Subvector Inference in Local Regression*

Ke-Li Xu†

Texas A&M University

January 4, 2015

Abstract

We consider estimation and inference of a subvector of parameters that are defined through local moment restrictions. The framework is useful for a number of econometric applications including those in policy evaluation based on discontinuities or kinks and in real-time financial risk management. We aim to provide approaches to inference that are generic (without requiring case-by-case standard error analysis) and are robust when regularity assumptions fail. These irregularities include non-differentiability, non-negligible bias and weak identification. We focus on the QLR criterion-function-based (in particular, empirical likelihood-based) inference, and establish conditions under which the test statistic has a pivotal asymptotic distribution. Confidence sets can be obtained by inverting the test. In the key step of eliminating nuisance parameters in the criterion function, we consider that based on concentration and Laplace-type plug-in estimation. The former is natural, and the latter does not require optimization and can be computationally attractive in applications using simulations. We provide the asymptotic analysis under the null and local/non-local alternatives, and illustrate the high-levels assumptions with several examples. Simulations and an empirical application illustrate the finite-sample performance.

Keywords: Bias correction; empirical likelihood; Laplace-type estimator; local moment restrictions; non-smooth criterion function; nonparametric and semiparametric inference; nuisance parameter; quantile regression discontinuity; weak identification.

JEL Classification: C12, C14, C21, C22.

*The author acknowledges the comments and suggestion from seminar participants at JSM 2013, Atlanta Econometrics Study Group 2013, NASM 2014, UBC, UIUC and Vanderbilt University.

†Department of Economics, Texas A&M University, 3063 Allen, 4228 TAMU, College Station, TX 77843-4228, USA. *Email:* keli.xu@tamu.edu.

1 Introduction

In this paper, we consider estimation and inference of a subvector of parameters that are defined through local moment restrictions. Local moment restrictions, which could be linear or nonlinear, do not require any kind of global specification as in traditional global moment restrictions models. The leading examples that motivate this research include cases when the object of interest is a function of several nonparametric regression estimates, and when the dependent variable in the nonparametric and semiparametric regression involves unknown quantities (nuisance parameters) in a nonseparable way that are also estimated in preliminary steps of local regressions. The framework includes applications when the delta method can be used and when it can not be directly used, e.g. when nuisance parameters enter estimating equations in a nonseparable way. The current work is related to the literature of local estimating equations in statistics (Carroll, Ruppert and Welsh, 1998) which assumes the full vector of parameters are of interest and other regularity conditions.

Examples of models in econometrics that fall in our framework are not rare. In the regression discontinuity (RD) design, the quantile treatment effect of interest is identified as a function of six quantities which are to be estimated nonparametrically in applications (Frandsen, Frölich, and Melly, 2012). Other models with discontinuous and kink incentive assignment mechanisms share the similar features (Imbens and Lemioux, 2008, and Card, Lee, Pei, Weber, 2012, Calonico, Cattaneo and Titiunik, 2014). In risk management, the coherent risk measure the expected shortfall depends on the value at risk that is also estimated (Artzner, Delbaen, Eber and Heath, 1999, Agarwal and Naik, 2004, Zhu and Fukushima, 2009), and recently nonparametric and semiparametric methods have been used (Linton and Xiao, 2013). In real-time forecasting, a multi-step forecast may depend on predicted values of covariates.

In such models, while consistency of the point estimator is preserved under regularity assumptions on the first step plug-in estimation of nuisance parameters, the standard error is more affected. A natural approach to conducting inference is based on properly constructed local estimating equations (the sample analogs of local moment restrictions) that connect the quantities of interest and auxiliary quantities. Joint asymptotic normality (and the asymptotic variance) follows from the GMM framework (Hansen, 1982, Newey, 1984, Newey and McFadden, 1994). The standard error is then obtained by estimating the asymptotic variance, and it can be done by two approaches.

The first approach is based on separately estimating each piece of the population quantities (usually conditional/unconditional moments, quantiles or densities) involved in the asymptotic variance. This approach, referred to as the plug-in approach or the unconditional approach, is typically adopted in the literature (although not necessarily through estimating equations) by articles that focus on specific models; see references cited later in the paper for examples and discussions in Section 3. The performance of this approach in finite samples depends on the qualities of two approximations: the approximation of the asymptotic variance (to the finite sample variance) and the approximation of the standard error (to the asymptotic variance). The first approximation could be poor if a low-quality asymptotic theory is drawn, and is not improved even the true value of the asymptotic variance is known. The quality of the second approximation could be affected by components of the asymptotic variance that are inaccurately estimated (typically nonparametric density functions or when the design points are in data-sparse area or close to boundaries). When implemented, this standard error typically involves multiple bandwidth selections when estimating each part of the variance, and selection criteria (e.g. MSE-minimizing-based) usually target on optimizing the second approximation instead of the quality of overall distributional approximation (given the bandwidth used in the point estimator). Calculation of kernel-specific constants adds extra complications. Xu (2013) and Fan and Liu (2014) are motivated by similar concerns and consider alternative inferences in nonparametric and semiparametric quantile regression models.

The second approach is to directly estimate the sandwich form of the asymptotic variance. In the simple case of linear estimating equations, this approach essentially estimates the conditional variance instead of the asymptotic variance as in the unconditional approach. This generic standard error, which only uses the form of estimating equations and does not need the explicit variance formula, can be shown to better approximate the finite sample variance (Fan and Gijbels, 1996, Carroll et al., 1998). However, the implementation is not straightforward when the estimating equations are not smooth in parameters since the sandwich-form variance estimate generally depends on derivatives.

The standard error based approach, built up on estimation through either the conditional or unconditional approach, collapses when either parameters of interest or nuisance parameters are weakly identified. In these cases, parameters are inconsistently estimated and joint asymptotic normality fails. Marmer, Feir and Lemieux (2014) showcase potential issues of standard inference under weak identification in the fuzzy RD design.

It is thus worthwhile to consider the criterion function-based approach to inference which completely avoids variance estimation and does not require consistency when weak identification occurs. We focus on the empirical likelihood (EL hereinafter, Owen, 2001) which arises naturally in the framework of estimating equations, and leads to automatically pivotalized test statistics under mild conditions.

In our setting, the crucial step to form the EL statistic is the elimination of nuisance parameters. The usual concentrating-out procedure, which involves optimizing the profile EL over the space of nuisance parameters, requires caution in theoretical treatment when nuisance parameters enter estimating equations non-smoothly since first-order conditions can not be used. It also raises practical issues in computation for the same reason (i.e. the search for optima can not be based on derivatives), and extant remedies may confound global optima with local optima.[1] We then extend the simulation-based quasi-Bayesian procedure (Chernozhukov and Hong, 2003) to form inference in local moment restrictions models with nuisance parameters. It requires a symmetric loss function and can address such computational issues without requiring global optima for valid inference.

We extend the large literature on empirical likelihood based estimation and inference which is primarily focused on parameters in global moment restrictions (Qin and Lawless, 1994, Newey and Smith, 2004, Guggenberger and Smith, 2005) and conditional moment conditions (Donald, Imbens and Newey, 2003, Kitamura, Tripathi and Ahn, 2004). See also Gagliardini, Gourieroux and Renault (2011) for applications of local moment restrictions in financial derivative pricing. In the global moment conditions model with iid observations, Kitamura (2001; 2006, Section 4.2) showed that the EL ratio test, unlike other members in the GEL family (Newey and Smith, 2004), achieves some optimality property (i.e. large deviation minimax optimality) due to its distinctive interpretation as the Kullback-Leibler divergence.

There are recent work on EL with nonsmooth estimating equations in different contexts. Molanes Lopez, van Keilegom and Veraverbeke (2009) focus on iid marginal non-regression models and allow the criterion function to be nonsmooth in nuisance parameters. Otsu (2008) considers efficient estimation in quantile regression under exogeneity by exploring the conditional moment restrictions. Chernozhukov and Hong (2003) and Parente and Smith (2011) consider unconditional moment re-

[1] Even with smooth estimating equations, computational issues of the concentrated EL have received considerable attention. Antoine, Bonnal and Renault (2007) and Fan, Gentry and Li (2011) proposed alternative estimators that are computationally less demanding and preserve certain properties of EL estimators.

striction models which include the instrumental variables quantile regression. These papers are mainly concerned about aspects of marginal distributions or partial effects, therefore exclude non-parametric regression estimands considered in this paper.

The paper is organized as follows. In Section 2, we introduce local moment restrictions and local estimating equations, and provide examples of models in the recent literature in which inference can be analyzed in this framework. A generic standard error for smoothing estimating equations is given in Section 3. We then introduce the empirical likelihood for local estimating equations, with two methods of dealing with nuisance parameters covered in Sections 4 and 5. Asymptotic theories and high-level assumptions are also given. Power analysis under local and non-local alternatives is provided in Section 6. In Sections 7 and 8, we verify for the examples given earlier that high-level assumptions do not require much more than standard ones which are typically imposed in the literature. Robustness to weak identification in part of the parameter space is discussed in Section 9. Sections 10 and 11 contain Monte Carlo simulations and an empirical example of heterogeneous effects of academic probation under the RD design, and Section 12 concludes. Technical details are contained in five appendices.

2 Local moment restrictions

Let $Y \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$ contains outcome variables and $X \in \overline{\mathcal{X}} \subset \mathbb{R}^{d_X}$ is the set of covariates. Suppose the true parameter values $\beta_0(\mathcal{X})$ and $\theta_0(\mathcal{X})$ satisfy d local moment restrictions

$$Mg_0(\beta, \theta) = 0, \tag{1}$$

where $g_0(\beta, \theta) = (\mathbb{E}[g_1(Y, \beta, \theta)|X \in \mathcal{X}_1], ..., \mathbb{E}[g_{d_g}(Y, \beta, \theta)|X \in \mathcal{X}_{d_g}])'$ is $d_g \times 1$ with $\mathcal{X} = \cup_{j=1}^{d_g} \mathcal{X}_j \subset \overline{\mathcal{X}}$, and M is a $d \times d_g$ matrix of constants. In all our applications below, $\mathcal{X}$ is a zero-measure set. For notational simplicity we write $\beta_0(\mathcal{X})$ and $\theta_0(\mathcal{X})$ as β_0 and θ_0 respectively. The moment functions (or residual functions) $g_1, ..., g_{d_g}$ are constructed for a specific application under investigation, having known functional forms up to unknown parameters $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\beta \in \mathcal{B} \subset \mathbb{R}^{d_\beta}$, where $d = d_\beta + d_\theta$. The matrix M reflects that each moment restriction in (1) (each row in (1)) may involve expectations

over different subpopulations. We assume all random variables in X are continuous.[2]

We are mainly interested in θ, treating β as the nuisance parameter.[3] We allow the $\mathbb{R}^{d_g}-$valued function $g(Y,\beta,\theta) = (g_1(Y,\beta,\theta),...,g_{d_g}(Y,\beta,\theta))'$ to be smooth or non-smooth in (β,θ), and the latter happens, e.g. if any element of (β,θ) enters an indicator function. The true values β_0 and θ_0 are typically functions of conditional moments or quantiles of outcome variables at given values of covariates (instead of parameters such as marginal effects in traditional moment restriction models). The paper focuses on estimation and in particular, inference of θ.

The parameters are usually estimated by $\widehat{\beta}$ and $\widehat{\theta}$ which solve following estimating equations

$$\sum_{i=1}^{n} M w_i(\mathcal{X}) g_i(\widehat{\beta},\widehat{\theta}) = 0, \tag{2}$$

where the weight function $w_i(\mathcal{X})$ and the estimating function $g_i(\beta,\theta)$ define estimators. In baseline cases (Examples 1, 3 and 4 below), $g_i(\beta,\theta) = g(Y_i,\beta,\theta)$, while in other cases $g_i(\beta,\theta)$ might depend on $\{X_i, Y_i : 1 \leq i \leq n\}$, e.g. when bias correction is used or the conditional set is estimated (Examples 2 and 3). The $d_g \times d_g$ diagonal matrix $w_i(\mathcal{X}) = \text{diag}(w_{1i}(\mathcal{X}_1),...,w_{d_g i}(\mathcal{X}_{d_g}))$ is such that $\sum_{i=1}^{n} w_i(\mathcal{X}) = \mathcal{I}_{d_g}$, where $\mathcal{I}_{d_g}$ is the d_g-dim identity matrix. It contains weights (usually the local polynomial weights or their variants) that are assigned to observations in the neighborhood of $\mathcal{X}$. The weights $w_i(\mathcal{X})$ generally depend on the whole sample $\{X_i, 1 \leq i \leq n\}$ or $\{X_i, Y_i : 1 \leq i \leq n\}$.

The weights $w_i(\mathcal{X})$ depend on design sets $\mathcal{X}$ to reflect that we are mainly interested in local regressions. Allowing different weights across different expectations (i.e. $w_i(\mathcal{X})$ being a non-scalar matrix) will be useful.

It is of course necessary when design sets are different across expectations (Examples 1, 2 and 4). When only one design set is of interest, the weights could be the same (the baseline case in Example 3) or different across expectations. The latter case occurs when, across expectations, different dimension-reduction parameters (Example 3), local neighborhoods (varying bandwidths across expectations), kernel functions or local specifications (less likely), are used. However, we emphasize that the results developed below apply to general estimating equations (which are not necessarily local).

[2] When the covariates X are discrete (i.e. $\mathbb{P}(X = x) > 0$), the conditional moment restrictions in (1) can be rewritten as unconditional moment restrictions, thus fitting in the traditional GMM framework.

[3] In a specific application, β contains all nuisance parameters that have to estimated in order to estimate θ.

The framework in (1) applies trivially to the case when there is no nuisance parameters (i.e. $d_\beta = 0$).

We provide below a few examples that fit in the framework (1) and (2).

Example 1 (Quantile regression discontinuity (RD) design). Let Y be the outcome and D be the binary treatment indicator.[4] The interest is in the effects on the outcome of receiving the treatment, which is usually endogenous. In the RD design, the treatment is determined at least in part by a (scalar and continuous) forcing variable X exceeding the threshold $X = c$.

A useful quantity for policy evaluation identified from this design is the local quantile treatment effect (for compliers, denoted by $\mathcal{C}$). It reads $\theta_0 = Q_{Y^1|\mathcal{C},X=c}(\tau) - Q_{Y^0|\mathcal{C},X=c}(\tau)$, where τ is the probability level. The two quantile functions above (for potential outcomes Y^1 and Y^0 respectively[5]) are the inverse of the corresponding CDFs, which are identified as

$$\begin{aligned} F_{Y^1|\mathcal{C},X=c}(y) &= \frac{\mathbb{E}[\mathbb{I}_{(Y\le y)}D|X=c+] - \mathbb{E}[\mathbb{I}_{(Y\le y)}D|X=c-]}{\mathbb{E}[D|X=c+] - \mathbb{E}[D|X=c-]}, \\ F_{Y^0|\mathcal{C},X=c}(y) &= \frac{\mathbb{E}[\mathbb{I}_{(Y\le y)}(1-D)|X=c+] - \mathbb{E}[\mathbb{I}_{(Y\le y)}(1-D)|X=c-]}{\mathbb{E}[(1-D)|X=c+] - \mathbb{E}[(1-D)|X=c-]}, \end{aligned}$$

where $\mathbb{E}[\mathbb{I}_{(Y\le y)}D|X=c+] = \lim_{x\to c+}\mathbb{E}[\mathbb{I}_{(Y\le y)}D|X=x]$ and others are defined similarly. Cast in our framework, $g_i(\beta,\theta) = g(Y_i,\beta,\theta)$ with

$$\begin{aligned} g(Y_i,\beta,\theta) &= (\mathbb{I}(Y_i < \beta_1+\theta)D_i - \tau\beta_2,\ \mathbb{I}(Y_i<\beta_1)(1-D_i) + \tau\beta_2,\ D_i - \beta_2, \\ \mathbb{I}(Y_i &< \beta_1+\theta)D_i,\ \mathbb{I}(Y_i<\beta_1)(1-D_i),\ D_i)'. \end{aligned} \tag{3}$$

Then (1) is satisfied at true values $\theta_0, \beta_{10} = Q_{Y^0|\mathcal{C},X=c}(\tau)$, $\beta_{20} = \mathbb{E}[(1-D)|X=c+] - \mathbb{E}[(1-D)|X=c-]$, with $Y = (Y,D)'$, $(d_\beta, d_\theta, d, d_g) = (2,1,3,6)$, $M = (\mathcal{I}_3, -\mathcal{I}_3)$ and $\mathcal{X}$ being infinitesimal right and left neighborhoods of c. Note that g is non-smooth in β_1 and θ, and smooth in β_2.

The weight functions in (2) are defined by local polynomial fitting. Let $I_i = \mathbb{I}(X_i \ge c)$, and $\widehat{S}_k^+ = \sum_{i=1}^n [(X_i-c)/h]^k I_i K((X_i-c)/h)$ for $k = 0,1,\cdots,2p$, where $K(\cdot)$ is a kernel function and h is the bandwidth parameter. Let $\widehat{S}^+$ be the $(p+1)$ by $(p+1)$ matrix with the $(i,j)-$th element $\widehat{S}^+_{i+j-2}$. Similarly, let $\widehat{S}_k^- = \sum_{i=1}^n [(X_i-c)/h]^k (1-I_i) K((X_i-c)/h)$, and $\widehat{S}^-$ is similarly defined. Then $W_p^+(u) = e_1'(\widehat{S}^+)^{-1}\varpi(u)I_i$ and $W_p^-(u) = e_1'(\widehat{S}^-)^{-1}\varpi(u)(1-I_i)$, with $\varpi(u) = (1,u,\cdots,u^p)'K(u)$

[4] For unit i, $D_i = \mathbb{I}\{\text{The unit } i \text{ is treated}\}$.

