Oryshchenko, Vitaliy; Smith, Richard J.

Working Paper

Improved density and distribution function estimation

cemmap working paper, No. CWP47/18

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

Suggested Citation: Oryshchenko, Vitaliy; Smith, Richard J. (2018) : Improved density and distribution function estimation, cemmap working paper, No. CWP47/18, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2018.4718

This Version is available at:
https://hdl.handle.net/10419/189775

cemmap
centre for microdata methods and practice

Improved density and distribution function estimation

Vitaliy Oryshchenko
Richard J. Smith

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP47/18

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

Improved Density and Distribution Function Estimation*

Vitaliy Oryshchenko†
University of Manchester
vitaliy.oryshchenko@manchester.ac.uk

Richard J Smith‡
c*e*mmap, U.C.L and I.F.S.
University of Cambridge
University of Melbourne
ONS Economic Statistics Centre of Excellence
rjs27@econ.cam.ac.uk

This Draft: July 2018

Abstract

Given additional distributional information in the form of moment restrictions, kernel density and distribution function estimators with implied generalised empirical likelihood probabilities as weights achieve a reduction in variance due to the systematic use of this extra information. The particular interest here is the estimation of the density or distribution functions of (generalised) residuals in semi-parametric models defined by a finite number of moment restrictions. Such estimates are of great practical interest, being potentially of use for diagnostic purposes, including tests of parametric assumptions on an error distribution, goodness-of-fit tests or tests of overidentifying moment restrictions. The paper gives conditions for the consistency and describes the asymptotic mean squared error properties of the kernel density and distribution estimators proposed in the paper. A simulation study evaluates the small sample performance of these estimators.

Keywords: Moment conditions, residuals, mean squared error, bandwidth.

JEL Classifications: C14.

*Address for correspondence: V. Oryshchenko, 2.068 Arthur Lewis Building, Department of Economics, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom.

†2.068 Arthur Lewis Building, Department of Economics, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom.

‡Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge CB3 9DD, UK.

1 Introduction

In many statistical and economic applications, additional distributional information about the data observation d_z-vector z may be available in the form of moment restrictions on its distribution. These constraints may arise from a particular economic or physical law, e.g., Chen (1997, Section 5), be implied by estimating equations, Qin and Lawless (1994, Example 1), or correspond to known population moments of another observable random vector correlated with z, e.g., in survey samples with auxiliary population information available from census data, e.g., Chen and Qin (1993) and Qin and Lawless (1994, Example 2). The primary purpose of the paper is to explore the advantages of this additional information for the estimation of the density and distribution functions of a scalar residual-like function of z which may depend on unknown parameters.

To this end, let $g(z, \beta)$ denote a d_g-vector of known functions of the data observation $z \in \mathcal{Z}$ and the d_β-vector $\beta \in \mathcal{B}$ of parameters where the sample space $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ and parameter space $\mathcal{B} \subset \mathbb{R}^{d_\beta}$ with $d_\beta \leq d_g$. The moment indicator vector $g(z, \beta)$ constitutes the basis for improved inference in the following discussion and analysis. In particular, it will be assumed that the true value β_0 taken by β uniquely satisfies the population unconditional moment equality condition

$$\mathrm{E}[g(z, \beta_0)] = 0, \tag{1.1}$$

where $\mathrm{E}[\cdot]$ denotes expectation taken with respect to the true population probability law of z. The true parameter value β_0 is generally unknown, but can also be fully or partially known in particular applications.

Models specified in the form of unconditional moment restrictions (1.1) convey partial information about the distribution F^z of z and are ubiquitous in areas such as economics; see, e.g., the monographs Hall (2005) and Mátyás (1999). Many other commonly used models lead to estimators that can be reformulated as solutions to a set of moment restrictions. Clearly, models given by conditional moment restrictions imply (1.1). Traditionally, such models are estimated by the generalised method of moments (GMM); see Hansen (1982). However, the performance of GMM estimators and associated test statistics is often poor in finite samples, which has lead to the development of a number of (information-theoretic) alternatives to GMM.

This paper focuses on the class of generalised (G) empirical likelihood (EL) estimators, which has attractive large sample properties; see, e.g., Newey and Smith (2004), Smith (1997, 2011), and Parente and Smith (2014) for a recent review. Special cases of GEL include EL, (Owen, 1988, 1990), Qin and Lawless (1994), exponentially tilting (ET), Corcoran (1998), Kitamura and Stutzer (1997), Imbens et al. (1998), and continuous-updating (GMM) estimators (CUE), Hansen et al. (1996); see also Euclidean EL, Antoine et al. (2007). Of these estimators, EL has the attractive property of being Bartlett-correctable; see Chen and Cui (2007).

When the parameter vector β_0 is overidentified by the moment restriction (1.1), i.e., $d_\beta < d_g$, these constraints generally carry useful additional information about F^z. Given a random sample z_i, $i = 1, \ldots, n$, of observations on z, such information is captured by the associated (G)EL implied probabilities π_i, $i = 1, \ldots, n$, which enable a nonparametric description of F^z satisfying the moment condition (1.1) given by the estimator $F^z_\pi(z) = \sum_{i=1}^n \pi_i \mathbb{I}\{z_i \leq z\}$, where $\mathbb{I}\{\cdot\}$ denotes

the indicator function, Back and Brown (1993), Qin and Lawless (1994). In the absence of the moment information (1.1) or when β_0 is just identified, $d_\beta = d_g$, $F^z_\pi(z)$ reduces to the empirical distribution function (EDF) $F^z_n(z) = n^{-1}\sum_{i=1}^n \mathbb{I}\{z_i \leq z\}$. In general, if $d_\beta < d_g$, $F^z_\pi(z)$ is a more efficient estimator of F^z than the EDF $F^z_n(z)$ reflecting the value of the overidentifying information in (1.1). This observation suggests therefore that estimation of the functionals of F^z, $T(F^z)$, by $T(F^z_\pi)$ rather than $T(F^z_n)$ will be similarly more efficient. Indeed this is the case when estimating expectations of certain known functions of z, see Brown and Newey (1998). A similar advantage is apparent for EL estimation of quantile functions with known β_0, e.g., Chen and Qin (1993) and Zhang (1995), general EL-based quantile estimation, Yuan et al. (2014), and EL-based kernel estimation of a univariate density function, e.g., Chen (1997) and Zhang (1998).

The concern of this paper is with efficient kernel estimation of the probability density (p.d.f.) and distribution (c.d.f.) functions of a scalar-valued function $u(z, \beta_0)$ of the data observation z with either known or unknown parameter vector β_0. The former case, when β_0 is known, is the classical situation briefly mentioned above. The central case of interest, when β_0 is unknown, is estimation of the p.d.f and c.d.f. of an error term based on the estimated residuals. Such estimates are routinely computed by practitioners and are used for both visual diagnostics, e.g., potentially revealing omitted structure such as multimodality or other features of interest, and formal diagnostic tests, e.g., goodness-of-fit and tests of parametric assumptions on the error distribution. The importance of obtaining residual density estimates with good (higher order) properties can hardly be understated. Yet, as discussed below, simply applying standard kernel estimators with default bandwidths to estimated residuals may result in an inconsistent p.d.f. or c.d.f. estimators as further conditions on the kernel function and bandwidth are generally required. Similar conclusions have been reached elsewhere in related literature on residual density estimation in nonparametric regression and other settings; see, e.g., Ahmad (1992), Cheng (2004), Kiwitt et al. (2008), Györfi and Walk (2012) and the discussion and references in Bott et al. (2013).

When β_0 is known, kernel density and distribution function estimators exploiting the (G)EL implied probabilities instead of the uniform EDF n^{-1} weights achieve a reduction of higher order variance due to the systematic use of the extra moment information in (1.1). The efficiency gains are first order asymptotically in the c.d.f. case and second order for p.d.f. estimation. In contradistinction, for residual p.d.f. and c.d.f. estimation, such gains will not always be realised. One can, however, expect efficiency gains from the knowledge that the mean of residuals is zero.

The outline of the paper is as follows. Section 2 briefly describes (G)EL estimation and the associated (G)EL implied probabilities. The main results concerning p.d.f. and c.d.f. estimators are given in Sections 3 and 4 for both known and unknown β_0 cases. The finite sample performance of the proposed estimators is evaluated via a simulation study reported in Section 5. Section 6 concludes. Supplements P: and E: in the Supplementary Information respectively detail some additional assumptions for and the proofs of the results in the main text and analyse a number of examples to illustrate the properties of the estimators developed in the paper.

2 Generalised Empirical Likelihood

The GEL class of estimators for β_0 is defined in terms of a real valued scalar carrier function $\rho : \mathcal{V} \mapsto \mathbb{R}$ that is concave on an open interval $\mathcal{V}$ containing zero with derivatives $\rho_j(v) = \mathrm{d}^j\rho(v)/\mathrm{d}v^j$ and $\rho_j = \rho_j(0)$, $j = 1, 2, \ldots$, normalized without loss of generality such that $\rho_1 = \rho_2 = -1$. The special cases $\rho(v) = \log(1-v)$ for $\mathcal{V} = (-\infty, 1)$, $\rho(v) = -\exp(v)$ and $\rho(v) = -v^2/2 - v$ correspond to EL, ET and CUE respectively and are all members of the Cressie and Read (1984) family where $\rho(v) = -(1+\gamma v)^{(\gamma+1)/\gamma}/(\gamma+1)$.

Given a random sample z_i, $i = 1, \ldots, n$, of size n of observations on the d_z-dimensional vector z, let $g_i(\beta) = g(z_i, \beta)$, $g_i = g_i(\beta_0)$, and $G_i(\beta) = \partial g(z_i, \beta)/\partial\beta^\top$, $G_i = G_i(\beta_0)$, $i = 1, \ldots, n$. Also let $\Lambda_n(\beta) = \{\lambda : \lambda^\top g_i(\beta) \in \mathcal{V}, i = 1, \ldots, n\}$. The GEL criterion $P_n^\rho(\beta, \lambda)$ is defined by $P_n^\rho(\beta, \lambda) = n^{-1}\sum_{i=1}^n \rho(\lambda^\top g_i(\beta)) - \rho(0)$, with λ a d_g-vector of auxiliary parameters, each element of which corresponding to an element of the moment function vector $g(z, \beta)$; for members of the Cressie and Read (1984) family of power divergence criteria λ is the Lagrange multiplier vector associated with imposition of the moment restriction (1.1). The GEL estimator $\hat\beta$ is the solution to the saddle point problem

$$\hat\beta = \operatorname*{argmin}_{\beta\in\mathcal{B}} \sup_{\lambda\in\Lambda_n(\beta)} P_n^\rho(\beta, \lambda). \tag{2.1}$$

If Supplement P: Assumptions P.1 and P.2 are satisfied, in particular, the population Jacobian $G = \mathrm{E}[\partial g(z, \beta_0)/\partial\beta^\top]$ and variance $\Omega = \mathrm{E}[g(z, \beta_0)g(z, \beta_0)^\top]$ matrices are full column rank and positive definite respectively, then all GEL estimators share the same first order large sample properties, see, e.g., Newey and Smith (2004, Theorems 3.1 and 3.2), i.e., $n^{1/2}(\hat\beta - \beta_0) \xrightarrow{d} N(0, \Sigma)$, achieving the semiparametric efficiency lower bound $\Sigma = (G^\top\Omega^{-1}G)^{-1}$, Chamberlain (1987, Theorem 2). Furthermore, if the additional Supplement P: Assumption P.3 is imposed, defining $H = \Sigma G^\top\Omega^{-1}$ and $P = \Omega^{-1} - \Omega^{-1}G\Sigma G^\top\Omega^{-1}$, the second order bias of $\hat\beta$ is

$$\mathrm{E}[\hat\beta] - \beta_0 = n^{-1}H\zeta_\lambda + O(n^{-2}),$$

where

$$\zeta_\lambda = -a + \mathrm{E}[G_iHg_i] + c_\rho\,\mathrm{E}[g_ig_i^\top Pg_i], \tag{2.2}$$

with $c_\rho = 1 + \rho_3/2$ and a a d_g-vector with elements $a^j = \mathrm{tr}(\Sigma\,\mathrm{E}[\partial^2 g^j(z, \beta_0)/\partial\beta\partial\beta^\top])$, $j = 1, \ldots, d_g$; see Newey and Smith (2004, Theorem 4.2).

Remark 2.1. The validity of the higher order bias and variance calculations, and hence the validity of the results reported below can be formally justified by that of an Edgeworth expansion of order $o(n^{-1})$ for the distribution of GEL parameter estimators. If z is continuously distributed, appropriate conditions may be found in Bhattacharya and Ghosh (1978) for general smooth functions of sample moments and Kundhi and Rilstone (2012) for Edgeworth expansions for (G)EL estimators. If some of the elements of z are discretely distributed, Jensen (1989) provides appropriate conditions.

For given β, the auxiliary parameter estimator is defined by $\lambda(\beta) = \operatorname{argmax}_{\lambda\in\Lambda_n(\beta)} P_n^\rho(\beta, \lambda)$. Whenever the constraint in $\lambda \in \Lambda_n(\beta)$ is not binding, $\lambda(\beta)$ solves the first-order conditions

$n^{-1}\sum_{i=1}^n \rho_1(\lambda(\beta)^\top g_i(\beta))g_i(\beta) = 0$. The GEL implied probabilities are then

$$\pi_i(\beta) = \frac{\rho_1(\lambda(\beta)^\top g_i(\beta))}{\sum_{j=1}^n \rho_{(1)}(\lambda(\beta)^\top g_j(\beta))}, \quad i = 1, \ldots, n.$$

The sample moment constraint $\sum_{i=1}^n \pi_i(\beta)g_i(\beta) = 0$ holds whenever the first order conditions for $\lambda(\beta)$ hold. In what follows, $\hat{\pi}_i = \pi_i(\hat{\beta})$, $i = 1, \ldots, n$, corresponds to the solution $\hat{\lambda} = \lambda(\hat{\beta})$, and, if β_0 is known, $\tilde{\pi}_i = \pi_i(\beta_0)$, $i = 1, \ldots, n$, with auxiliary parameter estimator $\tilde{\lambda} = \lambda(\beta_0)$. The generic notation π_i, $i = 1, \ldots, n$, is used whenever the distinction is unnecessary.

REMARK 2.2. Properties of the GEL implied probabilities relevant to the subsequent developments are summarized in Supplement P: Lemmas P.1 and P.2. Although $\pi_i(\beta)$, $i = 1, \ldots, n$, sum to unity and are positive if $\lambda(\beta)^\top g_i(\beta)$ is small uniformly in i, they are not guaranteed to be non-negative. The shrinkage estimator $\pi_i^\varepsilon = (\pi_i + \varepsilon_n)/\sum_{j=1}^n(\pi_j + \varepsilon_n)$, $i = 1, \ldots, n$, where $\varepsilon_n = -\min[\min_{1\leq i\leq n} \pi_i, 0]$, see Antoine et al. (2007), Smith (2011), ensures non-negativity $\pi_i^\varepsilon \geq 0$, $i = 1, \ldots, n$, and $\sum_{i=1}^n \pi_i^\varepsilon = 1$. Alternative solutions relevant to p.d.f. and c.d.f. estimation respectively are discussed in Sections 3 and 4.

REMARK 2.3. The implied probabilities were given for EL by Owen (1988), for ET by Kitamura and Stutzer (1997), for quadratic $\rho(\cdot)$ by Back and Brown (1993), and for the general case in the 1992 working paper version of Brown and Newey (2002); see also Smith (1997). For any function $a(z, \beta)$ and GEL estimator $\hat{\beta}$ the implied probabilities can be used to form a semi-parametrically efficient estimator $\sum_{i=1}^n \hat{\pi}_i a(z_i, \hat{\beta})$ of $\mathrm{E}[a(z, \beta_0)]$ as in Brown and Newey (1998).

3 GEL-Based Density Estimation

Suppose the p.d.f. $f(\cdot)$ of the scalar random variable $u = u(z, \beta_0)$ is of interest, where the scalar function $u : \mathcal{Z} \times \mathcal{B} \mapsto \mathcal{U} \subseteq \mathbb{R}$ is known up to the parameter vector β_0.

Let $\mathcal{N}$ denote an open neighbourhood of β_0.

Assumption 3.1. *For all $\beta \in \mathcal{N}$ there exists a function $v : \mathcal{Z} \times \mathcal{B} \mapsto \mathcal{V} \subseteq \mathbb{R}^{d_z - 1}$ such that the vector of functions $(u(z, \beta),\ v(z, \beta)^\top)^\top$ is a bijection between $\mathcal{Z}$ and $\mathcal{U} \times \mathcal{V}$.*

REMARK 3.1. Assumption 3.1 may be restated equivalently as requiring that for every $\beta \in \mathcal{N}$ there exists a bijection between z and some d_z-vector $w = w(z, \beta)$ such that, given $w^j(z, \beta)_{j=2}^{d_z}$, $u(z, \beta)$ and $w^1(z, \beta)$ are bijective. That is to say, z may be solved uniquely given values for u, v and β.

REMARK 3.2. A function $u(z, \beta)$ satisfying Assumption 3.1 may be thought of as defining a generalised residual in the sense of Cox and Snell (1968) and Loynes (1969), with $\hat{u}_i = u(z_i, \hat{\beta})$, $i = 1, \ldots, n$, the estimated residuals. Of course, other possibilities of interest are also included, e.g., estimating the density of an element of z subject to the extra information available in the moment condition (1.1).

3.1 Known β_0

Suppose that $u_i = u(z_i, \beta_0)$, $i = 1, \ldots, n$, are observed. Then the classical kernel density estimator for the p.d.f. f of $u = u(z, \beta_0)$ can be employed; *viz.*

$$\tilde{f}(u) = n^{-1} \sum_{i=1}^{n} k_b(u - u_i), \tag{3.1}$$

where $k_b(x) = k(x/b)/b$, $k(\cdot)$ is a kernel function and $b = b_n > 0$ is a bandwidth sequence; see Rosenblatt (1956) and Parzen (1962). The estimator $\tilde{f}$ (3.1) will serve as a benchmark for later comparisons.

The properties of $\tilde{f}$ are well known and can be formally established under different combinations of smoothness and integrability conditions on the kernel function k and p.d.f. f; see, e.g., Rao (1983, Section 2.1). A standard set of such conditions is given in Assumption 3.2 below. If k is square integrable, but not absolutely integrable, as is the case for the sinc kernel, conditions such as those in Tsybakov (2009, Theorem 1.5) can be imposed.

Let $R(k) = \int_{-\infty}^{\infty} k(x)^2 \mathrm{d}x$ for any square integrable function k; the limits of integration are omitted whenever there is little scope for confusion. Also let $f^{(j)}(u) = \mathrm{d}^j f(u)/\mathrm{d}u^j$ for any jth order differentiable function f.

Assumption 3.2. ***(a) (i)*** $\sup_{|x|<\infty}|k(x)| < \infty$, $\int |k(x)|\mathrm{d}x < \infty$, $\int k(x)\mathrm{d}x = 1$, *and* $\lim_{|x|\to\infty}|xk(x)| = 0$; ***(ii)*** *$k$ is a $(2r)$th order kernel, i.e., an even function such that, for some $r \geq 1$, $\mu_0(k) = 1$, $\mu_j(k) = 0$, $j = 1, \ldots, 2r-1$, and $\mu_{2r}(k) < \infty$, where $\mu_j(k) = \int x^j k(x)\mathrm{d}x$;* ***(iii)*** *$R(k) < \infty$;* ***(b)*** *$f(\cdot)$ is s times continuously differentiable and $R(f^{(j)}) < \infty$, $j = 0, 1, \ldots, s$;* ***(c)*** *as $n \to \infty$, $b \to 0$ and $nb \to \infty$.*

Remark 3.3. If Assumption 3.2(a)(i) holds, then by Supplement P: Lemma P.3, $\mathrm{E}[\tilde{f}(u)] \to f(u)$ as $b \to 0$ at all points u of continuity of f and if, in addition, Assumption 3.2(c) holds, then the mean squared error (MSE), $\mathrm{MSE}[\tilde{f}(u)] = \mathrm{E}[(\tilde{f}(u) - f(u))^2] \to 0$ as $n \to \infty$; see, e.g., Parzen (1962).

Remark 3.4. Higher order approximations to $\mathrm{MSE}[\tilde{f}(u)]$ can be obtained if f is sufficiently smooth. See, e.g. Rao (1983, Theorem 2.1.5), Wand and Jones (1995, Section 2.8) or Pagan and Ullah (1999, Section 2.4.3). The idea of using higher order kernels as a bias reduction technique originates at least as far back as Bartlett (1963).

Let $1 \leq r < \infty$. Suppose that Assumptions 3.2(a)(ii), 3.2(b) with $s = 2r + 2$, 3.2(c) together with $\mu_{2r+2}(k) < \infty$ and $\int x^2 k(x)^2 \mathrm{d}x < \infty$ hold. Then

$$\begin{aligned} \mathrm{E}[\tilde{f}(u)] &= f(u) + (2r)!^{-1}\mu_{2r}(k) f^{(2r)}(u) b^{2r} + O(b^{2r+2}), \\ \mathrm{Var}[\tilde{f}(u)] &= (nb)^{-1} R(k) f(u) - n^{-1} f(u)^2 + O(n^{-1}b). \end{aligned}$$

Hence, the mean squared error

$$\mathrm{MSE}[\tilde{f}(u)] = (nb)^{-1} R(k) f(u) + (2r)!^{-2}\mu_{2r}(k)^2 f^{(2r)}(u)^2 b^{4r} - n^{-1} f(u)^2 + O(b^{4r+2} \vee n^{-1}b). \tag{3.2}$$

REMARK 3.5. If k is a $(2r)$th order kernel and Assumption 3.2(b) holds with $s = 2r$, the remainder term in $\mathrm{E}[\hat{f}(u)]$ is $o(b^{2r})$. The $\sim n^{-1}$ term is kept explicit with O remainder for reasons that will become apparent below.

The mean integrated squared error (MISE), $\mathrm{MISE}[\tilde{f}] = \mathrm{E}[\int(\tilde{f}(u) - f(u))^2 \mathrm{d}u]$, is a commonly used global measure of performance. The optimal bandwidth is then defined as that value of $b > 0$ minimising MISE, or an approximation thereof. In particular, the asymptotically optimal bandwidth is defined as the value b^* minimising the two leading terms in the expansion

$$\mathrm{MISE}[\tilde{f}(\cdot; b)] = (nb)^{-1} R(k) + (2r)!^{-2} \mu_{2r}(k)^2 R(f^{(2r)}) b^{4r} - n^{-1} R(f) + O(b^{4r+2} \vee n^{-1} b), \quad (3.3)$$

i.e., $b^* = cn^{-1/(4r+1)}$ where $c = [(2r)!^2 R(k)/(4r\mu_{2r}(k)^2 R(f^{(2r)})]^{1/(4r+1)}$. The asymptotically optimal MISE is thereby

$$\mathrm{MISE}[\tilde{f}(\cdot; b^*)] = n^{-4r/(4r+1)} c^{-1} R(k)[1 + (4r)^{-1}] - n^{-1} R(f) + O(n^{-1-1/(4r+1)}).$$

REMARK 3.6. If k is of order greater than two, it necessarily takes negative values. Hence $\tilde{f}$ (3.1) itself need not be a density function. Note, however, that the positive part estimator, $\tilde{f}^+(u) = \max[\tilde{f}(u), 0]$ has MSE at most equal to $\mathrm{MSE}[\tilde{f}(u)]$. Further modifications that ensure integration to unity can be applied as described in Glad et al. (2003).

The GEL-based kernel density estimator incorporates the information embedded in the moment restriction (1.1) replacing the sample EDF weights n^{-1} in the construction of $\tilde{f}(u)$ (3.1) by the implied probabilities $\tilde{\pi}_i$, $i = 1, \ldots, n$; *viz.*

$$\tilde{f}_\rho(u) = \sum_{i=1}^{n} \tilde{\pi}_i k_b(u - u_i) \quad (3.4)$$

REMARK 3.7. The GEL-based kernel density estimator $\tilde{f}_\rho(u)$ (3.4) is the estimator of $f(u)$ obtained from the revised GEL criterion $\sum_{i=1}^n [\rho(\eta(f(u) - k_b(u - u_i)) + \lambda^\top g_i(\beta)) - \rho(0)]/n$ with the implicit moment condition $\mathrm{E}[k_b(u - u_i)] = f(u)$ and associated auxiliary parameter η; see Smith (2011, Section 3).

REMARK 3.8. If the validity of the moment restriction (1.1) is in doubt, a pre-test can be conducted using the GEL-based criterion paralleling the classical likelihood ratio test; see, e.g., Kitamura and Stutzer (1997), Imbens et al. (1998) and Smith (1997, 2011). For example, under the null hypothesis that (1.1) holds for some unique $\beta_0 \in \mathcal{B}$, the normalised GEL criterion (2.1) evaluated at the estimated parameters, $2nP_n^\rho(\hat{\beta}, \hat{\lambda})$, is asymptotically chi-square distributed with $d_g - d_\beta$ degrees of freedom. The parametric null hypothesis of known $\beta_0 = \beta^0$ can be tested at the α level using the critical region $\{2nP_n^\rho(\beta^0, \tilde{\lambda}) \geq \chi^2_{d_\beta}(\alpha)\}$.

