q_j + V_j\,, \quad W_j = X_j + U_j\,,$$

where $X_j \sim N(5, 9)$, $Q_j \sim \mathrm{Uniform}(0, 5)$, $V_j \sim t(5)$ and $\epsilon_j \sim N(0, \tau_j)$. Rather than choosing the τ_j's, for $j = 1, \ldots, n$, arbitrarily by hand, we generated them from a gamma distribution with mean 8 and standard deviation 4. Finally, we considered two types of measurement errors: $U_j \sim N(0, 3/4)$ and $U_j \sim \mathrm{Laplace}(\sqrt{3/8})$.

Our procedure relies on knowing F_{V/σ_V} and F_ϵ. To examine robustness against misspecified distributions, in each case we compared our prediction interval $\widehat{F}_{T\,|\,Q,W,Y}$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

at (2.4) constructed using the correct F_{V/σ_V} or pretending that it was equal to the standard normal distribution, where we used $\widehat{F}_{T\,|\,Q,W,Y}$ at (2.7). To demonstrate the importance of taking the noise into account, we also computed the naive prediction intervals obtained when ignoring the measurement errors U_j.

We generated 100 datasets for each combination of n and F_U and, in each case, we constructed the intervals $\hat{\mathcal{I}}_\alpha$ for three nominal levels, $\alpha = 0.99$, 0.95 and 0.90. For each generated sample, and for $i = 1, \ldots, n$, we constructed each prediction interval for T_i using the data $\mathcal{X}_{-i} = \{(Q_j^{\mathrm{T}}, W_j, Y_j) : 1 \le j \le n, j \ne i\}$. Let $\hat{\mathcal{I}}_\alpha^i$ denote a prediction interval for T_i obtained in this way, using either approach described above. For each approach we calculated the coverage rate of $\hat{\mathcal{I}}_\alpha$, $n^{-1}\sum_{i=1}^n I\big(T_i \in \hat{\mathcal{I}}_\alpha^i\big)$, and we averaged this number over the 100 generated samples to obtain an empirical measure of the coverage probability of the prediction intervals, which we denote below by ECP.

Table 1: Empirical coverage probabilities of the prediction interval (average length of the prediction intervals) computed from simulated data, using the correctly specified F_{V/σ_V} (True), erroneously pretending it is a normal distribution (Wrong), or using the naive prediction interval (Naive) which ignores the errors U_i.

U	n	F_{V/σ_V}	$1-\alpha$		
			0.99	0.95	0.90
Normal	30	True	0.987 (10.14)	0.943 (7.69)	0.889 (6.44)
		Wrong	0.985 (10.25)	0.944 (7.79)	0.890 (6.54)
		Naive	0.866 (8.67)	0.805 (6.60)	0.753 (5.54)
	50	True	0.991 (9.92)	0.949 (7.52)	0.903 (6.31)
		Wrong	0.990 (10.05)	0.953 (7.65)	0.905 (6.42)
		Naive	0.846 (7.98)	0.788 (6.07)	0.732 (5.10)
Laplace	30	True	0.988 (10.27)	0.943 (7.60)	0.888 (6.29)
		Wrong	0.987 (10.42)	0.945 (7.74)	0.893 (6.42)
		Naive	0.826 (8.03)	0.766 (6.11)	0.719 (5.12)
	50	True	0.990 (10.01)	0.947 (7.41)	0.900 (6.14)
		Wrong	0.990 (10.14)	0.949 (7.54)	0.901 (6.26)
		Naive	0.832 (7.74)	0.776 (5.89)	0.727 (4.94)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

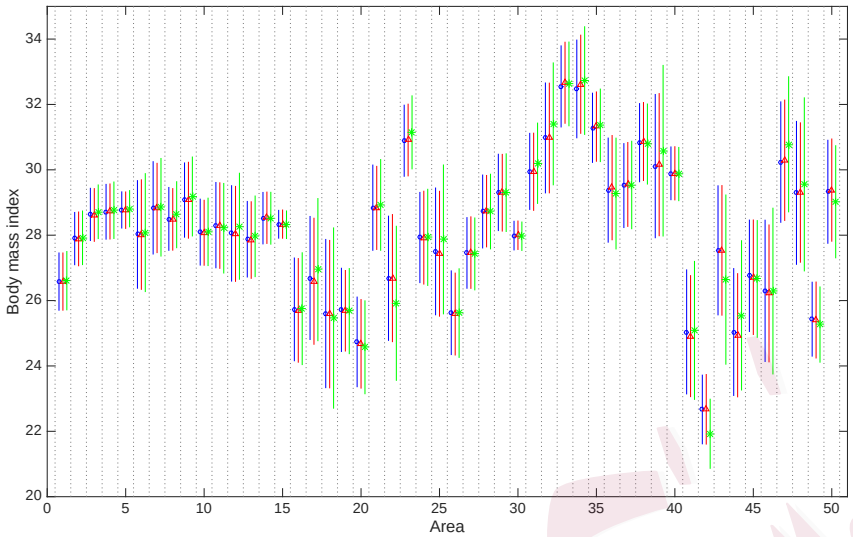

Figure 1: Each line represents a 95% prediction interval of the population mean BMI for one of 50 demographic subgroups from Example 1, assuming that f_{V/σ_V} is a standard normal density (—○—: naive method; —△—: our method) or estimating it nonparametrically as in Section 5 (—∗—).

We report the ECP in Table 1 for each configuration. The closer ECP is to $1-\alpha$, the more accurate the prediction interval is. These results suggest that our method is relatively robust against error misspecification, the main effect of misspecification being to increase the interval length. In Section 5 we also report simulation results for the case where F_V is estimated from the data. The latter approach is much more complex to implement, and while it improves slightly the level of the interval for small sample sizes, this comes at the cost of a significant increase in interval length. Those simulation results suggest that the rough parametric approximations of F_{V/σ_V} are preferable to a complex full nonparametric procedure. We also see that ignoring the errors completely leads to shorter intervals, but with very poor coverage.