[5] The observed outcome is then $Y = Y^0(1-D) + Y^1 D$.

and e_1 being the $(p+1)-$ dimension vector $(1, 0, \cdots, 0)'$. The weight functions in (2) are then

$$w_i(\mathcal{X}) = \text{diag}(W_p^+((X_i - c)/h)\mathcal{I}_3, W_p^-((X_i - c)/h)\mathcal{I}_3). \tag{4}$$

When $p = 0$, the weights reduce to $W_p^+((X_i - c)/h) = [\sum_{i=1}^n I_i K((X_i - c)/h)]^{-1} I_i K((X_i - c)/h)$ and similarly for $W_p^-(\cdot)$.

Frandsen, Frolich and Melly (2012) provided the identification conditions of the quantile treatment effect θ_0, and showed through direct calculation (instead of using estimating equations) that the asymptotic variance of the local linear estimator ($p = 1$) $\widehat{\theta}$ assumes a complex form.

A special case is the sharp RD design, in which $D = \mathbb{I}(X \geq c)$. $\mathbb{E}[D|X = c+] = 1$ and $\mathbb{E}[D|X = c-] = 0$. In this case $F_{Y^1|X=c}(y) = \mathbb{P}[Y \leq y|X = c+]$, $F_{Y^0|X=c}(y) = \mathbb{P}[Y \leq y|X = c-]$ and $\theta_0 = Q_{Y^1|X=c}(\tau) - Q_{Y^0|X=c}(\tau)$. Estimating functions are

$$g(Y_i, \beta, \theta) = (\mathbb{I}(Y_i < \beta + \theta) - \tau, \mathbb{I}(Y_i < \beta) - \tau)'. \tag{5}$$

Then (1) is satisfied with $\beta_0 = Q_{Y^0|X=c}(\tau)$, $M = \mathcal{I}_2$, and weight functions are

$$w_i(\mathcal{X}) = \text{diag}(W_p^+((X_i - c)/h), W_p^-((X_i - c)/h)). \tag{6}$$

Example 2 (QRD with bias correction). Continuing with the setting in Example 1, we now consider inference based on bias correction. The simple construction of estimating functions as in Example 1 requires an undersmoothing condition for valid tests and confidence intervals for θ; see Section 7. The motivation of using bias correction is to avoid such a condition thereby allowing the optimal (MSE-minimizing) bandwidth. Using the notations in the framework (2), this can be achieved by modifying the estimating functions in (3) and (5) as, for fuzzy and sharp designs,

respectively,

$$g_i(\beta,\theta)=\begin{pmatrix}\mathbb{I}(Y_i<\beta_1+\theta)D_i-\tau\beta_2-\widehat{\varrho}_{Y_1,+}(\beta,\theta)\\ \mathbb{I}(Y_i<\beta_1)(1-D_i)+\tau\beta_2-\widehat{\varrho}_{Y_0,+}(\beta)\\ D_i-\beta_2-\widehat{\varrho}_{D,+}\\ \mathbb{I}(Y_i<\beta_1+\theta)D_i-\widehat{\varrho}_{Y_1,-}(\beta,\theta)\\ \mathbb{I}(Y_i<\beta_1)(1-D_i)-\widehat{\varrho}_{Y_0,-}(\beta)\\ D_i-\widehat{\varrho}_{D,-}\end{pmatrix} \tag{7}$$

and

$$g_i(\beta,\theta)=\begin{pmatrix}\mathbb{I}(Y_i<\beta+\theta)-\tau-\widehat{\varrho}_{+}(\beta,\theta)\\ \mathbb{I}(Y_i<\beta)-\tau-\widehat{\varrho}_{-}(\beta)\end{pmatrix}. \tag{8}$$

The details of bias correction terms in (7) and (8) (like $\widehat{\varrho}_{Y_1,+}(\beta,\theta)$) are given in Section 7. Calonico, Cattaneo and Titiunik (2014) proposed drawing inference for the (local) average treatment effect in RD designs that allows optimal bandwidth (i.e. without using the undersmoothing condition). In this example we extend the idea to infer about the quantile treatment effect.

Example 3 (Expected shortfall). Suppose Y_t is the log return (or profit and loss, P&L) of an asset at time t. In risk management, a risk measure that receives substantial attention is the expected shortfall (or conditional VaR), defined as $\theta_0=\mathbb{E}(Y_{t+H}\mathbb{I}(Y_{t+H}\leq\beta_0)|X_t=x)/\tau$, where τ is the probability level, β_0 is the level-τ VaR (i.e. β_0 is such that $\mathbb{P}(Y_{t+H}<\beta_0|X_t=x)=\tau$) and H is the forecast horizon. The covariates X_t could contain lags of a function of Y_t (e.g. Y_t^2 or $|Y_t|$), and other exogenous variables (e.g. market indices). Estimating θ_0 and quantifying the estimation uncertainty at the forecast origin $X_t=x$ is important for real-time risk management. In our framework (1),

$$g(Y,\beta,\theta)=[\mathbb{I}(Y\leq\beta)-\tau,Y\mathbb{I}(Y\leq\beta)-\tau\theta]',$$

with $(d_\beta,d_\theta,d,d_g)=(1,1,2,2)$, $M=\mathcal{I}_2$ and $\mathcal{X}=\{x\}$. g is non-smooth in β and is smooth in θ.

Given a sample $\{(X_t,Y_t),t=1,...,T\}$, we consider p-th order ($p\geq 0$) local polynomial smoothing $w_t(\mathcal{X})=W_p((X_t-x)/h)\mathcal{I}_2$, where $t=1,...,T-H$, $W_p(u)=e_1'\widehat{S}^{-1}(1,u,\cdots,u^p)'K(u)$. Here $\widehat{S}$ is the $(p+1)$ by $(p+1)$ matrix with the (i,j)-th element $\widehat{S}_{i+j-2}$, where $\widehat{S}_k=\sum_{t=1}^{T-H}[(X_t-x)/h]^kK((X_t-x)/h)$, for $k=0,1,\cdots,2p$.

We can also consider the semiparametric single-index model which is especially useful when the

covariates are multiple: $\theta_0 = \mathbb{E}(Y_{t+H}\mathbb{I}(Y_{t+H} \leq \beta_0)|X_t'\gamma_{\theta 0} = x'\gamma_{\theta 0})/\tau$, i.e. the covariates predict the outcome through the index $X_t'\gamma_\theta$. The VaR β_0 is such that $\mathbb{E}(\mathbb{I}(Y_{t+H} \leq \beta_0)|X_t'\gamma_{\beta 0} = x'\gamma_{\beta 0}) = \tau$. Let $\widehat{\gamma}_\beta$ and $\widehat{\gamma}_\theta$ be $n^{1/2}$-consistent. Under this model, the weights are $w_t^{SP}(\mathcal{X}) = \text{diag}(W_p^\beta((X_t - x)'\widehat{\gamma}_\beta/h), W_p^\theta((X_t - x)'\widehat{\gamma}_\theta/h))$, where $W_p^\beta(u)$ is similarly defined as $W_p(u)$ above in the nonparametric model except that the elements in $\widehat{S}$ are $\widehat{S}_k = \sum_{t=1}^{T-H}[(X_t - x)'\widehat{\gamma}_\beta/h]^k K((X_t - x)'\widehat{\gamma}_\beta/h)$, and $W_p^\theta(u)$ is similarly defined as $W_p^\beta(u)$.

Example 4 (Regression kink design (RK design)). Consider the similar setting as in the Example 1 except that now D is a continuous variable. In the so-called RK design, $\mathbb{E}[D|X = x]$ is a kinked function of x at $x = c$ (i.e. non-differentiable at $x = c$). Card, Lee, Pei and Weber (2012) consider the effects of unemployment insurance benefits on unemployment duration when the unemployment insurance benefit, as a policy variable, is a kinked function (potentially with imperfect implementation or measurement errors) of previous earnings. In this (fuzzy) design, the identified effect of the continuous treatment[6] is

$$\theta_0 = \frac{\nabla\mathbb{E}[Y|X = c+] - \nabla\mathbb{E}[Y|X = c-]}{\nabla\mathbb{E}[D|X = c+] - \nabla\mathbb{E}[D|X = c-]},$$

where $\nabla\mathbb{E}[Y|X = c+] = \lim_{x\to c+}\partial\mathbb{E}[Y|X = x]/\partial x$ and others are defined similarly. The estimating functions can then be set as

$$g(Y_i, \beta, \theta) = (Y_i - \beta_1 - \theta\beta_2, D_i - \beta_2, Y_i - \beta_1, D_i)'.$$

Then (1) is satisfied at true values $\theta_0, \beta_{10} = \nabla\mathbb{E}[Y|X = c-]$, $\beta_{20} = \nabla\mathbb{E}[D|X = c+] - \nabla\mathbb{E}[D|X = c-]$, with $Y = (Y, D)'$, $(d_\beta, d_\theta, d, d_g) = (2, 1, 3, 4)$, and $M = (\mathcal{I}_3, M_1)$, where $M_1 = (0, -1, 0)'$.

To define the weight functions in (2), using the notations in Example 1, denote $\check{W}_p^+(u) = h^{-1}e_2'(\widehat{S}^+)^{-1}\varpi(u)I_i$ and $\check{W}_p^-(u) = h^{-1}e_2'(\widehat{S}^-)^{-1}\varpi(u)(1 - I_i)$, with $e_2 = (0, 1, 0, ..., 0)'$ and $\varpi(u) = (1, u, \cdots, u^p)'K(u)$. Then $w_i(\mathcal{X}) = \text{diag}(\check{W}_p^+((X_i - c)/h)\mathcal{I}_2, \check{W}_p^-((X_i - c)/h)\mathcal{I}_2)$.

In a special case of the sharp design, D is a known (perfect implementation of the policy rule) but kinked function of X; that is, $D = \kappa(X)$, where κ is a deterministic function with a kink at $X = c$. Let $\kappa_+ = \lim_{x\to c+}\nabla\kappa(x)$ and $\kappa_- = \lim_{x\to c-}\nabla\kappa(x)$. The estimating functions are simplified as $g(Y_i, \beta, \theta) = (Y_i - \beta - \theta(\kappa_+ - \kappa_-), Y_i - \beta)$. Then (1) is satisfied with $M = \mathcal{I}_2$. The weight functions are $w_i(\mathcal{X}) = \text{diag}(\check{W}_p^+((X_i - c)/h), \check{W}_p^-((X_i - c)/h))$.

[6] See Card et al. (2012, Proposition 2) for the interpretation of the identified effect.

Remark. In the examples above, estimating equations utilize directly or indirectly the closed-forms of local polynomial estimators (instead of being built on first-order conditions of locally weighted least squares). The framework also applies to the local estimators that are implicitly defined as local extremum estimators and solve local first-order conditions, e.g. local nonlinear least squares (Gozalo and Linton, 2000) and local GMM estimators (Lewbel, 2007), and local likelihood density estimators (Otsu, Xu and Matsushita, 2013).

3 Wald approach

The first product of the local estimating equations framework is the standard error of $\widehat{\theta}$. It is possible to extend the classical asymptotic theory in the GMM framework (Newey and McFadden, 1994) to cover estimators defined in (2), although usual GMM estimators are typically defined under global moment restrictions.[7] Throughout the paper the convergence is always as $n \to \infty$.

Define the Jacobian matrix $G(\beta, \theta) = M\nabla g_0(\beta, \theta) \equiv M\partial g_0(\beta, \theta)/\partial(\beta', \theta')$, and $G = G(\beta_0, \theta_0)$. Let Ω be the asymptotic variance matrix of sample local moments (which is made precise in Assumption AN in Section 4). Assume both G and Ω are non-singular. Under high-level assumptions (similar to Newey and McFadden, 1994, Sections 2 and 7),[8] we can show that $\widehat{\theta} \xrightarrow{p} \theta_0$ and

$$c_n(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Gamma), \tag{9}$$

where Γ is the lower right $d_\theta \times d_\theta$ submatrix of $G^{-1}\Omega G^{-1\prime}$. The asymptotic variance Γ is generally different from the one that assumes β_0 is known.[9]

The Wald statistic requires estimation of Γ. A generic consistent estimator of Γ is available when g is smooth, under which G can be estimated by $\widehat{G} = \sum_{i=1}^n Mw_i(\mathcal{X})\nabla g_i(\widehat{\beta}, \widehat{\theta})$, and Ω can be estimated by sample second moments (with $(\widehat{\beta}, \widehat{\theta})$ plugged in; see Assumption UC (ii) below). Such a variance estimator is generic in that it is immediate after estimating equations are determined for a specific

[7] See also Lewbel (2007) for an extension to local GMM and its relevance in applications.

[8] Modifications that adjust to local moment restrictions will be clear in Section 4 and later.

[9] For example, consider when β and θ are estimated sequentially, as in the examples illustrated above. We can then rewrite (1) as $(M_1', M_2')'g_0(\beta, \theta) = 0$, where $M = (M_1', M_2')'$, and M_2 is $d_\beta \times d_g$ such that $M_2 g_0(\beta, \theta)$ does not depend on θ. The "ideal" estimator θ that is solely based on the moment restriction $M_1 g_0(\beta_0, \theta) = 0$ (that is, treating the true value β_0 as known) has the asymptotic variance $[M_1\nabla_\theta g_0(\beta_0, \theta_0)]^{-1}\Omega_{11}[M_1\nabla_\theta g_0(\beta_0, \theta_0)]^{-1\prime}$, where Ω_{11} is the upper left $d_\theta \times d_\theta$ submatrix of Ω. Such variance that ignores the first-step estimation error (of β_0) is generally incorrect (comparing with Γ in (9)) if $M_1\nabla_\beta g_0(\beta_0, \theta_0) \neq 0$.

application, avoiding the case-by-case analysis of the variance formula.

This variance estimator has appeared earlier in the literature, under stronger assumptions and mostly without nuisance parameters. For classical local polynomial conditional mean estimator (thus linear estimating equations) under local homoskedasticity, Fan and Gijbels (1996, Section 4.3) proposed using the conditional variance estimator (which coincides with the sandwich-form variance estimator above under linearity), and argued that it stays closer to the finite sample variance than the direct plug-in approach which separately estimates each piece in the asymptotic variance formula. The latter approach, however, still dominates for practical recommendation.[10] Carroll, Ruppert and Welsh (1998) studied a setting (without nuisance parameters) that is similar to ours, and also advocated using the sandwich-form standard error of this sort.

However, estimation of G is not easy when g is nonsmooth since the analytical gradient is not available. One approach is to calculate the numerical derivative using finite-difference approximations. The choice of a step size parameter introduces noise and might affect non-trivially the asymptotic properties of the numerical derivative estimate (Hong, Mahajan and Nekipelov, 2012). In the next section, we propose an alternative method of inference that does not require variance estimation.

4 Concentrated empirical likelihood

We now consider the criterion-function-based inference and focus on empirical likelihood due to its attractive theoretical properties. In what follows we consider testing the null hypothesis $H_0 : \theta_0 = \theta_\dagger$.[11]

Let $m_i(\beta,\theta) = Mw_i(\mathcal{X})g_i(\beta,\theta)$, where $g_i(\beta,\theta) = g(Y_i,\beta,\theta)$. Let $\widehat{m}(\beta,\theta) = \sum_{i=1}^n m_i(\beta,\theta)$. For a given $(\beta,\theta) \in \mathcal{B}\times\Theta$, the empirical likelihood (EL) $\overline{L}(\beta,\theta)$ solves the constrained optimization problem $\overline{L}(\beta,\theta) = \max_{\{\pi_i:1\le i\le n\}} \Pi_{i=1}^n \pi_i$ subject to

$$\sum_{i=1}^n \pi_i m_i(\beta,\theta) = 0,\ \sum_{i=1}^n \pi_i = 1 \text{ and } \pi_i \ge 0. \tag{10}$$

The EL test statistic is defined as $\mathcal{L}_n(\beta,\theta) = -2\log[n^n\overline{L}(\beta,\theta)]$. Using the method of Lagrange

[10] For example, see Porter (2003, Section 3.5) Imbens and Lemioux (2008, Section 6), Imbens and Kalyanaraman (2012, Section 5.1), Frandsen et al. (2012, p. 387) and Marmer et al. (2014, Section 2) for variance estimation in RD designs. Card et al. (2012) is an exception which uses the conditional variance estimator.

[11] Throughout the paper, we use (β_0,θ_0) to denote true parameter values, and $(\beta_\dagger,\theta_\dagger)$ to denote the parameter values specified under the null when hypothesis testing is considered.

multiplier, the test statistic can be written as $\mathcal{L}_n(\beta,\theta) = 2\sup_\lambda P_n(\lambda,\beta,\theta)$, where $P_n(\lambda,\beta,\theta) = \sum_{i=1}^n \log(1-\lambda' m_i(\beta,\theta))$. See Owen (2001) and Kitamura (2006) for the motivation of empirical likelihood and its applications in econometrics.

The nuisance parameter β has to be estimated to form a test statistic for H_0. We will discuss two estimators in this section and the next. The first one is based on the concentrated EL estimator. Define

$$\widetilde{\beta}_C = \arg \min_{\beta\in\mathcal{B}\subset\mathbb{R}^{d_\beta}} \mathcal{L}_n(\beta,\theta_\dagger),$$

where $\theta_\dagger$ is the value in the null.[12] The behavior of $\widetilde{\beta}_C$ will be evaluated in both null and alternative hypotheses. The following assumptions are needed for asymptotic results. Denote $|\cdot|$ as the norm of a matrix or a vector.