To describe the properties of the GEL-based kernel p.d.f. estimator $\tilde{f}_\rho(u)$ (3.4), the shorthand notation, e.g., $\mathrm{E}[g_i|u] = \mathrm{E}[g(z, \beta_0)|z : u(z, \beta_0) = u]$, for conditional expectation given u is adopted.

Theorem 3.1. *If Supplement P: Assumptions P.1–P.3 and 3.2(a)(i) and (c) are satisfied, then* $\tilde{f}_\rho(u) = \tilde{f}(u) + o_p(1)$ *for all* u *such that* $f(u) < \infty$. *If, in addition, Assumption 3.1 is satisfied,*

then

$$\mathrm{E}[\tilde{f}_\rho(u)] = \mathrm{E}[\tilde{f}(u)] + n^{-1}c_\rho(-\mathrm{E}[g_i^\top \Omega^{-1} g_i|u] + \mathrm{E}[g_i^\top \Omega^{-1} g_i g_i^\top]\Omega^{-1}\,\mathrm{E}[g_i|u] + d_g)f(u) + o(n^{-1}), \tag{3.5}$$

$$\mathrm{Var}[\tilde{f}_\rho(u)] = \mathrm{Var}[\tilde{f}(u)] - n^{-1}\,\mathrm{E}[g_i|u]^\top \Omega^{-1}\,\mathrm{E}[g_i|u] f(u)^2 + o(n^{-1}). \tag{3.6}$$

Thus, the estimators $\tilde{f}$ and $\tilde{f}_\rho$ are asymptotically first-order equivalent, and the asymptotically optimal bandwidth for $\tilde{f}_\rho$ is identical to that of $\tilde{f}$, i.e., b^*.

Whenever $c_\rho = 0$, as is the case for (G)EL with $\rho_3 = -2$, e.g., EL, the n^{-1} bias term in (3.5) vanishes. In general, provided the bandwidth does not go to zero faster than $n^{-1/(2r)}$, and certainly when $b = b^* \sim n^{-1/(4r+1)}$, this bias term is at most third order. Its contribution to MISE is via the integrated squared bias (ISB)

$$\begin{aligned}\mathrm{ISB}[\tilde{f}_\rho] = \mathrm{ISB}[\tilde{f}] + n^{-1}b^{2r}c_\rho 2(2r)!^{-1}\mu_{2r}(k) \int (&-\mathrm{E}[g_i^\top \Omega^{-1} g_i|u] \\ &+ \mathrm{E}[g_i^\top \Omega^{-1} g_i g_i^\top]\Omega^{-1}\,\mathrm{E}[g_i|u] + d_g) f^{(2r)}(u) f(u)\mathrm{d}u + o(n^{-1}b^{2r} \vee n^{-2}),\end{aligned}$$

with the $O(n^{-1}b^{2r})$ term generally non-zero and either positive or negative. With the asymptotically optimal bandwidth, $n^{-1}(b^*)^{2r} \sim n^{-3/2+1/(8r+2)}$, which approaches $n^{-3/2}$ arbitrarily closely as r increases, whereas the leading terms in $\mathrm{MISE}[\tilde{f}(\cdot; b^*)]$ become arbitrarily close to n^{-1}.

As long as $\mathrm{E}[g_i|u] \neq 0$, the GEL-based estimator $\tilde{f}_\rho$ enjoys a second-order reduction in variance due to the n^{-1} term in (3.6), which does not depend on the choice of GEL carrier function $\rho(\cdot)$. Hence

$$\mathrm{MISE}[\tilde{f}_\rho] = \mathrm{MISE}[\tilde{f}] - n^{-1} \int \mathrm{E}[g_i|u]^\top \Omega^{-1}\,\mathrm{E}[g_i|u] f(u)^2 \mathrm{d}u + o(n^{-1}).$$

While this reduction is negligible asymptotically, the leading term in $\mathrm{MISE}[\tilde{f}]$ approaches zero only a little more slowly than n^{-1}. Hence the effect could be substantial in small samples.

3.2 Unknown β_0

Suppose now that β_0 is unknown. Then, after substitution of the estimator $\hat{u}_i = u(z_i, \hat{\beta})$ for u_i, $i = 1, \ldots, n$, in $\tilde{f}$ and $\tilde{f}_\rho$ in (3.1) and (3.4), the analogous estimators of $f(u)$ are

$$\hat{f}(u) = n^{-1} \sum_{i=1}^{n} k_b(u - \hat{u}_i), \tag{3.7}$$

$$\hat{f}_\rho(u) = \sum_{i=1}^{n} \hat{\pi}_i k_b(u - \hat{u}_i), \tag{3.8}$$

respectively. Because u_i, $i = 1, \ldots, n$, are now not directly observable, additional restrictions need to be imposed on k and b to describe the behaviour of the estimation error $\hat{u}_i - u_i$, $i = 1, \ldots, n$. Assumption 3.3 gives a set of mild sufficient conditions, see, e.g., Van Ryzin (1969) and Ahmad (1992); similar conditions have also been considered in, e.g., Cheng (2005) and Kiwitt et al. (2008).

Assumption 3.3. ***(a)*** *k is Hölder continuous with exponent $0 < \tau \leq 1$; **(b)** there exists $d(z) \geq 0$*

with $\mathrm{E}[d(z)^\tau] < \infty$ *such that, for some* $0 < \alpha \leq 1$, $|u(z,\beta) - u(z,\beta_0)| \leq d(z)\|\beta - \beta_0\|^\alpha$ *for all* z *and for all* $\beta \in \mathcal{N}$; ***(c)*** $b \to 0$ *and* $n^{\alpha\tau/2}b^{1+\tau} \to \infty$ *as* $n \to \infty$.

The uniform α-Hölder condition Assumption 3.3(b) on $u(z,\beta)$, also known as a Lipschitz condition of order α, is an appropriate way to quantify the 'degree of continuity' of $u(z,\beta)$; see Zygmund (2003, pp.42–45). Many kernels used in practice are Lipschitz continuous, and hence satisfy Assumption 3.3(a) with $\tau = 1$. For example, a kernel that satisfies Assumption 3.3(a) is $k(x) = (1+\gamma)(1-|x|)^\gamma/2$ if $|x| \leq 1$ and 0 otherwise for any $0 < \tau \leq \gamma$ yielding the Bartlett (triangular) kernel if $\gamma = 1$. Assumption 3.3(c) is important as it prevents the bandwidth from being too small. Intuitively, if b is very small, the kernel $k_b(u - \hat{u}_i)$ is very narrowly centered around the incorrect value $\hat{u}_i$ potentially excluding the true value u_i; see, e.g., Silverman (1986, Figure 2.5) for a generic illustration. Assumption 3.3(c) requires $nb^4 \to \infty$ regardless of the values of τ and α and $b = n^{-1/4}$ is the fastest rate achievable when $\alpha = \tau = 1$. Note that the optimal bandwidth b^* is excluded if $[\alpha(4r+1) - 2]\tau < 2$.

Under these conditions, Theorem 3.2 establishes that the differences between the kernel density estimators $\hat{f}$ (3.7) and $\hat{f}_\rho$ (3.8) and their counterparts $\tilde{f}$ (3.1) and $\tilde{f}_\rho$ (3.4) based on observable u_i, $i = 1, \ldots, n$, are negligible asymptotically.

Theorem 3.2. *If Supplement P: Assumptions P.1–P.3 and 3.3 are satisfied, then* $\hat{f}(u) = \tilde{f}(u) + o_p(1)$ *and* $\hat{f}_\rho(u) = \sum_{i=1}^n \hat{\pi}_i k_b(u - u_i) + o_p(1)$ *for all* u. *If, in addition, Assumption 3.2(a)(i) holds,* $\hat{f}_\rho(u) = \tilde{f}(u) + o_p(1)$ *a.e.*

To obtain higher order expansions for the mean and variance of $\hat{f}(u)$ (3.7) and $\hat{f}_\rho(u)$ (3.8) requires a further strengthening of the assumptions. Let $\nabla u(z,\beta)$ and $\nabla^2 u(z,\beta)$ denote respectively the d_β-vector and $d_\beta \times d_\beta$ matrix of the first and second derivatives of $u(z,\beta)$ with respect to β. Also let $\nabla u_i = \nabla u(z_i, \beta_0)$ and $\nabla^2 u_i = \nabla^2 u(z_i, \beta_0)$.

Assumption 3.4. ***(a)*** k *is twice differentiable and* $k^{(2)}$ *is Hölder continuous with exponent* $0 < \tau \leq 1$, k, $k^{(1)}$, *and* $k^{(2)}$ *are absolutely integrable;* $\lim_{|x|\to\infty}|x^s k^{(s-1)}(x)| = 0$, $s = 1, 2, 3$, *and* $\int k(x)\mathrm{d}x = 1$; ***(b)*** $u(z,\beta)$ *is twice differentiable for all* $\beta \in \mathcal{N}$, $\mathrm{E}[\|\nabla u_i\|^4] < \infty$, $\mathrm{E}[\|\nabla^2 u_i\|^4] < \infty$, *and there exists* $d(z) \geq 0$ *with* $\mathrm{E}[d(z)^4] < \infty$ *such that, for some* $0 < \alpha \leq 1$, $\|\nabla^2 u(z,\beta) - \nabla^2 u(z,\beta_0)\| \leq d(z)\|\beta - \beta_0\|^\alpha$ *for all* z *and for all* $\beta \in \mathcal{N}$; ***(c)*** $b \to 0$ *as* $n \to \infty$, $n^{\tau/2}b^{3+\tau} \to \infty$, *and* $n^{\alpha/2}b^{5/4} \to \infty$; ***(d)(i)*** f *is twice differentiable;* ***(ii)*** $\mathrm{E}[\nabla u_i|u]$, $\mathrm{E}[\nabla^\top u_i H g_i|u]$, *and* $\mathrm{E}[\nabla^2 u_i|u]$ *are differentiable in* u *and* $\mathrm{E}[\nabla u_i \nabla^\top u_i|u]$ *is twice differentiable in* u; ***(iii)*** $\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u$, $\mathrm{d}\{\mathrm{E}[\nabla^\top u_i H g_i|u]f(u)\}/\mathrm{d}u$, $\mathrm{d}\{\mathrm{E}[\nabla^2 u_i|u]f(u)\}/\mathrm{d}u$, *and* $\mathrm{d}^2\{\mathrm{E}[\nabla u_i \nabla^\top u_i|u]f(u)\}/\mathrm{d}u^2$ *are absolutely integrable functions of* u.

Assumptions 3.4(a)(b) imply Assumptions 3.3(a)(b) hold with $\alpha = \tau = 1$ with the requirement in Assumption 3.3(c) rendered as $n^{1/2}b^2 \to \infty$. Note that Assumption 3.4(a) also implies Assumption 3.2(a)(i). Assumption 3.4(d) imposes additional smoothness and integrability conditions on f and $u(z,\beta)$. Assumption 3.4(c) is much stronger than Assumption 3.3(c) requiring $nb^8 \to \infty$ regardless of the values of τ and α thereby prohibiting the asymptotically optimal bandwidth b^* when k is a second order kernel. For $r \geq 2$, b^* is permissible as long as $\tau > 6/(4r-1)$ and $\alpha > 5/(8r+2)$. Note that, if $\alpha > 5/16$, $n^{\tau/2}b^{3+\tau} \to \infty$ implies $n^{\alpha/2}b^{5/4} \to \infty$.

Theorem 3.3. *If Supplement P: Assumptions P.1–P.3, 3.1, and 3.4 are satisfied, then* $\mathrm{E}[\hat{f}(u)] = \mathrm{E}[\tilde{f}(u)] + n^{-1}\delta(u) + o(n^{-1})$ *and* $\mathrm{E}[\hat{f}_\rho(u)] = \mathrm{E}[\tilde{f}(u)] + n^{-1}\delta(u) + n^{-1}\delta_\rho(u) + o(n^{-1})$*, where*

$$\begin{aligned}\delta(u) = {} & \mathrm{d}\{\mathrm{E}[\nabla^\top u_i H g_i|u]f(u)\}/\mathrm{d}u - \zeta_\lambda^\top H^\top[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u] \\ & + \frac{1}{2}\operatorname{tr}(\Sigma[\mathrm{d}^2\{\mathrm{E}[\nabla u_i \nabla^\top u_i|u]f(u)\}/\mathrm{d}u^2 - \mathrm{d}\{\mathrm{E}[\nabla^2 u_i|u]f(u)\}/\mathrm{d}u]) \end{aligned} \tag{3.9}$$

and

$$\delta_\rho(u) = (-c_\rho \mathrm{E}[g_i^\top P g_i|u] + c_\rho(d_g - d_\beta) + \zeta_\lambda^\top P \mathrm{E}[g_i|u])f(u). \tag{3.10}$$

Also

$$\begin{aligned}\operatorname{Var}[\hat{f}(u)] = {} & \operatorname{Var}[\tilde{f}(u)] + n^{-1}[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top \Sigma[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u] \\ & + n^{-1}2[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top H \mathrm{E}[g_i|u]f(u) + o(n^{-1}), \end{aligned} \tag{3.11}$$

$$\operatorname{Var}[\hat{f}_\rho(u)] = \operatorname{Var}[\hat{f}(u)] - n^{-1}\mathrm{E}[g_i|u]^\top P \mathrm{E}[g_i|u]f(u)^2 + o(n^{-1}). \tag{3.12}$$

Remark 3.9. The general conclusion of Theorem 3.3 for both bias and variance is identical to that of Theorem 3.1, i.e., the estimation effects of substituting $\hat{u}_i$ for u_i, $i = 1, \ldots, n$, and the GEL implied probabilities $\hat{\pi}_i$ for $\tilde{\pi}_i$, $i = 1, \ldots, n$, are both of order n^{-1}. The bias term in $\hat{f}$ induced by estimation is similar to that for $\tilde{f}$ in Theorem 3.1 except that P in (3.10) replaces Ω^{-1} in (3.5) and two extra terms enter via ζ_λ, *viz.* $-a$ and $\mathrm{E}[G_i H g_i]$ in (2.2). These latter terms appear in the higher order asymptotic bias $n^{-1}H(-a + \mathrm{E}[G_i H g_i])$ for the infeasible GEL estimator based on the optimal moment indicator vector $G^\top \Omega^{-1} g(z, \beta)$, see Newey and Smith (2004, Theorem 4.2), and are inherited by all GEL estimators. Unlike Theorem 3.1 for the known β_0 case, this term no longer vanishes for a particular choice of a carrier function ρ. The replacement of Ω^{-1} by P represents the loss of information occasioned by the estimation of β_0. In a number of cases, the term $\mathrm{E}[g_i|u]^\top P \mathrm{E}[g_i|u]$ may vanish, see, e.g., Supplement E: Example E.3. This of course always occurs for an exactly identified model $d_g = d_\beta$ since $\hat{\pi}_i = n^{-1}$ and $\hat{f}_\rho$ (3.8) and $\hat{f}$ (3.7) are identical. However, see Supplement E: Example E.4, in general $\hat{f}_\rho$ may still enjoy a second-order reduction in variance due to the systematic use of overidentifying moment information (1.1).

The extra bias term $\delta(u)$ (3.9) for $\hat{f}_\rho$ and those terms appearing in $\operatorname{Var}[\hat{f}(u)]$ (3.11) primarily arise due to the substitution of $\hat{u}_i$ for u_i, $i = 1, \ldots, n$. Supplement E: Examples E.2 and E.3 examine these terms in more detail for regression on a constant and (G)EL with a constant and zero mean condition respectively. Here, although $\int [\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top \Sigma [\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]\mathrm{d}u$ is non-negative, the term $\int [\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top H \mathrm{E}[g_i|u]f(u)\mathrm{d}u$ can be negative, as can be the ISB term due to the additional $\delta(u)$ (3.9).

3.3 Bias Correction

While the contribution from the n^{-1} bias terms to MISE is of a lower order than the contribution from the variance terms, the effect of bias can be substantial in small and moderate samples, potentially offsetting any reduction in variance. The direction of the bias cannot of course be

known *a priori*. Hence it may be advisable to bias-correct the density estimates by estimating and subtracting the n^{-1} bias term.

To be more specific, the bias-corrected estimates are defined as

$$\hat{f}^{bc}(u) = \hat{f}(u) - n^{-1}\hat{\delta}(u)$$

and

$$\hat{f}_{\rho}^{bc}(u) = \hat{f}_{\rho}(u) - n^{-1}\hat{\delta}(u) - n^{-1}\hat{\delta}_{\rho}(u),$$

where $\hat{\delta}(u)$ and $\hat{\delta}_{\rho}(u)$ are suitable (asymptotically) unbiased estimators of $\delta(u)$ (3.9) and $\delta_{\rho}(u)$ (3.10). The implied probabilities $\hat{\pi}_i$, $i = 1, \ldots, n$, can be used to obtain efficient estimators of the component quantities entering $\delta(u)$ and $\delta_{\rho}(u)$ with the modifications described in Glad et al. (2003) applied to ensure that the bias-corrected estimate is a density.

REMARK 3.10. When β_0 is known, bias-correction requires the estimation of the n^{-1} term in (3.5) unless $c_{\rho} = 0$, i.e., $\rho_3 = -2$.

4 GEL-Based Distribution Function Estimation

The results for distribution function estimation parallel those given in Section 3 for density estimation but can be shown to hold under much weaker conditions, and so are given here separately.

4.1 Known β_0

When u_i, $i = 1, \ldots, n$, are observed, the c.d.f. F of $u(z, \beta_0)$ can be estimated by

$$\tilde{F}(u) = n^{-1} \sum_{i=1}^{n} K((u - u_i)/b), \tag{4.1}$$

with $K(u) = \int_{-\infty}^{u} k(x)\mathrm{d}x$; see Nadaraya (1964) and Watson and Leadbetter (1964). The kernel distribution function estimator (4.1) can be obtained by integrating (3.1) or motivated as a smoothed version of the EDF.

Assumption 3.2(a)(i) is sufficient for $\widetilde{F}$ to be an asymptotically unbiased and consistent estimator of F at all continuity points of F if $b \to 0$ as $n \to \infty$. In addition, if F is continuous then $\tilde{F}$ converges to F uniformly with probability 1 (w.p.1.); see Yamato (1973). If k satisfies Assumption 3.2(a)(ii) with $\mu_{2r+2}(k) < \infty$ for some $r \geq 1$, f satisfies Assumption 3.2(b) with $s = 2r + 1$, and $b \to 0$ as $n \to \infty$ (Assumption 3.2(c) is not required here), then

$$\begin{aligned}
\mathrm{E}[\tilde{F}(u)] &= F(u) + (2r)!^{-1}\mu_{2r}(k)f^{(2r-1)}(u)b^{2r} + O(b^{2r+2}), \\
\mathrm{Var}[\widetilde{F}(u)] &= n^{-1}F(u)(1 - F(u)) - n^{-1}bf(u)\psi(k) + O(n^{-1}b^{2+\mathbb{I}\{r>1\}}),
\end{aligned}$$

where $\psi(k) = 2\int xK(x)k(x)\mathrm{d}x$. Hence

$$\mathrm{MISE}[\tilde{F}(\cdot;b)] = n^{-1}V_F - n^{-1}b\psi(k) + (2r)!^{-2}\mu_{2r}(k)^2 R(f^{(2r-1)})b^{4r} + O(n^{-1}b^{2+\mathbb{I}\{r>1\}} \vee b^{4r+2}), \quad (4.2)$$

where $V_F = \int F(u)(1-F(u))\mathrm{d}u$.

Provided $\psi(k) > 0$, the asymptotically optimal bandwidth minimising the leading terms in (4.2) is $b^* = \varsigma n^{-1/(4r-1)}$, where $\varsigma = [(2r)!^2\psi(k)/(4r\mu_{2r}(k)^2 R(f^{(2r-1)}))]^{1/(4r-1)}$, and the asymptotically optimal MISE is

$$\mathrm{MISE}[\tilde{F}(\cdot;b^*)] = n^{-1}V_F - \varsigma\psi(k)[1-(4r)^{-1}]n^{-4r/(4r-1)} + O(n^{-(4r+1+\mathbb{I}\{r>1\})/(4r-1)}).$$

REMARK 4.1. The leading term $n^{-1}V_F$ in (4.2) is the integrated variance and, hence, the MISE of EDF. Thus, whenever $\psi(k) > 0$ and b approaches zero at least as fast as $n^{-1/(4r-1)}$, kernel smoothing provides a second order asymptotic improvement in MISE relative to the EDF. Smoothness of the kernel estimates and the reduction in MISE are the two main reasons to prefer the kernel distribution function estimator (4.1) over the EDF. The condition $\psi(k) > 0$ is satisfied if k is a symmetric second order kernel, since in this case $\psi(k) = \int K(x)(1-K(x))\mathrm{d}x > 0$. Although $\psi(k)$ need not be positive in general, this property holds for certain classes of kernels, including Gaussian kernels of arbitrary order; see Oryshchenko (2017).

REMARK 4.2. If k is of order greater than two, K is not monotone, and the resultant estimates may not themselves be distribution functions. However, if necessary, the estimates can be corrected by rearrangement; see Chernozhukov et al. (2009). The MISE of the rearranged estimator can be at most equal to, and is often strictly smaller, than the MISE of the original estimator.

The modified GEL-based kernel c.d.f. estimator corresponding to $\tilde{f}_\rho$ (3.4) which incorporates the information embedded in the moment restrictions (1.1) is

$$\tilde{F}_\rho(u) = \sum_{i=1}^{n} \tilde{\pi}_i K((u-u_i)/b) \quad (4.3)$$

Theorem 4.1. *If Supplement P: Assumptions P.1–P.3 and 3.2(a)(i) are satisfied and $b \to 0$ as $n \to \infty$, then $\tilde{F}_\rho(u) = \tilde{F}(u) + o_p(1)$ at all points of continuity of F. If, in addition, Assumption 3.1 is satisfied, then*

$$\mathrm{E}[\tilde{F}_\rho(u)] = \mathrm{E}[\tilde{F}(u)] + n^{-1}c_\rho \int_{-\infty}^{u} (-\mathrm{E}[g_i^\top \Omega^{-1} g_i | t] + \mathrm{E}[g_i^\top \Omega^{-1} g_i g_i^\top]\Omega^{-1}\mathrm{E}[g_i|t] + d_g)\mathrm{d}F(t) + o(n^{-1}). \quad (4.4)$$

If also $\lim_{|x|\to\infty}|x^2 k(x)| = 0$, then

$$\mathrm{Var}[\tilde{F}_\rho(u)] = \mathrm{Var}[\tilde{F}(u)] - n^{-1}[\int_{-\infty}^{u} \mathrm{E}[g_i|t]\mathrm{d}F(t)]^\top \Omega^{-1}[\int_{-\infty}^{u} \mathrm{E}[g_i|t]\mathrm{d}F(t)] + o(n^{-1}b). \quad (4.5)$$

These results are qualitatively similar to Theorem 3.1, the important difference being that the reduction in variance is now first-order asymptotically, whereas the contribution from the n^{-1} bias term in (4.4) to MISE is of order $n^{-1}b^{2r}$. *Ceteris paribus*, the asymptotically optimal c.d.f.

bandwidth converges to zero at a faster rate than that for density estimation. Hence the additional bias effect can be expected to be of less importance.

4.2 Unknown β_0

When β_0 is unknown, the analogues of $\tilde{F}$ and $\tilde{F}_\rho$ are respectively

$$\hat{F}(u) = n^{-1}\sum_{i=1}^{n} K((u-\hat{u}_i)/b), \tag{4.6}$$

$$\hat{F}_\rho(u) = \sum_{i=1}^{n} \hat{\pi}_i K((u-\hat{u}_i)/b). \tag{4.7}$$

Theorem 4.2. *If Supplement P: Assumptions P.1–P.3 and 3.2(a)(i) are satisfied, Assumption 3.3(b) holds with $\tau = 1$ for some $0 < \alpha \leq 1$, and $b \to 0$ and $n^{\alpha/2}b \to \infty$ as $n \to \infty$, then $\hat{F}(u) = \tilde{F}(u) + o_p(1)$, $\hat{F}_\rho(u) = \tilde{F}(u) + o_p(1)$ and $\hat{F}_\rho(u) = \sum_{i=1}^n \hat{\pi}_i K((u-u_i)/b) + o_p(1)$ for all u.*

Similar to Theorem 3.2, Theorem 4.2 establishes that the differences between $\hat{F}$ (4.6) and $\hat{F}_\rho$ (4.7) and their counterparts based on observable u_i, $i = 1, \ldots, n$, are negligible asymptotically. No additional requirements are placed on k beyond the standard conditions in 3.2(a)(i) and the restriction on the bandwidth is thus weaker than Assumption 3.3(c).

Higher order expansions similar to those in Theorem 3.3 may be obtained under the following conditions.