4.3 Data examples

Example 1. We considered an application to two survey datasets: one from the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

Table 2: Example 1: average and median length of prediction intervals for 50 areas.

f_{V/σ_V}	Average length	Median length
normal-naive	2.69	2.56
normal	2.69	2.52
estimated as in Section 5	2.95	2.69

2003-2004 US NHANES, and the other from the 2004 US NHIS. The small areas in this study are 50 demographic subgroups classified by race and ethnicity (Ybarra and Lohr (2008)). Our goal was to construct prediction intervals for the population mean body mass index (BMI) for the demographic subgroups (small areas), using the NHIS as auxiliary information. The datasets were combined according to the small areas.

In the NHANES, the height and weight for each respondent were measured carefully to calculate the BMI = weight(kg)/ height(m)^2, but in the NHIS, measurements of height and weight are provided by the interviewers during the interview. Thus, the auxiliary variable (reported BMI) is prone to measurement error. For the jth demographic subgroup, let Y_j be the mean BMI from the NHANES. Let X_j be the true value of the mean BMI from the NHIS. We considered the model $Y_j = T_j + \epsilon_j$, $T_j = \beta_0 + \beta_1 X_j + V_j$. Denoting by W_j the mean of reported BMI from the NHIS, we assumed that the measurement error was additive through $W_j = X_j + U_j$.

We constructed the 95% prediction intervals for the population mean BMI for all demographic subgroups using $\widehat{F}_{T\,|\,Q,W,Y}$ at (2.7), with f_{V/σ_V} the standard normal density. For comparison, we also computed the prediction intervals based on the approach introduced in Section 5, where f_{V/σ_V} is estimated nonparametrically. Finally, we computed the naive prediction intervals which ignore the errors U_i. For $j = 1, \ldots, 50$, Figure 1 shows the resulting 95% prediction intervals of T_j conditional on (W_j, Y_j). Table 2 reports the average and median length of those 50 prediction intervals. The intervals obtained using both corrected approaches are relatively similar although,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

Table 3: Example 2: average and median length of prediction intervals for 51 areas.

f_{V/σ_V}	Coverage	Average length	Median length
normal-naive	78.43%	6.55e+03	6.65e+03
normal	84.31%	7.48e+03	7.41e+03
estimated as in Section 5	86.27%	1.04e+04	1.01e+04

as in our simulations, the interval lengths using the approach from Section 5 tend to be larger than those obtained under the normality assumption. Our simulation results suggest that the latter are preferable, since generally shorter while not much less accurate. In this example, the naive prediction intervals are close to the corrected ones, but since we do not know the truth, we do not know which method gives the most accurate prediction intervals.

Example 2. The US Department of Health and Human Services administers a program of energy assistance to low-income families. An important variable that determines the eligibility of a family for benefits from the program is an estimate of the median income of four-person families in the state. Through the years prior to 2000, the basic data, also known as the direct estimates of the state median incomes, came from the March Annual Demographic Supplement (ADS), which collects the income data of the three-, four-, and five-person households statewide. Due to smallness of sample sizes for all 51 states (50 states and the District of Columbia), the direct estimates are subject to considerable sampling variability, and the US Bureau of the Census (BOC) annually provides estimates of the state median income for four-person families by using small area estimation methods (Fay (1987); Datta et al. (1991)).

The 1989 CPS three-person households state median income estimates are strongly correlated with the corresponding 1989 CPS four-person ones, and we use them as the W_j's in our prediction model (they are subject to measurement error due to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

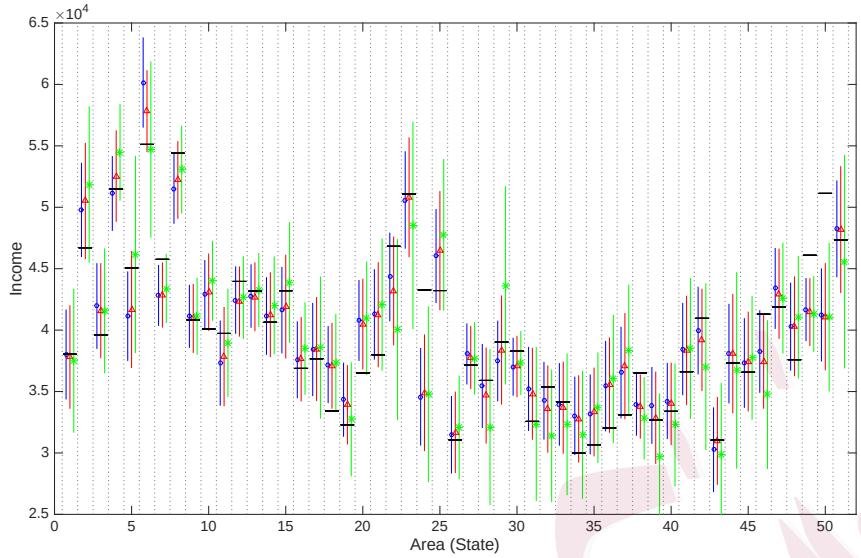

Figure 2: Each line represents a 95% prediction interval of the median income for one of the 51 states from Example 2, assuming that f_{V/σ_V} is a standard normal density (—○—: naive estimator; —△—: our method) or estimating it nonparametrically as in Section 5 (—∗—). The black flat line (–) shows year 1989 four-person family median incomes from the 1990 Census records, which can be regarded as the "true values".

their large sampling variability). For $j = 1, \ldots, 51$, we constructed 95% prediction intervals for the four-person family median incomes for the year 1989, conditional on observed values of (W_j, Y_j); see Figure 2. Numbers for this variable are available from the 1990 Census records, and we can compute the coverage of our intervals by treating these numbers as the "true values". Table 3 reports the coverage of the prediction intervals for 51 states using the three methods used in Example 1, and the average and median length of those prediction intervals. The prediction intervals constructed using the naive method which completely ignores the measurement errors provided the worst coverage rate (78.43%). With our methods that take measurement errors into account, the coverage rates were 84.31% and 86.27% when assuming that f_{V/σ_V} is a standard normal density or estimating it nonparametrically as in Section 5, respectively, but for the latter the intervals were again much longer.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