Assumption T (True value). The true parameter satisfies $(\beta_0,\theta_0) \in int(\mathcal{B}\times\Theta)$ where $\mathcal{B}\subset\mathbb{R}^{d_\beta}$ and $\Theta\subset\mathbb{R}^{d_\theta}$ are compact and convex.

Assumption ID (Identification). β_0 uniquely solves $Mg_0(\beta,\theta)=0$, for any θ such that $\theta\to\theta_0$.

Assumption CD (Continuous differentiability). The function $(\beta,\theta)\mapsto g_0(\beta,\theta)$ is twice differentiable in a neighborhood of (β_0,θ_0), and the derivative $\nabla g_0(\beta,\theta) \equiv \partial g_0(\beta,\theta)/\partial(\beta',\theta')$ is uniformly continuous in the neighborhood. The second derivative is uniformly bounded. Assume that $G_\beta = M\nabla_\beta g_0(\beta_0,\theta_0)$ has full rank.

Assumption AN (Asymptotic normality at the true value). There exists a sequence c_n, which satisfies $c_n\to\infty$ such that $c_n\widehat{m}(\beta_0,\theta_0)\xrightarrow{d}\mathcal{N}(0,\Omega)$, where Ω is nonsingular.

Assumption UC (Uniform convergence). (i). For any θ such that $\theta\to\theta_0$, $\sup_{\beta\in\mathcal{B}}|\widehat{m}(\beta,\theta) - Mg_0(\beta,\theta)| = o_p(1)$. (ii). For any θ such that $\theta\to\theta_0$, there exists a $d\times d$ matrix $V(\beta,\theta)$ such that for any $\delta_n\to 0$, $\sup_{|\beta-\beta_0|\le\delta_n}|c_n^2\sum_{i=1}^n m_i(\beta,\theta)m_i(\beta,\theta)' - V(\beta,\theta)| = o_p(1)$, where $V(\beta,\theta)$ is continuous at (β_0,θ_0). Furthermore, assume

$$V = \Omega \tag{11}$$

where $V = V(\beta_0,\theta_0)$.

Assumption SE (Stochastic equi-continuity). $v_n(\beta,\theta)$ is stochastically equi-continuous at (β_0,θ_0), where $v_n(\beta,\theta) = c_n[\widehat{m}(\beta,\theta) - Mg_0(\beta,\theta)]$. That is, for any $\delta_n\to 0$, $\sup_{|(\beta,\theta)-(\beta_0,\theta_0)|\le\delta_n}|v_n(\beta,\theta)-$

[12] In this paper $\mathcal{L}_n(\widetilde{\beta}_C,\theta)$ (for a given θ) is referred to as the concentrated (empirical) likelihood instead of the profile (empirical) likelihood. The latter stands for $\mathcal{L}_n(\beta,\theta)$ (as standard in the EL literature, with π_i's profiled out).

$v_n(\beta_0, \theta_0)|/(1 + c_n|(\beta, \theta) - (\beta_0, \theta_0)|) = o_p(1)$.

Assumption M (Moment condition). For any θ such that $\theta \to \theta_0$, $\sup_{\beta \in \mathcal{B}, 1 \leq i \leq n} |m_i(\beta, \theta)| = o_p(c_n^{-1})$.

We comment on these assumptions. Assumption CD holds even when g is discontinuous in β. Assumptions T, ID, CD and UC(i) deliver consistency of $\widetilde{\beta}_C$ under H_0 (and the local alternative considered in Section 6). We only require a local (around β_0) uniform convergence to the second moment of g in Assumption UC(ii), in contrast to a global assumption in UC(i). Assumption UC(ii), in particular, the equality (11), is imposed to obtain the pivotal limit distribution of the test statistic. It holds for iid and weakly dependent data, as illustrated in Sections 7 and 8, in contrast to the EL test based on global unconditional or conditional moment restrictions, which typically requires the blocking technique to handle serial dependence (Kitamura, 1997, Smith, 2011). Assumption M is usually satisfied by imposing existence of moments for the outcome variable.

Assumption SE is used to derive the asymptotic distribution (in particular, for non-smooth estimating equations), and is stated slightly different from the traditional stochastic equi-continuity assumption (where the empirical process $v_n(\beta, \theta)$ is centered around the expectation; see Andrews, 1994). The convergence rate c_n is tailored to allow different applications, typically slower than $n^{1/2}$. Assumptions UC, SE and AN are high-level assumptions on data that can allow iid and time series applications. Verification of these assumptions for a specific application may require substantial work. We provide lower-level sufficient conditions for the econometric examples above in Sections 7 and 8.

Theorem 1 *Suppose the assumptions listed above hold. Then under* H_0, $\mathcal{L}_n(\widetilde{\beta}_C, \theta_\dagger) \xrightarrow{d} \chi^2(d_\theta)$.

Theorem 1 only requires Assumptions ID, UC and M hold at $\theta = \theta_0$, and CD and SE hold marginally in β (at $\theta = \theta_0$). Slightly stronger assumptions stated above facilitate the local power analysis in Section 6. Theorem 1 together with the results below under alternatives (Theorems 3 and 4) shows that the valid confidence set of θ can be obtained by inverting the concentrated EL test. The confidence set such constructed is never empty and it always includes $\widehat{\theta}$.

An intermediate result in proving Theorem 1 is

$$\mathcal{L}_n(\beta_0, \theta_0) = \widehat{m}(\beta_0, \theta_0)' \left[\sum_{i=1}^n m_i(\beta_0, \theta_0) m_i(\beta_0, \theta_0)' \right]^{-1} \widehat{m}(\beta_0, \theta_0) + o_p(1) \xrightarrow{d} \chi^2(d). \tag{12}$$

Comparing with the t-test (based on (9)), the (square of) self-normalized-sum feature of $\mathcal{L}_n$ sidesteps variance estimation (especially appearance of the derivative G), which makes it particularly desirable for testing.

A key step of showing Theorem 1 from (12) handles the approximation error $\widehat{m}(\beta,\theta_0)-\widehat{m}(\beta_0,\theta_0)$, for β in a shrinking neighborhood of β_0, by using the stochastic equi-continuity assumption (Assumption SE) instead of the non-stochastic Taylor expansion as in the classical approach when estimating equations are sufficiently smooth in β.

In the next section we show $\widetilde{\beta}_C$ is not the only estimator for $\mathcal{L}_n(\beta,\theta_0)$ to reach the chi-square limit.

5 EL with plug-in estimation

In this section we consider an alternative way of dealing with nuisance parameters, using a plug-in estimator of β, Laplace type estimator (LTE, Chernozhukov and Hong, 2003), in the test statistic $\mathcal{L}_n$. It is based on simulations instead of optimization, and is most useful when the concentrating estimator is infeasible in finite samples or requires heavy computation. These happen most often when estimating equations are discontinuous or nonconvex in β, in which the global minimum is potentially unidentifiable from the local minimum,[13] or when the dimension of β is not small, in which multivariate optimization can be computationally costly.

To define the LTE, let

$$p_n(\beta)=\frac{\exp(-\mathcal{L}_n(\beta,\theta_\dagger))\pi(\beta)}{\int_{\mathcal{B}}\exp(-\mathcal{L}_n(\beta,\theta_\dagger))\pi(\beta)\mathrm{d}\beta},$$

for a given β, be the quasi-posterior, where π is a continuous and uniformly positive density function (the quasi-prior density). Let $Q_n(\beta)=\int_{\mathcal{B}}\ell(\beta-\xi)p_n(\xi)d\xi$, where ℓ is a loss function. Define $\widetilde{\beta}_{LTE}=\arg\min_{\beta\in\mathcal{B}}Q_n(\beta)$.

The loss function $\ell:\mathcal{B}\to\mathbb{R}_+\cup\{0\}$ is convex such that $\ell(u)=0$ if and only if $u=0$, $\ell(u)\leq 1+|u|^\delta$ for some $\delta\geq 1$. The loss function is assumed to be symmetric, so that a key condition for the resulted EL statistic to obey a pivotal limit distribution is satisfied (i.e. the condition (18) in Appendix A, which is also satisfied by $\widetilde{\beta}_C$). An asymmetric loss would introduce an asymptotic bias for the LTE.

[13] Gan and Jiang (1999) proposed a test for global optima within the likelihood framework, however, their approach relies on existence of derivatives.

If the quadratic loss or the absolute deviation loss is used, $\widetilde{\beta}_{LTE}$ is then the mean or the median, respectively, of the quasi-posterior density.

Computation of $\widetilde{\beta}_{LTE}$ is based on simulations, due to its formal resemblance to the Bayesian estimator when the nonparametric likelihood (instead of the classical parametric likelihood) is used.[14] A Markov chain can be generated, using standard MCMC sampling techniques, with the stationary density approaching the quasi-posterior density $p_n(\beta)$. Dropping a sufficiently long burn-in period, the marginal density of the chain can be used to approximate $p_n(\beta)$, then $\widetilde{\beta}_{LTE}$ is calculated, e.g. as the sample mean or median of the chain. To generate each point in the chain, only evaluation of $\mathcal{L}_n$ at a given β is needed (instead of global optimization). See Chib (2001) and Chernozhukov and Hong (2003) for details.

The justification of using the LTE plug-in estimation in Theorem 2 below relies on the uniform quadratic expansion of $\mathcal{L}_n$ in a larger shrinking neighborhood of β_0 (Lemma 2 in Appendix B) than is required for the concentrating estimator (Lemma 3), so we impose the following stronger assumption than Assumption M.

Assumption M′. For any θ such that $\theta \to \theta_0$, $\sup_{\beta\in\mathcal{B},1\leq i\leq n} |m_i(\beta,\theta)| = O_p(c_n^{-2})$.

The following result shows that using the LTE plug-in estimation in $\mathcal{L}_n$ delivers the same asymptotic distribution as the concentrated EL.

Theorem 2 *Suppose the assumptions in Theorem 1 and Assumption M′ hold. Then under* H_0, $\mathcal{L}_n(\widetilde{\beta}_{LTE},\theta_\dagger) = \mathcal{L}_n(\widetilde{\beta}_C,\theta_\dagger) + o_p(1)$.

Chernozhukov and Hong (2003) introduced the LTE in the framework of extremum estimators which nests the classical empirical likelihood criterion function. The main differences with the current work are in the following. First, we build the EL upon local estimating equations which permits the analysis of nonparametric models, instead of global unconditional moment restrictions as in theirs. Correspondingly the objects of interests are different and thus applications are different,[15] as highlighted in the introduction. Second, driven by our focus on inference in the presence of nuisance parameters, we consider the QLR and LM-type tests (which are missing in their treatment) using the LTE as the constrained estimator under the null. Consequently, we only need the central tendency

[14] The Bayesian perspective of this type of estimator has been pursued in the statistical literature by Lazar (2003), Schennach (2005) and Yang and He (2012).

[15] The leading example in Chernozhukov and Hong's (2003) framework is the parametric censored quantile regression model.

of the posterior density to have correct asymptotic distribution (i.e. to mimic the distribution of $\widetilde{\beta}_C$), without requiring the tails to match the quantiles of the asymptotic distribution, as in Chernozhukov and Hong (Theorem 3, p. 308, and assumptions associated). Third, in addition to standard situations, we also consider mis-specified moment conditions and weak identification (in Sections 6 and 9) that are of particular interest in inference. Fourth, we impose weaker assumptions on data that permit applications with serial dependence, and provide sufficient conditions for high-level assumptions in a few applications.

6 Power analysis

Now we consider the behavior of tests under the alternative $H_a : \theta_0 \neq \theta_\dagger$. Under H_a, the estimators of nuisance parameters β proposed in last sections are based on mis-specified local moment restrictions. Following the literature on moment restrictions with mis-specification (Newey 1985, Hall and Inoue, 2003), we consider both local and non-local alternatives, $H_{a-loc} : \theta_0 \equiv \theta_\dagger - c_n^{-1}\xi$ v.s. $H_{a-nloc} : \theta_0 \neq \theta_\dagger$ where ξ is a d_θ-dim non-zero constant, and θ_0 is a fixed value under H_{a-nloc}. Under H_{a-nloc}, there does not exist (even asymptotically) $\beta \in \mathcal{B}$ such that $Mg_0(\beta, \theta_\dagger) = 0$.

We consider H_{a-loc} first. Denote $G(\beta, \theta) = (G_\beta(\beta, \theta), G_\theta(\beta, \theta))$, where $G_\beta(\beta, \theta) = M\nabla_\beta g_0(\beta, \theta)$ is $d \times d_\beta$, and $G_\theta(\beta, \theta) = M\nabla_\theta g_0(\beta, \theta)$ is $d \times d_\theta$. Let $G_\beta = G_\beta(\beta_0, \theta_0)$ and $G_\theta = G_\theta(\beta_0, \theta_0)$.

Theorem 3 *(i). Suppose the assumptions in Theorem 1 hold. Under H_{a-loc}, $\mathcal{L}_n(\widetilde{\beta}_C, \theta_\dagger) \xrightarrow{d} \chi^2(\mu_\xi^2, d_\theta)$, where the non-centrality parameter $\mu_\xi^2 = \xi' G_\theta' V^{-1} G_\beta \left(G_\beta' V^{-1} G_\beta\right)^{-1} G_\beta' V^{-1} G_\theta \xi$. (ii). Suppose the assumptions in Theorem 2 hold. Under H_{a-loc}, $\mathcal{L}_n(\widetilde{\beta}_{LTE}, \theta_\dagger) = \mathcal{L}_n(\widetilde{\beta}_C, \theta_\dagger) + o_p(1)$.*

The test has non-trivial local power for any $\xi \neq 0$ if G_θ has full rank (which ensures μ_ξ^2 to be of positive definite quadratic form). More assumptions are needed to establish the asymptotic behavior under H_{a-nloc}.

Assumption AL1. There exists a function $P(\lambda, \beta, \theta)$ such that for any $\theta \in \Theta$, $(\lambda, \beta) \mapsto P(\lambda, \beta, \theta)$ is continuous and differentiable, and $\sup_{\lambda,\beta} |c_n^{-2} P_n(\lambda, \beta, \theta) - P(\lambda, \beta, \theta)| = o_p(1)$.

Assumption AL2. For $P(\lambda, \beta, \theta)$ defined in Assumption AL1, there exists a unique solution (λ^*, β^*) to the saddle point problem $\min_\beta \max_\lambda P(\beta, \theta, \lambda)$ for any $\theta \in \Theta$.

Assumptions AL1 and AL2 are also used in Chen, Hong and Shum (2007).

Theorem 4 *Suppose Assumptions T, ID, AL1 and AL2 hold. (i) Under* H_{a-nloc}, $c_n^{-2}\mathcal{L}_n(\widetilde{\beta}_C, \theta_\dagger) \xrightarrow{p} 2P(\lambda^*, \beta^*, \theta_\dagger)$ *and* $P(\lambda^*, \beta^*, \theta_\dagger) > 0$. *(ii) Under* H_{a-nloc}, $c_n^{-2}\mathcal{L}_n(\widetilde{\beta}_{LTE}, \theta_\dagger) = c_n^{-2}\mathcal{L}_n(\widetilde{\beta}_C, \theta_\dagger) + o_p(1)$.

7 Quantile regression discontinuity

In this section and the next, we re-examine examples in Section 2 and consider sufficient conditions for the high-levels assumptions introduced above. We focus on the quantile RD design and the expected shortfall for asset returns (Examples 1-3). Both examples have nuisance parameters that enter the estimating equations nonsmoothly. They are different in important aspects in dealing with independent and time series data, design points that are in the interior or on the boundaries, allowing expectations across different subpopulations, and bounded or unbounded estimating equations.

In the fuzzy quantile RD design, we need the following conditions. All conditions are innocuous, and some of them (Assumptions QRD (ii) and (iii)) are also used for identification of the quantile causal effect (Assumption I, Frandsen et al., 2012).

Assumption QRD. (i). $\{X_i, Y_i, D_i\}$ are iid.

(ii). $\overline{x} \longmapsto \mu(\overline{x})$ is continuous at c, and $\mu(c) > 0$, where $\mu(\cdot)$ is the density function of X_i.

(iii). $\mathbb{E}[D|X = c+] \neq \mathbb{E}[D|X = c-]$.

(iv). Both $y \mapsto F_{Y^1|C,X=c}(y)$ and $y \mapsto F_{Y^0|C,X=c}(y)$ are strictly increasing in a neighborhood of y such that $F_{Y^1|C,X=c}(y) = \tau$ and $F_{Y^0|C,X=c}(y) = \tau$ respectively.

(v). The following functions are continuously differentiable in a neighborhood of β_{10} : $y \mapsto \mathbb{P}(Y^0 < y, D = 0|X = c+)$ and $y \mapsto \mathbb{P}(Y^0 < y, D = 0|X = c-)$. The following functions are continuously differentiable in a neighborhood of $\beta_{10} + \theta_0$: $y \mapsto \mathbb{P}(Y^1 < y, D = 1|X = c+)$ and $y \mapsto \mathbb{P}(Y^1 < y, D = 1|X = c-)$.