Assumption 4.1. *Suppose Assumption 3.4(b) holds.* ***(a)*** *k is differentiable and $k^{(1)}$ is Hölder continuous with exponent $0 < \tau \leq 1$, k and $k^{(1)}$ are absolutely integrable, $\lim_{|x|\to\infty}|x^2 k(x)| = 0$, $\lim_{|x|\to\infty}|x^2 k^{(1)}(x)| = 0$ and $\int k(x)\mathrm{d}x = 1$;* ***(b)*** *$b \to 0$ as $n \to \infty$, $n^{\tau/2}b^{2+\tau} \to \infty$, and $n^{\alpha/2}b^{1/4} \to \infty$;* ***(c)(i)*** *$f(u)$ and $\mathrm{E}[\nabla u_i \nabla^\top u_i | u]$ are differentiable in u;* ***(ii)*** *$\mathrm{d}\{\mathrm{E}[\nabla u_i \nabla^\top u_i | u] f(u)\}/\mathrm{d}u$ is an absolutely integrable function of u.*

Theorem 4.3. *If Supplement P: Assumptions P.1–P.3, 3.1, and 4.1 are satisfied, then as $n \to \infty$, $\mathrm{E}[\hat{F}(u)] = \mathrm{E}[\tilde{F}(u)] + n^{-1}\Delta(u) + o(n^{-1})$ and $\mathrm{E}[\hat{F}_\rho(u)] = \mathrm{E}[\tilde{F}(u)] + n^{-1}\Delta(u) + n^{-1}\Delta_\rho(u) + o(n^{-1})$, where*

$$\begin{aligned}\Delta(u) = {} & \mathrm{E}[\nabla^\top u_i H g_i | u] f(u) - \zeta_\lambda^\top H^\top \mathrm{E}[\nabla u_i | u]^\top f(u) \\ & + \frac{1}{2}\operatorname{tr}(\Sigma[\mathrm{d}\{\mathrm{E}[\nabla u_i \nabla^\top u_i | u] f(u)\}/\mathrm{d}u - \mathrm{E}[\nabla^2 u_i | u] f(u)])\end{aligned} \tag{4.8}$$

and

$$\begin{aligned}\Delta_\rho(u) &= \int_{-\infty}^{u} (-c_\rho \mathrm{E}[g_i^\top P g_i | t] + c_\rho(d_g - d_\beta) + \zeta_\lambda^\top P \mathrm{E}[g_i | t])\mathrm{d}F(t) \\ &= \int_{-\infty}^{u} \delta_\rho(t)\mathrm{d}t.\end{aligned} \tag{4.9}$$

Also

$$\mathrm{Var}[\hat{F}(u)] = \mathrm{Var}[\tilde{F}(u)] + n^{-1}\,\mathrm{E}[\nabla u_i|u]^\top \Sigma\,\mathrm{E}[\nabla u_i|u] f(u)^2 + 2n^{-1}\,\mathrm{E}[\nabla u_i|u]^\top H[\int_{-\infty}^{u} \mathrm{E}[g_i|t]\mathrm{d}F(t)] f(u) + o(n^{-1}), \tag{4.10}$$

$$\mathrm{Var}[\hat{F}_\rho(u)] = \mathrm{Var}[\hat{F}(u)] - n^{-1}[\int_{-\infty}^{u} \mathrm{E}[g_i|t]\mathrm{d}F(t)]^\top P[\int_{-\infty}^{u} \mathrm{E}[g_i|t]\mathrm{d}F(t)] + o(n^{-1}b). \tag{4.11}$$

If, in addition, $\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u$ *is absolutely integrable, the remainder term of* $\mathrm{Var}[\hat{F}(u)]$ *is* $o(n^{-1}b)$.

REMARK 4.3. If $\delta(u)$ in Theorem 3.3 is defined, then $\Delta(u) = \int_{-\infty}^{u} \delta(t)\mathrm{d}t$, but there is no requirement that $\Delta(u)$ is absolutely continuous in Theorem 4.3. Otherwise, the interpretation is exactly the same as in Theorem 3.3. In particular, the main qualitative conclusions in Supplement E: Examples E.3 and E.4 still hold.

5 Simulation Evidence

5.1 Preliminaries

Consider the inverse hyperbolic sine (IHS) transformation model

$$\mathrm{arsinh}(\theta_0 y)/\theta_0 = \delta_0 + \gamma_0 x + u, \qquad \mathrm{E}[u|x] = 0; \tag{5.1}$$

here $\beta = (\delta, \gamma, \theta)^\top$ and $z = (y, x)^\top$. The IHS transformation was proposed in Johnson (1949, p.158) as an alternative to the Box-Cox power transform, $(y^\lambda - 1)/\lambda$, $y \geq 0$, and developed in Burbidge et al. (1988) and MacKinnon and Magee (1990); see also, e.g., Ramirez et al. (1994), Brown et al. (2015) and the references therein for recent applications in statistics and econometrics, and Tsai et al. (2017) for comparisons with other transformations. When $\theta = 0$, the IHS transform is defined as the limiting value, $\lim_{\theta\to 0} \mathrm{arsinh}(\theta y)/\theta = y$, which corresponds to the Box-Cox transform with $\lambda = 1$; when $\theta \neq 0$, the shapes of the IHS transforms are similar to those of the Box-Cox with $\lambda < 1$. The advantage of the IHS transform is that it is a smooth function of $y \in \mathbb{R}$ and $\theta \in \mathbb{R}$ with values at $\theta = 0$ defined as the corresponding limits.

The infeasible optimal instruments in the IHS transformation model (5.1) are

$$S(x, \beta_0) = (-1,\ -x,\ \mathrm{E}[\tanh(\theta_0(u + \delta_0 + \gamma_0 x))|x]/\theta_0^2 - (\delta_0 + \gamma_0 x)/\theta_0)^\top;$$

see Robinson (1991). The last element of $S(x;\beta_0)$, $s^3(x;\beta_0)$, depends on the conditional distribution of u given x, and, in general, there is little reason to argue for a particular scalar function of x as a good approximant. For example, if $u|x \sim N(0, \sigma^2)$, based on $\tanh(x) \simeq 2\Phi((\pi/2)^{1/2}x) - 1$ twice, $s^3(x;\beta_0)$ is approximately $\tanh(\theta_0(\delta_0 + \gamma_0 x)/(\pi\theta_0^2\sigma^2/2 + 1)^{1/2})/\theta_0^2 - (\delta_0 + \gamma_0 x)/\theta_0$ which suggests the use of odd degree polynomials in x as instruments; other approximations are of course possible.

In all cases the true parameters are $\delta_0 = 1$, $\gamma_0 = 2$ and $\theta_0 = 0.08$ which yield a signal-to-noise

ratio of $\gamma_0^2/(1+\gamma_0^2) = 4/5 = 0.8$ somewhat more stringent than that of $16/17 = 0.941$ in Robinson (1991, Section 7).

5.2 Design

Since the conditional distribution $u|x$ is unknown p.d.f. and c.d.f. estimators are compared based on moment condition $\mathrm{E}[g(z,\beta_0)] = 0$ (1.1) where

$$g(z,\beta) = u(z,\beta)(1,\, x,\, \ldots, x^{d_g-1})^\top,$$

for $d_g = 3$ (exactly identified), 4 and 5 (over-identified).

Three data generating processes for (x,u) are considered.

Scenario 1. x and u are independent standard normal $N(0,1)$ distributed; cf. Robinson (1991, Section 7, case (ii)).

Remark 5.1. Scenario 1 satisfies the conditions of Supplement E: Example E.3. Hence $\mathrm{IVar}[\hat{f}_\rho] = \mathrm{IVar}[\hat{f}] + o(n^{-1})$ and the relative integrated variance (IVar)

$$\mathrm{IVar}[\hat{f}]/\,\mathrm{IVar}[\tilde{f}] = 1 - \frac{b}{4\pi^{1/2}R(k)} + \frac{b}{\tau^\top D\tau R(k)}\int(\mathrm{d}\{(\tau_{0|u}(u)\tau_0)f(u)\}/\mathrm{d}u)^2\mathrm{d}u + o(b), \qquad (5.2)$$

where $\tau_{0|u}(u) = \mathrm{E}[\tanh(\theta_0(u+\delta_0+\gamma_0 x))|u]/\theta_0^2 - (\delta_0+u)/\theta_0$, $\tau_j = \mathrm{E}[x^j s^3(x,\beta_0)]$, $j = 0,1,2,\ldots$, $\tau = (\tau_0,\tau_1,\ldots,\tau_{d_g-1})^\top$, and $D = M^{-1} - \mathrm{diag}(I_2,0)$ with $M = \{M_{ij}\}_{i,j=1}^{d_g}$, $M_{ij} = \mathrm{E}[x^{i+j-2}]$, $i,j = 1,\ldots,d_g$. The term $-b/(4\pi^{1/2}R(k))$ does not depend on the number of moment conditions d_g and is the asymptotic reduction in integrated variance due to the constraint that the mean of u is zero; see also Supplement E: Example E.2. The second term in b is non-negative and represents the increase in integrated variance due to estimation of γ_0 and θ_0; it decreases as the number of moment condition increases; e.g. for $d_g = 4$, 5, 10, 20, $\tau^\top D\tau = 9.8092$, 9.8514, 9.9857 and 9.9859, respectively.

Scenarios 2 and 3. x and u have joint density $f_{ux}(u,x) = xf_{NM}(ux)f_x(x)$ where x is a generalised gamma random variable, Stacy (1962), with parameters $p = 2$, $d = \nu$ and $a = (2/\nu)^{1/2}$ for some $\nu > 4$ and f_{NM} is the normal mixture density with m components, *viz.* $f_{NM}(w) = \sum_{j=1}^m \omega_j\phi_{\sigma_j}(w-\mu_j)$, $-\infty < \mu_j < \infty$, $\sigma_j > 0$, $j = 1,\ldots,m$, $\sum_{j=1}^m \omega_j = 1$, and $\sum_{j=1}^m \omega_j\mu_j = 0$, i.e., $\mathrm{E}[w] = 0$. Here $\phi(x)$ denotes the standard normal p.d.f. and $\phi_\sigma(x) = \phi(x/\sigma)/\sigma$. The joint density f_{ux} is the density of $u = w/x$ and x where w and x are independent. The conditional density of u given x is $f_{u|x}(u|x) = xf_{NM}(ux) = \sum_{j=1}^m \omega_j\phi_{\sigma_j/x}(u-\mu_j/x)$. Hence, $\mathrm{E}[u|x] = 0$ and $\mathrm{E}[u_i^2|x] = \sum_{j=1}^m \omega_j(\sigma_j^2+\mu_j^2)/x^2$. The marginal density of u is a mixture of noncentral t densities $f_u(u) = \sum_{j=1}^m \omega_j t_\nu(u/\sigma_j;\, \mu_i/\sigma_i)/\sigma_j$ where $t_\nu(\cdot;\eta)$ is the density of a noncentral t-distributed random variable with ν degrees of freedom and noncentrality parameter η allowing a wide variety of shapes for f_u by varying the mixture f_{NM}. The skewed unimodal and bimodal densities shown in Figure 1 describe the NM densities for Scenarios 2 and 3 respectively, i.e., the mixture densities Marron and Wand (1992, #2 and #8) centered to have zero mean.

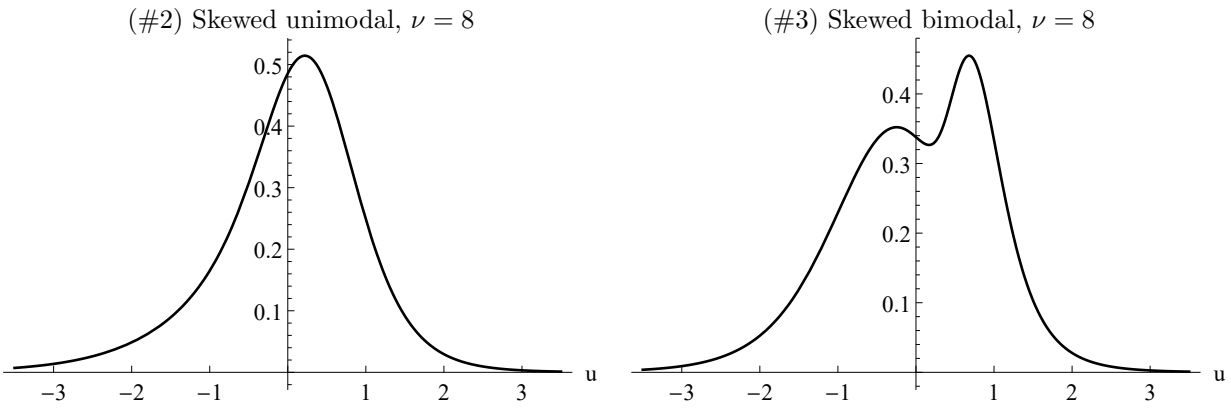


Figure 1: Selected mixture densities (scaled)

5.3 Kernel Functions and Bandwidths

Fourth order Gaussian-based kernels, $k(x) = (3-x^2)\phi(x)/2$ and $K(x) = \Phi(x)+x\phi(x)/2$, $\Phi(x) = \int_{-\infty}^{x}\phi(u)\mathrm{d}u$, are employed; see Wand and Schucany (1990, Section 2) and Oryshchenko (2017) respectively. Thus the choices of the asymptotically optimal bandwidths $(27/4\sqrt{\pi})^{1/9}R(f^{(4)})^{-1/9}n^{-1/9}$ and $(7/2\sqrt{\pi})^{1/7}R(f^{(3)})^{-1/7}n^{-1/7}$ for p.d.f. and c.d.f. estimation respectively are permitted, thereby satisfying Assumptions 3.4(c) and 4.1(b). The practical issue of estimating the derivatives of f required for the computation of $R(f^{(j)})$, $j = 3, 4$, is ignored and the respective true values used. For the standard normal distribution these are $R(\phi^{(3)}) = 15/(16\sqrt{\pi})$ and $R(\phi^{(4)}) = 105/(32\sqrt{\pi})$; for the mixture distributions with approximate values shown in Figure 1.

5.4 Results

The study compares the performance of GEL-based kernel density p.d.f. and c.d.f. estimators. The GEL parameter estimators are CUE, EL and ET, the most notable special cases of the GEL family. For each estimator the mean and variance were computed on a grid 1000 of points between -5 and 5 and are reported as the integrated squared bias and integrated variance relative to those of the corresponding infeasible estimator based on the true u, i.e., $\tilde{f}$ and $\tilde{F}$.

Tables 1, 2 and 3 report results for Scenarios 1, 2 and 3 respectively. The ISB, IVar and MISE (all $\times 10^5$) for the infeasible $\tilde{f}$ and $\tilde{F}$ are presented. Rows ISB, IVar, and MISE are the ISB, IVar, and MISE of $\hat{f}$, $\hat{f}_\rho$ ($\hat{F}$, $\hat{F}_\rho$) relative to the infeasible $\tilde{f}$ ($\tilde{F}$), respectively; row 'vs $d_g = 3$' is the MISE of $\hat{f}$, $\hat{f}_\rho$ ($\hat{F}$, $\hat{F}_\rho$) relative to the corresponding value for $d_g = 3$; row 'w. vs unw.' is the MISE of $\hat{f}_\rho$ ($\hat{F}_\rho$) relative to $\hat{f}$ ($\hat{F}$). Rows MISE, 'vs $d_g = 3$', and 'w. vs unw.' examine the significance of the paired t-statistics in a two-sided test for equality of the respective ISE means, e.g., $\int(\hat{f}(u) - f(u))^2\mathrm{d}u$; the symbol † indicates that the p-value is between 0.01 and 0.05 whereas ‡ that it is less than 0.01 and in all other cases the p-value is greater than 0.05. Values of relative MISE less than 1 are emphasised in bold.

Sample sizes $n = 100$, 500, 1000, and 2000 are examined.

All computations were carried out in MATLAB; the relevant code and additional results, including the properties of GEL estimators, are available from the first named author upon request. All results are based on $10,000$ random draws.

5.4.1 Scenario 1

The first $\sim b$ term in eq. (5.2) is approximately $-0.321n^{-1/9}$, which for $n = 100$, 500, 1000, and 2000 is approximately -0.192, -0.161, -0.149, and -0.138 respectively. The second $\sim b$ term is approximately $0.04728n^{-1/9}$ for $d_g = 4$ and $0.04708n^{-1/9}$ for $d_g = 5$, which offsets the reduction in variance slightly. The predicted relative IVar of $\hat{f}$ and $\hat{f}_\rho$ up to order $o(b)$ is thus 0.836, 0.863, 0.873 and 0.882 for $n = 100$, 500, 1000, and 2000 respectively and is identical within three digit precision for $d_g = 4$ and 5.

The results reported in Table 1 confirm these predictions. In fact, the reduction in variance is even larger than expected in small and medium samples due to the $o(b)$ effects. Furthermore, estimators $\hat{f}$ and $\hat{f}_\rho$ have smaller ISB relative to $\tilde{f}$. A comparison of $\hat{f}$ and $\hat{f}_\rho$ between $d_g = 3$ (just-identified) and $d_g = 4, 5$ (over-identified) for moderate and larger sample sizes emphasises further the contribution of additional moment information. Hence $\hat{f}$ and $\hat{f}_\rho$ enjoy a reduction in MISE of as much as 21% for $n = 100$ and 10% for $n = 2000$ relative to $\tilde{f}$. The benefits are even more pronounced for c.d.f. estimation, where the reduction in MISE can be as much as 56% for $n = 100$ and around 53% in moderate samples. There are also small but statistically significant benefits to re-weighting which are mostly due to the smaller biases of $\hat{f}_\rho$ and $\hat{F}_\rho$ relative to $\hat{f}$ and $\hat{F}$ at moderate and larger sample sizes. There is some deterioration in ISB, IVar and, thus, MISE with increases in d_g which can be contributed to the increased importance of outliers.

Finally, while in moderate and large samples the performances of CUE, EL, and ET are virtually identical, in small samples ET can be unstable with larger d_g.

5.4.2 Scenarios 2 and 3

Scenarios 2 and 3 with densities of (x, u) which are heavy-tailed and also, e.g., skewed and bimodal, illustrate the many difficulties for both GEL estimation and kernel p.d.f. and c.d.f. estimation which are absent in the relatively benign Scenario 1.

The performance of CUE in small samples is generally worse than that of EL and ET. It ranks last by MSE in both scenarios with $n = 100$ and 500, except Scenario 3 with $n = 100$ when ET underperforms. In a number of cases increasing with d_g the optimisation routine for ET failed. Somewhat surprisingly, although it is known to be sensitive to outliers, EL appears to deliver good results in the simulation experiments. It ranks first by MSE in Scenario 3 with $d_g = 5$ and alternates with ET otherwise. These differences become very small with $n = 1,000$ and greater.

The conclusion about the inferior performance of CUE in small samples holds for CUE-based kernel density p.d.f. and c.d.f. estimators as well; see Tables 2 and 3, in particular, the ISBs of $\hat{f}$ and $\hat{f}_\rho$ with $d_g = 4, 5$ in Table 2. However, the ranking of EL and ET-based kernel density p.d.f. and c.d.f. estimators by MISE does not always correspond to the ranking of the underlying EL and ET estimators of β_0 by MSE.

Table 1: Performance of GEL-based residual density and distribution function estimators in the IHS transformation model, $\text{arsinh}(0.08y)/0.08 = 1 + 2x + u$, in Scenario 1

				$d_g = 3$		$d_g = 4$				$d_g = 5$			
		$\tilde{f}$	$\tilde{F}$	$\hat{f}$	$\hat{F}$	$\hat{f}$	$\hat{f}_\rho$	$\hat{F}$	$\hat{F}_\rho$	$\hat{f}$	$\hat{f}_\rho$	$\hat{F}$	$\hat{F}_\rho$
						$n = 100$							
CUE	ISB	31.3	8.9	0.70	3.41	0.13	0.20	0.47	1.12	0.12	0.37	0.56	1.85
	IVar	449.0	402.3	1.29	0.71	0.88	0.88	0.46	0.43	0.87	0.86	0.47	0.42
	MISE	480.2	411.2	1.26‡	**0.78‡**	**0.83‡**	**0.83‡**	**0.46‡**	**0.45‡**	**0.82‡**	**0.83‡**	**0.47‡**	**0.45‡**
	vs $d_g = 3$					**0.66‡**	**0.66‡**	**0.60‡**	**0.58‡**	**0.65‡**	**0.66‡**	**0.61‡**	**0.59‡**
	w. vs unw.						**0.999**		**0.965‡**		1.010‡		**0.964‡**
EL	ISB					0.16	0.17	0.56	0.64	0.17	0.20	0.71	0.87
	IVar					0.84	0.84	0.45	0.42	0.87	0.89	0.51	0.45
	MISE					**0.80‡**	**0.80‡**	**0.45‡**	**0.42‡**	**0.83‡**	**0.85‡**	**0.51‡**	**0.46‡**
	vs $d_g = 3$					**0.64‡**	**0.64‡**	**0.58‡**	**0.55‡**	**0.66‡**	**0.68‡**	**0.66‡**	**0.60‡**
	w. vs unw.						1.001		**0.930‡**		1.024		**0.906‡**
ET	ISB					0.15	0.20	0.55	1.01	0.14	0.34	0.66	1.70
	IVar					0.83	0.89	0.43	0.47	0.85	0.86	0.48	0.53
	MISE					**0.79‡**	**0.85‡**	**0.44‡**	**0.48‡**	**0.81‡**	**0.84‡**	**0.49‡**	**0.88**
	vs $d_g = 3$					**0.63‡**	**0.68‡**	**0.56‡**	**0.62‡**	**0.64‡**	**0.67‡**	**0.64‡**	1.16
	w. vs unw.						1.071†		1.092		1.037		1.789
						$n = 500$							
CUE	ISB	10.0	1.8	0.37	1.53	0.46	0.36	0.29	0.28	0.40	0.27	0.27	0.33
	IVar	119.8	88.3	0.99	0.61	0.87	0.87	0.46	0.45	0.88	0.88	0.47	0.46
	MISE	129.8	90.2	**0.94‡**	**0.63‡**	**0.84‡**	**0.83‡**	**0.45‡**	**0.45‡**	**0.84‡**	**0.83‡**	**0.47‡**	**0.46‡**
	vs $d_g = 3$					**0.89‡**	**0.88‡**	**0.72‡**	**0.71‡**	**0.90‡**	**0.89‡**	**0.74‡**	**0.73‡**
	w. vs unw.						**0.991‡**		**0.988‡**		**0.988‡**		**0.986‡**
EL	ISB					0.45	0.45	0.29	0.30	0.41	0.40	0.28	0.29
	IVar					0.87	0.87	0.46	0.45	0.88	0.88	0.47	0.46
	MISE					**0.83‡**	**0.84‡**	**0.45‡**	**0.45‡**	**0.84‡**	**0.84‡**	**0.47‡**	**0.46‡**
	vs $d_g = 3$					**0.89‡**	**0.89‡**	**0.72‡**	**0.72‡**	**0.89‡**	**0.90‡**	**0.74‡**	**0.73‡**
	w. vs unw.						1.002‡		**0.993‡**		1.003†		**0.979‡**
ET	ISB					0.45	0.39	0.29	0.28	0.40	0.31	0.27	0.30
	IVar					0.87	0.87	0.46	0.45	0.88	0.88	0.47	0.46
	MISE					**0.83‡**	**0.83‡**	**0.45‡**	**0.45‡**	**0.84‡**	**0.83‡**	**0.46‡**	**0.46‡**
	vs $d_g = 3$					**0.89‡**	**0.88‡**	**0.72‡**	**0.71‡**	**0.89‡**	**0.89‡**	**0.74‡**	**0.73‡**
	w. vs unw.						**0.996‡**		**0.991‡**		**0.994‡**		**0.986‡**
						$n = 1000$							
CUE	ISB	6.1	0.9	0.48	1.03	0.62	0.55	0.41	0.33	0.58	0.46	0.36	0.28
	IVar	66.1	45.6	0.99	0.62	0.89	0.89	0.48	0.48	0.90	0.90	0.49	0.49
	MISE	72.2	46.5	**0.95‡**	**0.63‡**	**0.87‡**	**0.86‡**	**0.48‡**	**0.47‡**	**0.87‡**	**0.86‡**	**0.49‡**	**0.48‡**
	vs $d_g = 3$					**0.91‡**	**0.91‡**	**0.76‡**	**0.75‡**	**0.92‡**	**0.91‡**	**0.78‡**	**0.77‡**
	w. vs unw.						**0.992‡**		**0.990‡**		**0.988‡**		**0.988‡**
EL	ISB					0.62	0.62	0.40	0.40	0.59	0.58	0.37	0.36
	IVar					0.89	0.89	0.48	0.48	0.89	0.89	0.49	0.48
	MISE					**0.86‡**	**0.86‡**	**0.48‡**	**0.48‡**	**0.87‡**	**0.87‡**	**0.49‡**	**0.48‡**
	vs $d_g = 3$					**0.91‡**	**0.91‡**	**0.76‡**	**0.76‡**	**0.91‡**	**0.91‡**	**0.77‡**	**0.77‡**
	w. vs unw.						1.001†		**0.996‡**		1.001		**0.989‡**
ET	ISB					0.62	0.58	0.40	0.36	0.58	0.50	0.36	0.30
	IVar					0.89	0.89	0.48	0.48	0.89	0.89	0.49	0.48
	MISE					**0.86‡**	**0.86‡**	**0.48‡**	**0.47‡**	**0.87‡**	**0.86‡**	**0.48‡**	**0.48‡**
	vs $d_g = 3$					**0.91‡**	**0.91‡**	**0.76‡**	**0.76‡**	**0.91‡**	**0.91‡**	**0.77‡**	**0.76‡**
	w. vs unw.						**0.996‡**		**0.993‡**		**0.993‡**		**0.990‡**
						$n = 2000$							
CUE	ISB	3.5	0.4	0.55	0.62	0.74	0.69	0.53	0.45	0.71	0.62	0.49	0.37
	IVar	36.6	23.0	1.02	0.65	0.92	0.92	0.52	0.52	0.93	0.93	0.53	0.52
	MISE	40.1	23.5	**0.98‡**	**0.65‡**	**0.90‡**	**0.90‡**	**0.52‡**	**0.51‡**	**0.91‡**	**0.90‡**	**0.53‡**	**0.52‡**
	vs $d_g = 3$					**0.92‡**	**0.92‡**	**0.80‡**	**0.79‡**	**0.93‡**	**0.92‡**	**0.81‡**	**0.80‡**
	w. vs unw.						**0.994‡**		**0.994‡**		**0.990‡**		**0.992‡**
EL	ISB					0.74	0.74	0.52	0.52	0.71	0.71	0.48	0.48
	IVar					0.92	0.92	0.52	0.52	0.92	0.92	0.52	0.52
	MISE					**0.90‡**	**0.90‡**	**0.52‡**	**0.52‡**	**0.90‡**	**0.90‡**	**0.52‡**	**0.52‡**
	vs $d_g = 3$					**0.92‡**	**0.92‡**	**0.80‡**	**0.80‡**	**0.92‡**	**0.93‡**	**0.81‡**	**0.80‡**
	w. vs unw.						1.000		**0.999**		1.001		**0.996‡**
ET	ISB					0.74	0.71	0.53	0.48	0.71	0.65	0.49	0.41
	IVar					0.92	0.92	0.52	0.52	0.92	0.92	0.52	0.52
	MISE					**0.90‡**	**0.90‡**	**0.52‡**	**0.52‡**	**0.90‡**	**0.90‡**	**0.52‡**	**0.52‡**
	vs $d_g = 3$					**0.92‡**	**0.92‡**	**0.80‡**	**0.79‡**	**0.93‡**	**0.92‡**	**0.80‡**	**0.80‡**
	w. vs unw.						**0.997‡**		**0.996‡**		**0.994‡**		**0.994‡**

Notes: see text.