5 Estimating $F_{T\,|\,Q,W,Y}$ when F_V is unknown

The known distributions assumptions used in Section 2.2 can be relaxed, although since only (Q^{T}, W, Y) is observed, the distributions of V/σ_V and ϵ are confounded in the model at (2.1), so neither of the distributions can be estimated without knowing at least one of them. In this section, we show how to estimate $F_{T\,|\,Q,W,Y}$ from data $(Q_j^{\mathrm{T}}, W_j, Y_j)$ when the distribution of V is unknown and that of ϵ is known; similar ideas can be used if it is the distribution of ϵ that is unknown. We assume that f_ϵ is $\tau^{-1/2}\, f_1(\,\cdot\,/\tau^{1/2})$, where $\tau = \mathrm{var}(\epsilon)$ and f_1 is known.

Recall that, by the Fourier inversion theorem, we have

$$f_V(v) = (2\pi)^{-1} \int e^{-itv} \phi_V(t)\, dt\,. \tag{5.1}$$

To estimate f_V, we estimate ϕ_V, and plug a regularised version of it in (5.1). We start by expressing ϕ_V in terms of quantities that are either known, or which can be estimated directly from the $(Q_j^{\mathrm{T}}, W_j, Y_j)$'s. It follows from (2.1) that $Y_j = \beta_0 + \beta_1\, X_j + \beta_2^{\mathrm{T}}\, Q_j + V_j + \epsilon_j$. Therefore,

$$\phi_{Y_j}(t) = \exp(it\beta_0)\, \phi_X(\beta_1 t)\, \phi_{\beta_2^{\mathrm{T}} Q}(t)\, \phi_{\epsilon_j}(t)\, \phi_V(t)\,, \tag{5.2}$$

so that we can write $\phi_V(t) = \exp(-it\beta_0)\phi_{Y_j}(t) \big/ \big\{\phi_X(\beta_1 t)\, \phi_{\beta_2^{\mathrm{T}} Q}(t)\, \phi_{\epsilon_j}(t)\big\}$. Here, ϕ_{ϵ_j} is known and we can estimate $\phi_{\beta_2^{\mathrm{T}} Q}(t)$ and $\phi_X(\beta_1 t) = \phi_W(\beta_1 t)/\phi_U(\beta_1 t)$ by

$$\widehat{\phi_{\beta_2^{\mathrm{T}} Q}}(t) = n^{-1} \sum_{j=1}^{n} \exp\left(it\, \hat\beta_2^{\mathrm{T}} Q_j\right), \quad \widehat{\phi_X(\beta_1 t)} = \widehat{\phi_W(\beta_1 t)} \big/ \phi_U(\hat\beta_1 t)\,, \tag{5.3}$$

respectively, with $\hat\beta_1$ and $\hat\beta_2$ as in Section B and $\widehat{\phi_W(\beta_1 t)} = n^{-1} \sum_j \exp(it\, \hat\beta_1 W_j)$. However, the Y_j's are not identically distributed and we only have one observation, Y_j, to estimate $\phi_{Y_j}(t)$. To overcome this difficulty, (5.2) also implies that $\phi_{Y'}(t) = \exp(it\, \beta_0)\, \phi_X(\beta_1 t)\, \phi_{\beta_2^{\mathrm{T}} Q}(t)\, \phi_{\epsilon'}(t)\, \phi_V(t)\,$, where we used the notation $\phi_{Y'}(t) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

$n^{-1}\sum_{j=1}^{n}\phi_{Y_j}(t)$ and $\phi_{\epsilon'}(t) = n^{-1}\sum_{j=1}^{n}\phi_{\epsilon_j}(t)$. Therefore, we can write

$$\phi_V(t) = \exp(-it\beta_0)\,\phi_{Y'}(t)\big/\big\{\phi_X(\beta_1 t)\,\phi_{\beta_2^{\mathrm{T}}Q}(t)\,\phi_{\epsilon'}(t)\big\}\,, \tag{5.4}$$

where $\phi_{Y'}(t)$ can be estimated by $\hat{\phi}_{Y'}(t) = n^{-1}\sum_{j=1}^{n} e^{itY_j}$, and $\phi_{\beta_2^{\mathrm{T}}Q}(t)$ and $\phi_X(\beta_1 t)$ can be estimated as above.

As in standard nonparametric errors-in-variable problems, substituting this estimator and (5.3) into (5.4), and plugging the resulting estimator $\hat{\phi}_V$ of ϕ_V directly into (5.1) needs to be done in combination with some regularisation. Several approaches can be taken, such as one based on kernel regularisation as in (2.3). However, the denominator of $\hat{\phi}_V$ can vanish, and a more suitable approach consists in replacing the denominator, when it gets too small, by a ridge parameter $\rho > 0$. For a complex number a, we have $1/a = \bar{a}/|a|^2$, and as the ϵ_j's are symmetric around zero, so that $\phi_{\epsilon'}$ is real, we suggest using

$$\hat{f}_V(v) = \frac{1}{2\pi}\int e^{-it\,(v+\hat{\beta}_0)}\hat{\phi}_{Y'}(t)\phi_U(t\,\hat{\beta}_1)\hat{\phi}_{W;\,Q;\,\rho}(t)\;dt,$$

where

$$\hat{\phi}_{W;\,Q;\,\rho}(t) = \frac{\overline{\widehat{\phi_W(\beta_1 t)}}\;\overline{\widehat{\phi_{\beta_2^{\mathrm{T}}Q}}}(t)}{\max\big\{\rho, \big|\widehat{\phi_W(\beta_1 t)}\big|^2\big\}\max\big\{\rho, \big|\widehat{\phi_{\beta_2^{\mathrm{T}}Q}}(t)\big|^2\big\}\,\phi_{\epsilon'}(t)}$$