(vi). The following functions are $(p+1)$-th continuously differentiable in right and left neighborhoods of $\overline{x} = c$: $\overline{x} \mapsto \mathbb{P}(Y^1 < y, D = 1|X = \overline{x})$, $\overline{x} \mapsto \mathbb{P}(Y^0 < y, D = 0|X = \overline{x})$ and $\overline{x} \mapsto \mathbb{P}(D = 1|X = \overline{x})$. The $(p+1)$-th derivative $\nabla^{(p+1)}\mathbb{P}(Y^1 < y, D = 1|X = c+)$ is uniformly bounded in a neighborhood of $y = \beta_{10}+\theta_0$. The $(p+1)-th$ derivative $\nabla^{(p+1)}\mathbb{P}(Y^0 < y, D = 0|X = c-)$ is uniformly bounded in a neighborhood of $y = \beta_{10}$.

Assumption K. $K(\cdot)$ has bounded support $\mathcal{K} \subset \mathcal{R}^1$ such that $\int_{\mathcal{K}} |u^k K(u)|^\alpha du < \infty$ for $k = 0, 1, ..., 2p+1$ and some $\alpha > 0$.

Assumption BW. $(nh)^{-1} + nh^{2p+3} \to 0$.

In Appendix C we provide details of verifying the high-levels assumptions used in Sections 3-5. We briefly summarize below.

Assumption ID holds under Assumptions QRD (iii) and (iv). Assumption CD holds under Assumption QRD (v), Nonsingularity of G requires Assumption QRD (iii). Assumption UC (i) holds by Assumption K, compactness and the uniform law of large numbers, and $h \to 0$. Assumption UC (ii) holds Assumption QRD (ii). The equality (11) holds by iid data. Assumption SE holds with $c_n = (nh)^{1/2}$, by adapting the standard results (Andrews, 1994) for stochastic equi-continuity to the context of local estimating equations using Assumptions K and BW. Assumption AN follows from the asymptotic normality of the local polynomial estimators at boundary design points (Fan and Gijbels, 1996). The formula for matrices Ω and G are contained in Appendix C.

7.1 QRD with bias correction

We use the notations in Example 1 (writing $\widehat{S}^+$ as $\widehat{S}^+(h)$). For the fuzzy design, the bias correction terms are defined as

$$\begin{array}{cc}
\widehat{\varrho}_{Y_1,+}(\beta,\theta) = h^{p+1}\widehat{\psi}_{Y_1,+}(\beta,\theta;b)\widehat{B}_+/(p+1)!, & \widehat{\varrho}_{Y_1,-}(\beta,\theta) = h^{p+1}\widehat{\psi}_{Y_1,-}(\beta,\theta;b)\widehat{B}_-/(p+1)!, \\
\widehat{\varrho}_{Y_0,+}(\beta) = h^{p+1}\widehat{\psi}_{Y_0,+}(\beta;b)\widehat{B}_+/(p+1)!, & \widehat{\varrho}_{Y_0,-}(\beta) = h^{p+1}\widehat{\psi}_{Y_0,-}(\beta;b)\widehat{B}_-/(p+1)!, \\
\widehat{\varrho}_{D,+} = h^{p+1}\widehat{\psi}_{D,+}(b)\widehat{B}_+/(p+1)!, & \widehat{\varrho}_{D,-} = h^{p+1}\widehat{\psi}_{D,-}(b)\widehat{B}_-/(p+1)!,
\end{array} \tag{13}$$

where the notations are clarified as follows. We only define the quantities that use observations $X_i \geq c$, like $\widehat{B}_+$, and recognize that those using observations $X_i < c$, like $\widehat{B}_-$, are defined in the obviously similar way. $\widehat{\psi}_{Y_1,+}(\beta,\theta;b)$, $\widehat{\psi}_{Y_0,+}(\beta;b)$ and $\widehat{\psi}_{D,+}(b)$ are the local $(p+1)$-th polynomial estimators of the $(p+1)$-th right derivatives (at $x=c$) $\psi_{Y_1,+}(\beta,\theta) = \nabla_x^{p+1}\mathbb{E}(\mathbb{I}(Y_i < \beta_1+\theta)D_i|x=c+)$, $\psi_{Y_0,+}(\beta) = \nabla_x^{p+1}\mathbb{E}(\mathbb{I}(Y_i < \beta_1)(1-D_i)|x=c+)$ and $\psi_{D,+} = \nabla_x^{p+1}\mathbb{E}(D_i|x=c+)$ respectively.[16] The bandwidth $b \to 0$, which is in general different from h, is used in the derivative estimation above. The other quantity $\widehat{B}_+$ (which does not depend on outcome variables or parameters β or θ) is defined as $\widehat{B}_+ = e_1'[\widehat{S}^+(h)/(nh)]^{-1}X_p(h)'\Phi_+(h)s_{p+1}(h)/(nh)$, where $X_p(h)$ is the $n \times (p+1)$ matrix with (i,j)-element $((X_i-c)/h)^{j-1}$, $\Phi_+(h)$ is the $n \times n$ diagonal matrix with elements $I_iK((X_i-c)/h)$,

[16] Card et al. (2012) illustrated the reason of not using the local $(p+2)$-th polynomial to estimate the $(p+1)$-th derivative at a boundary point when p is odd.

and $s_{p+1}(h) = [((X_1 - c)/h)^{p+1}, ..., ((X_n - c)/h)^{p+1}]'$.

For the sharp design, the bias correction terms are

$$\widehat{\varrho}_+(\beta, \theta) = h^{p+1}\widehat{\psi}_{Y_1}(\beta, \theta; b)\widehat{B}_+/(p+1)!,\ \widehat{\varrho}_-(\beta) = h^{p+1}\widehat{\psi}_{Y_0}(\beta; b)\widehat{B}_-/(p+1)!,$$

where $\widehat{\psi}_+(\beta, \theta; b)$ and $\widehat{\psi}_{Y_0}(\beta; b)$ are the local $(p+1)$-th polynomial estimators of the $(p+1)$-th right derivatives (at $x = c$) $\psi_{Y_1}(\beta, \theta) = \nabla_x^{p+1}\mathbb{P}(Y_i < \beta + \theta | x = c+)$ and $\psi_{Y_0}(\beta) = \nabla_x^{p+1}\mathbb{P}(Y_i < \beta | x = c-)$ respectively, and $\widehat{B}_+$ and $\widehat{B}_-$ are defined as above.

Assumption BW′. $(nh)^{-1} + nh^{2p+3}b^2 + h/b \to 0$.

We can show that high-level assumptions that validate Theorems 1 and 2 are satisfied when bias correction is incorporated, under Assumptions QRD, K and BW′ (which is weaker than Assumption BW on h). We provide details in Appendix D.

In Assumption BW′, the conditions $nh^{2p+3}b^2 \to 0$ and $h/b \to 0$ are needed for bias correction terms not to introduce additional non-negligible bias and variance respectively (so that (11) holds).[17]

Suppose $p = 1$ (local linear estimation of θ) and $b \sim n^{-1/7}$ (of the optimal order for local quadratic estimators of the second derivatives). Then Assumption BW′ requires that $h \sim n^r$, where $r \in (-1, -1/7)$, and the usual optimal bandwidth ($r = -1/5$) is allowed.

Calonico, Cattaneo and Titiunik (2014) studied inference of local average treatment effect in RD and RK designs, and also aimed to resolve the undersmoothing condition. They use the similar bias correction as in (13) and proposed robust standard errors which are valid for a wide range of b. In particular, when local linear estimation of θ is used, they allow the optimal bandwidth $h \sim n^{-1/5}$, and the simple choice $b = h$ (which leads to inconsistent estimation of second derivatives). They also recognize the MSE-optimal joint selection of h and b requires $h/b \to 0$.

The conditional arguments used in Calonico, et al. (2014) are not easily extendable to the quantile effect due to the implicit form (no-closed-form) of the estimator. The concern about additional variability induced by bias correction is reduced (compared to the traditional Wald approach), as in the local-estimating-equation approach we take, the values of θ and β in derivative estimators (like $\widehat{\psi}_{Y_1}(\beta, \theta; b)$) in bias-correction terms are adopted from the null and concentrated out respectively (instead of both being estimated).

[17] Note that Assumption BW′ implies $nh^{2p+5} \to 0$ which ensures the smaller-order bias (which we did not correct) to be asymptotically negligible.

The estimating function $g_i(\beta,\theta)$ as in (7) or (8) is not the only way to incorporate bias correction. For the sharp design (similarly for the fuzzy design), we could consider the implicit bias correction with

$$g_i(\beta,\theta)=\begin{pmatrix}\mathbb{I}(Y_i<\beta+\theta)-\tau-[\widehat{\varrho}_{Y_1}(X_i;\beta,\theta)-\widehat{\varrho}_{Y_1}(c;\beta,\theta)]\\ \mathbb{I}(Y_i<\beta)-\tau-[\widehat{\varrho}_{Y_0}(X_i;\beta)-\widehat{\varrho}_{Y_0}(c;\beta)]\end{pmatrix},$$

where $\widehat{\varrho}_{Y_1}(x;\beta,\theta)$ is the local p-th polynomial estimator (using the bandwidth h) of $\varrho_{Y_1}(x;\beta,\theta)=\mathbb{P}[Y_i<\beta+\theta|X=x]$ for $x\in[c,c+h]$ using the observations such that $X_i\geq c$, and $\widehat{\varrho}_{Y_0}(x;\beta)$ is the local p-th polynomial estimator of $\varrho_{Y_0}(x;\beta)=\mathbb{P}[Y_i<\beta|X=x]$ for $x\in[c-h,c]$ using the observations such that $X_i<c$. This approach does not need to estimate the derivatives (thus no need for an extra bandwidth), and was followed by Xue and Zhu (2007) and Xu (2013) in different settings in which, though, nuisance parameters were neither considered or removed efficiently.[18] Although this implicit approach is nicely motivated by aiming to include the term $\varrho_{Y_1}(X_i;\beta,\theta)-\varrho_{Y_1}(c;\beta,\theta)$ in estimating equations (so that estimating equations are unbiased if the term is known), negligible effects of estimating this term require $nh^{2p+3}=O(1)$, which is stronger than Assumption BW′. It is also computationally more taxing than the approach based on (8) (involving plug-in of infinite-dimensional nuisance parameters).

8 Expected shortfall

We now turn to verifying high-level assumptions for Example 3. Denote $F(y|x)$ as the conditional CDF of Y_{t+H} given $X_t=x$, and $f(y|x)$ as the corresponding density function. Assume the conditions for the kernel and the bandwidth as in Assumptions K and BW hold.

Assumption ES.

(i) $\{X_t,Y_t\}$ is stationary and $\phi-$mixing with mixing coefficients decaying at an exponential rate.

(ii) There exists $a>2$ and $B(x)$, a neighborhood of x, such that $\sup_{\beta\in\mathcal{B},\overline{x}\in B(x)}\mathbb{E}(|Y_{t+H}|^a|X_t=\overline{x})<C$.

(ii') $\sup_{1\leq t\leq T-H}|Y_tK((X_i-x)/h)|<C$ for any $h>0$.

(iii) $\overline{x}\longmapsto\mu(\overline{x})$ is continuous at x, and $\mu(x)>0$, where $\mu(\cdot)$ is the density function of X_t.

[18] Xue and Zhu (2007) applied the local constant approach to the varying coefficient model, and in forming the test statistic (in their Section 4.2) they replaced nuisance parameters by the corresponding point estimates (like $\widehat{\beta}$ in our setting) instead of using concentration. Xu (2013) considered the nonparametric quantile regression for time series with no nuisance parameters.

(iv) $y \mapsto F(y|x)$ is strictly increasing.

(v) $y \mapsto F(y|x)$ is continuously differentiable, and $f(y|x) > 0$ for any $y \in \mathbb{R}^1$. $\overline{x} \mapsto F(y|\overline{x})$ is $(p+1)-th$ continuously differentiable at x, and its $(p+1)-th$ derivative $F^{(p+1)}(y|x)$ is uniformly bounded in y.

Assumption ID holds by (iv). Assumption CD holds by (v). Assumptions M and M' hold under Assumption ES (ii) and ES (ii') (which implies ES (ii)) respectively. Assumption SE holds with $c_n = (nh)^{1/2}$ using the results in Andrews (1993) under the mixing condition (Assumption ES (i)). The equality (11) holds by the mixing condition in Assumption ES (i). The intuition behind the result that the chi-square limit still holds under weak dependence without appealing to the blocking or smoothing technique (Kitamura, 1997, Anatolyev, 2005, Smith, 2011) is that local smoothing in the state domain weakens serial dependence (Fan and Yao, 2003). Appendix E provides details of verifying Assumptions UC, SE and M.

9 Weak identification

Now we consider the behavior of the tests when the parameter of interest θ is weakly identified, which is made precise in Assumption CD_W below. This might happen in certain applications, and when it does, θ_0 is vaguely distinguished from other values in Θ and $g_0(\beta, \theta)$ is relatively flat in the neighborhood of θ_0. It leads the Jacobian matrix G to be close to singular, and the generic standard error-based t-test induced from (9) is problematic.

As an example, in the fuzzy regression discontinuity design, the local average treated effect is weakly identified when the jump in the treatment probability is close to zero (Marmer, Feir and Lemieux, 2014).[19] Similarly, in the regression kink design (Example 3), θ_0 is weakly identified when β_{20} is close to zero.

The EL test statistics (with concentration or LTE plug-in) preserve the same asymptotic distribution under the null under weak identification; Theorems 1 and 2 still hold since their assumptions do not require strong identification of θ_0. The tests, however, have lower (possibly trivial) power under the local alternative since G_θ (which enters the local power function in Theorem 3) is of a smaller magnitude under weak identification. We thus consider a larger deviation from the null (but

[19] Marmer et al. (2014) proposed a robust t-test of the local mean effect in the fuzzy RD design to weak identification based on the explicit variance formula (see the Introduction and Section 3).

is not too large for the strongly identified parameters to uniquely solve the estimating equations approximately). To establish useful power properties, we make the following assumptions.

In what follows, we consider a more general setting which allows part of nuisance parameters in β is also weakly identified. Partition $\beta = (\beta_s', \beta_w')'$ where β_s and β_w are $d_{\beta_s} \times 1$ and $d_{\beta_w} \times 1$, respectively. We assume β_s is strongly identified while β_w and θ are weakly identified. Partition G accordingly as $G = (G_{\beta_s}', G_{\beta_w,\theta}')'$.

Assumption ID_W. β_{s0} uniquely solves $Mg_0(\beta_s, \beta_w, \theta) = 0$, for any $(\beta_w, \theta) \in \mathcal{B}_w \times \Theta$.

Assumption CD_W. Assumption CD holds except that $G_{\beta_w,\theta}$ (thus G) is nearly singular as $n \to \infty$; i.e. there exists $\overline{G}_{\beta_w,\theta}$ with full rank such that $G_{\beta_w,\theta} = \alpha_n^{-1}\overline{G}_{\beta_w,\theta}$, where $\alpha_n \to \infty$ and $\lim_{n\to\infty} c_n^{-1}\alpha_n < \infty$.

Assumption UC_W. (i). For any $(\beta_w, \theta) \in \mathcal{B}_w \times \Theta$, $\sup_{\beta_s \in \mathcal{B}_s} |\widehat{m}(\beta, \theta) - Mg_0(\beta, \theta)| = o_p(1)$. (ii). For any $(\beta_w, \theta) \in \mathcal{B}_w \times \Theta$, and the sequence c_n in Assumption AN, there exists a $d \times d$ matrix $V(\beta, \theta)$ such that for any $\delta_n \to 0$, $\sup_{|\beta_s - \beta_{s0}| \le \delta_n} |c_n^2 \sum_{i=1}^n m_i(\beta, \theta) m_i(\beta, \theta)' - V(\beta, \theta)| = o_p(1)$, where $V(\beta, \theta)$ is continuous at (β_0, θ_0). Furthermore, assume $V = \Omega$ where $V = V(\beta_0, \theta_0)$.

Assumption SE_W. $v_n(\beta, \theta)$ is stochastically equi-continuous at (β_0, θ_0), where $v_n(\beta, \theta) = c_n[\widehat{m}(\beta, \theta) - Mg_0(\beta, \theta)]$. That is, for any $\delta_n \to 0$, $\sup_{|\beta_s - \beta_{s0}| \le \delta_n, (\beta_w, \theta) \in \mathcal{B}_w \times \Theta} |v_n(\beta, \theta) - v_n(\beta_0, \theta_0)|/(1 + c_n|(\beta, \theta) - (\beta_0, \theta_0)|) = o_p(1)$.

Assumption M_W. For any $(\beta_w, \theta) \in \mathcal{B}_w \times \Theta$, $\sup_{\beta_s \in \mathcal{B}_s, 1 \le i \le n} |m_i(\beta, \theta)| = o_p(c_n^{-1})$.

Assumption M′_W. For any $(\beta_w, \theta) \in \mathcal{B}_w \times \Theta$, $\sup_{\beta_s \in \mathcal{B}_s, 1 \le i \le n} |m_i(\beta, \theta)| = O_p(c_n^{-2})$.

In Assumption CD_W, α_n reflects the degree of weak identification. It is usually satisfied by assuming the parameter that determines the identification strength to be in a α_n^{-1} neighborhood of the point that causes weak identification. In the regression kink design, it is achieved by $\beta_{20} = \alpha_n^{-1}\zeta_c$, where ζ_c is a non-zero constant. Like Assumption CD_W, other assumptions above also have flavor of weak identification as they hold in entire parameter spaces for parameters that are weakly identified (which is assumed for simplicity, although not completely necessary if $c_n^{-1}\alpha_n \to 0$), not only in shrinking neighborhoods of true values as in their counterparts in Sections 4-6. As in the strongly identified case, Assumption M_W is reinforced as M′_W for the results for the EL test with LTE plug-in.