Table 2: Performance of GEL-based residual density and distribution function estimators in the IHS transformation model, $\text{arsinh}(0.08y)/0.08 = 1 + 2x + u$, in Scenario 2

				$d_g = 3$		$d_g = 4$				$d_g = 5$			
		$\tilde{f}$	$\tilde{F}$	$\hat{f}$	$\hat{F}$	$\hat{f}$	$\hat{f}_\rho$	$\hat{F}$	$\hat{F}_\rho$	$\hat{f}$	$\hat{f}_\rho$	$\hat{F}$	$\hat{F}_\rho$
$n = 100$													
CUE	ISB	29.3	3.8	0.73	5.71	0.85	1.13	2.65	5.14	0.99	1.52	3.65	7.65
	IVar	822.0	427.7	1.12	0.75	0.93	0.94	0.54	0.50	0.93	0.94	0.55	0.49
	MISE	852.5	432.5	1.11‡	**0.79**‡	**0.93**‡	**0.94**‡	**0.56**‡	**0.54**‡	**0.93**‡	**0.95**‡	**0.57**‡	**0.55**‡
	vs $d_g = 3$					**0.84**‡	**0.85**‡	**0.70**‡	**0.68**‡	**0.84**‡	**0.86**‡	**0.72**‡	**0.70**‡
	w. vs unw.						1.013‡		**0.963**‡		1.026‡		**0.966**‡
EL	ISB					0.71	0.78	1.95	2.59	0.70	0.83	1.99	3.03
	IVar					0.92	0.93	0.53	0.50	0.93	0.96	0.56	0.51
	MISE					**0.91**‡	**0.93**‡	**0.54**‡	**0.52**‡	**0.92**‡	**0.95**‡	**0.57**‡	**0.54**‡
	vs $d_g = 3$					**0.82**‡	**0.84**‡	**0.68**‡	**0.66**‡	**0.83**‡	**0.86**‡	**0.72**‡	**0.68**‡
	w. vs unw.						1.015‡		**0.956**‡		1.029‡		**0.934**‡
ET	ISB					0.77	0.97	2.37	4.19	0.86	1.28	2.99	6.22
	IVar					0.91	0.95	0.51	0.59	0.92	0.94	0.55	0.63
	MISE					**0.90**‡	**0.95**	**0.53**‡	**0.62**‡	**0.92**‡	**0.96**	**0.58**‡	**0.80**
	vs $d_g = 3$					**0.82**‡	**0.86**‡	**0.67**‡	**0.80**	**0.83**‡	**0.87**‡	**0.73**‡	1.03
	w. vs unw.						1.054†		1.174		1.053‡		1.390
$n = 500$													
CUE	ISB	10.0	0.8	0.80	4.25	0.88	0.86	0.96	1.36	0.87	0.84	1.15	1.92
	IVar	210.2	89.9	1.08	0.82	0.93	0.93	0.53	0.52	0.93	0.93	0.53	0.51
	MISE	220.4	90.9	1.07‡	**0.85**‡	**0.93**‡	**0.93**‡	**0.54**‡	**0.52**‡	**0.93**‡	**0.93**‡	**0.54**‡	**0.53**‡
	vs $d_g = 3$					**0.87**‡	**0.87**‡	**0.63**‡	**0.62**‡	**0.87**‡	**0.87**‡	**0.63**‡	**0.62**‡
	w. vs unw.						1.001		**0.980**‡		1.001		**0.980**‡
EL	ISB					0.87	0.89	0.87	0.94	0.84	0.85	0.86	0.98
	IVar					0.93	0.98	0.53	0.72	0.93	0.93	0.53	0.52
	MISE					**0.93**‡	**0.97**	**0.53**‡	**0.72**	**0.92**‡	**0.93**‡	**0.53**‡	**0.52**‡
	vs $d_g = 3$					**0.87**‡	**0.91**†	**0.63**‡	**0.85**	**0.87**‡	**0.87**‡	**0.63**‡	**0.62**‡
	w. vs unw.						1.049		1.343		1.005‡		**0.978**‡
ET	ISB					0.88	0.87	0.93	1.16	0.85	0.82	1.03	1.50
	IVar					0.93	0.93	0.53	0.52	0.92	0.93	0.52	0.51
	MISE					**0.93**‡	**0.93**‡	**0.53**‡	**0.52**‡	**0.92**‡	**0.92**‡	**0.53**‡	**0.52**‡
	vs $d_g = 3$					**0.87**‡	**0.87**‡	**0.63**‡	**0.62**‡	**0.86**‡	**0.87**‡	**0.62**‡	**0.61**‡
	w. vs unw.						1.001‡		**0.982**‡		1.002‡		**0.984**‡
$n = 1000$													
CUE	ISB	6.5	0.5	0.84	2.78	0.94	0.92	0.89	0.99	0.91	0.87	0.94	1.19
	IVar	115.0	45.7	1.09	0.81	0.94	0.94	0.54	0.53	0.94	0.94	0.54	0.52
	MISE	121.7	46.3	1.07‡	**0.83**‡	**0.94**‡	**0.94**‡	**0.54**‡	**0.53**‡	**0.94**‡	**0.94**‡	**0.54**‡	**0.53**‡
	vs $d_g = 3$					**0.88**‡	**0.88**‡	**0.65**‡	**0.64**‡	**0.88**‡	**0.87**‡	**0.65**‡	**0.64**‡
	w. vs unw.						**0.999**‡		**0.982**‡		**0.999**‡		**0.980**‡
EL	ISB					0.94	0.95	0.87	0.89	0.92	0.92	0.84	0.88
	IVar					0.94	0.94	0.54	0.53	0.94	0.94	0.54	0.53
	MISE					**0.94**‡	**0.94**‡	**0.54**‡	**0.54**‡	**0.94**‡	**0.94**‡	**0.54**‡	**0.53**‡
	vs $d_g = 3$					**0.88**‡	**0.88**‡	**0.65**‡	**0.64**‡	**0.87**‡	**0.88**‡	**0.65**‡	**0.64**‡
	w. vs unw.						1.001‡		**0.982**‡		1.003‡		**0.984**‡
ET	ISB					0.94	0.93	0.88	0.93	0.91	0.88	0.89	1.01
	IVar					0.94	0.94	0.54	0.53	0.94	0.94	0.53	0.52
	MISE					**0.94**‡	**0.94**‡	**0.54**‡	**0.53**‡	**0.94**‡	**0.94**‡	**0.54**‡	**0.53**‡
	vs $d_g = 3$					**0.88**‡	**0.88**‡	**0.65**‡	**0.64**‡	**0.87**‡	**0.87**‡	**0.65**‡	**0.63**‡
	w. vs unw.						**1.000**		**0.982**‡		1.000		**0.983**‡
$n = 2000$													
CUE	ISB	4.2	0.3	0.93	1.88	0.99	0.97	0.96	0.97	0.96	0.92	0.94	1.01
	IVar	64.9	23.7	1.09	0.81	0.95	0.95	0.55	0.54	0.95	0.95	0.55	0.54
	MISE	69.1	24.0	1.08‡	**0.82**‡	**0.95**‡	**0.95**‡	**0.56**‡	**0.55**‡	**0.95**‡	**0.95**‡	**0.55**‡	**0.54**‡
	vs $d_g = 3$					**0.88**‡	**0.88**‡	**0.68**‡	**0.67**‡	**0.88**‡	**0.87**‡	**0.68**‡	**0.66**‡
	w. vs unw.						**0.999**‡		**0.983**‡		**0.998**‡		**0.979**‡
EL	ISB					0.99	0.99	0.95	0.95	0.97	0.97	0.91	0.93
	IVar					0.95	0.95	0.56	0.55	0.95	0.95	0.55	0.54
	MISE					**0.95**‡	**0.95**‡	**0.56**‡	**0.55**‡	**0.95**‡	**0.95**‡	**0.56**‡	**0.54**‡
	vs $d_g = 3$					**0.88**‡	**0.88**‡	**0.68**‡	**0.67**‡	**0.88**‡	**0.88**‡	**0.68**‡	**0.66**‡
	w. vs unw.						**1.000**		**0.983**‡		**1.000**		**0.978**‡
ET	ISB					0.99	0.98	0.96	0.95	0.95	0.93	0.92	0.94
	IVar					0.95	0.95	0.55	0.54	0.95	0.95	0.55	0.54
	MISE					**0.95**‡	**0.95**‡	**0.56**‡	**0.55**‡	**0.95**‡	**0.94**‡	**0.55**‡	**0.54**‡
	vs $d_g = 3$					**0.88**‡	**0.88**‡	**0.68**‡	**0.67**‡	**0.87**‡	**0.87**‡	**0.67**‡	**0.66**‡
	w. vs unw.						**0.999**‡		**0.983**‡		**0.999**‡		**0.980**‡

Notes: see text.

Table 3: Performance of GEL-based residual density and distribution function estimators in the IHS transformation model, $\text{arsinh}(0.08y)/0.08 = 1 + 2x + u$, in Scenario 3

				$d_g = 3$		$d_g = 4$				$d_g = 5$			
		$\tilde{f}$	$\tilde{F}$	$\hat{f}$	$\hat{F}$	$\hat{f}$	$\hat{f}_\rho$	$\hat{F}$	$\hat{F}_\rho$	$\hat{f}$	$\hat{f}_\rho$	$\hat{F}$	$\hat{F}_\rho$
$n = 100$													
CUE	ISB	23.9	1.3	7.34	41.60	4.25	5.01	14.61	24.59	4.68	6.19	17.12	33.44
	IVar	1546.0	485.4	1.19	0.76	1.01	1.01	0.53	0.50	1.01	1.02	0.55	0.50
	MISE	1570.5	487.0	1.28‡	**0.88‡**	1.06‡	1.08‡	**0.57‡**	**0.56‡**	1.07‡	1.10‡	**0.59‡**	**0.59‡**
	vs $d_g = 3$					**0.82‡**	**0.84‡**	**0.65‡**	**0.65‡**	**0.83‡**	**0.86‡**	**0.67‡**	**0.68‡**
	w. vs unw.						1.018‡		**0.986‡**		1.033‡		**0.999**
EL	ISB					4.08	4.08	13.99	15.96	4.28	4.38	13.51	17.07
	IVar					0.99	1.01	0.51	0.49	1.00	1.04	0.54	0.51
	MISE					1.04‡	1.05‡	**0.55‡**	**0.53‡**	1.05‡	1.09‡	**0.57‡**	**0.55‡**
	vs $d_g = 3$					**0.81‡**	**0.82‡**	**0.62‡**	**0.61‡**	**0.82‡**	**0.85‡**	**0.65‡**	**0.64‡**
	w. vs unw.						1.014‡		**0.972‡**		1.034‡		**0.972**
ET	ISB					4.14	4.71	14.22	22.01	4.53	5.69	14.89	27.85
	IVar					0.99	1.01	0.50	0.49	1.00	1.02	0.66	0.84
	MISE					1.03‡	1.06‡	**0.54‡**	**0.54‡**	1.05‡	1.09‡	**0.78**	1.91
	vs $d_g = 3$					**0.81‡**	**0.83‡**	**0.62‡**	**0.62‡**	**0.82‡**	**0.85‡**	**0.90**	2.25
	w. vs unw.						1.027‡		1.006		1.040‡		2.459
$n = 500$													
CUE	ISB	9.6	0.4	2.39	13.99	1.54	1.58	2.61	3.94	1.60	1.69	2.84	5.19
	IVar	379.2	100.3	1.10	0.76	1.00	1.00	0.51	0.50	1.00	1.00	0.51	0.50
	MISE	388.9	100.8	1.13‡	**0.81‡**	1.01‡	1.01‡	**0.52‡**	**0.51‡**	1.01‡	1.02‡	**0.52‡**	**0.51‡**
	vs $d_g = 3$					**0.89‡**	**0.90‡**	**0.64‡**	**0.63‡**	**0.90‡**	**0.90‡**	**0.65‡**	**0.64‡**
	w. vs unw.						1.002‡		**0.981‡**		1.004‡		**0.982‡**
EL	ISB					1.54	1.55	2.55	2.65	1.57	1.59	2.46	2.59
	IVar					1.00	1.00	0.51	0.50	1.00	1.07	0.51	0.80
	MISE					1.01‡	1.01‡	**0.51‡**	**0.50‡**	1.01‡	1.09	**0.52‡**	**0.81**
	vs $d_g = 3$					**0.89‡**	**0.89‡**	**0.64‡**	**0.62‡**	**0.89‡**	**0.96**	**0.64‡**	**1.00**
	w. vs unw.						1.002‡		**0.979‡**		1.074		1.563
ET	ISB					1.53	1.56	2.59	3.40	1.58	1.64	2.65	4.11
	IVar					1.00	1.00	0.51	0.49	1.00	1.00	0.50	0.49
	MISE					1.01‡	1.01‡	**0.51‡**	**0.50‡**	1.01‡	1.01‡	**0.51‡**	**0.51‡**
	vs $d_g = 3$					**0.89‡**	**0.89‡**	**0.64‡**	**0.62‡**	**0.89‡**	**0.90‡**	**0.64‡**	**0.63‡**
	w. vs unw.						1.002‡		**0.982‡**		1.004‡		**0.985‡**
$n = 1000$													
CUE	ISB	6.6	0.2	1.86	8.15	1.33	1.34	1.66	2.16	1.37	1.39	1.80	2.79
	IVar	206.9	50.4	1.09	0.74	1.00	1.00	0.51	0.50	1.00	1.00	0.51	0.50
	MISE	213.5	50.6	1.12‡	**0.77‡**	1.01‡	1.01‡	**0.52‡**	**0.51‡**	1.01‡	1.01‡	**0.52‡**	**0.51‡**
	vs $d_g = 3$					**0.90‡**	**0.90‡**	**0.67‡**	**0.66‡**	**0.91‡**	**0.91‡**	**0.67‡**	**0.66‡**
	w. vs unw.						1.001‡		**0.980‡**		1.001‡		**0.978‡**
EL	ISB					1.34	1.35	1.65	1.64	1.36	1.37	1.68	1.74
	IVar					1.00	1.13	0.51	1.09	1.00	1.00	0.51	0.50
	MISE					1.01‡	1.14	**0.52‡**	1.09	1.01‡	1.01‡	**0.52‡**	**0.50‡**
	vs $d_g = 3$					**0.90‡**	1.02	**0.67‡**	1.41	**0.90‡**	**0.90‡**	**0.67‡**	**0.65‡**
	w. vs unw.						1.124		2.108		1.001†		**0.973‡**
ET	ISB					1.34	1.34	1.66	1.91	1.36	1.37	1.76	2.32
	IVar					1.00	1.00	0.51	0.50	1.00	1.00	0.51	0.49
	MISE					1.01‡	1.01‡	**0.51‡**	**0.50‡**	1.01‡	1.01‡	**0.51‡**	**0.50‡**
	vs $d_g = 3$					**0.90‡**	**0.90‡**	**0.67‡**	**0.65‡**	**0.90‡**	**0.90‡**	**0.66‡**	**0.65‡**
	w. vs unw.						1.001‡		**0.979‡**		1.001‡		**0.979‡**
$n = 2000$													
CUE	ISB	4.0	0.1	1.60	4.53	1.24	1.23	1.34	1.54	1.25	1.25	1.41	1.86
	IVar	113.2	25.6	1.10	0.74	1.00	1.00	0.52	0.51	1.00	1.00	0.52	0.51
	MISE	117.2	25.8	1.11‡	**0.76‡**	1.01‡	1.01‡	**0.53‡**	**0.52‡**	1.01‡	1.01‡	**0.52‡**	**0.51‡**
	vs $d_g = 3$					**0.91‡**	**0.91‡**	**0.69‡**	**0.68‡**	**0.91‡**	**0.91‡**	**0.69‡**	**0.67‡**
	w. vs unw.						**0.999‡**		**0.981‡**		**0.999‡**		**0.977‡**
EL	ISB					1.25	1.25	1.34	1.35	1.25	1.26	1.36	1.38
	IVar					1.00	1.00	0.52	0.51	1.00	1.11	0.52	0.91
	MISE					1.01‡	1.01‡	**0.53‡**	**0.52‡**	1.01‡	1.12	**0.53‡**	**0.92**
	vs $d_g = 3$					**0.91‡**	**0.91‡**	**0.69‡**	**0.68‡**	**0.91‡**	1.00	**0.69‡**	1.20
	w. vs unw.						**0.999‡**		**0.980‡**		1.105		1.743
ET	ISB					1.24	1.24	1.34	1.43	1.24	1.25	1.39	1.62
	IVar					1.00	1.00	0.52	0.51	1.00	1.00	0.52	0.51
	MISE					1.01‡	1.01‡	**0.53‡**	**0.52‡**	1.01‡	1.01‡	**0.52‡**	**0.51‡**
	vs $d_g = 3$					**0.91‡**	**0.91‡**	**0.69‡**	**0.68‡**	**0.91‡**	**0.91‡**	**0.69‡**	**0.67‡**
	w. vs unw.						**0.999‡**		**0.980‡**		**0.999‡**		**0.977‡**

Notes: see text.

In particular, the sensitivity of EL to outliers adversely affects the estimators $\hat{f}_\rho$ and $\hat{F}_\rho$ via the implied probabilities in Scenario 3 with $n = 500$ and greater; see Table 3. ET and CUE perform better in those cases.

Unlike Scenario 1, in Scenario 3 none of the feasible kernel density estimators have smaller MISE than their infeasible counterparts for the sample sizes considered. In Scenario 2, with less complicated distributional features, these estimators do achieve a reduction in MISE with $d_g = 4, 5$. The same is true for the feasible kernel c.d.f. estimators in Scenario 2 with $d_g = 3, 4, 5$, and more often than not in Scenario 3 as well, with the few exceptions mentioned above. Importantly, it is generally beneficial to increase the number of moment conditions beyond those necessary to identify the parameters except when stability of GEL estimators of β_0 is likely to deteriorate.

Finally, the benefits of re-weighting are present, but not universal, and as expected, are quite small; cf. Supplement E: Example E.4.

6 Summary and Conclusions

Large sample results and simulation evidence reported in this paper suggest that it is generally sensible to apply either standard or re-weighted kernel estimators to estimate the p.d.f. or c.d.f. of a scalar residual $u(z, \beta_0)$ in a variety of situations, provided the error associated with the estimation of β_0 satisfies some mild regularity conditions and care is taken to ensure the bandwidth is not too small. If the assumptions on $u(z, \beta)$ prove difficult to verify in practice, using fourth or higher order kernels and the corresponding asymptotically optimal bandwidths will generally assist with ensuring the appropriate regularity conditions hold.

Incorporating information from overidentifying moment conditions by re-weighting the estimators using GEL implied probabilities offers efficiency gains which are realised in regular situations. However, if the model is highly nonlinear and the distribution of the data is heavy-tailed or contaminated with outliers, the methods proposed in this paper, including GEL, should be applied with some caution in very small samples. Robust hybrid estimators such as exponentially tilted empirical likelihood, see, e.g., Schennach (2007), may prove useful in these circumstances.

While the results in this paper are presented only for scalar-valued $u(z, \beta)$, generalisations to the vector case are relatively straightforward provided an analogue of the bijection Assumption 3.1 holds.

An issue for future research is the construction of tests for overidentifying moment conditions or parametric restrictions based on the differences between the kernel p.d.f. estimators $\hat{f}_\rho$ and $\hat{f}$ or $\tilde{f}_\rho$ and $\tilde{f}$ for known β_0. Test statistics of the Bickel-Rosenblatt type based on the integrated squared difference $\int(\hat{f}_\rho(u) - \hat{f}(u))^2 \mathrm{d}u$, Bickel and Rosenblatt (1973), Fan (1994, 1998), or the integrated absolute difference, Cao and Lugosi (2005), would be of interest. Alternatively, Kolmogorov-Smirnov or Cramér-von Mises-type tests could be constructed based on the differences between kernel c.d.f. estimators.

References

Ahmad, I. A. (1992), 'Residuals density estimation in nonparametric regression', *Statistics & Probability Letters* **14**(2), 133–139. doi: 10.1016/0167-7152(92)90077-I

Antoine, B., Bonnal, H. and Renault, E. (2007), 'On the efficient use of the informational content of estimating equations: Implied probabilities and Euclidean empirical likelihood', *Journal of Econometrics* **138**(2), 461–487. doi: 10.1016/j.jeconom.2006.05.005

Back, K. and Brown, D. P. (1993), 'Implied probabilities in GMM estimators', *Econometrica* **61**(4), 971–975. doi: 10.2307/2951771

Bartlett, M. S. (1963), 'Statistical estimation of density functions', *Sankhyā: The Indian Journal of Statistics, Series A* **25**(3), 245–254. URL: `www.jstor.org/stable/25049271`

Bhattacharya, R. N. and Ghosh, J. K. (1978), 'On the validity of the formal Edgeworth expansion', *The Annals of Statistics* **6**(2), 434–451. doi: 10.1214/aos/1176344134

Bickel, P. J. and Rosenblatt, M. (1973), 'On some global measures of the deviations of density function estimates', *The Annals of Statistics* **1**(6), 1071–1095. doi: 10.1214/aos/1176342558

Bochner, S. (1955), *Harmonic analysis and the theory of probability*, University of California Press.