if $\inf_t \phi_{\epsilon'}(t) \geq 0$, and

$$\hat{\phi}_{W;\,Q;\,\rho}(t) = \frac{\overline{\widehat{\phi_W(\beta_1 t)}}\;\overline{\widehat{\phi_{\beta_2^{\mathrm{T}}Q}}}(t)\,\phi_{\epsilon'}(t)}{\max\big\{\rho, \big|\widehat{\phi_W(\beta_1 t)}\big|^2\big\}\max\big\{\rho, \big|\widehat{\phi_{\beta_2^{\mathrm{T}}Q}}(t)\big|^2\big\}\max\big\{\rho, \phi_{\epsilon'}^2(t)\big\}}$$

otherwise. Since $\mathrm{var}(\epsilon) = \tau$, we can then estimate $f_{V+\epsilon}(s)$ by $\hat{f}_{V+\epsilon}(s) = \int \tau^{-1/2}\,f_1\big(v/\tau^{1/2}\big)\,\hat{f}_V(s-v)\,dv$. Finally, using (2.2), we can estimate the distribution function of T, given that $(Q, W, Y) = (q, w, y)$, by

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q, w, y)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

Table 4: Empirical coverage probabilities of the prediction interval (average length of the prediction intervals) computed from simulated data when f_V is estimated non-parametrically.

U	n	$1-\alpha$		
		0.99	0.95	0.90
Normal	30	0.985 (13.59)	0.948 (10.49)	0.895 (8.80)
	50	0.986 (13.03)	0.950 (9.99)	0.904 (8.41)
Laplace	30	0.987 (14.06)	0.949 (10.9)	0.894 (9.17)
	50	0.987 (13.47)	0.946 (10.44)	0.892 (8.80)

$$= \frac{\int_{-\infty}^{t} f_1\{(s-y)/\tau^{1/2}\} \int \hat{f}_V(s-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q)\, \hat{f}_X(x)\, f_U(w-x)\, dx\, ds}{\int f_U(w-x)\, \hat{f}_X(x) \int f_1(v/\tau^{1/2})\, \hat{f}_V(y-\hat{\beta}_0-\hat{\beta}_1 x-\hat{\beta}_2^{\mathrm{T}} q-v)\, dv\, dx}\,, \tag{5.5}$$

where unqualified integrals are over the entire real line.

It can be proved that this estimator is consistent under sufficient regularity conditions, but estimating all these unknowns is challenging. Here f_V and f_ϵ only appear in an indirect way in the expression for $\widehat{F}_{T\,|\,Q,W,Y}$, and our simulation results in Section 4.2 indicated that our method was somewhat robust to error misspecification, with the main effect being to increase the length of our prediction interval. Therefore, it is not clear that a purely nonparametric approach is worth the additional complexity it incurs, since its main effect is to increase the variability of the predictors, which too will have the effect of increasing the interval length.

To investigate this, we constructed prediction intervals based on (5.5), using the same simulation settings as in Section 4.2. To compute them, we need to choose h used in $\hat{f}_X$ at (2.3) and the ridge parameter ρ. We used the approach described in Section 4.1, replacing there the minimisation of $J(h)$ at (4.1) by a minimisation of $J(H)$, where $H = (h, \rho)$ and $J(H)$ is the version of $J(h)$ obtained when replacing $\widehat{F}_{T\,|\,Q,W,Y}$ at (2.4) by $\widehat{F}_{T\,|\,Q,W,Y}$ at (5.5). We searched for h in the grid described in Section 4.1, and for ρ on an equispaced grid of ten values between $0.2/n$ and $5/n$,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

which was time consuming given the bivariate grid search.

The results are summarised in Table 4, where, as in Table 1, we report the EPC of the prediction intervals and the mean interval lengths, both computed over 100 simulated samples. Comparing with Table 1, the results confirm that making a rough guess for the unknown densities, as in Section 4.2, seems preferable to using a purely nonparametric approach. The former is also much faster to compute than the latter.

6 Technical arguments

6.1 Proof of Theorem 1

We need some notation. Put

$$\tau_{0k} = \int \psi_k(t, y, q, w, x)\, f_X(x)\, dx\,, \quad c_{0k} = \frac{1}{\ell!}\,\kappa_\ell\,\lambda_k^{[\ell]}(0 \,|\, s, y, q, w)\,, \tag{6.1}$$

where the subscript 0 on τ_{0k} and c_{0k} indicates that, in this instance, the true values of the parameters β_0, β_1, β_2, and σ_V^2 are used to construct ψ_k. In this notation,

$$F_{T\,|\,Q,W,Y}(t \,|\, q, w, y) = \tau_{01}(t, y, q, w)\big/\tau_{02}(t, y, q, w)\,. \tag{6.2}$$

6.1.1 Proof of part (i) of Theorem 1

Step 1: Approximation of $\hat{\psi}_k$ by ψ_k. Let $\mathcal{S}$, $\mathcal{Y}$, and $\mathcal{Q}$ be compact sets in the respective domains of s, y, and q. Using a Taylor expansion it can be proved that

$$\hat{\psi}_k(s, y, q, w, x) = \psi_k(s, y, q, w, x) + \Delta(s, y, q, w, x)\,, \tag{6.3}$$

where ψ_k and $\hat{\psi}_k$ are as at (2.6) and (2.8), respectively, ψ_k is at (2.6) and, for positive constants C_1, C_2 depending on $\mathcal{S}$, $\mathcal{Y}$, and $\mathcal{Q}$, and the true values of the parameters β_0, β_1, β_2 and σ_V^2 (which we denote as here), but not on n,

$$|\Delta(s, y, q, w, x)| \;\leq\; C_1\,(1 + |x| + \|q\|)\, f_U(w - x)\,\Big\{\big|\hat{\beta}_0 - \beta_0\big|$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

$$+\big|\hat\beta_1-\beta_1\big|\,|x|+\big\|\hat\beta_2-\beta_2\big\|\,\|q\|+\big|\hat\sigma_V^2-\sigma_V^2\big|\Big\} \quad (6.4)$$

whenever

$$\max\Big\{\big|\hat\beta_0-\beta_0\big|,\big|\hat\beta_1-\beta_1\big|,\big\|\hat\beta_2-\beta_2\big\|,\big|\hat\sigma_V^2-\sigma_V^2\big|\Big\}\le C_2\,. \quad (6.5)$$

The bound at (6.4) holds uniformly in all $s\in\mathcal{S}$, all $y\in\mathcal{Y}$, all $q\in\mathcal{Q}$ and all real x, provided (6.5) holds.