Consider the joint hypothesis $\overline{H}_0$ and its alternative: $\overline{H}_0 : \beta_{w0} = \beta_{w\dagger}$, $\theta_0 = \theta_\dagger$ v.s. $\overline{H}_a : \beta_{w0} = \beta_{w\dagger} - c_n^{-1}\alpha_n\zeta_w$, $\theta_0 = \theta_\dagger - c_n^{-1}\alpha_n\xi$. The following result gives the asymptotic distribution under $\overline{H}_a$.

Theorem 5 *Let $\widetilde{\beta} = (\widetilde{\beta}_s', \widetilde{\beta}_w')'$ be either $\widetilde{\beta}_C$ or $\widetilde{\beta}_{LTE}$. Suppose that Assumptions T, AN, and the assumptions stated in this section hold. Under $\overline{H}_a$, we have*

$$\mathcal{L}_n(\widetilde{\beta}_s, \beta_{w\dagger}, \theta_\dagger) \xrightarrow{d} \chi^2(\zeta_{w,\theta}' \overline{G}_{\beta_w,\theta}' V^{-1} G_{\beta_s} (G_{\beta_s}' V^{-1} G_{\beta_s})^{-1} G_{\beta_s}' V^{-1} \overline{G}_{\beta_w,\theta} \zeta_{w,\theta}, d_{\beta_w} + d_\theta),$$

where $\zeta_{w,\theta} = (\zeta_w, \xi)'$.

Consider the leading case when all nuisance parameters are strongly identified (i.e. $d_{\beta_w} = 0$). The EL test (with concentration or LTE plug-in) with the same critical value as in the strong identification case still has the correct null rejection probability. Under the fixed alternative, the EL test is consistent when $\alpha_n/c_n \to 0$, and has non-unity (while non-trivial) power when $\alpha_n = c_n$. It has trivial power in the entire parameter space when $\alpha_n/c_n \to \infty$ (identification is too weak). The inverted confidence interval is consequently longer than the strong identification case. The practically relevant implication is that the EL-based inference is robust to the identification strength (of θ) and provides confidence sets with covering areas that automatically reflect the identification strength.

In the setting of global and smooth moment restrictions, Guggenberger and Smith (2005) and Otsu (2006) established the similar robust results of EL using Stock and Wright (2000)-type separable weak identification assumption.[20] As noted in Andrews and Cheng (2012), such separable weak identification assumption is not directly applicable to the case when the parameter that determines the identification strength also enters the criterion function, which we want to allow for in our context. Another difference from the earlier work on EL under weak identification is that the statistic used in our setting is a likelihood ratio (not merely an LM-type statistic, due to the exact identification nature of the setting), which is crucial for empirical likelihood based inference (Kitamura, 2006, Section 6.4).

If $d_{\beta_w} \geq 1$, the EL test is generally invalid, since part of nuisance parameters, β_w, is inconsistently estimated (incorrectly eliminated) due to weak identification. A simple confidence set for θ that controls the coverage probability can be formed by the projection method, i.e. forming a joint confidence set for (β_w, θ) by inverting $\mathcal{L}_n(\widetilde{\beta}_s, \beta_{w\dagger}, \theta_\dagger)$ and then projecting it to $\mathbb{R}^{d_\theta}$. The corresponding projection-based ϵ-level test rejects $H_0 : \theta_0 = \theta_\dagger$ if $\inf_{\beta_w} \mathcal{L}_n(\widetilde{\beta}_s, \beta_w, \theta_\dagger) > \chi^2_{1-\epsilon}(d_{\beta_w} + d_\theta)$.

[20] Guggenberger and Smith (2005) also proposed a score-type test statistic (similar to Kleibergen, 2005) that is based the derivative of the EL criterion function. It can be extended to our setting and has $\chi^2(d_{\beta_w})$ under $\overline{H}_0$ if estimating equations are smooth in β, although it is not obvious without smoothness.

10 Monte Carlo simulations

We consider simulation experiments to illustrate the finite-sample performance of the non-smooth EL test with bias correction in the sharp quantile regression discontinuity design. The data generating processes (DGP) are those in Calonico et al. (2014, Section 6, Models 1 and 3), where we use the identical intercepts from the left and the right corresponding to the null hypothesis $\theta_0 = 0$. They are called DGP 1 and DGP 2 respectively in this section. DGP 1 was also used in Imbens and Kalyanaraman (2012) among others, and was obtained by fitting piecewise fifth-order global polynomials for $X_i > 0$ and $X_i < 0$ (so $c = 0$) to Lee's (2008) U.S. House elections data. DGP 2 adjusts the global curvature of DGP 1 (by adjusting coefficients) and thus increases estimation biases when a large bandwidth is used.

We are interested in testing for zero local median treatment effect ($\tau = 0.5$). Both EL tests based on bias-corrected local estimating equations and the uncorrected ones (i.e. (8) and (5) respectively) are considered. We use the local linear fit ($p = 1$) to construct the point estimator (thus the weights $w_i(\mathcal{X})$), and the local quadratic fit to estimate the second derivative in the bias term. To investigate the effects of using different bandwidths, we set $h = C_{bw}\widehat{\sigma}_X n^{-1/5}$, and for derivative estimation in bias correction terms $b = C_{bw}\widehat{\sigma}_X n^{-1/7}$, where $C_{bw} \in \{1, 2, 3, 3.5\}$ and $\widehat{\sigma}_X$ is the standard deviation of $\{X_i\}$. The ranges for h and b are large enough to include the finite-sample MSE-optimal bandwidths with and without bias correction. The sample size is $n = 500$. To remove the nuisance parameter in EL statistic, we consider both the concentrating-out and LTE procedures. When evaluating the minimum in the concentration step, we consider two searching algorithms that can handle non-smooth criterion function, the Nelder-Mead simplex method (with the initial value $\widehat{\beta}$ as in (2)) which may settle at a local minimum, and the grid search which can find the global minimum.

Figures 1 and 2 show the finite-sample null rejection rates at various nominal levels 0.5%-10% of the uncorrected and bias-corrected EL tests (both coupled with the grid search). Both uncorrected and corrected tests under-reject for $C_h = 1$ and $C_h = 2$, and over-reject when $C_h = 3$ and $C_h = 3.5$, for DGPs 1 and 2.

The uncorrected test works fine in some cases but can have large size distortion in other cases, and is relatively sensitive to the smoothing bandwidth. The bias-corrected test has less size distortion in almost all cases considered in experiments, and is overall fairly robust to the bandwidth used.

The most striking results are for DGP 2 when $C_h = 3$ and $C_h = 3.5$, in which the 5%-level uncorrected test, for example, has the actual size about 14.5% and 30.0% respectively, while the size for the bias-corrected test is 7.0% and 10.8% respectively (Figure 2 (c) and (d)). We find that the large size distortion of the uncorrected test is mostly attributed to the estimation bias (Table 1). Table 1 also gives the bias, standard deviation and root mean square error (RMSE) for the bias corrected point estimator and the uncorrected estimator. It shows the bias correction works properly in reducing the estimation bias (especially for DGP 2), thus improves the null rejection rate of the associated test.

Figures 3 and 4 focus on the bias-corrected test with different ways to remove the nuisance parameter. Along with the grid search, we also implement the Nelder-Mead search (using $\widehat{Q}_{Y^0|X=c}(\tau)$ as the initial value), the LTE and the stochastic search, where the last ones are MCMC-based. The random walk Metropolis-Hasting algorithm we implement generates the $(i+1)$-th draw $\beta^{(i+1)}$ from the transition density $\mathcal{N}(\beta^{(i)}, (1/15)^2)$, where the standard deviation 1/15 is selected so that the acceptance rate in generating the Markov chain is kept within the reasonable range (about 20%-40% on average; see Table 1). For each realization, a Markov chain with 1000 observations is generated starting from the initial value, which is set as the Nelder-Mead estimator of β. The first 15% observations are treated as the burn-in period. The LTE estimator of β is obtained as the posterior median. The stochastic search uses the minimum value which the Markov chain travels through.

We find that the test based on the Nelder-Mead search over-rejects (sometimes seriously) in all cases which is consistent with the fact that the search settles at a local minimum for a portion of replications, and is very sensitive to the initial value. The performance is improved by replacing the search with the LTE. The LTE-based test generally over-rejects more often than the test with the grid search which is hardly surprising since the latter test is based on a smaller statistic. The stochastic search works very well in finding the global minimum (also indistinguishable with the grid search). We also alter the length of Markov chains generated and find that the rejection rates settle down very quickly and converge as the length of the chain increases. This finding is robust and is still true when the initial value is set as an biased estimator of β (e.g. the unconditional quantile for the controlled group).

11 An empirical example

In this section we consider an empirical application to the effects of being placed on the academic probation status for college students on their subsequent performance. Lindo, Sanders and Oreopoulos (2010) explore the rule that students whose first-year GPAs are below a cutoff point are required to be placed on academic probation under the sharp RD design, and we use the same dataset for the large Canadian university. We follow Lindo et al. (2010) to only use observations on students whose first-year GPAs fall within $h = 0.6$ and $h = 0.3$ windows of the cutoff point, which contain 11258 (with 4166 male and 7092 female students) and 5489 observations (with 2039 male and 3450 female students) respectively. The outcome variable is the GPA in the next session which can be in the summer or in the second year. In their Section D (Table 5, p. 110), Lindo et al. report the average treatment effect estimates and find they are all statistically highly significant.

The methodological contributions of this section include examination of the effects in different quantiles of the population and implementation of bias correction in inference as described in Examples 1 and 2 (Section 7). The data in the neighborhood of $h = 0.6$ are shown in Figure 5.

Main results are shown in Figures 6 and 7. Overall, being placed on academic probation benefits students in lower quantiles (the low-ability group) more than those in higher quantiles (the high-ability group); see Figure 6. The estimated quantile effects decrease monotonically (as the rank τ increases) from about 0.32 to 0.19 grade points in next sessions (when $h = 0.6$ is used), which integrate to an average effect estimate[21] close to 0.23, as reported in Lindo et al. (2010). The downslope pattern of quantile treatment effects tends to reflect the incentive (for the group of students around the cutoff) is mainly driven by passing the cutoff GPA point for the probation status to be released in the subsequential term instead of doing their best to achieve high grades. The quantile effects are about 0.02-0.05 lower when $h = 0.3$ is used. The estimates are highly significant when $h = 0.6$, and are most significant for the middle quantiles. A smaller bandwidth yields the statistic that is less significant. The test statistics reported in this section are based on the bias-corrected local moment restriction (as in (8), with the Epanechnikov kernel), and the concentrated nonsmooth empirical likelihood (i.e. $\mathcal{L}_n(\widetilde{\beta}_C, \theta_\dagger)$) coupled with one-dimensional grid search in the concentration step.

Now we consider two groups for male and female students. While the pattern for quantile effects

[21] This estimate (the so-called composite quantile estimate) is generally different from the usual average effect estimate. Kai et al. (2010) and Zhao and Xiao (2014) argued it could be more efficient than the usual estimate for non-normal data when the number of quantiles and their weights are properly chosen.

and their significance for female students is similar (with even sharper monotonicity in the rank τ) to the overall population, the results for male students are quite different; see Figure 7. The effects of being put in academic probation is the lowest for the mid-range group, and remain at the similar level for groups at two ends. The significance is also much lower than the overall and the female counterpart, and we find the effects at almost all quantiles are insignificant at 5% level when a smaller bandwidth $h = 0.3$ is used. The findings echo the literature on gender differences in response to educational incentives (e.g. Angrist et al., 2009; see also Curto and Fryer, 2015), in that men are less responsive than women to such an academic warning and have a mixed (less pronounced) pattern of heterogenous effects.

12 Conclusion

In this paper we demonstrate the versatility and generality of the framework of local moment restrictions and their sample analogs local estimating equations using several recent popular models in policy evaluation based on discontinuities or kinks and in real-time risk forecast. We consider the standard error-based Wald-type and the criterion function-based QLR/LM-type approaches to inference, and the focus is given to the empirical likelihood. We establish general conditions which lead to asymptotically pivotal statistics, and break them down for a few applications. The non-standard issues that we are able to handle under high-level assumptions include presence of nuisance parameters, non-differentiability of the criterion function, non-negligible bias due to local smoothing, and weak identification when a model assumption is barely satisfied. The method we advocate has advantages in certain aspects over those more classical and in wide use in the literature as shown in details in the paper, and comes with a computational cost (e.g. when obtaining a confidence set) that also applies to other criterion function-based approaches.

13 Appendix A: Proofs for the main results

In this section, C and C_1 are generic bounded positive constants. In places when both β and θ are arguments of a function, we do not write θ explicitly when $\theta = \theta_0$, e.g. we write $m_i(\beta) = m_i(\beta, \theta_0)$, $\widehat{m}(\beta) = \widehat{m}(\beta, \theta_0)$, etc.

Let $\rho(v) = \log(1-v)$. Then $P_n(\lambda, \beta, \theta) = \sum_{i=1}^n [\rho(\lambda' m_i(\beta, \theta)) - \rho(0)]$ and $\mathcal{L}_n(\beta, \theta) = 2 \sup_{\lambda \in \Lambda_n(\beta,\theta)} P_n(\lambda, \beta, \theta)$.

Let $\rho_1(v)$ and $\rho_2(v)$ be the first and second derivatives of ρ. Note that $\rho_1(0) = \rho_2(0) = -1$.

Two expressions below in (14) and (15) will be useful, which are true for any $\beta \in \mathcal{B}$ and $\theta \in \Theta$. Denote $\lambda(\beta,\theta) = \arg\max_\lambda P_n(\lambda,\beta,\theta)$. Then $\lambda(\beta,\theta)$ satisfies the FOC $0 = \sum_{i=1}^n \rho_1(\lambda(\beta,\theta)' m_i(\beta,\theta)) m_i(\beta,\theta) = \sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta) - \widehat{V}_1(\beta,\theta)[c_n^{-2}\lambda(\beta,\theta) - \lambda_0]$, or

$$c_n^{-2}\lambda(\beta,\theta) - \lambda_0 = \widehat{V}_1(\beta,\theta)^{-1} \sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta), \tag{14}$$

where $\widehat{V}_1(\beta,\theta) = -c_n^2 \sum_{i=1}^n \rho_2(\dot{\lambda}' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta) m_i(\beta,\theta)'$. We used a Taylor expansion of ρ_1 at λ_0, and $\dot{\lambda}$ is a point between λ_0 and $c_n^{-2}\lambda(\beta,\theta)$.

A second-order Taylor expansion of $\mathcal{L}_n(\beta,\theta) = 2[\sum_{i=1}^n \rho(\lambda(\beta,\theta)' m_i(\beta,\theta)) - n\rho(0)]$ at λ_0 yields

$$\begin{aligned}
& \mathcal{L}_n(\beta,\theta) \\
= \;& 2[\sum_{i=1}^n \rho(\lambda_0' c_n^2 m_i(\beta,\theta)) - n\rho(0)] + 2\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta))[c_n^{-2}\lambda(\beta,\theta) - \lambda_0]' c_n^2 m_i(\beta,\theta) \\
& -c_n^2 [c_n^{-2}\lambda(\beta,\theta) - \lambda_0]' \widehat{V}_2 [c_n^{-2}\lambda(\beta,\theta) - \lambda_0] \\
\overset{(14)}{=} \;& 2[\sum_{i=1}^n \rho(\lambda_0' c_n^2 m_i(\beta,\theta)) - n\rho(0)] \\
& +2c_n^2 [\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta)]' \widehat{V}_1^{-1} [\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta)] \\
& -c_n^2 [\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta)]' \widehat{V}_1^{-1} \widehat{V}_2 \widehat{V}_1^{-1} \cdot [\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta)] \\
= \;& 2[\sum_{i=1}^n \rho(\lambda_0' c_n^2 m_i(\beta,\theta)) - n\rho(0)] + c_n [\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta)]' \cdot \\
& \cdot \left[2\widehat{V}_1^{-1} - \widehat{V}_1^{-1}\widehat{V}_2\widehat{V}_1^{-1}\right] c_n [\sum_{i=1}^n \rho_1(\lambda_0' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta)],
\end{aligned} \tag{15}$$

where $\widehat{V}_2(\beta,\theta) = -c_n^2 \sum_{i=1}^n \rho_2(\ddot{\lambda}' c_n^2 m_i(\beta,\theta)) m_i(\beta,\theta) m_i(\beta,\theta)'$ and $\ddot{\lambda}$ is a point between λ_0 and $c_n^{-2}\lambda(\beta,\theta)$.

In particular, if $\lambda_0 = 0$, (14) and (15) reduce to

$$\lambda(\beta,\theta) = -c_n^2 \widehat{V}_1(\beta,\theta)^{-1} \widehat{m}(\beta,\theta), \tag{16}$$

$$\mathcal{L}_n(\beta,\theta) = c_n \widehat{m}(\beta,\theta)' \left[2\widehat{V}_1^{-1} - \widehat{V}_1^{-1}\widehat{V}_2\widehat{V}_1^{-1}\right] c_n \widehat{m}(\beta,\theta), \tag{17}$$

which will be useful when (β,θ) is close to (β_0,θ_0).

Proof of Theorem 1. It follows from Theorem 3 (i) when $\xi = 0$.

Proof of Theorem 2. It follows from Theorem 3 (ii) when $\xi = 0$.