Bott, A.-K., Devroye, L. and Kohler, M. (2013), 'Estimation of a distribution from data with small measurement errors', *Electronic Journal of Statistics* **7**, 2457–2476. doi: 10.1214/13-EJS850

Brown, B. W. and Newey, W. K. (1998), 'Efficient semiparametric estimation of expectations', *Econometrica* **66**(2), 453–464. doi: 10.2307/2998566

Brown, B. W. and Newey, W. K. (2002), 'Generalized method of moments, efficient bootstrapping, and improved inference', *Journal of Business & Economic Statistics* **20**(4), 507–517. doi: 10.1198/073500102288618649

Brown, S., Greene, W. H., Harris, M. N. and Taylor, K. (2015), 'An inverse hyperbolic sine heteroskedastic latent class panel tobit model: An application to modelling charitable donations', *Economic Modelling* **50**, 228–236. doi: 10.1016/j.econmod.2015.06.018

Burbidge, J. B., Magee, L. and Robb, A. L. (1988), 'Alternative transformations to handle extreme values of the dependent variable', *Journal of the American Statistical Association* **83**(401), 123–127. doi: 10.1080/01621459.1988.10478575

Cao, R. and Lugosi, G. (2005), 'Goodness-of-fit tests based on the kernel density estimator', *Scandinavian Journal of Statistics* **32**(4), 599–616. doi: 10.1111/j.1467-9469.2005.00471.x

Chamberlain, G. (1987), 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Econometrics* **34**(3), 305–334. doi: 10.1016/0304-4076(87)90015-7

Chen, J. and Qin, J. (1993), 'Empirical likelihood estimation for finite populations and the effective usage of auxiliary information', *Biometrika* **80**(1), 107–116. doi: 10.1093/biomet/80.1.107

Chen, S. X. (1997), 'Empirical likelihood-based kernel density estimation', *Australian and New Zealand Journal of Statistics* **39**(1), 47–56. doi: 10.1111/j.1467-842X.1997.tb00522.x

Chen, S. X. and Cui, H. (2007), 'On the second-order properties of empirical likelihood with moment restrictions', *Journal of Econometrics* **141**(2), 492–516. doi: 10.1016/j.jeconom.2006.10.006

Cheng, F. (2004), 'Weak and strong uniform consistency of a kernel error density estimator in nonparametric regression', *Journal of Statistical Planning and Inference* **119**(1), 95–107. doi: 10.1016/S0378-3758(02)00417-2

Cheng, F. (2005), 'Asymptotic distributions of error density estimators in first-order autoregressive models', *Sankhyā: The Indian Journal of Statistics* **67**(3), 553–567. URL: `http://www.jstor.org/stable/25053449`

Chernozhukov, V., Fernández-Val, I. and Galichon, A. (2009), 'Improving point and interval estimators of monotone functions by rearrangement', *Biometrika* **96**(3), 559–575. doi: 10.1093/biomet/asp030

Corcoran, S. A. (1998), 'Bartlett adjustment of empirical discrepancy statistics', *Biometrika* **85**(4), 967–972. doi: 10.1093/biomet/85.4.967

Cox, D. R. and Snell, E. J. (1968), 'A general definition of residuals', *Journal of the Royal Statistical Society. Series B* **30**(2), 248–275. URL: `www.jstor.org/stable/2984505`

Cressie, N. and Read, T. R. C. (1984), 'Multinomial goodness-of-fit tests', *Journal of the Royal Statistical Society. Series B* **46**(3), 440–464. URL: `http://www.jstor.org/stable/2345686`

Fan, Y. (1994), 'Testing the goodness of fit of a parametric density function by kernel method', *Econometric Theory* **10**(2), 316–356. doi: 10.1017/S0266466600008434

Fan, Y. (1998), 'Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters', *Econometric Theory* **14**(5), 604–621. doi: 10.1017/s0266466698145036

Glad, I. K., Hjort, N. L. and Ushakov, N. G. (2003), 'Correction of density estimators that are not densities', *Scandinavian Journal of Statistics* **30**(2), 415–427. doi: 10.1111/1467-9469.00339

Györfi, L. and Walk, H. (2012), 'Strongly consistent density estimation of the regression residual', *Statistics & Probability Letters* **82**(11), 1923–1929. doi: 10.1016/j.spl.2012.06.021

Hall, A. R. (2005), *Generalized method of moments*, Oxford University Press.

Hansen, L. P. (1993), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**(1), 1029–1054. doi: 10.1093/1912775

Hansen, L. P., Heaton, J. and Yaron, A. (1996), 'Finite-sample properties of some alternative GMM estimators', *Journal of Business & Economic Statistics* **14**(3), 262–280. doi: 10.2307/1392442

Imbens, G. W., Spady, R. H. and Johnson, P. (1998harvardyearright , 'Information theoretic approaches to inference in moment condition models', *Econometrica* **66**(2), 333–357. doi: 10.2307/2998561

Jensen, J. L. (1989), 'Validity of the formal Edgeworth expansion when the underlying distribution is partly discrete', *Probability Theory and Related Fields* **81**(4), 507–519. doi: 10.1007/BF00367300

Johnson, N. L. (1949), 'Systems of frequency curves generated by methods of translation', *Biometrika* **36**(1–2), 149–176. doi: 10.1093/biomet/36.1-2.149

Kitamura, Y. and Stutzer, M. (1997), 'An information-theoretic alternative to generalized method of moments estimation', *Econometrica* **65**(4), 861–874. doi: 10.2307/2171942

Kiwitt, S., Nagel, E. and Neumeyer, N. (2008), 'Empirical likelihood estimators for the error distribution in nonparametric regression models', *Mathematical Methods of Statistics* **17**(3), 241–260. doi: 10.3103/S1066530708030058

Kundhi, G. and Rilstone, P. (2012), 'Edgeworth expansions for GEL estimators', *Journal of Multivariate Analysis* **106**, 118–146. doi: 10.1016/j.jmva.2011.11.005

Loynes, R. M. (1969), 'On Cox and Snell's general definition of residuals', *Journal of the Royal Statistical Society. Series B* **31**(1), 103–106. URL: `www.jstor.org/stable/2984331`

MacKinnon, J. G. and Magee, L. (1990), 'Transforming the dependent variable in regression models', *International Economic Review* **31**(2), 315–339. doi: 10.2307/2526842

Marron, J. S. and Wand, M. P. (1992), 'Exact mean integrated squared error', *The Annals of Statistics* **20**(2), 712–736. doi: 10.1214/aos/1176348653

Mátyás, L., ed. (1999), *Generalized method of moments estimation*, Cambridge University Press.

Muhsal, B. and Neumeyer, N. (2010), 'A note on residual-based empirical likelihood kernel density estimation', *Electronic Journal of Statistics* **4**, 1386–1401. doi: 10.1214/10-EJS586

Nadaraya, E. A. (1964), 'Some new estimates for distribution functions', *Theory of Probability and its Applications* **9**(3), 497–500. doi: 10.1137/1109069

Newey, W. K. and Smith, R. J. (2004), 'Higher order properties of GMM and generalized empirical likelihood estimators', *Econometrica* **72**(1), 219–255. doi: 10.1111/j.1468-0262.2004.00482.x

Oryshchenko, V. (2017), 'Exact mean integrated squared error and bandwidth selection for kernel distribution function estimators', Working paper, University of Manchester. URL: `https://arxiv.org/abs/1606.06993`

Owen, A. (1988), 'Empirical likelihood ratio confidence intervals for a single functional', *Biometrika* **75**(2), 237–249. doi: 10.1093/biomet/75.2.237

Owen, A. (1990), 'Empirical likelihood ratio confidence regions', *The Annals of Statistics* **18**(1), 90–120. doi: 10.1214/aos/1176347494

Pagan, A. and Ullah, A. (1999), *Nonparametric econometrics*, Cambridge University Press.

Parente, P. M. and Smith, R. J. (2014), 'Recent developments in empirical likelihood and related methods', *Annual Review of Economics* **6**, 77–102. doi: 10.1146/annurev-economics-080511-110925

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* **33**(3), 1065–1076. doi: 10.1214/aoms/1177704472

Qin, J. and Lawless, J. (1994), 'Empirical likelihood and general estimating equations', *The Annals of Statistics* **22**(1), 300–325. doi: 10.1214/aos/1176325370

Ramirez, O. A., Moss, C. B. and Boggess, W. G. (1994), 'Estimation and use of the inverse hyperbolic sine transformation to model non-normal correlated random variables', *Journal of Applied Statistics* **21**(4), 289–304. doi: 10.1080/757583872

Rao, B. L. S. P. (1983), *Nonparametric functional estimation*, Academic Press.

Robinson, P. M. (1991), 'Best nonlinear three-stage least squares estimation of certain econometric models', *Econometrica* **59**(3), 755–786. doi: 10.2307/2938227

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function', *The Annals of Mathematical Statistics* **27**(3), 832–837. doi: 10.1214/aoms/1177728190

Schennach, S. M. (2007), 'Point estimation with exponentially tilted empirical likelihood', *The Annals of Statistics* **35**(2), 634–672. doi: 10.1214/009053606000001208

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall.

Smith, R. J. (1997), 'Alternative semi-parametric likelihood approaches to generalised method of moments estimation', *The Economic Journal* **107**(441), 503–519. doi: 10.1111/j.0013-0133.1997.174.x

Smith, R. J. (2011), 'GEL criteria for moment condition models', *Econometric Theory* **27**(6), 1192–1235. doi: 10.1017/S026646661100003X

Stacy, E. W. (1962), 'A generalization of the gamma distribution', *The Annals of Mathematical Statistics* **33**(3), 1187–1192. doi: 10.1214/aoms/1177704481

Tsai, A. C., Liou, M., Simak, M. and Cheng, P. E. (2017), 'On hyperbolic transformations to normality', *Computational Statistics & Data Analysis* **115**, 250–266. doi: 10.1016/j.csda.2017.06.001

Tsybakov, A. B. (2009), *Introduction to Nonparametric Estimation*, (Springer Series in Statistics), Springer. doi: 10.1007/b13794

Van Ryzin, J. (1969), 'On strong consistency of density estimates', *The Annals of Mathematical Statistics* **40**(5), 1765–1772. doi: 10.1214/aoms/1177697388

Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, Chapman & Hall.

Wand, M. P. and Schucany, W. R. (1990), 'Gaussian-based kernels', *Canadian Journal of Statistics* **18**(3), 197–204. doi: 10.2307/3315450

Watson, G. S. and Leadbetter, M. R. (1964), 'Hazard analysis II', *Sankhyā: The Indian Journal of Statistics, Series A* **26**(1), 101–116. URL: `http://www.jstor.org/stable/25049316`

Yamato, H. (1973), 'Uniform convergence of an estimator of a distribution function', *Bulletin of Mathematical Statistics* **15**(3-4), 69–78. URL: `http://ci.nii.ac.jp/naid/120001036895/`

Yuan, A., Xu, J. and Zheng, G. (2014), 'On empirical likelihood statistical functions', *Journal of Econometrics* **178**(3), 613–623. doi: 10.1016/j.jeconom.2013.08.037

Zhang, B. (1995), 'M-estimation and quantile estimation in the presence of auxiliary information', *Journal of Statistical Planning and Inference* **44**(1), 77–94. doi: 10.1016/0378-3758(94)00040-3

Zhang, B. (1998), 'A note on kernel density estimation with auxiliary information', *Communications in Statistics—Theory and Methods* **27**(1), 1–11. doi: 10.1080/03610929808832647

Zygmund, A. (2003), *Trigonometric series*, 3rd edn, Cambridge University Press.

Supplement P: Proofs for 'Improved Density and Distribution Function Estimation'

Vitaliy Oryshchenko
University of Manchester
vitaliy.oryshchenko@manchester.ac.uk

Richard J Smith
cemmap, U.C.L and I.F.S.
University of Cambridge
University of Melbourne
ONS Economic Statistics Centre of Excellence
rjs27@econ.cam.ac.uk

This Draft: July 2018

Throughout the Appendix, $0 < C < \infty$ and $0 \leq \omega \leq 1$ will denote generic constants that may be different in different uses. CS, T, and H refer to the Cauchy-Schwarz, triangle, and Hölder inequalities, respectively with LIE and WLLN the law of iterated expectations and Khintchine's i.i.d. weak law of large numbers.

In addition, int$(\cdot)$ denotes the interior of $\cdot$, w.p.(a.)1 with probability (approaching) 1, and $\mathcal{N}$ is an open neighbourhood of β_0.

P.1 GEL Stochastic Expansions

The following identification and regularity conditions are imposed.

Assumption P.1. ***(a)*** *$\beta_0 \in \mathcal{B}$ is the unique solution to* $\mathrm{E}[g(z,\beta)] = 0$*;* ***(b)*** *$\mathcal{B}$ is compact;* ***(c)*** *$g(z,\beta)$ is continuous at each $\beta \in \mathcal{B}$ w.p.1;* ***(d)*** $\mathrm{E}[\sup_{\beta\in\mathcal{B}}\|g(z,\beta)\|^2]| < \infty$*;* ***(e)*** *$\Omega$ is nonsingular;* ***(f)*** *$\rho(v)$ is twice continuously differentiable in a neighbourhood of zero.*

Assumption P.1 is Newey and Smith (2004, Assumption 1) and is sufficient for the consistency of $\hat{\beta}$ for β_0. Moreover, $\hat{\lambda} = \arg\max_{\lambda\in\Lambda_n(\hat{\beta})} P_n^{\rho}(\hat{\beta},\lambda)$ exists w.p.a.1 and $\hat{\lambda} = O_p(n^{-1/2})$; see Newey and Smith (2004, Theorem 3.1).

Let $\nabla g(z,\beta)$ denote the vector of first order partial derivatives of $g(z,\beta)$ with respect to β.

Assumption P.2. ***(a)*** *$\beta_0 \in int(\mathcal{B})$;* ***(b)*** *$g(z,\beta)$ is continuously differentiable for $\beta \in \mathcal{N}$ and* $\mathrm{E}[\sup_{\beta\in\mathcal{N}}\|\nabla g(z,\beta)\|] < \infty$*;* ***(c)*** $\mathrm{rank}(G) = d_\beta$.

Assumption P.2 is Newey and Smith (2004, Assumption 2). If Assumptions P.1 and P.2 hold then $n^{1/2}((\hat{\beta}-\beta_0)^\top,\ \hat{\lambda})^\top \xrightarrow{d} N(0, \mathrm{diag}(\Sigma, P))$; see Newey and Smith (2004, Theorem 3.2).

Let $\nabla^2 g(z,\beta)$ denote a vector of all distinct second order partial derivatives of $g(z,\beta)$ with respect to β.

Assumption P.3. ***(a)*** $\mathrm{E}[\|g(z,\beta_0)\|^6] < \infty$; ***(b)*** *$g(z,\beta)$ is twice differentiable for $\beta \in \mathcal{N}$,* $\mathrm{E}[\|\nabla g(z,\beta_0)\|^4] < \infty$, $\mathrm{E}[\|\nabla^2 g(z,\beta_0)\|^2] < \infty$; ***(c)*** *there exists $d(z) \geq 0$ with $\mathrm{E}[d(z)^2] < \infty$ such that $\|\nabla^2 g(z,\beta) - \nabla^2 g(z,\beta_0)\| \leq d(z)\|\beta - \beta_0\|$ for all z and $\beta \in \mathcal{N}$;* ***(d)*** *$\rho(v)$ is four times differentiable with Lipschitz fourth derivative in a neighbourhood of zero.*

Cf. Newey and Smith (2004, Assumption 3).

Write $\tilde{g} = n^{-1}\sum_{i=1}^n g_i$, $\tilde{G} = n^{-1}\sum_{i=1}^n G_i - G$, and $\tilde{\Omega} = n^{-1}\sum_{i=1}^n g_i g_i^\top - \Omega$. Also let $g_i^j = \partial g(z_i,\beta_0)/\partial\beta_j$ and $G_i^j = \partial^2 g(z_i,\beta_0)/\partial\beta_j\partial\beta^\top$, $j = 1,\ldots,d_\beta$. From the Proof of Theorem 3.4 in Newey and Smith (2004), GEL estimators satisfy the following stochastic expansion

$$\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\lambda}\end{pmatrix} = -\begin{pmatrix}H\\ P\end{pmatrix}\tilde{g} + \begin{pmatrix}-\Sigma & H\\ H^\top & P\end{pmatrix}\tilde{\zeta} + O_p(n^{-3/2}), \tag{P.1}$$

where

$$\begin{aligned}\tilde{\zeta} = \Bigg\{ &\begin{pmatrix}0 & \tilde{G}^\top\\ \tilde{G} & \tilde{\Omega}\end{pmatrix} - \frac{1}{2}\sum_{j=1}^{d_\beta}[H\tilde{g}]_j \begin{pmatrix}0 & \mathrm{E}[G_i^j]^\top\\ \mathrm{E}[G_i^j] & \mathrm{E}[g_i^j g_i^\top + g_i g_i^{j\top}]\end{pmatrix}\\ &-\frac{1}{2}\sum_{j=1}^{d_g}[P\tilde{g}]_j \begin{pmatrix}\mathrm{E}[\partial^2 g_{ij}/\partial\beta\partial\beta^\top] & \mathrm{E}[G_i^\top e_j g_i' + g_{ij}G_i^\top]\\ \mathrm{E}[g_i e_j^\top G_i + g_{ij}G_i] & -\rho_3\mathrm{E}[g_{ij}g_i g_i^\top]\end{pmatrix}\Bigg\}\begin{pmatrix}H\\ P\end{pmatrix}\tilde{g}.\end{aligned}$$

Remark P.1. Write $\tilde{\zeta} = (\tilde{\zeta}_\beta^\top, \tilde{\zeta}_\lambda^\top)^\top$ partitioned conformably with β and λ. Then $\mathrm{E}[\tilde{\zeta}_\beta] = 0$ and $\mathrm{E}[\tilde{\zeta}_\lambda] = \zeta_\lambda$ given in eq. (2.2). If β_0 is known, the stochastic expansion for $\tilde{\lambda}$ is identical to that in eq. (P.1) except H is set to zero and Ω^{-1} replaces P, i.e., $\tilde{\lambda} = -\Omega^{-1}\tilde{g} + \Omega^{-1}\tilde{\zeta}_\lambda + O_p(n^{-3/2})$, where $\tilde{\zeta}_\lambda = \tilde{\Omega}\Omega^{-1}\tilde{g} + \rho_3\sum_{j=1}^{d_\beta}[\Omega^{-1}\tilde{g}]_j\mathrm{E}[g_{ij}g_i g_i^\top]\Omega^{-1}\tilde{g}/2$. Thus, in expectation, the first two terms in eq. (2.2) are eliminated and $\mathrm{E}[\tilde{\zeta}_\lambda] = n^{-1}c_\rho\mathrm{E}[g_i g_i^\top\Omega^{-1}g_i]$.

Remark P.2 When β_0 is known, Assumptions P.3(b)(c) can be relaxed to $g(z,\beta)$ is continuously differentiable for $\beta \in \mathcal{N}$, $\mathrm{E}[\sup_{\beta\in\mathcal{N}}\|\nabla g(z,\beta_0)\|] < \infty$, and there exists $d(z) \geq 0$ with $\mathrm{E}[d(z)] < \infty$ such that $\|\nabla g(z,\beta) - \nabla g(z,\beta_0)\| \leq d(z)\|\beta - \beta_0\|$ for all z and $\beta \in \mathcal{N}$. The Lipschitz condition in Assumptions P.3(b)(c)(d) can also be relaxed to α-Hölder for some $0 < \alpha \leq 1$ and changing the remainder terms from $O(n^{-3/2})$ to $O(n^{-1-\alpha/2})$.

Remark P.3 The two-step GMM estimator is defined as $\hat{\beta}_{GMM} = \arg\min_{\beta\in\mathcal{B}}\hat{g}(\beta)^\top\hat{\Omega}(\tilde{\beta})^{-1}\hat{g}(\beta)$ where $\tilde{\beta}$ is a $\sqrt{n}$-consistent preliminary estimator of β_0. If the preliminary estimator $\tilde{\beta}$ is first order efficient, i.e., $\tilde{\beta} - \beta_0 = -H\tilde{g} + O_p(n^{-1})$, then, if Assumptions P.1–P.3 hold, all GMM estimators $\hat{\beta}_{GMM}$ admit the same expansion to order $O_p(n^{-3/2})$; see Newey and Smith (2004, Section 3). Moreover, defining $\hat{\lambda}_{GMM} = -\hat{\Omega}(\tilde{\beta})^{-1}\hat{g}(\hat{\beta}_{GMM})$, the expansion is

$$\begin{pmatrix}\hat{\beta}_{GMM}-\beta_0\\ \hat{\lambda}_{GMM}\end{pmatrix} = -\begin{pmatrix}H\\ P\end{pmatrix}\tilde{g} + \begin{pmatrix}-\Sigma & H\\ H^\top & P\end{pmatrix}\tilde{\zeta}^{GMM} + O_p(n^{-3/2}),$$

where

$$\tilde{\zeta}^{GMM} = \left\{ \begin{pmatrix} 0 & \tilde{G}^\top \\ \tilde{G} & \tilde{\Omega} - \sum_{j=1}^{d_\beta} \mathrm{E}[g_i^j g_i^\top + g_i g_i^{j\top}] e_j^\top H\tilde{g} \end{pmatrix} - \frac{1}{2}\sum_{j=1}^{d_\beta} [H\tilde{g}]_j \begin{pmatrix} 0 & \mathrm{E}[G_i^j]^\top \\ \mathrm{E}[G_i^j] & 0 \end{pmatrix} \right.$$
$$\left. -\frac{1}{2}\sum_{j=1}^{d_g} [P\tilde{g}]_j \begin{pmatrix} \mathrm{E}[\partial^2 g_{ij}/\partial\beta\partial\beta^\top] & 0 \\ 0 & 0 \end{pmatrix} \right\} \begin{pmatrix} H \\ P \end{pmatrix} \tilde{g}.$$

Writing $\tilde{\zeta}^{GMM} = (\tilde{\zeta}_\beta^{GMM\top}, \tilde{\zeta}_\lambda^{GMM\top})^\top$ partitioned conformably with β and λ, $\zeta_\beta^{GMM} = \mathrm{E}[\tilde{\zeta}_\beta^{GMM}] = \mathrm{E}[G_i^\top P g_i]$ and $\zeta_\lambda^{GMM} = \mathrm{E}[\tilde{\zeta}_\lambda] = -a + \mathrm{E}[G_i H g_i] + \mathrm{E}[g_i g_i^\top P g_i]$. Hence, the second order bias of $\hat{\beta}_{GMM}$, Newey and Smith (2004, Theorem 4.2), is given by

$$\mathrm{E}[\hat{\beta}_{GMM}] - \beta_0 = -n^{-1}\Sigma\zeta_\beta^{GMM} + n^{-1}H\zeta_\lambda^{GMM} + O(n^{-3/2}),$$

the notable difference with GEL being the additional term $-n^{-1}\Sigma\zeta_\beta^{GMM}$ with the term $n^{-1}H\zeta_\lambda^{GMM}$ identical to that of CUE.

P.2 Preliminary Lemmas

Lemma P.1. *If Assumptions P.1–P.3 are satisfied, then*

$$n\hat{\pi}_i = 1 - g_i^\top P\tilde{g} - \frac{1}{2}\rho_3(g_i^\top P\tilde{g})^2 + g_i^\top (H^\top, P)\tilde{\zeta} + \tilde{g}^\top P G_i H\tilde{g} + c_\rho \tilde{g}^\top P\tilde{g} + o_p(n^{-1}) \tag{P.2}$$

uniformly $i = 1, \ldots, n$.

Proof: Let $\hat{v}_i = \hat{\lambda}^\top g(z_i, \hat{\beta})$. A third order Taylor expansion of $\rho_1(\hat{v}_i)$ around 0 yields

$$\rho_1(\hat{v}_i) = -1 - \hat{v}_i + \frac{1}{2}\rho_3\hat{v}_i^2 + \frac{1}{6}\rho_4\hat{v}_i^3(1 + o_p(1))$$

noting $|\hat{v}_i| \xrightarrow{p} 0$ uniformly $i = 1, \ldots, n$ by Newey and Smith (2004, Lemma A1). A Taylor expansion from eq. (P.1) of $g(z_i, \hat{\beta})$ about β_0 yields $g(z_i, \hat{\beta}) = g_i - G_i H\tilde{g} + o_p(n^{-1/2})$ uniformly $i = 1, \ldots, n$ by Owen (1990, Lemma 3). Hence, substituting, using eq. (P.1),

$$\rho_1(\hat{v}_i) = -1 + g_i^\top P\tilde{g} - g_i^\top (H^\top, P)\tilde{\zeta} - \tilde{g}^\top H^\top G_i^\top P\tilde{g} + \frac{1}{2}\rho_3(g_i^\top P\tilde{g})^2 + o_p(n^{-1}).$$

From a similar expansion, using $n^{-1}\sum_{i=1}^n g(z_i, \hat{\beta}) = \Omega P\tilde{g} + O_p(n^{-1})$, eq. (P.1), and $P\Omega P = P$,

$$n^{-1}\sum_{j=1}^n \rho_1(\hat{v}_j) = -1 - \hat{\lambda}^\top \Omega P\tilde{g} + \frac{1}{2}\rho_3\hat{\lambda}^\top\Omega\hat{\lambda} + O_p(n^{-3/2}) = -1 + c_\rho\tilde{g}^\top P\tilde{g} + O_p(n^{-3/2}).$$

Hence, $[n^{-1}\sum_{j=1}^n \rho_1(\hat{v}_j)]^{-1} = -1 - c_\rho\tilde{g}^\top P\tilde{g} + O_p(n^{-3/2})]$ and

$$n\hat{\pi}_i = 1 - g_i^\top P\tilde{g} + g_i^\top (H^\top, P)\tilde{\zeta} + \tilde{g}^\top H^\top G_i^\top P\tilde{g} - \frac{1}{2}\rho_3(g_i^\top P\tilde{g})^2 + c_\rho\tilde{g}^\top P\tilde{g} + o_p(n^{-1})$$

uniformly $i = 1, \ldots, n$. ■

Corollary P.1 (Known β_0)**.** *If Assumptions P.1–P.3 are satisfied, then*

$$n\tilde{\pi}_i = 1 - g_i^\top \Omega^{-1}\tilde{g} - \frac{1}{2}\rho_3 (g_i^\top \Omega^{-1}\tilde{g})^2 + g_i^\top \Omega^{-1}\tilde{\zeta}_\lambda + c_\rho \tilde{g}^\top \Omega^{-1}\tilde{g} + o_p(n^{-1}) \tag{P.3}$$

uniformly $i = 1, \ldots, n$.