Step 2: Mean and variance of numerator and denominator on right-hand side of (2.7).

From the definition of ψ_k at (2.6), let $\psi=\psi_k$ for either $k=1$ or $k=2$, and let

$$J_1=\int\psi(s,y,q,w,W+hu)\,K_U(u;h)\,du\,,\quad \tau_j=E\big(J_1^j\big)\,, \quad (6.6)$$

for $j=1,2$, with

$$\begin{aligned} J_2 &\equiv \int\psi(s,y,q,w,x)\,\hat f_X(x)\,dx \quad (6.7)\\ &= \frac{1}{nh}\sum_{j=1}^n\int\psi(s,y,q,w,x)\,K_U\Big(\frac{x-W_j}{h};h\Big)\,dx\\ &= \frac{1}{n}\sum_{j=1}^n\int\psi(s,y,q,w,W_j+hu)\,K_U(u;h)\,du\,, \end{aligned}$$

$$n\,\mathrm{var}(J_2)=\mathrm{var}\Big\{\int\psi(s,y,q,w,W+hu)\,K_U(u;h)\,du\Big\}=\tau_2-\tau_1^2\,, \quad (6.8)$$

$$\int e^{itu}\,\psi(s,y,q,w,W+hu)\,du=h^{-1}\,\exp(-it\,W/h)\,\chi(t/h)\,,$$

where $\chi(t)=\chi(t\,|\,s,y,q,w)=\int e^{itu}\,\psi(s,y,q,w,u)\,du$. Therefore, by Parseval's identity,

$$\begin{aligned} &\int\psi(s,y,q,w,W+hu)\,K_U(u;h)\,du\\ =\;&\frac{1}{2\pi h}\int\exp(-it\,W/h)\,\chi(t/h)\,\frac{\phi_K(t)}{\phi_U(t/h)}\,dt=\frac{1}{2\pi}\int\exp(-it\,W)\,\chi(t)\,\frac{\phi_K(ht)}{\phi_U(t)}\,dt\\ =\;&\frac{1}{2\pi}\int\Re\big\{\exp(-it\,W)\,\chi(t)\big\}\,\frac{\phi_K(ht)}{\phi_U(t)}\,dt \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

$$= \frac{1}{2\pi} \int \left\{ \cos(t\,W)\,\chi_1(t) - \sin(t\,W)\,\chi_2(t) \right\} \frac{\phi_K(ht)}{\phi_U(t)}\,dt\,,$$

where $\chi_1 = \Re\,\chi$, $\chi_2 = \Im\,\chi$, and we have used the fact that $\phi_K(ht)/\phi_U(t)$ is a symmetric function of t. Hence, with $\rho_j = \chi_j/\phi_U$,

$$\begin{aligned}
&(2\pi)^2\,\tau_2 \\
&\quad = \int\int E\Big\{ \cos(t_1 W)\,\cos(t_2 W)\,\chi_1(t_1)\,\chi_1(t_2) + \sin(t_1 W)\,\sin(t_2 W)\,\chi_2(t_1)\,\chi_2(t_2) \\
&\qquad - \cos(t_1 W)\,\sin(t_2 W)\,\chi_1(t_1)\,\chi_2(t_2) - \sin(t_1 W)\,\cos(t_2 W)\,\chi_2(t_1)\,\chi_1(t_2) \Big\} \\
&\qquad \times \frac{\phi_K(ht_1)\,\phi_K(ht_2)}{\phi_U(t_1)\,\phi_U(t_2)}\,dt_1\,dt_2 \\
&\quad = \tfrac{1}{2} \int\int \Big\{ E\Big[\cos\{(t_1+t_2)\,W\} + \cos\{(t_1-t_2)\,W\} \Big]\,\chi_1(t_1)\,\chi_1(t_2) \\
&\qquad + E\Big[\cos\{(t_1-t_2)\,W\} - \cos\{(t_1+t_2)\,W\} \Big]\,\chi_2(t_1)\,\chi_2(t_2) \\
&\qquad - E\Big[\sin\{(t_1+t_2)\,W\} + \sin\{(t_2-t_1)\,W\} \Big]\,\chi_1(t_1)\,\chi_2(t_2) \\
&\qquad - E\Big[\sin\{(t_1+t_2)\,W\} + \sin\{(t_1-t_2)\,W\} \Big]\,\chi_2(t_1)\,\chi_1(t_2) \Big\} \\
&\qquad \times \frac{\phi_K(ht_1)\,\phi_K(ht_2)}{\phi_U(t_1)\,\phi_U(t_2)}\,dt_1\,dt_2 \\
&\quad = \tfrac{1}{2} \int\int \Big[\Re\left\{ \phi_W(t_1+t_2) + \phi_W(t_1-t_2) \right\} \rho_1(t_1)\,\rho_1(t_2) \\
&\qquad + \Re\left\{ \phi_W(t_1-t_2) - \phi_W(t_1+t_2) \right\} \rho_2(t_1)\,\rho_2(t_2) \\
&\qquad - \Im\left\{ \phi_W(t_1+t_2) + \phi_W(t_2-t_1) \right\} \rho_1(t_1)\,\rho_2(t_2) \\
&\qquad - \Im\left\{ \phi_W(t_1+t_2) + \phi_W(t_1-t_2) \right\} \rho_2(t_1)\,\rho_1(t_2) \Big]\,\phi_K(ht_1)\,\phi_K(ht_2), dt_1\,dt_2\,. \qquad (6.9)
\end{aligned}$$