Proof of Theorem 3. The theorem is true for any plug-in estimator $\widetilde{\beta}$ which is consistent, and satisfies the following condition under H_{a-loc} :

$$c_n(\widetilde{\beta} - \beta_0) = -\left(G_\beta' V^{-1} G_\beta\right)^{-1} G_\beta' V^{-1}[c_n \widehat{m}\left(\beta_0, \theta_0\right) + G_\theta \xi] + o_p\left(1\right). \tag{18}$$

We write $V = V^{1/2} V^{1/2\prime}$. Since $\widetilde{\beta}$ is c_n-consistent by (18), the results (33), (34) and (35) in Lemma 3 hold. Checking the term $\widehat{m}(\widetilde{\beta}, \theta_\dagger)$, we have

$$\begin{aligned}
c_n \widehat{m}(\widetilde{\beta}, \theta_\dagger) &\overset{(33)}{=} c_n G_\beta(\widetilde{\beta} - \beta_0) + G_\theta \xi + c_n \widehat{m}(\beta_0) + o_p(1) \\
&\overset{(18)}{=} [-G_\beta \left(G_\beta' V^{-1} G_\beta\right)^{-1} G_\beta' V^{-1} + I_d][c_n \widehat{m}(\beta_0) + G_\theta \xi] + o_p(1) \\
&\overset{NT}{=} \underbrace{[-G_\beta \left(G_\beta' V^{-1} G_\beta\right)^{-1} G_\beta' V^{-1} + I_d]}_{=A} \Phi_d + o_p(1) \\
&: \quad = A\Phi_d + o_p(1),
\end{aligned} \tag{19}$$

where Φ_d is a $d-$dim $\mathcal{N}(G_\theta \xi, V)$ random variable. Then $\mathcal{L}_n(\widetilde{\beta}, \theta_\dagger) \overset{(34)}{=} c_n^2 \widehat{m}(\widetilde{\beta}, \theta_\dagger)' V^{-1} \widehat{m}(\widetilde{\beta}, \theta_\dagger) + o_p(1) \overset{(19)}{=} \Phi_d' A' V^{-1} A \Phi_d + o_p(1)$. Theorem 3 follows from the fact that $A' V^{-1} A V$ is idempotent with rank d_θ.

(i). $\widetilde{\beta}_C$ is consistent by Lemma 4. The proof then proceeds by showing that (18) holds for $\widetilde{\beta}_C$. It is based on Lemmas 1, 3 and 4 listed and proved in the next section.

Recall $\widetilde{\beta}_C = \arg\min_\beta \mathcal{L}_n(\beta, \theta_\dagger)$ and define $\check{\beta} = \arg\min_\beta A_n(\beta, \theta_\dagger)$, where A_n is the quadratic approximation defined in Lemma 2. Noticing the quadratic form of $A_n(\beta, \theta_\dagger)$, we have $c_n(\check{\beta} - \beta_0) = -\left(G_\beta' V^{-1} G_\beta\right)^{-1} G_\beta' V^{-1}[c_n \widehat{m}\left(\beta_0, \theta_0\right) + G_\theta \xi]$. By this and (38) in Lemma 4, both $\check{\beta}$ and $\widetilde{\beta}_C$ belong to the c_n^{-1} neighborhood of β_0 with probability approaching one. Rewrite $\widetilde{\beta}_C = \beta_0 + c_n^{-1}\widetilde{b}$ and $\check{\beta} = \beta_0 + c_n^{-1}\check{b}$. Then $\widetilde{b} = \arg\min_b \mathcal{L}_n(\beta_0 + c_n^{-1} b, \theta_\dagger)$ and $\check{b} = \arg\min_b A_n(\beta_0 + c_n^{-1} b, \theta_\dagger)$. By Lemma 1, $\widetilde{b} - \check{b} = o_p(1)$. Note that in the assumptions of Lemma 1, (21) holds by (35), and $b \mapsto A_n(\beta_0 + c_n^{-1} b, \theta_\dagger)$ is continuous in probability. So $\widetilde{\beta}_C$ satisfies (18).

(ii). Note that under H_{a-loc}, $p_n(\beta_0) = O_p(1)$ (by (26) in Lemma 2, setting $\zeta = 0$), and $p_n(\beta) \xrightarrow{p} 0$ for any $\beta \neq \beta_0$ (by (26), setting $\zeta = c_n(\beta - \beta_0)$ and letting $c_n \to \infty$). Thus $p_n(\beta)$ converges to zero in probability for any $\beta \in \mathcal{B} \backslash \beta_0$ while satisfying $\int_\mathcal{B} p_n(\beta) \mathrm{d}\beta = 1$. Thus $Q_n(\beta) = \int_\mathcal{B} \ell(\beta - \xi) p_n(\xi) d\xi \xrightarrow{p}$

$\ell(\beta-\beta_0)$. By the convexity of $\beta \mapsto \ell(\beta)$ (thus of $\beta \mapsto Q_n(\beta)$) and the convexity lemma (Pollard, 1991), $Q_n(\beta) \xrightarrow{p} \ell(\beta-\beta_0)$ uniformly in β. Then by the continuity of ℓ and the uniqueness of β_0, using Lemma 1, we have $\widetilde{\beta}_{LTE} \xrightarrow{p} \beta_0$.

Given the arguments above, we now only need to verify (18) for $\widetilde{\beta}_{LTE}$. It follows from Chernozhukov and Hong (2003, Theorems 1 and 2) when the symmetric loss function is used. By (28) in Lemma 2, we verify that the key assumption for their results (Assumption 4 in theirs) holds in our context. □

Proof of Theorem 4. (i) By Assumption AL1, under H_{a-nloc}, $\sup_\beta |\sup_\lambda c_n^{-2} P_n(\lambda,\beta,\theta_\dagger) - \sup_\lambda P(\lambda,\beta,\theta_\dagger)| \leq \sup_\beta \sup_\lambda |c_n^{-2} P_n(\lambda,\beta,\theta_\dagger) - P(\lambda,\beta,\theta_\dagger)| = o_p(1)$. Then by the continuity and the unique saddle point solution of P, using Lemma 1, we have $\widetilde{\beta} \xrightarrow{p} \beta^*$. So by Assumption AL1, for any λ, $c_n^{-2} P_n(\widetilde{\beta}_C, \theta_\dagger, \lambda) \xrightarrow{p} P(\beta^*, \theta_\dagger, \lambda)$. Noting that $\lambda \mapsto P_n(\beta,\theta,\lambda)$ is convex, by the convexity lemma (Pollard, 1991), the last convergence in probability holds uniformly in λ. So by Lemma 1, $\lambda(\widetilde{\beta}_C, \theta_\dagger) \xrightarrow{p} \lambda^*$.The convergence in Theorem 4 (i) then follows from Assumption AL1.

Noting that $P(\lambda^*,\beta^*,\theta_\dagger) \geq 0$, it remains to show $P(\lambda^*,\beta^*,\theta_\dagger) \neq 0$. Suppose $P(\lambda^*,\beta^*,\theta_\dagger) = 0$. Then $\lambda^* = 0$, since otherwise λ^* and 0 would be different saddle point solutions (given β^*) which violates Assumption AL2. $\lambda^* = 0$ leads to $M g_0(\beta^*,\theta_\dagger) = 0$ (FOC for λ^* given β^*) which is contradictory to Assumption ID. The proof is then complete.

(ii). We only need to show, under H_{a-nloc},

$$c_n^{-2} \mathcal{L}_n(\widetilde{\beta}_{LTE}, \theta_\dagger) \xrightarrow{p} 2P(\lambda^*,\beta^*,\theta_\dagger). \tag{20}$$

Note that

$$p_n(\beta) = \frac{\exp(-(\mathcal{L}_n(\beta,\theta_\dagger) - \mathcal{L}_n(\beta^*,\theta_\dagger)))\pi(\beta)}{\int_{\mathcal{B}} \exp(-(\mathcal{L}_n(\beta,\theta_\dagger) - \mathcal{L}_n(\beta^*,\theta_\dagger)))\pi(\beta)\mathrm{d}\beta},$$

and for any $\beta \neq \beta^*$, $p_n(\beta) \xrightarrow{p} 0$. Thus $p_n(\beta)$ converges zero in probability for any $\beta \in \mathcal{B}\backslash\beta^*$ while satisfying $\int_{\mathcal{B}} p_n(\beta)\mathrm{d}\beta = 1$. Thus $Q_n(\beta) = \int_{\mathcal{B}} \ell(\beta-\xi) p_n(\xi) d\xi \xrightarrow{p} \ell(\beta-\beta^*)$. By the convexity of $\beta \mapsto \ell(\beta)$ (thus of $\beta \mapsto Q_n(\beta)$) and the convexity lemma (Pollard, 1991), $Q_n(\beta) \xrightarrow{p} \ell(\beta-\beta^*)$ uniformly in β. Then by the continuity and the unique saddle point solution of P, using Lemma 1, we have $\widetilde{\beta}_{LTE} \xrightarrow{p} \beta^*$. Then $\lambda(\widetilde{\beta}_{LTE}, \theta_\dagger) \xrightarrow{p} \lambda^*$ follows from the similar arguments as in (i). Thus (20) follows from Assumption AL1. □

Proof of Theorem 5. It follows from the proof of Theorem 3. □

14 Appendix B: Lemmas

The following lemma extends van der Vaart (2000, Theorem 5.7) by allowing the approximating function and its minimizer to be random.

Lemma 1 *Let $\mathcal{L}_n(b)$ and $A_n(b)$ be random functions such that*

$$\sup_{b\in B}|A_n(b)-\mathcal{L}_n(b)|=o_p(1), \tag{21}$$

and $A_n(b)$ is continuous in probability (i.e. $\forall b_1,b_2\in B$, $b_1-b_2\to 0$ implies $A_n(b_1)-A_n(b_2)=o_p(1)$). Suppose $\check{b}$ uniquely minimizes $A_n(b)$ in B, and $\mathcal{L}_n(\widetilde{b})\leq\mathcal{L}_n(\check{b})+o_p(1)$. Then $\widetilde{b}-\check{b}=o_p(1)$.

Proof. Note that $0\overset{\text{def. }\check{b}}{\geq}A_n(\check{b})-A_n(\widetilde{b})\overset{(21)}{=}\mathcal{L}_n(\check{b})-A_n(\widetilde{b})+o_p(1)\overset{\text{def. }\widetilde{b}}{\geq}\mathcal{L}_n(\widetilde{b})-A_n(\widetilde{b})+o_p(1)=-[A_n(\widetilde{b})-\mathcal{L}_n(\widetilde{b})]+o_p(1)\geq-\sup_{b\in B}|A_n(b)-\mathcal{L}_n(b)|+o_p(1)\overset{(21)}{=}o_p(1)$. So

$$A_n(\check{b})-A_n(\widetilde{b})=o_p(1). \tag{22}$$

By continuity in probability and the unique $\check{b}$, for every $\varepsilon>0$ such that $|\widetilde{b}-\check{b}|<\varepsilon$, there exists $\eta(\varepsilon)$ such that $|A_n(\check{b})-A_n(\widetilde{b})|<\eta(\varepsilon)$. The probability of the event that such $\eta(\varepsilon)$ does not exist approaches zero. So $\mathbb{P}(|\widetilde{b}-\check{b}|<\varepsilon)=\mathbb{P}(|A_n(\widetilde{b})-A_n(\check{b})|<\eta(\varepsilon))+o(1)\overset{(22)}{\to}1$. □

Lemmas 2-4 below provide preliminary asymptotic results under the local alternative H_{a-loc} : $\theta_0=\theta_\dagger-c_n^{-1}\xi$.

Lemma 2 *Suppose the assumptions in Theorem 3 (ii) hold. Let $B_{\beta_0}(\delta_n)=\{\beta:|\beta-\beta_0|\leq\delta_n\}$, for any $\delta_n\to 0$. Then under H_{a-loc},*

$$\sup_{\beta\in B_{\beta_0}(\delta_n)}|\lambda(\beta,\theta_\dagger)| = o_p(c_n^2), \tag{23}$$

$$\sup_{\beta\in B_{\beta_0}(\delta_n)}\frac{|c_n\widehat{m}(\beta,\theta_\dagger)-[c_nG_\beta(\beta-\beta_0)+G_\theta\xi+c_n\widehat{m}(\beta_0)]|}{1+c_n|\beta-\beta_0|+|\xi|} = o_p(1), \tag{24}$$

$$\sup_{\beta\in B_{\beta_0}(\delta_n)}|\mathcal{L}_n(\beta,\theta_\dagger)-c_n^2\widehat{m}(\beta,\theta_\dagger)'V^{-1}\widehat{m}(\beta,\theta_\dagger)| = o_p(1), \tag{25}$$

$$\mathcal{L}_n(\beta_a,\theta_\dagger) \xrightarrow{d} \chi^2((G_\beta\zeta+G_\theta\xi)'V^{-1}(G_\beta\zeta+G_\theta\xi),d), \tag{26}$$

where $\beta_a = \beta_0 + c_n^{-1}\zeta$ *in (26). Moreover, for* $\beta \in B_{\beta_0}(\delta_n)$, $\mathcal{L}_n$ *has the expansion*

$$\mathcal{L}_n(\beta, \theta_\dagger) - \mathcal{L}_n(\beta_0, \theta_\dagger) = A_n(\beta, \theta_\dagger) + R_n, \tag{27}$$

where $A_n(\beta, \theta_\dagger) = (\beta - \beta_0)'\Delta_n(\beta_0, \theta_\dagger) + (\beta - \beta_0)' c_n^2 J(\beta_0)(\beta - \beta_0)/2$, *with* $\Delta_n(\beta_0, \theta_\dagger) = 2c_n^2 G_\beta' V^{-1}\widehat{m}(\beta_0, \theta_\dagger)$ *and* $J(\beta_0) = 2G_\beta' V^{-1} G_\beta$, *such that the remainder* R_n *satisfies*

$$\sup_{\beta \in B_{\beta_0}(\delta_n)} |R_n| / (1 + c_n^2|\beta - \beta_0|^2 + |\xi|^2) = o_p(1). \tag{28}$$

Proof. Note that for any $(\beta, \theta) \in \mathcal{B} \times \Theta$, $\lambda(\beta, \theta)$ satisfies

$$0 \overset{FOC}{=} \lambda(\beta, \theta)' \sum_{i=1}^n \frac{m_i(\beta, \theta)}{1 - \lambda(\beta, \theta)' m_i(\beta, \theta)} = \lambda(\beta, \theta)' \sum_{i=1}^n m_i(\beta, \theta)[1 + \frac{\lambda(\beta, \theta)' m_i(\beta, \theta)}{1 - \lambda(\beta, \theta)' m_i(\beta, \theta)}]$$

which gives

$$-\lambda(\beta, \theta)'\widehat{m}(\beta, \theta) = \lambda(\beta, \theta)' \sum_{i=1}^n \frac{m_i(\beta, \theta) m_i(\beta, \theta)'}{1 - \lambda(\beta, \theta)' m_i(\beta, \theta)} \lambda(\beta, \theta). \tag{29}$$

So

$$\begin{aligned}
& c_n^2 \lambda(\beta, \theta)' \sum_{i=1}^n m_i(\beta, \theta) m_i(\beta, \theta)' \lambda(\beta, \theta) \\
\leq \quad & \max_{1 \leq i \leq n} (1 - \lambda(\beta, \theta)' m_i(\beta, \theta)) \cdot c_n^2 \lambda(\beta, \theta)' \sum_{i=1}^n \frac{m_i(\beta, \theta) m_i(\beta, \theta)'}{1 - \lambda(\beta, \theta)' m_i(\beta, \theta)} \lambda(\beta, \theta) \\
\overset{(29)}{=} \quad & - \max_{1 \leq i \leq n} (1 - \lambda(\beta, \theta)' m_i(\beta, \theta)) \cdot c_n^2 \lambda(\beta, \theta)' \widehat{m}(\beta, \theta),
\end{aligned}$$

where the second line follows from $1 - \lambda(\beta, \theta)' m_i(\beta, \theta) > 0$ (since the solution of the constrained optimization (10) $\pi_i = [1 - \lambda(\beta, \theta)' m_i(\beta, \theta)]^{-1} > 0$), and the inequality $\leq$ holds in matrix sense (the difference is negative semi-definite). Taking norm on both sides,

$$|\lambda(\beta, \theta)| \cdot |c_n^2 \sum_{i=1}^n m_i(\beta, \theta) m_i(\beta, \theta)'| \leq [1 + |\lambda(\beta, \theta)| \max_{1 \leq i \leq n} |m_i(\beta, \theta)|] \cdot c_n^2 |\widehat{m}(\beta, \theta)|. \tag{30}$$

Note that (30) holds for any β, θ. Taking the supremum of (30) over $\beta \in B_{\beta_0}(\delta_n)$ and setting $\theta = \theta_\dagger$, $\sup_{\beta \in B_{\beta_0}(\delta_n)} |\lambda(\beta, \theta_\dagger)| \cdot \sup_{\beta \in B_{\beta_0}(\delta_n)} |c_n^2 \sum_{i=1}^n m_i(\beta, \theta_\dagger) m_i(\beta, \theta_\dagger)'| \leq [1 + \sup_{\beta \in B_{\beta_0}(\delta_n)} |\lambda(\beta, \theta_\dagger)| \cdot$

$\sup_{\beta\in B_{\beta_0}(\delta_n)} \max_{1\le i\le n} |m_i(\beta,\theta_\dagger)|] \cdot \sup_{\beta\in B_{\beta_0}(\delta_n)} c_n^2|\widehat{m}(\beta,\theta_\dagger)|$, by which we have,

$$\begin{aligned} & \sup_{\beta\in B_{\beta_0}(\delta_n)} |\lambda(\beta,\theta_\dagger)| \cdot [\sup_{\beta\in B_{\beta_0}(\delta_n)} |c_n^2 \sum_{i=1}^n m_i(\beta,\theta_\dagger) m_i(\beta,\theta_\dagger)'| \\ & - \sup_{\beta\in B_{\beta_0}(\delta_n)} \max_{1\le i\le n} |m_i(\beta,\theta_\dagger)| \sup_{\beta\in B_{\beta_0}(\delta_n)} c_n^2|\widehat{m}(\beta,\theta_\dagger)|] \\ \le & \sup_{\beta\in B_{\beta_0}(\delta_n)} c_n^2|\widehat{m}(\beta,\theta_\dagger)|. \end{aligned} \tag{31}$$

The right-hand side of (31) is $o_p(c_n^2)$ by Assumption UC(i), while the left-hand side is $\sup_{\beta\in B_{\beta_0}(\delta_n)} |\lambda(\beta,\theta_\dagger)| \cdot [V + o_p(1)]$ by Assumptions M′, UC(i) and UC(ii). This forces (23) to be true.