Let $a(z)$ denote a real scalar function of z such that $\mathrm{E}[a(z)^2] < \infty$. Write $a_i = a(z_i)$, $i = 1, \ldots, n$.

Lemma P.2. *If Assumptions P.1–P.3 are satisfied, then*

$$\mathrm{E}[(n\hat{\pi}_i - 1)a_i] = n^{-1}(-c_\rho \mathrm{E}[a_i g_i^\top P g_i] + \mathrm{E}[a_i g_i^\top] P\zeta_\lambda + c_\rho(d_g - d_\beta)\mathrm{E}[a_i]) + o(n^{-1}) \tag{P.4}$$

uniformly $i = 1, \ldots, n$. *For* $i \neq j$,

$$\mathrm{E}[(n\hat{\pi}_i - 1)a_i a_j] = \mathrm{E}[(n\hat{\pi}_i - 1)a_i]\mathrm{E}[a_j] - n^{-1}\mathrm{E}[a_i g_i^\top] P \mathrm{E}[g_j a_j] + O(n^{-2}), \tag{P.5}$$

$$\mathrm{E}[(n\hat{\pi}_i - 1)(n\hat{\pi}_j - 1)a_i a_j] = n^{-1}\mathrm{E}[a_i g_i^\top] P \mathrm{E}[g_j a_j] + O(n^{-2}). \tag{P.6}$$

Let $\bar{a} = n^{-1}\sum_{i=1}^n a_i$ *and* $\hat{a} = \sum_{i=1}^n \hat{\pi}_i a_i$. *Then,*

$$\mathrm{Var}[\hat{a}] = \mathrm{Var}[\bar{a}] - n^{-1}\mathrm{E}[a_i g_i^\top] P \mathrm{E}[g_j a_j] + O(n^{-2}). \tag{P.7}$$

Proof: The first result follows from the expansion for $\hat{\pi}_i$ in Lemma P.1. In particular, noting $\mathrm{E}[g_i] = 0$ and $\mathrm{E}[a_i o_p(n^{-1})] = o(n^{-1})$ by uniformity of $o_p(n^{-1})$, then, by independence,

$$\begin{aligned}
\mathrm{E}[(n\hat{\pi}_i - 1)a_i] &= -n^{-1}\mathrm{E}[a_i g_i^\top P g_i] - \frac{1}{2}\rho_3 n^{-1}\mathrm{E}[a_i g_i^\top P \mathrm{E}[g_j g_j^\top] P g_i] + \mathrm{E}[a_i g_i^\top](H^\top, P)\mathrm{E}[\tilde{\zeta}] \\
&\quad + n^{-1}\,\mathrm{tr}(\mathrm{E}[a_i G_i] H \mathrm{E}[g_j g_j^\top] P) + c_\rho n^{-1}\mathrm{E}[a_i]\,\mathrm{tr}(\mathrm{E}[g_j g_j^\top] P) + o(n^{-1}) \\
&= -n^{-1}c_\rho \mathrm{E}[a_i g_i^\top P g_i] + n^{-1}\mathrm{E}[a_i g_i^\top] P\zeta_\lambda + n^{-1}c_\rho(d_g - d_\beta)\mathrm{E}[a_i] + o(n^{-1})
\end{aligned}$$

uniformly $i = 1, \ldots, n$, using $\mathrm{E}[\tilde{\zeta}] = (0^\top, n^{-1}\zeta_\lambda^\top)^\top$, $P\Omega P = P$, $H\Omega P = 0$, and $\mathrm{tr}(\Omega P) = d_g - d_\beta$. Eqs. (P.5) and (P.6) follow by a similar argument.

Finally note that $\hat{a} - \bar{a} = n^{-1}\sum_{i=1}^n (n\hat{\pi}_i - 1)a_i$. Hence, $\mathrm{Var}[\hat{a}] = \mathrm{Var}[\bar{a}] + \mathrm{Var}[\hat{a} - \bar{a}] + 2\,\mathrm{Cov}[\hat{a} - \bar{a}, \bar{a}]$. Now, from above, $\mathrm{E}[\hat{a} - \bar{a}] = O(n^{-1})$. Hence,

$$\mathrm{Var}[\hat{a} - \bar{a}] = \mathrm{E}_{i\neq j}[(n\hat{\pi}_i - 1)(n\hat{\pi}_j - 1)a_i a_j] + O(n^{-2}) = n^{-1}\mathrm{E}[a_i g_i^\top] P \mathrm{E}[g_j a_j] + O(n^{-2}).$$

Also,

$$\begin{aligned}
\mathrm{Cov}[\hat{a} - \bar{a}, \bar{a}] &= n^{-1}\mathrm{E}[(n\hat{\pi}_i - 1)a_i^2] + (1 - n^{-1})\mathrm{E}_{i\neq j}[(n\hat{\pi}_i - 1)a_i a_j] - \mathrm{E}[(n\hat{\pi}_i - 1)a_i)]\mathrm{E}[a_j] \\
&= -n^{-1}\mathrm{E}[a_i g_i^\top] P \mathrm{E}[g_j a_j] + O(n^{-2}).
\end{aligned}$$

■

Corollary P.2 (Known β_0)**.** *If Assumptions P.1–P.3 are satisfied, then*

$$\mathrm{E}[(n\tilde{\pi}_i - 1)a_i] = n^{-1}c_\rho(-\mathrm{E}[g_i^\top \Omega^{-1} g_i a_i] + \mathrm{E}[g_i^\top \Omega^{-1} g_i g_i^\top]\Omega^{-1}\mathrm{E}[g_i a_i] + d_g \mathrm{E}[a_i]) + o(n^{-1}) \tag{P.8}$$

uniformly $i = 1, \ldots, n$. *Lemma P.2 remains valid with* Ω^{-1} *replacing* P.

Repeated use is made of the following lemma; see Bochner (1955, Theorem 1.1.1) and Parzen (1962, Theorem 1A). See also Pagan and Ullah (1999, App.A.2.6).

Lemma P.3. *Suppose that* $f : \mathbb{R} \mapsto \mathbb{R}$ *and* $k : \mathbb{R} \mapsto \mathbb{R}$ *are Borel functions satisfying* ***(a)*** $\int_{-\infty}^{\infty} |f(x)|\mathrm{d}x < \infty$ *and* ***(b)*** $\sup_{-\infty<x<\infty} |k(x)| < \infty$, $\int_{-\infty}^{\infty} |k(x)|\mathrm{d}x < \infty$, *and* $\lim_{|x|\to\infty} |xk(x)| = 0$. *Then* $\int_{-\infty}^{\infty} b^{-1}|k((y-x)/b)||f(x)|\mathrm{d}x < \infty$ *a.e. and*

$$\lim_{b\downarrow 0} \left| \int_{-\infty}^{\infty} b^{-1}k((y-x)/b)f(x)\mathrm{d}x - f(y)\int_{-\infty}^{\infty} k(t)\mathrm{d}t \right| = 0 \tag{P.9}$$

at every continuity point y *of* f*; if* f *is uniformly continuous, then convergence is uniform. Under the same conditions* $\lim_{b\downarrow 0} |\int_{-\infty}^{\infty} b^{-1}k((y-x)/b)^r f(x)\mathrm{d}x - f(y)\int_{-\infty}^{\infty} k(t)^r\mathrm{d}t| = 0$ *at every continuity point* y *of* f *for any* $r \geq 1$. *If* $\sup_{-\infty<x<\infty} |f(x)| < \infty$, $\int_{-\infty}^{\infty} |k(x)|\mathrm{d}x < \infty$ *is sufficient for* (P.9).

REMARK P.4 If k is Hölder continuous with exponent $0 < \tau \leq 1$ and, thus, uniformly continuous, and absolutely integrable, then it is bounded.

P.3 Proofs of Theorems

PROOF OF THEOREM 3.1: Write $\tilde{f}_\rho(u) = \tilde{f}(u) + n^{-1}\sum_{i=1}^n (n\tilde{\pi}_i - 1)k_b(u-u_i)$. By Corollary P.1 and Owen (1990, Lemma 3), $\max_{1\leq i\leq n} |n\tilde{\pi}_i - 1| = o_p(1)$. By Lemma P.3, $\mathrm{E}[|k_b(u-u_i)|] < \infty$ whenever $|f(u)| < \infty$ which holds a.e. Thus, $k_b(u-u_i)$, $i = 1,\ldots,n$, satisfies the conditions for WLLN. Hence, the first conclusion follows.

From Assumption 3.2(a)(i), $b\mathrm{E}[k_b(u-u_i)^2] < \infty$ a.e. By CS, invoking Assumptions P.1(e) and P.3(a), $\mathrm{E}[|g_i b^{1/2}k_b(u-u_i)|] < \infty$ and $\mathrm{E}[|g_i^\top \Omega^{-1} g_i b^{1/2}k_b(u-u_i)|] < \infty$. Hence, by Corollary P.2, setting $a_i = b^{1/2}k_b(u-u_i)$,

$$\begin{aligned}\mathrm{E}[(n\tilde{\pi}_i - 1)k_b(u-u_i)] = n^{-1}c_\rho(&-\mathrm{E}[g_i^\top\Omega^{-1}g_i k_b(u-u_i)] + \mathrm{E}[g_i^\top\Omega^{-1}g_i g_i^\top]\Omega^{-1}\mathrm{E}[g_i k_b(u-u_i)] \\ &+ d_g\mathrm{E}[k_b(u-u_i)]) + o(n^{-1}).\end{aligned}$$

Under Assumption 3.2(a)(i), $\mathrm{E}[k_b(u-u_i)] = f(u) + o(1)$. Invoking Assumption 3.1 and the change of variables $z \mapsto (u, v^\top)^\top$, then, by LIE and Lemma P.3, $\mathrm{E}[g_i k_b(u-u_i)] = \int \mathrm{E}[g_i|t]f(t)k_b(u-t)\mathrm{d}t = \mathrm{E}[g_i|u]f(u) + o(1)$. Similarly, $\mathrm{E}[g_i^\top\Omega^{-1}g_i k_b(u-u_i)] = \mathrm{E}[g_i^\top\Omega^{-1}g_i|u]f(u) + o(1)$. The final result is a direct consequence of Lemma P.2 and the same argument. ■

Set

$$\hat{\delta}_1(u) = n^{-1}\sum_{i=1}^n [k_b(u-\hat{u}_i) - k_b(u-u_i)]; \tag{P.10}$$

$$\hat{\delta}_2(u) = n^{-1}\sum_{i=1}^n (n\hat{\pi}_i - 1)[k_b(u-\hat{u}_i) - k_b(u-u_i)]; \tag{P.11}$$

$$\hat{\delta}_3(u) = n^{-1}\sum_{i=1}^n (n\hat{\pi}_i - 1)k_b(u-u_i). \tag{P.12}$$

Note $\hat{f}(u) = \tilde{f}(u) + \hat{\delta}_1(u)$ and $\hat{f}_\rho(u) = \hat{f}(u) + \hat{\delta}_2(u) + \hat{\delta}_3(u)$.

PROOF OF THEOREM 3.2. Under Assumptions P.1 and P.2, $\hat{\beta} \in \mathcal{N}$ w.p.a.1 and $n^{1/2}(\hat{\beta} - \beta_0) = O_p(1)$. First, by Assumption 3.3(a)(b), from eq. (P.10),

$$|\hat{\delta}_1(u)| \leq n^{-1} \sum_{i=1}^{n} |k_b(u - \hat{u}_i) - k_b(u - u_i)| \leq \frac{C}{nb^{1+\tau}} \sum_{i=1}^{n} |\hat{u}_i - u_i|^\tau$$
$$\leq \frac{C}{n^{\alpha\tau/2} b^{1+\tau}} \|n^{1/2}(\hat{\beta} - \beta_0)\|^{\alpha\tau} n^{-1} \sum_{i=1}^{n} d(z_i)^\tau = o_p(1)$$

since $n^{-1} \sum_{i=1}^{n} d(z_i)^\tau = O_p(1)$ by WLLN and $n^{\alpha\tau/2} b^{1+\tau} \to \infty$ from Assumption 3.3(c). Next, $\max_{1 \leq i \leq n} |n\hat{\pi}_i - 1| = o_p(1)$ by Lemma P.1 and Owen (1990, Lemma 3), from eq. (P.11),

$$|\hat{\delta}_2(u)| \leq n^{-1} \sum_{i=1}^{n} |(n\hat{\pi}_i - 1)[k_b(u - \hat{u}_i) - k_b(u - u_i)]|$$
$$\leq \frac{C}{n^{\alpha\tau/2} b^{1+\tau}} \|n^{1/2}(\hat{\beta} - \beta_0)\|^{\alpha\tau} (\max_{1 \leq i \leq n} |n\hat{\pi}_i - 1|) n^{-1} \sum_{i=1}^{n} d(z_i)^\tau = o_p(1).$$

Hence, the first conclusion follows. The final result follows from eq. (P.12) by noting that also

$$|\hat{\delta}_3(u)| \leq n^{-1} \sum_{i=1}^{n} |(n\hat{\pi}_i - 1) k_b(u - u_i)|$$
$$\leq (\max_{1 \leq i \leq n} |n\hat{\pi}_i - 1|) n^{-1} \sum_{i=1}^{n} |k_b(u - u_i)| = o_p(1)$$

by WLLN since $\mathrm{E}[|k_b(u - u_i)|] < \infty$ a.e. by Lemma P.3. ■

PROOF OF THEOREM 3.3: **Preliminaries.** From a second order Taylor expansion around β_0,

$$k_b(u - \hat{u}_i) = k_b(u - u_i) - k_b^{(1)}(u - u_i) \nabla^\top u_i (\hat{\beta} - \beta_0)$$
$$+ \frac{1}{2} (\hat{\beta} - \beta_0)^\top [k_b^{(2)}(u - \bar{u}_i) \nabla \bar{u}_i \nabla^\top \bar{u}_i - k_b^{(1)}(u - \bar{u}_i) \nabla^2 \bar{u}_i] (\hat{\beta} - \beta_0),$$

where $k_b^{(j)}(x) = k^{(j)}(x/b)/b^{j+1}$, $j = 1, 2$, and $\bar{u}_i = u(z_i, \bar{\beta})$, $i = 1, \ldots, n$, with $\bar{\beta}$ on the line segment joining $\hat{\beta}$ and β_0 with $\nabla \bar{u}_i$ and $\nabla^2 \bar{u}_i$, $i = 1, \ldots, n$, defined analogously. Note that $\|\bar{\beta} - \beta_0\| \leq \|\hat{\beta} - \beta_0\| = O_p(n^{-1/2})$. Assumption 3.4(b) and twice differentiability of $u(z, \beta)$ for $\beta \in \mathcal{N}$ implies there exist $d_0(z) \geq 0$ with $\mathrm{E}[d_0(z)^4] < \infty$ and $d_1(z) \geq 0$ with $\mathrm{E}[d_1(z)^4] < \infty$ such that $|u(z, \beta) - u(z, \beta_0)| \leq d_0(z) \|\beta - \beta_0\|$ and $\|\nabla u(z, \beta) - \nabla u(z, \beta_0)\| \leq d_1(z) \|\beta - \beta_0\|$ for all z and $\beta \in \mathcal{N}$. Thus, by T, $\|\nabla \bar{u}_i \nabla^\top \bar{u}_i - \nabla u_i \nabla^\top u_i\| \leq 2 d_1(z_i) \|\nabla u_i\| \|\hat{\beta} - \beta_0\| + d_1(z_i)^2 \|\hat{\beta} - \beta_0\|^2$. By Owen (1990, Lemma 3), $\max_{1 \leq i \leq n} d_1(z_i)^2 = o_p(n^{1/2})$ and $\max_{1 \leq i \leq n} d_1(z_i) \|\nabla u_i\| = o_p(n^{1/2})$. Hence, $\|\nabla \bar{u}_i \nabla^\top \bar{u}_i - \nabla u_i \nabla^\top u_i\| \leq n^{-1/2} [d_1(z_i) \|\nabla u_i\| + o_p(1)] O_p(1)$. By CS, from Assumption 3.4(b), for $0 < \tau \leq 1$, $\mathrm{E}[d_0(z_i)^\tau \|\nabla u_i\|^2] < \infty$ and $\mathrm{E}[d_1(z_i)^2 \|\nabla u_i\|^2] < \infty$. Thus, $\mathrm{E}[|b^{-1/2} k^{(2)}((u - u_i)/b)| d_1(z_i) \|\nabla u_i\|] < \infty$ since $\mathrm{E}[b^{-1} k^{(2)}((u - u_i)/b)^2] < \infty$ also by CS and using Lemma P.3.

Hence, by T, and noting $n^{\tau/2}b^{3+\tau} \to \infty$, $0 < \tau \leq 1$, from Assumption 3.4(c),

$$\begin{aligned}
&\|n^{-1}\sum_{i=1}^{n}(k_b^{(2)}(u-\bar{u}_i)\nabla\bar{u}_i\nabla^\top\bar{u}_i - k_b^{(2)}(u-u_i)\nabla u_i\nabla^\top u_i)\| \\
&\quad\leq \frac{C}{n^{\tau/2}b^{3+\tau}}\|n^{1/2}(\hat{\beta}-\beta_0)\|^\tau n^{-1}\sum_{i=1}^{n} d_0(z_i)^\tau\|\nabla u_i\|^2 \\
&\qquad + n^{-1}\sum_{i=1}^{n}[\frac{1}{n^{1/2}b^{5/2}}|b^{-1/2}k^{(2)}((u-u_i)/b)| + \frac{o_p(1)}{n^{\tau/2}b^{3+\tau}}][d_1(z_i)\|\nabla u_i\| + o_p(1)]\,O_p(1) \\
&\quad= o_p(1).
\end{aligned}$$

Assumption 3.4(a) implies $k^{(1)}$ is Lipschitz, and hence, invoking Assumption 3.4(b), for all mean values $\bar{\beta}$ between $\hat{\beta}$ and β_0, $|k_b^{(1)}(u-\bar{u}_i) - k_b^{(1)}(u-u_i)| \leq b^{-3}Cd_0(z_i)\|\hat{\beta}-\beta_0\|$ w.p.a.1. By Assumption 3.4(a) and Lemma P.3, $\mathrm{E}[b^{-1}|k^{(1)}((u-u_i)/b)|^{4/3}] < \infty$ a.e., and as $\mathrm{E}[d(z_i)^4] < \infty$, $\mathrm{E}[|b^{-3/4}k^{(1)}((u-u_i)/b)|d(z_i)] < \infty$ using H with exponents 4/3 and 4. Therefore, by the same argument as above,

$$\begin{aligned}
&\|n^{-1}\sum_{i=1}^{n}(k_b^{(1)}(u-\bar{u}_i)\nabla^2\bar{u}_i - k_b^{(1)}(u-u_i)\nabla^2 u_i)\| \\
&\quad\leq \frac{C}{n^{1/2}b^3}\|n^{1/2}(\hat{\beta}-\beta_0)\|n^{-1}\sum_{i=1}^{n} d_0(z_i)\|\nabla^2 u_i\| \\
&\qquad + \frac{1}{n^{\alpha/2}b^{5/4}}\|n^{1/2}(\hat{\beta}-\beta_0)\|^\alpha n^{-1}\sum_{i=1}^{n}[|b^{-3/4}k^{(1)}((u-u_i)/b)| \\
&\qquad + \frac{C}{n^{1/2}b^{7/4}}d_0(z_i)\|n^{1/2}(\hat{\beta}-\beta_0)\|]d(z_i) \\
&\quad= o_p(1).
\end{aligned}$$

Using expansion eq. (P.1) and Lemma P.1 eq. (P.2), from eq. (P.10),

$$\begin{aligned}
\hat{\delta}_1(u) &= n^{-1}\sum_{i=1}^{n} k_b^{(1)}(u-u_i)\nabla^\top u_i H\tilde{g} - n^{-1}\sum_{i=1}^{n} k_b^{(1)}(u-u_i)\nabla^\top u_i(-\Sigma, H)\tilde{\zeta} \\
&\quad + \frac{1}{2}\tilde{g}^\top H^\top n^{-1}\sum_{i=1}^{n}[k_b^{(2)}(u-u_i)\nabla u_i\nabla^\top u_i - k_b^{(1)}(u-u_i)\nabla^2 u_i]H\tilde{g} + o_p(n^{-1}),
\end{aligned} \tag{P.13}$$

from eq. (P.11),

$$\hat{\delta}_2(u) = -n^{-1}\sum_{i=1}^{n} k_b^{(1)}(u-u_i)\nabla^\top u_i H\tilde{g}\tilde{g}^\top P g_i + o_p(n^{-1}), \tag{P.14}$$

and, from eq. (P.12),

$$\begin{aligned}
\hat{\delta}_3(u) &= n^{-1}\sum_{i=1}^{n}[-g_i^\top P\tilde{g} - \frac{1}{2}\rho_3(g_i^\top P\tilde{g})^2 + g_i^\top(H^\top, P)\tilde{\zeta} + \tilde{g}^\top P G_i H\tilde{g} \\
&\quad + c_\rho\tilde{g}^\top P\tilde{g}]k_b(u-u_i) + o_p(n^{-1}).
\end{aligned} \tag{P.15}$$

Expectation. Since $H\Omega H^\top = \Sigma$, from eq. (P.13),

$$\begin{aligned}\mathrm{E}[\hat{\delta}_1(u)] &= n^{-1}\mathrm{E}[k_b^{(1)}(u-u_i)\nabla^\top u_i H g_i] - n^{-1}\mathrm{E}[k_b^{(1)}(u-u_i)\nabla^\top u_i]H\zeta_\lambda \\ &\quad + \frac{1}{2}n^{-1}\operatorname{tr}(\Sigma\mathrm{E}[k_b^{(2)}(u-u_i)\nabla u_i\nabla^\top u_i - k_b^{(1)}(u-u_i)\nabla^2 u_i]) + o(n^{-1}).\end{aligned}$$

Assumption 3.4(a) states $\lim_{|x|\to\infty}|x^2k^{(1)}(x)| = 0$ and implies that $\int k^{(1)}(x)\mathrm{d}x = 0$, $\int xk^{(1)}(x)\mathrm{d}x = -1$, and $xk^{(1)}(x)$ satisfies the hypotheses of Lemma P.3, i.e., it is bounded and absolutely integrable. Thus, invoking Assumption 3.4(d), by the mean value theorem and Lemma P.3,

$$\begin{aligned}\mathrm{E}[k_b^{(1)}(u-u_i)\nabla u_i] &= \frac{1}{b}\int \mathrm{E}[\nabla u_i|u-bt]f(u-bt)k^{(1)}(t)\mathrm{d}t \\ &= \frac{1}{b}\mathrm{E}[\nabla u_i|u]f(u)\int k^{(1)}(t)\mathrm{d}t - \int(\mathrm{d}\{\mathrm{E}[\nabla u_i|u-\omega bt]f(u-\omega bt)\}/\mathrm{d}u)tk^{(1)}(t)\mathrm{d}t \\ &= \mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u + o(1). \qquad \text{(P.16)}\end{aligned}$$

Similarly, $\mathrm{E}[k_b^{(1)}(u-u_i)\nabla^\top u_i H g_i] = \mathrm{d}\{\mathrm{E}[\nabla^\top u_i H g_i|u]f(u)\}/\mathrm{d}u + o(1)$ and $\mathrm{E}[k_b^{(1)}(u-u_i)\nabla^2 u_i] = \mathrm{d}\{\mathrm{E}[\nabla^2 u_i|u]f(u)\}/\mathrm{d}u + o(1)$. Furthermore, Assumption 3.4(a) also implies that $\int k^{(2)}(x)\mathrm{d}x = 0$, $\int xk^{(2)}(x)\mathrm{d}x = 0$, $\int x^2k^{(2)}(x)\mathrm{d}x = 2$, and $x^2k^{(2)}(x)$ satisfies the hypotheses of Lemma P.3. Thus, by a second order Taylor expansion and a similar argument to eq. (P.16),

$$\begin{aligned}\mathrm{E}[k_b^{(2)}(u-u_i)\nabla u_i\nabla^\top u_i] &= \frac{1}{b^2}\int \mathrm{E}[\nabla u_i\nabla^\top u_i|u-bt]f(u-bt)k^{(2)}(t)\mathrm{d}t \\ &= \frac{1}{b^2}\mathrm{E}[\nabla u_i\nabla^\top u_i|u]f(u)\int k^{(2)}(t)\mathrm{d}t \\ &\quad - \frac{1}{b}\mathrm{d}\{\mathrm{E}[\nabla u_i\nabla^\top u_i|u]f(u)\}/\mathrm{d}u\int tk^{(2)}(t)\mathrm{d}t \\ &\quad + \frac{1}{2}\int(\mathrm{d}^2\{\mathrm{E}[\nabla u_i\nabla^\top u_i|u-\omega bt]f(u-\omega bt)\}/\mathrm{d}u^2)t^2k^{(2)}(t)\mathrm{d}t \\ &= \mathrm{d}^2\{\mathrm{E}[\nabla u_i\nabla^\top u_i|u]f(u)\}/\mathrm{d}u^2 + o(1).\end{aligned}$$

Since $H\Omega P = 0$, from eq. (P.14), $\mathrm{E}[\hat{\delta}_2(u)] = o(n^{-1})$. By Lemma P.2 eq. (P.4) and the same argument used in the proof of Theorem 3.1, $\mathrm{E}[\hat{\delta}_3(u)] = n^{-1} - c_\rho\mathrm{E}[g_i^\top Pg_i|u] + \mathrm{E}[g_i|u]^\top P\zeta_\lambda + c_\rho(d_g - d_\beta)f(u) + o(n^{-1})$.