For simplicity, take the support of ϕ_K to be $[-1,1]$. Then, taking ϕ_{W0} to be either $\Re\,\phi_W$ or $\Im\,\phi_W$, and, as in (3.1), we have, taking the $\pm$ signs respectively throughout:

$$\begin{aligned}
&\int\int \phi_{W0}(t_1 \pm t_2)\,\chi_{j_1}(t_1)\,\chi_{j_2}(t_2)\,\frac{\phi_K(ht_1)\,\phi_K(ht_2)}{\phi_U(t_1)\,\phi_U(t_2)}\,dt_1\,dt_2 \\
&\quad = \int_{-1/h}^{1/h} \int_{-1/h}^{1/h} \phi_{W0}(t_1 \pm t_2)\,\rho_{j_1}(t_1)\,\rho_{j_2}(t_2)\,\phi_K(ht_1)\,\phi_K(ht_2)\,dt_1\,dt_2
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

$$
\begin{aligned}
&= \int_{-1/h}^{1/h} \rho_{j_1}(t_1)\,\phi_K(ht_1)\,dt_1 \int_{t_1-1/h}^{t_1+1/h} \phi_{W0}(t)\,\rho_{j_2}\{\pm(t-t_1)\}\,\phi_K\{h\,(t-t_1)\}\,dt \\
&= \frac{1}{h}\int_{-1}^{1} \rho_{j_1}(t_1/h)\,\phi_K(t_1)\,dt_1 \int_{(t_1-1)/h}^{(t_1+1)/h} \phi_{W0}(t)\,\rho_{j_2}\{\pm(t-t_1/h)\} \\
&\quad \times\phi_K(ht-t_1)\,dt \\
&\equiv R_1(h)\,. \qquad (6.10)
\end{aligned}
$$

We prove in Appendix F that, uniformly in $s \in \mathcal{S}$, $y \in \mathcal{Y}$, and $q \in \mathcal{Q}$,

$$
R_1(h) = \begin{cases} O(1) & \text{if } w \neq 0 \\ C_1\,(s_k/b_1)^2\,h^{-1} + o(h^{-1}) & \text{if } w = 0 \text{ and } j_1 = j_2 = 1 \\ o(h^{-1}) & \text{if } w = 0 \text{ and } j_1 = j_2 = 1 \text{ fails}\,, \end{cases} \qquad (6.11)
$$

where s_k and b_1 are as in (3.1), and the positive constant C_1 depends only on K and $f_W(0)$.

Combining (6.9)–(6.11) we deduce that, if $w \neq 0$,

$$
\tau_2 = O(1)\,. \qquad (6.12)
$$

If $w = 0$ then, noting from (6.11) that $R_1(h) = o(h^{-1})$ if $j_1 = j_2 = 1$ fails, we have:

$$
\begin{aligned}
\tau_2 &= \tfrac{1}{2}\,(2\pi)^{-2} \iint \Re\left\{\phi_W(t_1+t_2) + \phi_W(t_1-t_2)\right\} \chi_1(t_1)\,\chi_1(t_2) \\
&\qquad \times \frac{\phi_K(ht_1)\,\phi_K(ht_2)}{\phi_U(t_1)\,\phi_U(t_2)}\,dt_1\,dt_2 + o\big(h^{-1}\big) \\
&= C_2\,(s_k/b_1)^2\,h^{-1} + o\big(h^{-1}\big)\,, \qquad (6.13)
\end{aligned}
$$

where on this occasion $\phi_{W0} = \Re\,\phi_W$ and we let $C_2 = C_1/(2\pi)^2$.

More simply, the definition of κ_ℓ below (3.5), and assuming that (3.4) holds, we can write:

$$
\begin{aligned}
\tau_1 &= \iint \psi(s,y,q,w,w_1+hu)\,f_W(w_1)\,K_U(u;h)\,du\,dw_1 \\
&= \iint \psi(s,y,q,w,x+hu)\,f_X(x)\,K(u)\,du\,dx = \int K(u)\,\lambda_k(hu\,|\,s,y,q,w)\,du
\end{aligned}
$$

$$= \int \psi(s,y,q,w,x)\, f_X(x)\, dx + \frac{1}{\ell!}\, \kappa_\ell\, h^\ell\, \lambda_k^{[\ell]}(0 \,|\, s,y,q,w) + o\big(h^\ell\big)\,. \tag{6.14}$$

Combining (6.8), (6.12), and (6.14) we deduce that if $w \neq 0$,

$$\mathrm{var}\bigg\{ \int \psi(s,y,q,w,x)\, \hat{f}_X(x)\, dx \bigg\} = O\big(n^{-1}\big)\,, \tag{6.15}$$

whereas if, using (6.13) and (6.14) we deduce that if $w = 0$,

$$n\,\mathrm{var}\bigg\{ \int \psi(s,y,q,w,x)\, \hat{f}_X(x)\, dx \bigg\} = C_2\, (s_k/b_1)^2\, h^{-1} + o\big(h^{-1}\big)\,. \tag{6.16}$$

Step 3: Completion. For simplicity we treat only the case $w = 0$. The quantities J_2 and τ_1 each have two forms, depending on whether we take $k = 1$ or $k = 2$ in the formula $\psi = \psi_k$ used in defining J_2 at (6.7) and τ_1 at (6.6). Write J_1, J_2, and τ_1 as J_{1k}, J_{2k} and τ_{1k}, respectively, to indicate these possibilities, and put

$$\Delta_k = J_{2k} - E(J_{2k}) = J_{2k} - E(J_{1k}) = J_{2k} - \tau_{1k}\,.$$

Let τ_{0k} denote the version of τ_{1k} when we take $h = 0$. From (6.14), $\tau_{1k} = \tau_{0k} + c_{0k}\, h^\ell + o(h^\ell)$, while J_2, and hence also J_{2k} and τ_{1k}, are functions of s, q, and w.