Now we consider (24). By the Assumption SE at (β_0,θ_0),

$$\begin{aligned} o_p(1) & \overset{SE}{=} \sup_{\beta\in B_{\beta_0}(\delta_n),|\theta|\le|\theta_\dagger|} \frac{|c_n\widehat{m}(\beta,\theta) - c_n[Mg_0(\beta,\theta) - Mg_0(\beta_0,\theta_0) + \widehat{m}(\beta_0,\theta_0)]|}{1 + c_n|\beta-\beta_0| + |\xi|} \\ & \ge \sup_{\beta\in B_{\beta_0}(\delta_n)} \frac{|c_n\widehat{m}(\beta,\theta_\dagger) - c_n[Mg_0(\beta,\theta_\dagger) - Mg_0(\beta_0,\theta_0) + \widehat{m}(\beta_0,\theta_0)]|}{1 + c_n|\beta-\beta_0| + |\xi|} \\ & \overset{CD}{=} \sup_{\beta\in B_{\beta_0}(\delta_n)} \frac{|c_n\widehat{m}(\beta,\theta_\dagger) - c_nM[\nabla_\beta g_0(\dot\beta)\cdot(\beta-\beta_0) + \nabla_\theta g_0(\beta_0,\dot\theta)\cdot c_n^{-1}\xi] - c_n\widehat{m}(\beta_0)|}{1 + c_n|\beta-\beta_0| + |\xi|} \\ & \overset{CD}{=} \sup_{\beta\in B_{\beta_0}(\delta_n)} \frac{|c_n\widehat{m}(\beta,\theta_\dagger) - c_nG_\beta(\beta-\beta_0) - G_\theta\xi - c_n\widehat{m}(\beta_0)| + o_p(1)}{1 + c_n|\beta-\beta_0| + |\xi|}, \end{aligned}$$

where $\dot\beta$ is a point between β_0 and β, $\dot\theta$ is a point between θ_0 and $\theta_\dagger$, and the last lines use the uniform continuity of $\nabla g_0(\beta,\theta)$ in the neighborhood of (β_0,θ_0). So (24) holds.

Now we consider (25), which will be shown using (17). For $\beta \in B_{\beta_0}(\delta_n)$, checking the term $\widehat{V}_1(\beta,\theta_\dagger)$ in (which is defined in (16), setting $\theta=\theta_\dagger$), we write $\widehat{V}_1(\beta,\theta_\dagger) = -c_n^2\sum_{i=1}^n[\rho_2(\dot\lambda' m_i(\beta,\theta_\dagger)) + 1]m_i(\beta,\theta_\dagger)m_i(\beta,\theta_\dagger)' + c_n^2\sum_{i=1}^n m_i(\beta,\theta_\dagger)m_i(\beta,\theta_\dagger)' := T_1 + T_2$. The term $\sup_{\beta\in B_{\beta_0}(\delta_n)}|T_2| \xrightarrow{p} V$ by Assumption UC(ii). Looking at the term T_1, $\sup_{\beta\in B_{\beta_0}(\delta_n)}|T_1| \le \sup_{\beta\in B_{\beta_0}(\delta_n)} \max_{1\le i\le n} |\rho_2(\dot\lambda' m_i(\beta,\theta_\dagger)) + 1| \cdot \sup_{\beta\in B_{\beta_0}(\delta_n)} c_n^2\sum_{i=1}^n |m_i(\beta,\theta_\dagger)m_i(\beta,\theta_\dagger)'|$. The second factor on the right hand side is bounded in probability by Assumption UC(ii) and the Cauchy-Schwarz inequality. The first factor $\xrightarrow{p} 0$, since $\rho_2(0) = -1$ and

$$\sup_{\beta\in B_{\beta_0}(\delta_n)} \sup_{1\le i\le n} |\lambda(\beta,\theta_\dagger)' m_i(\beta,\theta_\dagger)| = \sup_{\beta\in B_{\beta_0}(\delta_n)} |\lambda(\beta,\theta_\dagger)| \cdot \sup_{\beta\in B_{\beta_0}(\delta_n),1\le i\le n} |m_i(\beta,\theta_\dagger)| \overset{(23),\mathrm{M}}{=} o_p(c_n^2)O_p(c_n^{-2}) = o_p(1).$$

So $\sup_{\beta\in B_{\beta_0}(\delta_n)}|T_1| \xrightarrow{p} 0$, $\sup_{\beta\in B_{\beta_0}(\delta_n)}|\widehat{V}_1(\beta,\theta_\dagger)-V| \xrightarrow{p} 0$. Using the similar argument, we have $\sup_{\beta\in B_{\beta_0}(\delta_n)}|\widehat{V}_2(\beta,\theta_\dagger)-V| \xrightarrow{p} 0$ in (17) (setting $\theta=\theta_\dagger$). So (25) follows from (17).

(26) follows from (25) and (24) by setting $\beta=\beta_a$, and Assumption AN.

Finally, (28) follows from (24), (25) and (26). □

Lemma 3 *Suppose the assumptions in Theorem 3 (i) hold. Under H_{a-loc}, we have*

$$\begin{aligned}
\sup_{\beta\in B_{\beta_0}(c_n^{-1})}|\lambda(\beta,\theta_\dagger)| &= O_p(c_n), &(32)\\
\sup_{\beta\in B_{\beta_0}(c_n^{-1})}|c_n\widehat{m}(\beta,\theta_\dagger)-[c_nG_\beta(\beta-\beta_0)+G_\theta\xi+c_n\widehat{m}(\beta_0)]| &= o_p(1), &(33)\\
\sup_{\beta\in B_{\beta_0}(c_n^{-1})}|\mathcal{L}_n(\beta,\theta_\dagger)-c_n^2\widehat{m}(\beta,\theta_\dagger)'V^{-1}\widehat{m}(\beta,\theta_\dagger)| &= o_p(1). &(34)
\end{aligned}$$

Let R_n be defined in (27), then

$$\sup_{\beta\in B_{\beta_0}(c_n^{-1})}|R_n| = o_p(1). \quad (35)$$

Proof. It follows from the proof of Lemma 2. A weaker bound assumption (Assumption M) than Assumption M′ suffices here for the results in a smaller neighborhood $B_{\beta_0}(c_n^{-1})$ to hold. □

Lemma 4 *For the concentrating estimator $\widetilde{\beta}_C$ under H_{a-loc}, we have*

$$\begin{aligned}
c_n\widehat{m}(\widetilde{\beta}_C,\theta_\dagger) &= O_p(1), &(36)\\
\widetilde{\beta}_C &\xrightarrow{p} \beta_0, &(37)\\
c_n(\widetilde{\beta}_C-\beta_0) &= O_p(1). &(38)
\end{aligned}$$

Proof. Given the results for $\lambda(\beta_0,\theta_\dagger)$ (in (32)) and $\widehat{m}(\beta_0,\theta_\dagger)$ (in (33)), the assumption for the second sample moment at $(\beta_0,\theta_\dagger)$ (Assumption UC (ii)) and the global bounded condition (Assumption M), the bound in (36) can be proved following the arguments in Newey and Smith (2004, Lemma A3).

Now consider (37). Note that $|Mg_0(\widetilde{\beta}_C,\theta_\dagger)| \le |Mg_0(\widetilde{\beta}_C,\theta_\dagger)-\widehat{m}(\widetilde{\beta}_C,\theta_\dagger)|+|\widehat{m}(\widetilde{\beta}_C,\theta_\dagger)| \overset{\text{UC(i)}}{=} o_p(1)+|\widehat{m}(\widetilde{\beta}_C,\theta_\dagger)| \overset{(36)}{=} o_p(1)$. By the continuity of g_0 (Assumption CD), $Mg_0(\widetilde{\beta}_C,\theta_0)=Mg_0(\widetilde{\beta}_C,\theta_\dagger)+o_p(1)=o_p(1)$. Then (37) follows from the fact that β_0 uniquely solves $Mg_0(\beta,\theta_0)=0$, implied by

Assumption ID. Now consider (38). Note that

$$c_n|g_0(\widetilde{\beta}_C,\theta_\dagger)| \leq |v_n(\widetilde{\beta}_C,\theta_\dagger) - v_n(\beta_0)| + c_n|\widehat{m}(\widetilde{\beta}_C,\theta_\dagger)| + c_n|\widehat{m}(\beta_0)|. \tag{39}$$

By the mean value theorem and Assumption CD, $g_0(\widetilde{\beta}_C,\theta_\dagger) = g_0(\beta_0)+G_\beta(\dot{\beta})(\widetilde{\beta}_C-\beta_0)+c_n^{-1}G_\theta(\dot{\theta})\xi = G_\beta(\widetilde{\beta}_C-\beta_0)+o_p(1)$, where $\dot{\beta}$ is between $\widetilde{\beta}_C$ and β_0, and $\dot{\theta}$ is between $\theta_\dagger$ and θ_0. Evaluating stochastic orders of both sides of (39), $c_nC|\widetilde{\beta}_C - \beta_0| \leq o_p(1 + c_n|\widetilde{\beta}_C - \beta_0| + |\xi|) + O_p(1)$, by Assumption SE, (36) and Assumption AN. So $c_n(\widetilde{\beta}_C - \beta_0) = O_p(1)$, as asserted in (38). □

15 Appendix C: Details in the quantile regression discontinuity design

C.1. The following two facts about the weighting functions are useful. For $k = 0, 1, ..., p$,

$$\sum_{i=1}^n W_p^+((X_i - c)/h)(X_i - c)^k = \mathbb{I}_{\{k=0\}},\ \sum_{i=1}^n W_p^-((X_i - c)/h)(X_i - c)^k = \mathbb{I}_{\{k=0\}}, \tag{40}$$

$$e_1'(S^+)^{-1}\int_0^\infty z^k\varpi(z)dz = \mathbb{I}_{\{k=0\}},\ e_1'(S^-)^{-1}\int_0^\infty z^k\varpi(z)dz = \mathbb{I}_{\{k=0\}}. \tag{41}$$

Note that(40) follows from $h^{-k}\sum_{i=1}^n W_p^+((X_i-c)/h)(X_i-c)^k = e_1'(\widehat{S}^+)^{-1}\sum_{i=1}^n \varpi((X_i-c)/h)I_i[(X_i-c)/h]^k = e_1'(\widehat{S}^+)^{-1}\widehat{S}^+e_{k+1} = \mathbb{I}_{\{k=0\}}$, and the second equality follows similarly. (41) follows from $e_1'(S^+)^{-1}\int_0^\infty z^k\varpi(z)dz = e_1'(S^+)^{-1}S^+e_{k+1} = \mathbb{I}_{(k=0)}$, and the second equality follows similarly.

C.2. Asymptotic variance. Let S^+ and S^- be the $(p+1)$ by $(p+1)$ matrices with the $(i,j)-$th elements $\int_0^\infty u^{i+j-2}K(u)\mathrm{d}u$ and $\int_{-\infty}^0 u^{i+j-2}K(u)\mathrm{d}u$ respectively. Assume S^+ and S^- to be non-singular. Here we use the notation $\mathbb{E}_+(\cdot) = \mathbb{E}_+(\cdot|X = c+)$. Similar notations apply to $\mathbb{E}_-(\cdot)$, $f_+(\cdot)$, $f_-(\cdot), \mathbb{P}_+(\cdot), \mathbb{P}_-(\cdot)$. The matrix G in Assumption CD takes the form:

$$G = \begin{pmatrix} G_{11} & -\tau & G_{11} \\ G_{21} & \tau & 0 \\ 0 & -1 & 0 \end{pmatrix},$$

where $G_{11} = f_+(\beta_{10}+\theta_0|_{D=1})\mathbb{P}_+(D=1)-f_-(\beta_{10}+\theta_0|_{D=1})\mathbb{P}_-(D=1)$ and $G_{21} = f_+(\beta_{10}|_{D=0})\mathbb{P}_+(D=0) - f_-(\beta_{10}|_{D=0})\mathbb{P}_-(D=0)$. The matrix Ω in Assumption AN takes the form: $\Omega = (\Omega_{ij})_{i,j=1,2,3}$,

where $\Omega_{ij} = \Omega_{ji}$ and $\Omega_{11} = [(1 - 2\tau\beta_{20})\mathbb{E}_+\mathbb{I}(Y_i < \beta_{10} + \theta_0)D_i + \tau^2\beta_{20}^2]C^+/\mu(c) + \mathbb{E}_-\mathbb{I}(Y_i < \beta_{10} + \theta_0)D_iC^-/\mu(c), \Omega_{12} = [\tau\beta_{20}\mathbb{E}_+\mathbb{I}(Y_i < \beta_{10}+\theta_0)D_i - \tau\beta_{20}\mathbb{E}_+\mathbb{I}(Y_i < \beta_{10})(1-D_i) - \tau^2\beta_{20}^2]C^+/\mu(c), \Omega_{13} = [(1-\beta_{20})\mathbb{E}_+\mathbb{I}(Y_i < \beta_{10}+\theta_0)D_i - \tau\beta_{20}\mathbb{E}_+D_i + \tau\beta_{20}^2]C^r/\mu(c) + \mathbb{E}_-\mathbb{I}(Y_i < \beta_{10}+\theta_0)D_iC^-/\mu(c), \Omega_{22} = [(1+2\tau\beta_{20})\mathbb{E}_+\mathbb{I}(Y_i < \beta_{10})(1-D_i) + \tau^2\beta_{20}^2]C^+/\mu(c) + \mathbb{E}_-\mathbb{I}(Y_i < \beta_{10})(1-D_i)C^-/\mu(c), \Omega_{23} = [-\beta_{20}\mathbb{E}_+\mathbb{I}(Y_i < \beta_{10})(1 - D_i) + \tau\beta_{20}\mathbb{E}_+D_i - \tau\beta_{20}^2]C^+/\mu(c), \Omega_{33} = \mathbb{E}_+[(1 - 2\beta_{20})D_i + \beta_{20}^2]C^+/\mu(c) + \mathbb{E}_-D_iC^-/\mu(c)$ with the constants $C^+ = e_1'(S^+)^{-1}\int_0^\infty \varpi\varpi'(S^+)^{-1}e_1$ and $C^- = e_1'(S^-)^{-1}\int_{-\infty}^0 \varpi\varpi'(S^-)^{-1}e_1$.

Here Ω requires six nonparametric regressions to be estimated, and G involves four conditional densities.

C.3. Verification of Assumption UC (i). We only verify the convergence for the following element. Other elements follow similarly.

$$
\begin{aligned}
& \sum_{i=1}^n W_p^+((X_i - c)/h)[\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \tau\beta_2] \\
= \;& (nh)^{-1}e_1'((nh)^{-1}\widehat{S}^+)^{-1}\sum_{i=1}^n \varpi((X_i - c)/h)I_i[\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \tau\beta_2] \\
\overset{LLN}{=} \;& e_1'((nh)^{-1}\widehat{S}^+)^{-1}\mathbb{E}h^{-1}\varpi((X_i - c)/h)I_i[\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \tau\beta_2] + o_p(1) \\
\overset{LIE}{=} \;& e_1'((nh)^{-1}\widehat{S}^+)^{-1}\mathbb{E}h^{-1}\varpi((X_i - c)/h)I_i[\mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|X) - \tau\beta_2] + o_p(1) \\
= \;& e_1'((nh)^{-1}\widehat{S}^+)^{-1}\int_0^\infty \varpi(u)[\mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|c + hu) - \tau\beta_2]\mu(c + hu)du + o_p(1) \\
= \;& e_1'(S^+)^{-1}\int_0^\infty \varpi(u)[\mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|c+) - \tau\beta_2]du + o_p(1) \\
= \;& \mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|c+) - \tau\beta_2 + o_p(1)
\end{aligned}
$$

where we use the facts that $(nh)^{-1}\widehat{S}^+ \xrightarrow{p} S^+\mu(c+)$ and $e_1'(S^+)^{-1}\int_0^\infty \varpi(u)du = 1$ (by (41)). The convergence in probability is uniform in $(\beta_1, \beta_2, \theta)$ by the bounded kernel with a bounded support, the compactness of $\mathcal{B} \times \Theta$ and the ULLN.

C.4. Verification of Assumption UC (ii). Write $g = (g_+', g_-')'$ to separate estimating functions on the right and left of c. Denote $V^+(\beta, \theta) = \mathbb{E}[g_+(Y_i, \beta, \theta)g_+(Y_i, \beta, \theta)'|X = c+]$ and $V^-(\beta, \theta) = \mathbb{E}[g_-(Y_i, \beta, \theta)g_-(Y_i, \beta, \theta)'|X = c-]$. Then $V(\beta, \theta)$ takes the form $M\underline{V}(\beta, \theta)M'$ where $\underline{V}(\beta, \theta) = \text{diag}(V^+(\beta, \theta)C^+/\mu(c+), V^-(\beta, \theta)C^-/\mu(c-))$, with the constants C^+ and C^- being defined above. Uniform convergence in probability follows similarly as above.