Variance. Since $\mathrm{E}[\hat{\delta}_1(u)] = O(n^{-1})$, from eq. (P.13),

$$\begin{aligned}\operatorname{Var}(\hat{\delta}_1(u)) &= n^{-2}\sum_{i=1}^n\sum_{j=1}^n \mathrm{E}[k_b^{(1)}(u-u_i)\nabla^\top u_i H\tilde{g}\tilde{g}^\top H^\top\nabla u_j k_b^{(1)}(u-u_j)] + o(n^{-1}) \\ &= n^{-1}[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top\Sigma[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u] + o(n^{-1}).\end{aligned}$$

Similarly, noting $\mathrm{E}[\hat{\delta}_2(u)] = o(n^{-1})$, from Lemma P.2, it is straightforward to verify that $\operatorname{Var}[\hat{\delta}_2(u)] = o(n^{-1})$. Furthermore, also using Lemma P.2, as $\mathrm{E}[\hat{\delta}_3(u)] = O(n^{-1})$ and $\mathrm{E}[k_b(u-u_i)g_i] = \mathrm{E}[g_i|u]f(u)$, $\operatorname{Var}(\hat{\delta}_3(u)) = n^{-1}\mathrm{E}[g_i|u]^\top P\mathrm{E}[g_i|u]f(u)^2 + o(n^{-1})$. It is straightforward to verify that $\operatorname{Cov}[\hat{\delta}_1, \hat{\delta}_2] =$

$o(n^{-1})$, recalling $H\Omega P = 0$,

$$\begin{aligned}\mathrm{Cov}[\hat{\delta}_1(u)\hat{\delta}_3(u)] &= -n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}[k_b^{(1)}(u-u_i)\nabla^\top u_i H\tilde{g}\tilde{g}^\top P g_j k_b(u-u_j)] + O(n^{-2}) \\ &= O(n^{-2}),\end{aligned}$$

$$\begin{aligned}\mathrm{Cov}[\hat{\delta}_1(u), \tilde{f}(u)] &= n^{-1}\mathrm{E}[k_b^{(1)}(u-u_i)\nabla^\top u_i]H\mathrm{E}[g_j k_b(u-u_j)] + o(n^{-1}) \\ &= n^{-1}[\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top H\mathrm{E}[g_i|u]f(u) + o(n^{-1}),\end{aligned}$$

$\mathrm{Cov}[\hat{\delta}_2(u), \hat{\delta}_3(u)] = o(n^{-1})$, $\mathrm{Cov}[\hat{\delta}_2(u), \tilde{f}(u)] = o(n^{-1})$, noting again $H\Omega P = 0$, and finally,

$$\mathrm{Cov}[\hat{\delta}_3(u), \tilde{f}(u)] = -n^{-1}\mathrm{E}[g_i|u]^\top P\mathrm{E}[g_i|u]f(u)^2 + o(n^{-1}).$$

Combining these results gives eqs. (3.9)–(3.12). ■

Proof of Theorem 4.1: Since $\lim_{x\to-\infty}K(x) = 0$ and $\lim_{x\to\infty}K(x) = 1$, $2\int K(x)k(x)\mathrm{d}x = 1$, and $\int|K(x)k(x)|\mathrm{d}x < \infty$, $\mathrm{E}[K((u-u_i)/b)] = F(u) + \int k(t)[F(u-bt) - F(u)]\mathrm{d}t$ and $\mathrm{E}[K((u-u_i)/b)^2] = F(u) + 2\int K(t)k(t)[F(u-bt) - F(u)]\mathrm{d}t$. F as a c.d.f. is bounded and hence $\mathrm{E}[K((u-u_i)/b)^2] < \infty$, and $\mathrm{E}[K((u-u_i)/b)] = F(u)+o(1)$ and $\mathrm{E}[K((u-u_i)/b)^2] = F(u)+o(1)$ as $b \to 0$ and at all points of continuity of F. Therefore, cf. the Proof of Theorem 3.1, $|\tilde{F}_\rho(u) - \tilde{F}(u)| = o_p(1)$.

Equation (4.4) follows by Corollary P.2 with $a_i = K((u-u_i)/b)$, $i = 1,\ldots,n$. Assumptions 3.2(a)(i) and $\lim_{|x|\to\infty}|x^2k(x)| = 0$ imply that $xk(x)$ satisfies the conditions of Lemma P.3. Since $\int xk(x) = 0$ and $\mathrm{E}[|\mathrm{E}[g_i|u]|] < \infty$, integration by parts and an application of the mean value theorem give

$$\begin{aligned}\mathrm{E}[g_iK((u-u_i)/b)] &= \int_{-\infty}^{\infty}K((u-s)/b)\,\mathrm{E}[g_i|s]\mathrm{d}F(s) \\ &= [K((u-s)/b)\int_{-\infty}^{s}\mathrm{E}[g_i|t]\mathrm{d}F(t)]_{-\infty}^{\infty} + \int_{-\infty}^{u}\mathrm{E}[g_i|t]\mathrm{d}F(t) \\ &\quad - b\int_{-\infty}^{\infty}(\mathrm{E}[g_i|u-\omega bt]f(u-\omega bt))tk(t)\mathrm{d}t \\ &= \int_{-\infty}^{u}\mathrm{E}[g_i|t]\mathrm{d}F(t) + o(b). \qquad \text{(P.17)}\end{aligned}$$

Similarly, $\mathrm{E}[g_i^\top\Omega^{-1}g_iK((u-u_i)/b)] = \int_{-\infty}^{u}\mathrm{E}[g_i^\top\Omega^{-1}g_i|t]\mathrm{d}F(t)$. Eq. (4.5) follows by Corollary P.2

and eq. (P.17). ■

Set

$$\hat{\Delta}_1(u) = n^{-1}\sum_{i=1}^{n}[K((u-\hat{u}_i)/b) - K((u-u_i)/b)]; \tag{P.18}$$

$$\hat{\Delta}_2(u) = n^{-1}\sum_{i=1}^{n}(n\hat{\pi}_i - 1)[K((u-\hat{u}_i)/b) - K((u-u_i)/b)]; \tag{P.19}$$

$$\hat{\Delta}_3(u) = n^{-1}\sum_{i=1}^{n}(n\hat{\pi}_i - 1)K((u-u_i)/b). \tag{P.20}$$

Note $\tilde{F}(u) = \hat{F}(u) + \hat{\Delta}_1(u)$ and $\hat{F}_\rho(u) = \hat{F}(u) + \hat{\Delta}_2(u) + \hat{\Delta}_3(u)$.

Proof of Theorem 4.2: Since k is bounded, K is Lipschitz continuous and, by the Proof of Theorem 4.1, $\mathrm{E}[|K((u-u_i)/b)|] < \infty$ for all u. Then, as in the Proof of Theorem 3.2, invoking Assumptions P.1–P.3 and 3.3(b), from eqs. (P.18)–(P.20),

$$|\hat{\Delta}_1(u)| \leq \frac{C}{n^{\alpha/2}b}\|n^{1/2}(\hat{\beta}-\beta_0)\|^\alpha n^{-1}\sum_{i=1}^{n} d(z_i) = o_p(1);$$

$$|\hat{\Delta}_2(u)| \leq (\max_{1\leq i\leq n}|n\hat{\pi}_i - 1|)\frac{C}{n^{\alpha/2}b}\|n^{1/2}(\hat{\beta}-\beta_0)\|^\alpha n^{-1}\sum_{i=1}^{n} d(z_i) = o_p(1);$$

$$|\hat{\Delta}_3(u)| \leq (\max_{1\leq i\leq n}|n\hat{\pi}_i - 1|)n^{-1}\sum_{i=1}^{n}|K((u-u_i)/b)| = o_p(1).$$ ■

Proof of Theorem 4.3. **Preliminaries.** From a second order Taylor expansion around β_0,

$$\begin{aligned}K((u-\hat{u}_i)/b) = {} & K((u-u_i)/b) - k_b(u-u_i)\nabla^\top u_i(\hat{\beta}-\beta_0) \\ & + \frac{1}{2}(\hat{\beta}-\beta_0)^\top[k_b^{(1)}(u-\bar{u}_i)\nabla\bar{u}_i\nabla^\top\bar{u}_i - k_b(u-\bar{u}_i)\nabla^2\bar{u}_i](\hat{\beta}-\beta_0),\end{aligned}$$

where $\bar{u}_i = u(z_i,\bar{\beta})$, $i = 1,\ldots,n$, with $\bar{\beta}$ on the line segment joining $\hat{\beta}$ and β_0; $\nabla\bar{u}_i$ and $\nabla^2\bar{u}_i$, $i = 1,\ldots,n$, are defined analogously. By the same argument as in the Proof of Theorem 3.3, noting that Assumption 4.1(a) implies k is Lipschitz and $nb^6 \to \infty$ as $n^{\tau/2}b^{2+\tau} \to \infty$, invoking Assumption 3.4(b),

$$\begin{aligned}&\|n^{-1}\sum_{i=1}^{n}(k_b^{(1)}(u-\bar{u}_i)\nabla\bar{u}_i\nabla^\top\bar{u}_i - k_b^{(1)}(u-u_i)\nabla u_i\nabla^\top u_i)\| \\ &\leq \frac{C}{n^{\tau/2}b^{2+\tau}}\|n^{1/2}(\hat{\beta}-\beta_0)\|^\tau n^{-1}\sum_{i=1}^{n} d_0(z_i)^\tau\|\nabla u_i\|^2 \\ &\quad + n^{-1}\sum_{i=1}^{n}[\frac{1}{n^{1/2}b^{3/2}}|b^{-1/2}k^{(1)}((u-u_i)/b)| + \frac{o_p(1)}{n^{\tau/2}b^{2+\tau}}][d_1(z_i)\|\nabla u_i\| + o_p(1)]\,O_p(1) \\ &= o_p(1)\end{aligned}$$

and

$$\begin{aligned}
&\|n^{-1}\sum_{i=1}^{n}(k_b(u-\bar{u}_i)\nabla^2\bar{u}_i - k_b(u-u_i)\nabla^2 u_i)\| \\
&\quad\le \frac{C}{n^{1/2}b^2}\|n^{1/2}(\hat{\beta}-\beta_0)\|n^{-1}\sum_{i=1}^{n} d_0(z_i)\|\nabla^2 u_i\| \\
&\qquad + \frac{1}{n^{\alpha/2}b^{1/4}}\|n^{1/2}(\hat{\beta}-\beta_0)\|^{\alpha}n^{-1}\sum_{i=1}^{n}[|b^{-3/4}k((u-u_i)/b)| + \frac{C}{n^{1/2}b^{7/4}}d_0(z_i)\|n^{1/2}(\hat{\beta}-\beta_0)\|]d(z_i) \\
&\quad= o_p(1).
\end{aligned}$$

Therefore, using expansion eq. (P.1) and Lemma P.1,

$$\begin{aligned}
\hat{\Delta}_1(u) &= n^{-1}\sum_{i=1}^{n} k_b(u-u_i)\nabla^\top u_i H\tilde{g} - n^{-1}\sum_{i=1}^{n} k_b(u-u_i)\nabla^\top u_i(-\Sigma, H)\tilde{\zeta} \\
&\qquad + \frac{1}{2}\tilde{g}^\top H^\top n^{-1}\sum_{i=1}^{n}[k_b^{(1)}(u-u_i)\nabla u_i\nabla^\top u_i - k_b(u-u_i)\nabla^2 u_i]H\tilde{g} + o_p(n^{-1}), \qquad \text{(P.21)} \\
\hat{\Delta}_2(u) &= -n^{-1}\sum_{i=1}^{n} k_b(u-u_i)\nabla^\top u_i H\tilde{g}\tilde{g}^\top P g_i + O_p(n^{-3/2}), \qquad \text{(P.22)} \\
\hat{\Delta}_3(u) &= n^{-1}\sum_{i=1}^{n}[-g_i^\top P\tilde{g} - \frac{1}{2}\rho_3(g_i^\top P\tilde{g})^2 + g_i^\top(H^\top, P)\tilde{\zeta} + \tilde{g}^\top P G_i H\tilde{g} \\
&\qquad + c_\rho\tilde{g}^\top P\tilde{g}]K((u-u_i)/b) + o_p(n^{-1}). \qquad \text{(P.23)}
\end{aligned}$$

Expectation. Similarly to the Proof of Theorem 3.3, from eq. (P.21),

$$\begin{aligned}
\mathrm{E}[\hat{\Delta}_1(u)] &= n^{-1}\,\mathrm{E}[k_b(u-u_i)\nabla^\top u_i H g_i] - n^{-1}\,\mathrm{E}[k_b(u-u_i)\nabla^\top u_i]H\zeta_\lambda \\
&\qquad + \frac{1}{2}n^{-1}\,\mathrm{tr}(\Sigma\,\mathrm{E}[k_b^{(1)}(u-u_i)\nabla u_i\nabla^\top u_i - k_b(u-u_i)\nabla^2 u_i]) + o(n^{-1}).
\end{aligned}$$

Assumption 4.1(a) implies $k(x)$ satisfies the hypotheses of Lemma P.3. Hence $\mathrm{E}[k_b(u-u_i)\nabla u_i] = \mathrm{E}[\nabla u_i|u]f(u) + o(1)$, $\mathrm{E}[k_b(u-u_i)\nabla^\top u_i H g_i] = \mathrm{E}[\nabla^\top u_i H g_i|u]f(u) + o(1)$, and $\mathrm{E}[k_b(u-u_i)\nabla^2 u_i] = \mathrm{E}[\nabla^2 u_i|u]f(u) + o(1)$. Assumption 4.1(a) also implies $xk^{(1)}(x)$ satisfies the hypotheses of Lemma P.3. Hence, by the mean value theorem as in eq. (P.16), $\mathrm{E}[k_b^{(1)}(u-u_i)\nabla u_i\nabla^\top u_i] = \mathrm{d}\{\mathrm{E}[\nabla u_i\nabla^\top u_i|u]f(u)\}/\mathrm{d}u + o(1)$. Therefore, $\mathrm{E}[\hat{\Delta}_1(u)] = n^{-1}\Delta(u) + o(n^{-1})$ as required.

Likewise, as in the Proof of Theorem 3.3, from eq. (P.22), $\mathrm{E}[\hat{\Delta}_2(u)] = o(n^{-1})$. Finally, by Lemma P.2 and the Proof of Theorem 4.1, $\mathrm{E}[\hat{\Delta}_3(u)] = n^{-1}\Delta_\rho(u) + o(n^{-1})$.

Variance. Using expansions eqs. (P.21)–(P.23) for $\hat{\Delta}_j(u)$, $j = 1, 2, 3$, $\mathrm{Cov}[\hat{\Delta}_1(u), \hat{\Delta}_2(u)]$, $\mathrm{Cov}[\hat{\Delta}_1(u), \hat{\Delta}_3(u)]$,

$\mathrm{Cov}[\hat{\Delta}_2(u), \hat{\Delta}_3(u)]$, $\mathrm{Cov}[\tilde{F}(u), \hat{\Delta}_2(u)]$, and $\mathrm{Var}[\hat{\Delta}_2(u)]$ are all $O(n^{-2})$. Also,

$$\begin{aligned}
\mathrm{Var}[\hat{\Delta}_1(u)] &= n^{-1}\,\mathrm{E}[k_b(u-u_i)\nabla u_i]^\top \Sigma\,\mathrm{E}[k_b(u-u_j)\nabla u_j] + O(n^{-2}),\\
\mathrm{Cov}[\tilde{F}(u), \hat{\Delta}_1(u)] &= n^{-1}\,\mathrm{E}[k_b(u-u_i)\nabla u_i]^\top H\,\mathrm{E}[g_j K((u-u_j)/b)] + o(n^{-3/2}),\\
\mathrm{Var}[\hat{\Delta}_3(u)] &= n^{-1}\,\mathrm{E}[g_i K((u-u_i)/b)]^\top P\,\mathrm{E}[g_i K((u-u_i)/b)] + O(n^{-2}),\\
\mathrm{Cov}[\tilde{F}(u), \hat{\Delta}_3(u)] &= -n^{-1}\,\mathrm{E}[g_i K((u-u_i)/b)]^\top P\,\mathrm{E}[g_i K((u-u_i)/b)] + O(n^{-2}).
\end{aligned}$$

Eqs. (4.10) and (4.11) then follow immediately using eq. (P.17) and $\mathrm{E}[k_b(u-u_i)\nabla u_i] = \mathrm{E}[\nabla u_i|u]f(u) + o(1)$. If $\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u$ is absolutely integrable, then, using Lemma P.3, $\mathrm{E}[k_b(u-u_i)\nabla u_i] = \mathrm{E}[\nabla u_i|u]f(u) - b\int(\mathrm{d}\{\mathrm{E}[\nabla u_i|u-\omega bt]f(u-\omega bt)\}/\mathrm{d}u)tk(t)\mathrm{d}t = \mathrm{E}[\nabla u_i|u]f(u) + o(b)$. ■

References

Bochner, S. (1955), *Harmonic analysis and the theory of probability*, University of California Press.

Newey, W. K. and Smith, R. J. (2004), 'Higher order properties of GMM and generalized empirical likelihood estimators', *Econometrica* **72**(1), 219–255.

Owen, A. (1990), 'Empirical likelihood ratio confidence regions', *The Annals of Statistics* **18**(1), 90–120.

Pagan, A. and Ullah, A. (1999), *Nonparametric econometrics*, Cambridge University Press.

Parzen, E. (1962), 'On estimation of a probability density function and mode', *The Annals of Mathematical Statistics* **33**(3), 1065–1076.

Supplement E: Examples for 'Improved Density and Distribution Function Estimation'

Vitaliy Oryshchenko
University of Manchester
vitaliy.oryshchenko@manchester.ac.uk

Richard J Smith
cemmap, U.C.L and I.F.S.
University of Cambridge
University of Melbourne
ONS Economic Statistics Centre of Excellence
rjs27@econ.cam.ac.uk

This Draft: June 2018

Example E.1 u Not A Function Of β

When $u = u(z)$ is a function of z but not of β, u_i, $i = 1, \ldots, n$, are of course observable. Hence the estimators $\tilde{f}$ eq. (3.1) and $\hat{f}$ eq. (3.7) are identical and the terms $\hat{\delta}_1$ and $\hat{\delta}_2$ in the Proof of Theorem 3.3 are zero. The density estimators $\tilde{f}_\rho$ eq. (3.4) and $\hat{f}_\rho$ eq. (3.8) use different implied probabilities, $\tilde{\pi}_i$ versus $\hat{\pi}_i$, $i = 1, \ldots, n$. Thus, Theorem 3.1 with known β_0 is unchanged whereas, in Theorem 3.3 with estimated β_0, $\mathrm{E}[\hat{f}_\rho(u)] = \mathrm{E}[\tilde{f}(u)] + n^{-1}\delta_\rho(u) + o(n^{-1})$ with $\delta_\rho(u)$ defined in eq. (3.10). Eq. (3.12) also holds with $\tilde{f}$ replacing $\hat{f}$.

Classical examples, e.g., either fully or partially known mean, variance, or third moment of u, are included here. For example, symmetry may be imposed by the moment condition that the third moment around an unknown mean is known to be zero.

The situation in which interest concerns the density of $u(z_1)$, say, but the remaining $d_z - 1$ variates z_2 satisfy moment conditions $\mathrm{E}[g(z_2, \beta_0)] = 0$ is also permitted. Provided $u(z_1)$ and $g(z_2, \beta_0)$ are not independent, (G)EL-based estimators for f will generally enjoy a reduction in variance due to the extra information from the moment condition $\mathrm{E}[g(z_2, \beta_0)] = 0$.

Example E.2 Regression On A Constant

To explain the method behind the Proof of Theorem 3.3 and to provide the background for Example E.3 below, the estimation of the density of the residual u from a regression on a constant is

examined, *viz.*, $y = \beta_0 + u$, with β_0 estimated by the sample average $\hat{\beta} = \bar{y} = n^{-1}\sum_{i=1}^n y_i = \beta_0 + \bar{u}$. The estimated residuals are $\hat{u}_i = y_i - \hat{\beta} = u_i - \bar{u}$, $i = 1, \ldots, n$. If Assumption 3.4(a) holds, $\hat{f}(u) = \tilde{f}(u) + \hat{\delta}_1(u)$, where, for some $0 \leq \omega \leq 1$,

$$\hat{\delta}_1(u) = n^{-1}\sum_{i=1}^n k_b^{(1)}(u-u_i)\bar{u} + \frac{1}{2}n^{-1}\sum_{i=1}^n k_b^{(2)}(u-u_i)\bar{u}^2 + \frac{1}{2}n^{-1}\sum_{i=1}^n [k_b^{(2)}(u-u_i+\omega\bar{u}) - k_b^{(2)}(u-u_i)]\bar{u}^2.$$

By Hölder continuity of $k^{(2)}$, for some $0 < C < \infty$, $|k_b^{(2)}(u - u_i + \omega\ \bar{u}) - k_b^{(2)}(u - u_i)| \leq C|n^{1/2}\bar{u}|^\tau / n^{\tau/2}b^{3+\tau} \to 0$ in probability if $n^{\tau/2}b^{3+\tau} \to \infty$, and in mean square if $\mathrm{E}[u^4] < \infty$. Furthermore, for some $\epsilon > 0$, $n^{(1-\epsilon)/2}\bar{u}^2$ is essentially bounded w.p.1 as $n \to \infty$. To see this, suppose $\mathrm{E}[X_n^2] < \infty$. Then, for any $\epsilon > 0$ and $0 < B < \infty$, by the Chebyshev inequality, $\sum_{n=1}^\infty P(|X_n| \geq n^{(1+\epsilon)/2}B) \leq \mathrm{E}[X_n^2]\sum_{n=1}^\infty n^{-(1+\epsilon)}/B^2 < \infty$. Thus, by the first Borel-Cantelli Lemma, $P(n^{-(1+\epsilon)/2}|X_n| \geq B \quad \text{i.o.}) = 0$, i.e., $n^{-(1+\epsilon)/2}|X_n|$ is essentially bounded w.p.1 as $n \to \infty$. Since $\mathrm{E}[u^4] < \infty$ by assumption, for some $\epsilon > 0$, however small, $n^{(1-\epsilon)/2}\bar{u}^2 = n^{-(1+\epsilon)/2}(n^{1/2}\bar{u})^2$ is essentially bounded w.p.1 as $n \to \infty$. Next,

$$\begin{aligned}
\mathrm{E}[(n^{-1}\sum_{i=1}^n [k_b^{(2)}(u-u_i+\omega\bar{u}) - k_b^{(2)}(u-u_i)])^2\bar{u}^4] &\leq \mathrm{E}[(\max_{1\leq i\leq n}|k_b^{(2)}(u-u_i+\omega\bar{u}) - k_b^{(2)}(u-u_i)|)^2\bar{u}^4] \\
&\leq C^2(n^{\tau/2}b^{3+\tau})^{-2}n^{\tau(1+\epsilon)/2}\,\mathrm{E}[(n^{(1-\epsilon)/2}|\bar{u}|^2)^\tau\bar{u}^4] \\
&= o(n^{-2+\tau(1+\epsilon)/2}) = o(n^{-1}b^3) \quad \text{w.p.1.}
\end{aligned}$$

The first inequality follows from $n^{-1}\sum_i a_i^2 \leq \max_{1\leq i\leq n} a_i^2$, the second by Hölder continuity of $k^{(2)}$ as above and writing $|n^{1/2}\bar{u}|^{2\tau} = n^{\tau(1+\epsilon)/2}(|n^{(1-\epsilon)/2}\bar{u}|^2)^\tau$, the third as, by Assumption 3.4(c), $n^{\tau/2}b^{3+\tau} \to \infty$ and, by the extremal Hölder inequality with exponents ∞ and 1, $\mathrm{E}[(n^{(1-\epsilon)/2}|\bar{u}|^2)^\tau\bar{u}^4] \leq O(n^{-2})$ noting that $n^{(1-\epsilon)/2}|\bar{u}|^2$ is essentially bounded w.p.1 as $n \to \infty$ and $\mathrm{E}[\bar{u}^4] = O(n^{-2})$ and, finally, as $o(n^{-2+\tau(1+\epsilon)/2}) = o(n^{-1}b^3)n^{(\tau-1)/2+9(\tau-1)/[8(3+\tau)]+(4\tau\epsilon-1)/8}$ because $n^{-3\tau/[2(3+\tau)]}b^{-3} \to 0$ by Assumption 3.4(c), choosing $\epsilon \leq 1/4\tau$ gives the result.