Using (6.3), Theorem 4, and the definition of $\widehat{F}_{T\,|\,Q,W,Y}$ at (2.7), we have:

$$\begin{aligned}
\widehat{F}_{T\,|\,Q,W,Y}(t \,|\, q,w,y) &= \frac{J_{21}}{J_{22}} + O_p\big(n^{-1/2}\big) = \frac{\tau_{11} + \Delta_1}{\tau_{12} + \Delta_2} + O_p\big(n^{-1/2}\big) \\
&= \frac{\tau_{11}}{\tau_{12}} + \frac{\Delta_1}{\tau_{12}} - \frac{\tau_{11}}{\tau_{12}^2}\, \Delta_2 + o_p\big\{(nh)^{-1/2}\big\} \\
&= \frac{\tau_{01}}{\tau_{02}} + \bigg(\frac{c_{01}}{\tau_{02}} - \frac{\tau_{01}}{\tau_{02}^2}\, c_{02}\bigg)\, h^\ell + \tau_{02}^{-1}\, \Delta_1 - \tau_{01}\, \tau_{02}^{-2}\, \Delta_2 \\
&\qquad + o_p\big\{(nh)^{-1/2}\big\} + o\big(h^\ell\big)\,,
\end{aligned} \tag{6.17}$$

where the quantities J_{2k}, Δ_k, and τ_{1k}, earlier defined as functions of s, q, and w, are here computed for $(s,y,q,w) = (t,y,q,w)$.

As well,

$$\mathrm{var}\big(\tau_{02}^{-1}\, \Delta_1 - \tau_{01}\, \tau_{02}^{-2}\, \Delta_2\big) = \frac{1}{\tau_{02}^2}\, \mathrm{var}(\Delta_1) + \frac{\tau_{01}^2}{\tau_{02}^4}\, \mathrm{var}(\Delta_2) - \frac{2\,\tau_{01}}{\tau_{02}^3}\, \mathrm{cov}(\Delta_1, \Delta_2)\,. \tag{6.18}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

By (6.13), $\mathrm{var}(\Delta_k) = (nh)^{-1} C_2 \, (s_k/b_1)^2 + o\{(nh)^{-1}\}$, and similarly it can be proved that $\mathrm{cov}(\Delta_1, \Delta_2) = (nh)^{-1} C_2 \left(s_1 \, s_2 / b_1^2\right) + o\{(nh)^{-1}\}$. Hence, by (6.18),

$$nh \, \mathrm{var}\left(\tau_{02}^{-1} \, \Delta_1 - \tau_{01} \, \tau_{02}^{-2} \, \Delta_2\right) = \frac{C_2}{b_1^2} \left(\frac{s_1}{\tau_{02}} - \frac{s_2 \, \tau_{01}}{\tau_{02}^2}\right)^2 + o(1) \, . \tag{6.19}$$

Thus, part (i) of Theorem 1 follows from (6.2), (6.17) and (6.19).

6.1.2 Proof of part (ii) of Theorem 1

We treat only the case $w = 0$. With J_1 and J_2 as at (6.6) and (6.7), let $\tau_j = E(J_1^j)$ for each integer $j \geq 1$. In step 3 of the proof of part (i) of the theorem we showed that $\tau_4 = O(h^{-3})$, and more generally it can be proved that $\tau_j = O(h^{1-j})$. Hence, since $nh \to \infty$, we have for each integer $\nu \geq 2$, $\tau_{2\nu} = o(n^{\nu-1} \, h^{-\nu})$. (These convergence results, and the order of magnitude bounds below, hold uniformly in s, y, and q in their respective compact sets, while J_1 and J_2 are functions of (s, y, q).) Therefore, by Rosenthal's inequality,

$$n^{2\nu} \, E(J_2 - EJ_2)^{2\nu} = O\{(n\tau_2)^\nu + n \, \tau_{2\nu}\} = O\{(n/h)^\nu\} \, ,$$

whence $E(J_2 - EJ_2)^{2\nu} = O\{(nh)^{-\nu}\}$. Hence, by Markov's inequality,

$$P\Big\{\big|J_2(s,y,q) - EJ_2(s,y,q)\big| > n^{\varepsilon_3} \, (nh)^{-1/2}\Big\} = O\big(n^{-B_5}\big)$$

for all $B_5, \varepsilon_3 > 0$.

This bound applies uniformly in $(s, y, q) \in \mathcal{S} \times \mathcal{Y} \times \mathcal{Q}$. Therefore, if $\mathcal{S}'$, $\mathcal{Y}'$ and $\mathcal{Q}'$ are subsets of $\mathcal{S}$, $\mathcal{Y}$ and $\mathcal{Q}$, respectively, each representing a regular lattice and containing no more than $O(n^{B_6})$ points for some $B_6 > 0$, then for all $B_5, \varepsilon_3 > 0$,

$$P\Big\{\sup_{s \in \mathcal{S}'} \sup_{y \in \mathcal{Y}'} \sup_{q \in \mathcal{Q}'} \big|J_2(s,y,q) - EJ_2(s,y,q)\big| > n^{\varepsilon_3} \, (nh)^{-1/2}\Big\} = O\big(n^{-B_5}\big) \, , \tag{6.20}$$

Given $(s, y, q) \in \mathcal{S} \times \mathcal{Y} \times \mathcal{Q}$, let $(s', y', q') \in \mathcal{S}' \times \mathcal{Y}' \times \mathcal{Q}'$ minimise the distance from (s', y', q') to (s, y, q). Making use of continuity properties of J_2, as a function of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

s, y, q, and w, it can be proved that if B_6 is sufficiently large then

$$P\Big\{ \sup_{s\in\mathcal{S}} \sup_{y\in\mathcal{Y}} \sup_{q\in\mathcal{Q}} \big|J_2(s,y,q) - J_2(s',y',q')\big| > n^{-1} \Big\} = O\big(n^{-B_5}\big)\,, \tag{6.21}$$

$$\sup_{s\in\mathcal{S}} \sup_{y\in\mathcal{Y}} \sup_{q\in\mathcal{Q}} \big|E\{J_2(s,y,q)\} - E\{J_2(s',y',q')\}\big| = O\big(n^{-1}\big)\,. \tag{6.22}$$