If f is twice differentiable and $f^{(2)}(u)$ and $uf^{(1)}(u)$ are absolutely integrable, applying Supplement P: Lemma P.3,

$$\begin{aligned}
\mathrm{E}[\hat{\delta}_1(u)] &= n^{-1}\,\mathrm{E}[u_i k_b^{(1)}(u-u_i)] + \frac{1}{2}\sigma^2 n^{-1}\,\mathrm{E}[k_b^{(2)}(u-u_i)] + o(n^{-1}) \\
&= n^{-1}(f(u) + uf^{(1)}(u) + \frac{1}{2}\sigma^2 f^{(2)}(u)) + o(n^{-1}),
\end{aligned}$$

where $\sigma^2 = \mathrm{E}[u^2]$. Since $\mathrm{Var}[\tilde{f}(u)] \sim (nb)^{-1}$, the covariance between $\tilde{f}(u)$ and the remainder term in $\hat{\delta}_1(u)$ is of order $o(n^{-1}b)$, and, hence,

$$\begin{aligned}
\mathrm{Cov}[\tilde{f}(u), \hat{\delta}_1(u)] &= n^{-1}\,\mathrm{E}[k_b^{(1)}(u-u_i)]\,\mathrm{E}[k_b(u-u_j)u_j] + o(n^{-1}b) \\
&= n^{-1}uf^{(1)}(u)f(u) + o(n^{-1}b), \qquad \text{(E.1)} \\
\mathrm{Var}[\hat{\delta}_1(u)] &= n^{-1}\,\mathrm{E}[k_b^{(1)}(u-u_i)]^2\,\mathrm{E}[u_j^2] + o(n^{-1}b^3) \\
&= n^{-1}\sigma^2 f^{(1)}(u)^2 + o(n^{-1}b). \qquad \text{(E.2)}
\end{aligned}$$

Note that $\zeta_\lambda = 0$, $\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u = -f^{(1)}(u)$, $\mathrm{d}\{\mathrm{E}[\nabla^\top u_i H g_i|u]f(u)\}/\mathrm{d}u = f(u) + uf^{(1)}(u)$, $\mathrm{d}\{\mathrm{E}[\nabla^2 u_i|u]f(u)\}/\mathrm{d}u = 0$, and $\mathrm{d}^2\{\mathrm{E}[\nabla u_i \nabla^\top u_i|u]f(u)\}/\mathrm{d}u^2 = f^{(2)}(u)$ from the unbiasedness of $\hat{\beta}$

and linearity of $u(z,\beta)$; cf. Theorem 3.3.

Assuming $f^{(1)}(u)$ is square integrable, and if $\lim_{|u|\to\infty} uf(u)^2 = 0$, $\int uf^{(1)}(u)f(u)\mathrm{d}u = -\frac{1}{2}R(f)$ and, thus,

$$\mathrm{IVar}[\hat{f}] = \mathrm{IVar}[\tilde{f}] - n^{-1}(R(f) - \sigma^2 R(f^{(1)})) + o(n^{-1}).$$

Hence, whenever $R(f) > \sigma^2 R(f^{(1)})$, $\hat{f}$ achieves a second order reduction in variance relative to $\tilde{f}$. While this may appear as a costless reduction in variance, it is not so. Construction of $\hat{f}$ explicitly assumes that $\mathrm{E}[u]$ exists, and the validity of the above result requires the first four moments of u exist whereas $\tilde{f}$ makes no such assumptions.

When the mean $\mathrm{E}[u]$ is known, the (G)EL-reweighted estimator $\tilde{f}_\rho$ eq. (3.4) imposing the constraint $\mathrm{E}[u] = 0$ will achieve a second order reduction in variance of $n^{-1}\sigma^{-2}u^2 f(u)^2$, i.e., $\mathrm{IVar}[\tilde{f}_\rho] = \mathrm{IVar}[\tilde{f}] - n^{-1}\sigma^{-2}\int u^2 f(u)^2\mathrm{d}u + o(n^{-1})$; see, e.g., Chen (1997, eq. (13), p.56). In particular, for normally distributed u, $R(\phi_\sigma) - \sigma^2 R(\phi_\sigma^{(1)}) = 1/4\sqrt{\pi}\sigma$, which equals $\sigma^{-2}\int u^2\phi_\sigma(u)^2\mathrm{d}u$ exactly. For the Student t distribution with $\nu > 2$ degrees of freedom, $R(t_\nu) - \sigma^2 R(t_\nu^{(1)}) = R(t_\nu)(2\nu^2 - 3\nu - 17)/4(\nu^2 - 4)$, which is positive for $\nu > 4$, the condition for the first four moments of u to exist, whereas $\sigma^{-2}\int u^2 t_\nu(u)^2\mathrm{d}u = R(t_\nu)(\nu-2)/(2\nu-1)$ which is always larger than $R(t_\nu) - \sigma^2 R(t_\nu^{(1)})$. This difference may be interpreted as the cost of estimating the mean of u.

The same or similar terms appear in the expansions for the variance of $\hat{f}$ in other contexts (the $O(n^{-1})$ bias terms tend to be ignored as their contribution to MISE is $o(n^{-1})$); cf. Muhsal and Neumeyer (2010, eq.(3.5)). As the next example demonstrates, these same effects appear in a large class of parametric moment condition models.

Example E.3 (G)EL With A Constant And Zero Mean Restriction

Consider (G)EL estimation based on moment indicator functions of the form $g(z,\beta) = u(z,\beta)\alpha(w)$ where $u(z,\beta)$ is scalar, β a d_β-vector of parameters, and $\alpha(w)$ a d_g-vector of functions of w. Suppose that $u(z,\beta_0)$ is independent of w, Assumption 3.1 holds and the moment condition $\mathrm{E}[g(z,\beta_0)] = 0$ includes the restriction $\mathrm{E}[u(z,\beta_0)] = 0$. Furthermore, it is assumed that $u(z,\beta)$ contains a constant; the inclusion of an explicit constant is not essential as the results here continue to hold if $\mathrm{E}[\partial u(z,\beta_0)/\partial\beta^\top|w]\gamma = c$ for some non-zero vector γ and scalar c, in which case $\mathrm{E}[\alpha(w)] = G\gamma/c$. Without loss of generality let $\alpha_1(w) = 1$ and $\partial u(z,\beta_0)/\partial\beta_1 = -1$.

Since u and w are independent, $\mathrm{E}[g_i|u] = u\,\mathrm{E}[\alpha(w)]$, $\Omega = \sigma^2\,\mathrm{E}[\alpha(w)\alpha(w)^\top]$, where $\sigma^2 = \mathrm{E}[u^2|w] = \mathrm{E}[u^2]$. Then, because the first column of G is $-\mathrm{E}[\alpha(w)]$, as $PG = 0$, $\mathrm{E}[g_i|u]^\top P\,\mathrm{E}[g_i|u] = 0$. That is, there is no second order reduction in variance due to re-weighting.

Since the first column (and row) of Ω is $\sigma^2\,\mathrm{E}[\alpha(w)]$, $\Omega^{-1}\,\mathrm{E}[g_i|u] = u\sigma^{-2}e_1$, where e_j is the jth unit d_g-vector, $j = 1,\ldots,d_g$. For an $n\times m$ matrix A, let $A_{(s:t)}$, $1 \le s \le t \le m$, denote the $n\times(t-s+1)$ submatrix comprised of columns $j = s,\ldots,t$ of A. Noting that $\Omega^{-1}\,\mathrm{E}[\alpha(w)] = e_1/\sigma^2$

and $\mathrm{E}[\alpha(w)]^\top \Omega^{-1} \mathrm{E}[\alpha(w)] = 1/\sigma^2$, partition Σ and H as

$$\Sigma = \sigma^2 \begin{pmatrix} 1 + e_1^\top G_{(2:d_\beta)} Q^{-1} G_{(2:d_\beta)}^\top e_1 & e_1^\top G_{(2:d_\beta)} Q^{-1} \\ Q^{-1} G_{(2:d_\beta)}^\top e_1 & Q^{-1} \end{pmatrix}, H = - \begin{pmatrix} e_1^\top - e_1^\top G_{(2:d_\beta)} Q^{-1} G_{(2:d_\beta)}^\top (\sigma^2 \Omega^{-1} - e_1 e_1^\top) \\ -Q^{-1} G_{(2:d_\beta)}^\top (\sigma^2 \Omega^{-1} - e_1 e_1^\top) \end{pmatrix},$$

where $Q = G_{(2:d_\beta)}^\top (\sigma^2 \Omega^{-1} - e_1 e_1^\top) G_{(2:d_\beta)}$ and $e_1^\top G_{(2:d_\beta)} = (\mathrm{E}[\partial u(z, \beta_0)/\partial \beta_2], \ldots, \mathrm{E}[\partial u(z, \beta_0)/\partial \beta_{d_\beta}])$ is the first row of $G_{(2:d_\beta)}$. Thus, $H \mathrm{E}[g_i|u] = -u e_1$.

As the first element of $\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u$ is $-f^{(1)}(u)$, $[\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u]^\top H \mathrm{E}[g_i|u] f(u) = u f^{(1)}(u) f(u)$, the same as eq. (E.1). Partition $\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u$ as

$$\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u = (-f^{(1)}(u), [\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u]_{(2:d_\beta)})^\top.$$

Hence,

$$\begin{aligned} [\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u]^\top \Sigma [\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u] &= \sigma^2 f^{(1)}(u)^2 + \sigma^2 f^{(1)}(u)^2 e_1^\top G_{(2:d_\beta)} Q^{-1} G_{(2:d_\beta)}^\top e_1 \\ &\quad - 2\sigma^2 f^{(1)}(u) e_1^\top G_{(2:d_\beta)} Q^{-1} [\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u]_{(2:d_\beta)} \\ &\quad + \sigma^2 [\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u]_{(2:d_\beta)}^\top Q^{-1} [\mathrm{d}\{\mathrm{E}[\nabla u_i|u] f(u)\}/\mathrm{d}u]_{(2:d_\beta)} \end{aligned} \tag{E.3}$$

The first term in (E.3) is the same as the main term in (E.2). The remaining terms represent the additional increase in the variance of $\hat{f}(u)$ due to the estimation error in $\beta_2, \ldots, \beta_{d_\beta}$.

The independence of u and w is crucial to the above argument implying $\mathrm{E}[g_i|u] = u \mathrm{E}[\alpha(w)]$, and P annihilates $\mathrm{E}[\alpha(w)]$. The next example illustrates that these relationships need not hold in the dependent case.

Example E.4 Linear Regression Model With $\mathrm{E}[u|x] = 0$ But Dependent u And x

For simplicity, consider the linear regression model

$$y = \delta_0 + \gamma_0 x + u, \tag{E.4}$$

where $\mathrm{E}[u|x] = 0$. Here $\beta = (\delta, \gamma)^\top$ and $z = (y, x)^\top$.

Estimation of β_0 may be based on the unconditional moment restriction $\mathrm{E}[g(z, \beta_0)] = 0$ where

$$g(z, \beta) = u(z, \beta)(1,\ x,\ x^2,\ \ldots,\ x^{q-1})^\top, \quad q \geq 2 \tag{E.5}$$

Suppose that u and x are distributed with joint density

$$f_{U,X}(u, x) = \frac{2(\nu/2)^{\nu/2}}{\sqrt{2\pi}\omega\Gamma(\nu/2)} x^\nu \mathrm{e}^{-x^2(\nu + u^2/\omega^2)/2}, \quad x \geq 0,\ -\infty < u < \infty,\ \nu > 0,\ \omega > 0. \tag{E.6}$$

The marginal distributions of u and x are the non-standardized Student t distribution with ν degrees of freedom and scale parameter ω and the generalized gamma distribution (Stacy, 1962) with parameters $p = 2$, $d = \nu$, and $a = (2/\nu)^{1/2}$.[1] The moments of x are $m_k = \mathrm{E}[x^k] = (2/\nu)^{k/2}\Gamma((\nu+k)/2)/\Gamma(\nu/2)$, $k > -\nu$, and satisfy the recursion $m_{k+2} = (1 + k/\nu)m_k$. The odd moments of u of order $k < \nu$ are zero, while the even moments are $\mathrm{E}[u^{2k}] = \omega^{2k}\pi^{-1/2}\nu^k\Gamma(\nu/2 - k)\Gamma(k + 1/2)/\Gamma(\nu/2)$, $k < \nu/2$.

The conditional density of u given x is $f_{U|X}(u, x) = \phi_{\omega/x}(u)$ and, hence, $\mathrm{E}[u|x] = 0$, but u and x are dependent. If $\nu > 2$, $\mathrm{E}[u^2|x] = \omega^2/x^2$. The conditional moments of x given u are $m_{k|u}(u) = \mathrm{E}[x^k|u] = m_{k+1}/m_1(1 + (u/\omega)^2/\nu)^{k/2}$, $k > -\nu - 1$. The transformation in Assumption 3.1 has $v(z, \beta) = x$ and, hence, $\mathrm{E}[g_i|u] = u(1,\, m_{1|u}(u),\, m_{2|u}(u),\, \ldots,\, m_{q-1|u}(u))^\top$.

To describe the quantities involved, let ${}^q_s\mathrm{M} = \{m_{i+j-2-s}\}_{i,j=1}^q$ be a $q \times q$ matrix composed of the $(i + j - 2 - s)$th moments of x. Note that, if $q > 2$, then ${}^q_0\mathrm{M}^\top_{(s)}\ {}^q_2\mathrm{M}^{-1}\ {}^q_0\mathrm{M}_{(t)} = m_{s+t}$ for $s, t = 1, 2, \ldots, (s \wedge t) \leq q - 2$, and ${}^q_2\mathrm{M}^{-1}\ {}^q_0\mathrm{M}_{(t)} = e_{t+2}$ for $1 \leq t \leq q - 2$. The relevant (G)EL matrices are $\Omega = \omega^2\ {}^q_2\mathrm{M}$, $G = -{}^q_0\mathrm{M}_{(1:2)}$ and, if $q \geq 4$,

$$\Sigma = \frac{\omega^2\nu}{\nu(\nu+2) - (\nu+1)^2 m_1^2}\begin{pmatrix} \nu+2 & -(\nu+1)m_1 \\ -(\nu+1)m_1 & \nu \end{pmatrix}, \qquad H = -\frac{1}{\omega^2}\Sigma\begin{pmatrix} e_3^\top \\ e_4^\top \end{pmatrix},$$

and

$$P = \frac{1}{\omega^2}\,{}^q_2\mathrm{M}^{-1} - \frac{1}{\omega^2}\frac{\nu}{\nu(\nu+2) - (\nu+1)^2 m_1^2}[(\nu+2)e_3e_3^\top - (\nu+1)m_1(e_3e_4^\top + e_4e_3^\top) + \nu e_4e_4^\top].$$

Remark E.1 For the exactly identified case, $q = 2$, G is square and invertible. Hence, $\Sigma = G^{-1}\Omega G^{\top -1}$, $H = G^{-1}$, and $P = 0$. Closed form expressions for Σ, H, and P when $q = 3$ can be obtained in a straightforward fashion. That Σ remains unaltered as q increases above 4 is of course due to the special form of the conditional variance of u. Figure 1 displays the relative efficiency of $\hat{\beta}$ based on the first q compared with the first q' moment conditions, $[\det(\Sigma_q)/\det(\Sigma_{q'})]^{1/p}$, for various values of ν.

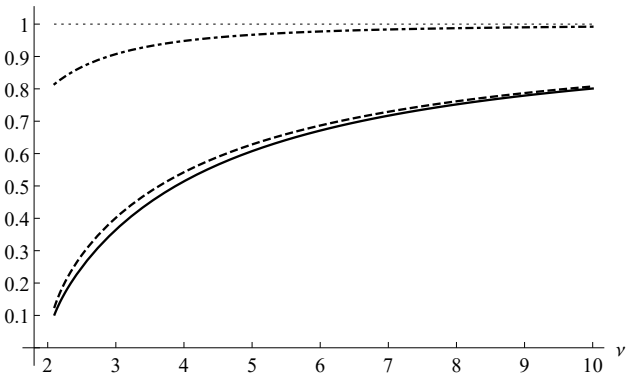


Figure 1: **Relative Efficiency of** $\hat{\beta}$: $q \geq 4$ vs. $q = 2$ (solid line); $q = 3$ vs. $q = 2$ (dashed line); $q \geq 4$ vs. $q = 3$ (dash-dotted line).

If $q \geq 4$, only the moment indicators $x^{j-1}u(z, \beta)$, $j = 3, 4$, are used to estimate β_0. Infor-

[1] x is distributed as $(w/\nu)^{1/2}$ where $w \sim \chi^2_\nu$ and, if $\nu = 1$, as a standard half-normal random variable. The joint density eq. (E.6) is that of $u = z/x$ and x, where $z \sim N(0, \omega^2)$, independent of x.

mation in the remaining moment conditions, however, can be usefully exploited to improve the efficiency of the density estimators $\hat{f}$ and $\hat{f}_\rho$. The quantities entering the integrated variance eqs. (3.11) and (3.12) can be computed as $\mathrm{tr}(\Sigma \int [\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u][\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top \mathrm{d}u)$, $\mathrm{tr}(H \int \mathrm{E}[g_i|u][\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top f(u)\mathrm{d}u)$, and $\mathrm{tr}(P \int \mathrm{E}[g_i|u]\,\mathrm{E}[g_i|u]^\top f(u)^2\mathrm{d}u)$, where

$$\int \frac{\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}}{\mathrm{d}u}\Big(\frac{\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}}{\mathrm{d}u}\Big)^\top \mathrm{d}u = \frac{\Gamma((\nu+3)/2)}{\omega^3\pi^{1/2}\nu^{3/2}\Gamma(\nu/2)} \begin{pmatrix} \frac{\nu\Gamma(\nu+3/2)\Gamma((\nu+3)/2)}{\Gamma(\nu/2+1)\Gamma(\nu+3)} & \frac{\nu^{1/2}\Gamma(\nu+3)}{2^{1/2}\Gamma(\nu+7/2)} \\ \frac{\nu^{1/2}\Gamma(\nu+3)}{2^{1/2}\Gamma(\nu+7/2)} & \frac{\Gamma(\nu+5/2)\Gamma(\nu/2+2)}{2\Gamma(\nu+2)\Gamma((\nu+5)/2)} \end{pmatrix},$$

the $q \times 2$ matrix $\int \mathrm{E}[g_i|u][\mathrm{d}\{\mathrm{E}[\nabla u_i|u]f(u)\}/\mathrm{d}u]^\top f(u)\mathrm{d}u$ has rows

$$\frac{1}{(2\pi)^{1/2}\omega}\Big(\frac{(2/\nu)^{i/2}\Gamma((\nu+3)/2)\Gamma(\nu+i/2)\Gamma((\nu+i)/2)}{\Gamma(\nu/2)^2\Gamma(\nu+(i+3)/2)}, \quad \frac{(2/\nu)^{(i-1)/2}(\nu+2)\Gamma((\nu+1)/2)\Gamma(\nu+(i+1)/2)}{2\Gamma(\nu/2)\Gamma(\nu+i/2+2)},\Big) \quad i = 1, \ldots, q,$$

and the $q \times q$ matrix $\int \mathrm{E}[g_i|u]\,\mathrm{E}[g_i|u]^\top f(u)^2\mathrm{d}u$ with (i,j)th element

$$\omega m_i m_j \frac{\nu^{3/2}\Gamma(\nu+(i+j-3)/2)}{4\pi^{1/2}\Gamma(\nu+(i+j)/2)}, \quad i,j = 1, \ldots, q.$$

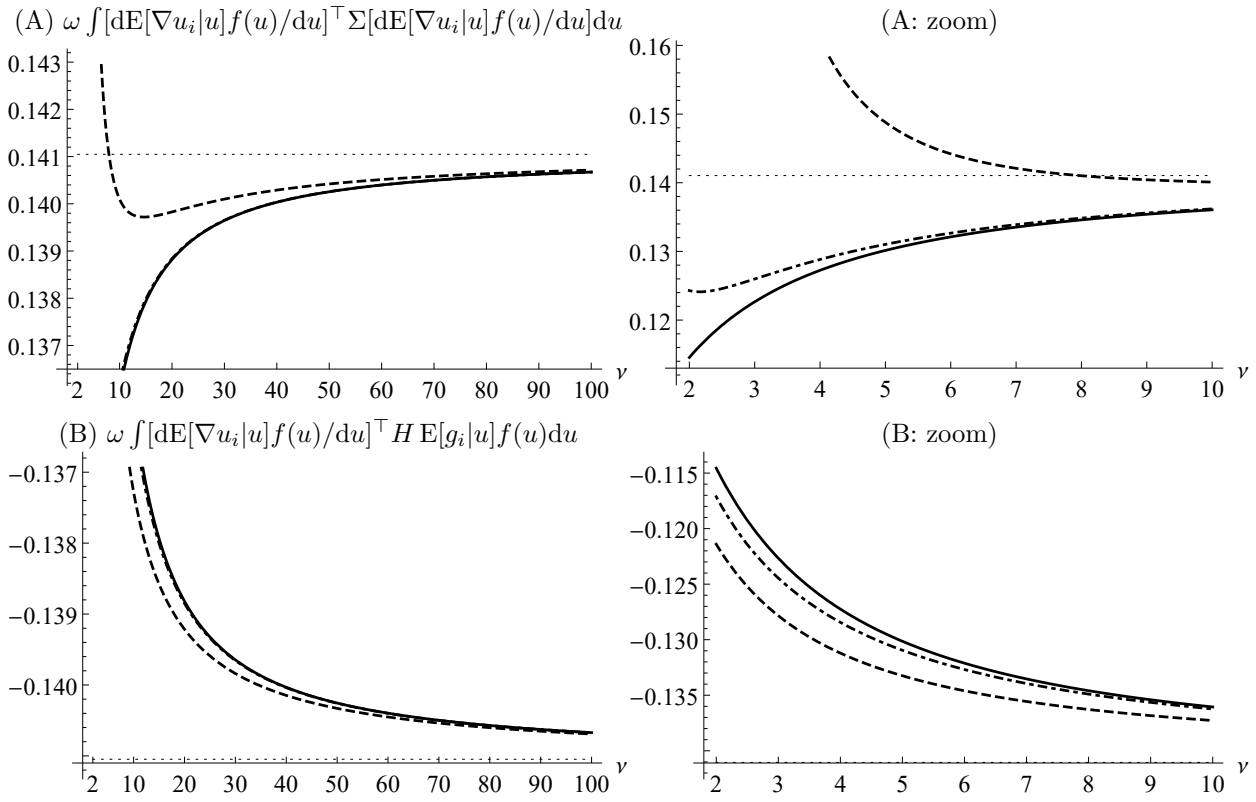
REMARK E.2 Figure 2 shows the values of the above quantities and the overall effect on the integrated variance for selected values of q and $\nu > 2$; note that the validity of the asymptotic expansions requires $\nu > 4$, but variance is defined for $\nu > 2$. While the main reduction in variance is still due to the zero mean restriction as in Example Example E.3 (Panels A and B), there are small additional gains due to re-weighting (Panel C). The latter do increase as more moment conditions are added.

References

Chen, S. X. (1997), 'Empirical likelihood-based kernel density estimation', *Australian and New Zealand Journal of Statistics* **39**(1), 47–56.

Muhsal, B. and Neumeyer, N. (2010), 'A note on residual-based empirical likelihood kernel density estimation', *Electronic Journal of Statistics* **4**, 1386–1401.

Stacy, E. W. (1962), 'A generalization of the gamma distribution', *The Annals of Mathematical Statistics* **33**(3), 1187–1192.



Legend to panels A and B: dashed line: $q = 2$; dash-dotted line: $q = 3$; solid line: $q \geq 4$.

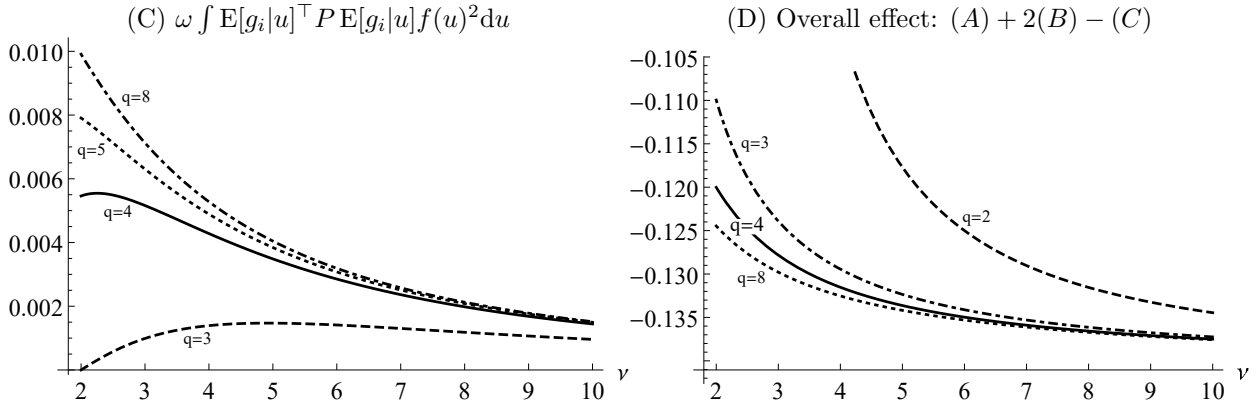


Figure 2: Quantities entering the integrated variance of $\hat{f}$ and $\hat{f}_\rho$ in Example E.4 ($\times n\omega$)