Combining (6.20)–(6.22) we deduce that for all $B_5, \varepsilon_3 > 0$,

$$P\Big\{ \sup_{s\in\mathcal{S}} \sup_{y\in\mathcal{Y}} \sup_{q\in\mathcal{Q}} \big|J_2(s,y,q) - EJ_2(s,y,q)\big| > n^{\varepsilon_3}\,(nh)^{-1/2} \Big\} = O\big(n^{-B_5}\big)\,. \tag{6.23}$$

There are two versions of J_2, depending on whether we take $\psi = \psi_1$ or ψ_2 at (6.7). Result (6.23) holds for both versions, and we distinguish them by writing J_{k2} for J_2 when $\psi = \psi_k$. In this notation, (6.14) implies that

$$\frac{E\{J_{21}(t,y,q)\}}{E\{J_{22}(t,y,q)\}} = F_{T\,|\,Q,W,Y}(t\,|\,q,w,y) + O\big(h^\ell\big)\,, \tag{6.24}$$

uniformly in $(t,y,q) \in \mathcal{S}\times\mathcal{Y}\times\mathcal{Q}$. More simply, (2.7) implies that

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \frac{J_{21}(t,y,q)}{J_{22}(t,y,q)}\,,$$

and so by (6.23), for each $\varepsilon_3 > 0$,

$$\widehat{F}_{T\,|\,Q,W,Y}(t\,|\,q,w,y) = \frac{E\{J_{21}(t,y,q)\}}{E\{J_{22}(t,y,q)\}} + O_p\big\{n^{\varepsilon_3}\,(nh)^{-1/2}\big\}\,, \tag{6.25}$$

uniformly in $(t,y,q) \in \mathcal{S}\times\mathcal{Y}\times\mathcal{Q}$. Part (ii) of Theorem 1 follows from (6.24) and (6.25).

7 Supplementary Materials

Additional results can be found in the online supplementary materials. In Appendix S1, we calculate the conditional distribution of T; in Appendix S2 we provide the details about estimating the unknown parameters in the model, as well as a theorem (Theorem 4) establishing asymptotic normality of the estimated parameters. In Appendix S3 we discuss the conditions of Section 3.1. In Appendix S4 we provide

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0416

a theorem (Theorem 5) which establishes theoretical properties of the conditional distribution of T and its derivatives. Appendices S5 to S8 contain the proofs of some of our results.

Acknowledgment

Delaigle and Hall's research was supported by grants and fellowships from the Australian Research Council.

References

Arima, S., Datta, G. and Liseo, B. (2012). Objective Bayesian analysis of a measurement error small area model. *Bayesian Analysis*, **7**, 363–384.

Arima, S., Datta, G. and Liseo, B. (2015). Bayesian estimators for small area models when auxiliary information is measured with error. *Scand. J. Statist.*, **42**, 518–529.

Buonaccorsi, J.P. (2010). *Measurement Error: Models, Methods, and Applications*. CRC Press.

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, **83**, 1184–1186.

Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Ann. Statist.*, **36**, 1221–1245.

Datta, G.S., Fay, R.E. and Ghosh, M. (1991). Hierarchical and empirical multivariate Bayes analysis in small area estimation. *Proceedings of the Seventh Annual Research Conference of the Bureau of the Census*. Arlington, VA, 63-79.

Datta, G.S., Ghosh, M., Smith, D.D. and Lahiri, P. (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical Bayes confidence intervals. *Scand. J. Statist.*, **29**, 139–152.

Datta, G.S., Rao, J.N.K. and Torabi, M. (2010), Pseudo-empirical Bayes estimation of small area means under a nested error linear regression model with functional measurement errors. *J. Statist. Plann. Inference*, **140**, 2952–2962.

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. Roy. Statist. Soc. Ser. B*, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2004). Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation. *Comput. Statist. Data Anal.*, **45**, 249–267.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.*, **103**, 280–287.

Delaigle, A. and Hall, P. (2011). Estimation of observation-error variance in errors-in-variables regression. *Statistica Sinica*, **21**, 1023–1063.

Diao, L., Smith, D.D., Datta, G.S., Maiti, T. and Opsomer, J.D. (2013). Accurate confidence interval estimation of small area parameters under the Fay-Herriot model. *Scand. J. Statist.*, **41**, 497–515.

Fay, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*, Eds.: R. Platek, J.N.K. Rao, C.E. Sarndal and M.P. Singh. New York: Wiley, 91-102.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *J. Amer. Statist. Assoc.*, **74**, 269–277.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Ghosh, M., Sinha, K. and Kim, D. (2006). Empirical and hierarchical Bayesian estimation in finite population sampling under structural measurement error models. *Scand. J. Statist.*, **33**, 591–608.

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., Santamaría, L. (2010). Small area estimation under Fay-Herriot models with nonparametric estimation of heteroscedasticity. *Statistical Modelling*, **10**, 197–214.

Hall, P. and Lahiri, S. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.*, **36**, 2110–2134.

Hall, P. and Maiti, T. (2006). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist.*, **34**, 1733–1750.

Otto, M.C. and Bell, W.R. (1995). Sampling error modeling of poverty and income statistics for states. *A*merican Statistical Association Proceedings of the Section on Government Statistics. American Statistical Association, Alexandria.

Pfeffermann, D. (2013). New important developments in small area estimation. *Stat. Sci.*, **28**, 40–68.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation, 2nd Ed.*. Hoboken, New Jersey: John Wiley & Sons.

Stefanski, L. and Carroll, R.J. (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184.

Torabi, M., Datta, G.S. and Rao, J.N.K. (2009). Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scand. J. Statist.*, **36**, 355–368.

Ybarra, L.M.R. and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, **95**, 919–931.