C.5. Verification of Assumption SE. We only verify stochastic equi-continuity for the fol-

lowing element in $v_n(\beta,\theta)$. Other elements follow similarly. Note that

$$
\begin{aligned}
& \sqrt{nh}[\sum_{i=1}^{n} W_p^+((X_i-c)/h)\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|X_i=c+)] \\
= \; & \sqrt{nh}[\sum_{i=1}^{n} e_1'(\widehat{S}^+)^{-1}\varpi((X_i-c)/h)I_i\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|X_i=c+)] \\
= \; & \underbrace{e_1'((nh)^{-1}\widehat{S}^+)^{-1}n^{-1/2}\sum_{i=1}^{n}[h^{-1/2}\varpi((X_i-c)/h)I_i\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \mathbb{E}h^{-1/2}\varpi((X_i-c)/h)I_i\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i]}_{=T_1} \\
& \underbrace{+(nh)^{-1/2}\sum_{i=1}^{n} e_1'((nh)^{-1}\widehat{S}^+)^{-1}\mathbb{E}\varpi((X_i-c)/h)I_i\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \sqrt{nh}\mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|X_i=c+)}_{=T_2} \\
: \; & = T_1+T_2.
\end{aligned}
$$

Stochastic equi-continuity (SE) of $(\beta,\theta)\mapsto v_n(\beta,\theta)$ follows by stochastic equi-continuity of T_1 and negligibility of T_2, as we will show below.

We first show that $T_2=o_p(1)$ uniformly in a neighborhood (β_0,θ_0). Let $F_{Y^1,D=1}(\cdot|X)=\mathbb{P}(Y^1<\cdot,D=1|X)$. For the first term of T_2,

$$
\begin{aligned}
& (nh)^{-1/2}\sum_{i=1}^{n} e_1'((nh)^{-1}\widehat{S}^+)^{-1}\mathbb{E}\varpi((X_i-c)/h)I_i\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i \\
\overset{LIE}{=} \; & (nh)^{-1/2}\sum_{i=1}^{n} e_1'((nh)^{-1}\widehat{S}^+)^{-1}\mathbb{E}\varpi((X_i-c)/h)I_iF_{Y^1,D=1}(\beta_1+\theta|X_i) \\
= \; & (nh)^{-1/2}ne_1'((nh)^{-1}\widehat{S}^+)^{-1}\mathbb{E}\varpi((X_i-c)/h)I_iF_{Y^1,D=1}(\beta_1+\theta|X_i) \\
= \; & (nh)^{-1/2}ne_1'((nh)^{-1}\widehat{S}^+)^{-1}\int_c^{\infty}\varpi((u-c)/h)F_{Y^1,D=1}(\beta_1+\theta|u)\mu(u)du \\
\overset{z=\frac{u-c}{h}}{=} \; & (nh)^{1/2}\mu(c+)e_1'((nh)^{-1}\widehat{S}^+)^{-1}\int_0^{\infty}\varpi(z)F_{Y^1,D=1}(\beta_1+\theta|c+zh)dz+o_p(1) \\
= \; & (nh)^{1/2}e_1'(S^+)^{-1}\int_0^{\infty}\varpi(z)\left[z^{p+1}h^{p+1}F^{(p+1)}_{Y^1,D=1}(\beta+\theta|\dot{c})/(p+1)!+F_{Y^1,D=1}(\beta_1+\theta|c+)\right]dz+ \\
& +o_p(1), \\
= \; & \underbrace{n^{\frac{1}{2}}h^{\frac{2p+3}{2}}e_1'(S^+)^{-1}F^{(p+1)}_{Y^1,D=1}(\beta+\theta|c+)\int_0^{\infty}z^{p+1}\varpi(z)dz/(p+1)!}_{=\overline{T}_2}+ \\
& +(nh)^{\frac{1}{2}}F_{Y^1,D=1}(\beta_1+\theta|c+)+o_p(1),
\end{aligned}
\tag{42}
$$

where the last two equalities use $(nh)^{-1}\widehat{S}^+\xrightarrow{p}S^+\mu(c+)$ and (41), and $\dot{c}$ is a point between 0 and c.

Thus $T_2 = \overline{T}_2 + o_p(1) = o_p(1)$ by the undersmoothing condition $nh^{2p+3} \to 0$. The uniformity in a neighborhood (β_0, θ_0) follows by Assumption QRD (ii).

Consider the term T_1. Note that (β, θ) enter the function g through an indicator function (thus having bounded variation). The class of functions $\{h^{-1/2}\varpi((X_i - c)/h)I_i g(Y_i, \beta, \theta) : (\beta, \theta) \in \mathcal{B} \times \Theta\}$ belongs to type I in Andrews (1994, Theorem 3) with the envelop function $h^{-1/2}||\varpi((X_i - c)/h)I_i|| \vee C$. So the class satisfies the Pollard's entropy condition (Andrews, 1994, Theorem 2) and is thus Donsker (which implies SE) by van der Vaart (2000, Theorem 19.14) and $\mathbb{E}[h^{-1/2}||\varpi((X_i - c)/h)I_i||]^2 < \infty$ under Assumption K and $h \to 0$. Then the term T_1 is SE by the following property (a). So Assumption SE is satisfied by the following property (b).

We have used the following two properties. Let $v_n(\beta)$ be an empirical process indexed by $\beta \in \mathcal{B}$, and is SE. Then (a) $c_n v_n(\beta)$ is SE, where $c_n = O_p(1)$ does not depend on β, (b) $v_n(\beta) + c_n(\beta)$ is also SE, where $\sup_{\beta \in \mathcal{B}} |c_n(\beta)| = o_p(1)$.

To prove (a), for any ε and η, there exists $C > 0$ such that

$$
\begin{aligned}
& \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |c_n v_n(\beta_1) - c_n v_n(\beta_2)| > \eta \right) \\
= \; & \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |c_n| \cdot |v_n(\beta_1) - v_n(\beta_2)| > \eta \text{ and } |c_n| \le C \right) \\
& + \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |c_n| \cdot |v_n(\beta_1) - v_n(\beta_2)| > \eta \text{ and } |c_n| > C \right) \\
\le \; & \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |v_n(\beta_1) - v_n(\beta_2)| > \eta/C \right) + \lim_{n\to\infty} \mathbb{P}(|c_n| > C) \\
\le \; & \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |v_n(\beta_1) - v_n(\beta_2)| > \eta/C \right) + \varepsilon/2 \le \varepsilon/2 + \varepsilon/2 = \varepsilon
\end{aligned}
$$

where the first inequality holds by $c_n = O_p(1)$, and the second inequality holds since $v_n(\beta)$ is SE. Thus there exists $\delta > 0$ for the derivation above to hold, as desired.

To prove (b), for any ε and η,

$$
\begin{aligned}
& \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |v_n(\beta_1) + c_n(\beta_1) - v_n(\beta_2) - c_n(\beta_2)| > \eta \right) \\
\le \; & \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |v_n(\beta_1) - v_n(\beta_2)| + \sup_{\beta \in \mathcal{B}} |c_n(\beta)| + \sup_{\beta \in \mathcal{B}} |c_n(\beta)| > \eta \right) \\
\le \; & \lim_{n\to\infty} \mathbb{P}\left(\sup_{|\beta_1 - \beta_2| < \delta} |v_n(\beta_1) - v_n(\beta_2)| > \eta/3 \right) + \lim_{n\to\infty} \mathbb{P}\left(\sup_{\beta \in \mathcal{B}} |c_n(\beta)| > \eta/3 \right) \\
& + \lim_{n\to\infty} \mathbb{P}\left(\sup_{\beta \in \mathcal{B}} |c_n(\beta)| > \eta/3 \right) \le \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon
\end{aligned}
$$

where the last inequality holds since $v_n(\beta)$ is SE and $c_n(\beta) = o_p(1)$ uniformly in $\beta \in \mathcal{B}$. Thus there exists $\delta > 0$ for the derivation above to hold, as desired.

C.6. Verification of Assumption AN. It follows from the asymptotic normality for the local polynomial estimator (Fan and Gijbels, 1996), combined with Cramér-Wold device.

C.7. Verification of Assumptions M and M′. Assumption M′ (which is stronger than Assumption M) is satisfied since $\sup_{(\beta,\theta),1\leq i\leq n} |m_i(\beta,\theta)| \leq C|(\widehat{S}^+)^{-1}| \sup_{1\leq i\leq n} |\varpi((X_i - c)/h)| \leq C_1(nh)^{-1}|((nh)^{-1}\widehat{S}^+)^{-1}| = O_p((nh)^{-1})$, where we have used compactness, the bounded $K(\cdot)$ with a bounded support and S^+ being non-singular.

16 Appendix D: QRD with bias correction

When bias correction is used, each element of $g_i(\beta,\theta)$ contains one more term (like $\widehat{\varrho}_{Y_1,+}(\beta,\theta)$) than the uncorrected version. The arguments used in this subsection are based on those above (in Appendix C) and focus on the effects of additional terms on validation of high-level assumptions.

D1. **Verification of Assumption UC (i)**. It follows from that the contribution from the bias-correction term approaches zero uniformly. Consider the first entry of $\widehat{m}(\beta,\theta)$. Noting that $\sum_{i=1}^n W_p^+((X_i - c)/h) = 1$ and $\sum_{i=1}^n W_p^-((X_i - c)/h) = 1$ (which follow from (40)), we only need to show

$$\widehat{\varrho}_{Y_1,+}(\beta,\theta) = o_p(1),\ \widehat{\varrho}_{Y_1,-}(\beta,\theta) = o_p(1) \tag{43}$$

uniformly in β and θ. It is known (in the term $\overline{T}_2$ in (42)) that $\widehat{B}_+ \xrightarrow{p} B_+ = e_1'(S^r)^{-1}\int_0^\infty z^{p+1}\varpi(z)dz$ and $\widehat{B}_- \xrightarrow{p} B_-$ where B_+ and B_- are bounded kernel-specific constants (which do not depend on β or θ) if $h + (nh)^{-1} \to 0$. It follows from standard results and a uniform LLN that $\widehat{\psi}_{Y_1,+}(\beta,\theta;b) \xrightarrow{p} \psi_{Y_1,+}(\beta,\theta)$ and $\widehat{\psi}_{Y_1,-}(\beta,\theta;b) \xrightarrow{p} \psi_{Y_1,-}(\beta,\theta)$ uniformly in β and θ, if $b + (nb^{2p+3})^{-1} \to 0$. Thus (43) holds. Similarly we can show uniform convergence for other entries of $\widehat{m}(\beta,\theta)$.

D2. Verification of Assumption UC (ii). Let $c_n = (nh)^{1/2}$. In Assumption UC(ii), write

$$c_n^2 \sum_{i=1}^n m_i(\beta_0,\theta_0)m_i(\beta_0,\theta_0)' = c_n^2 \sum_{i=1}^n M w_i(\mathcal{X})g_i(\beta_0,\theta_0)g_i(\beta_0,\theta_0)'w_i(\mathcal{X})M'. \tag{44}$$

Focusing on each element in the matrix,

$$\begin{aligned} & \text{the (1,1)-element of (44)} \\ = \; & c_n^2 \sum_{i=1}^{n} \begin{array}{l} \{W_p^+((X_i-c)/h)[\mathbb{I}(Y_i<\beta_1+\theta)D_i-\tau\beta_2-\widehat{\varrho}_{Y_1,+}(\beta,\theta)] \\ \quad -W_p^-((X_i-c)/h)[\mathbb{I}(Y_i<\beta_1+\theta)D_i-\widehat{\varrho}_{Y_1,-}(\beta,\theta)]\}^2 \end{array} = T^+ + T^-, \end{aligned}$$

where $T^+ = c_n^2 \sum_{i=1}^n W_p^+((X_i-c)/h)^2\{\mathbb{I}(Y_i<\beta_1+\theta)D_i-\tau\beta_2-h^{p+1}\psi_{Y_1,+}(\beta,\theta;b)\widehat{B}_+/(p+1)!-[\widehat{\psi}_{Y_1,+}(\beta,\theta;b)-\psi_{Y_1,+}(\beta,\theta;b)]h^{p+1}\widehat{B}_+/(p+1)!\}^2 = T_1^+ + T_2^+ + T_3^+$ with $T_1^+ = c_n^2\sum_{i=1}^n W_p^+((X_i-c)/h)^2[\mathbb{I}(Y_i<\beta_1+\theta)D_i-\tau\beta_2-h^{p+1}\psi_{Y_1,+}(\beta,\theta;b)\widehat{B}_+/(p+1)!]^2$, T_2^+ =the cross-product term, $T_3^+ = c_n^2\sum_{i=1}^n W_p^+((X_i-c)/h)^2[\widehat{\psi}_{Y_1,+}(\beta,\theta;b)-\psi_{Y_1,+}(\beta,\theta;b)]^2h^{2(p+1)}[\widehat{B}_+/(p+1)!]^2$, and T^- is similarly defined and decomposed. We have used above the cross-product term disappears. It can be shown (combined with arguments above when bias-uncorrected EL was under consideration) that $T_1^+ \xrightarrow{p} \Omega_{11}^+$ uniformly in β and θ, where Ω_{11}^+ is defined as in $\Omega_{11} = \Omega_{11}^+ + \Omega_{11}^-$. Note that $T_3^+ = O_p((b+n^{-1/2}b^{-(2p+3)/2})^2h^{2(p+1)}) = O_p(b^2h^{2(p+1)} + n^{-1}b^{-1}(h/b)^{2(p+1)}) = o_p(1)$, under Assumption BW′, where the last $o_p(1)$ term is uniform in β and θ. $T_2^+ = o_p(1)$ by Cauchy-Schwarz inequality. Thus $T^+ \xrightarrow{p} \Omega_{11}^+$ uniformly. Similarly we can show $T^- \xrightarrow{p} \Omega_{11}^-$. So the (1,1)-element of (44)$\xrightarrow{p} \Omega_{11}$. Using similar arguments for other elements of (44), we can show that the uniform convergence in Assumption UC(ii) holds. Then the equality (11) follows given that Assumption AN is satisfied.

D3. Verification of Assumption AN. Define $\varrho_{Y_1,+}(\beta,\theta) = h^{p+1}\psi_{Y_1,+}(\beta,\theta)\widehat{B}_+/(p+1)!$, and other quantities of ϱ are similarly defined. It follows from the standard asymptotic normality arguments for the local polynomial estimator at boundary points (Fan and Gijbels, 1996), combined with Cramér-Wold device, that Assumption AN is true for $g_i(\beta,\theta)$ as defined in (7) if we replace $\widehat{\varrho}$'s by ϱ's, if $(nh)^{-1} + nh^{2p+5} \to 0$. Now we verify that the "replacing" effect is asymptotically negligible under Assumption BW′. We establish the following result (the results for other elements follow similarly): $c_n\sum_{i=1}^n W_p^+((X_i-c)/h)[\widehat{\varrho}_{Y_1,+}(\beta,\theta)-\varrho_{Y_1,+}(\beta,\theta)] = c_n[\widehat{\psi}_{Y_1,+}(\beta,\theta;b)-\psi_{Y_1,+}(\beta,\theta)]h^{p+1}\widehat{B}_+/(p+1)! = O_p(n^{1/2}h^{1/2}h^{p+1}(b+n^{-1/2}b^{-(2p+3)/2})) = O_p(n^{1/2}h^{(2p+3)/2}b + (h/b)^{(2p+3)/2}) = o_p(1)$, by Assumption BW′.

D4. Verification of Assumption SE. To show stochastic equi-continuity of the first component of $v_n(\beta,\theta) = c_n[\widehat{m}(\beta,\theta) - Mg_0(\beta,\theta)]$, we only need to show $c_n[\widehat{\varrho}_{Y_1,+}(\beta,\theta) - \varrho_{Y_1,+}(\beta,\theta)]$ and $c_n[\widehat{\varrho}_{Y_1,-}(\beta,\theta) - \varrho_{Y_1,-}(\beta,\theta)]$ are SE, given the results in C.5 (for uncorrected estimating equations).

This is because slightly modifying the arguments in C.5 (regarding the term T_2, while keeping T_1 unchanged) shows that e.g. $(nh)^{1/2}\{\sum_{i=1}^n W_p^+((X_i-c)/h)[\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i - \tau\beta_2 - \varrho_{Y_1,+}(\beta,\theta)] - \mathbb{E}(\mathbb{I}_{(Y_i<\beta_1+\theta)}D_i|X_i=c+)\}$ is SE if $nh^{2p+5}\to 0$.

Note that $c_n[\widehat{\varrho}_{Y_1,+}(\beta,\theta) - \varrho_{Y_1,+}(\beta,\theta)] = (nh)^{1/2}[\widehat{\psi}_{Y_1,+}(\beta,\theta;b) - \psi_{Y_1,+}(\beta,\theta)]h^{p+1}\widehat{B}_+/(p+1)! = (h/b)^{-(2p+3)/2}(nb^{2p+3})^{1/2}[\widehat{\psi}_{Y_1,+}(\beta,\theta;b) - \psi_{Y_1,+}(\beta,\theta)]\widehat{B}_+/(p+1)!$. We can show that $(nh^{2p+3})^{1/2}[\widehat{\psi}_{Y_1,+}(\beta,\theta;b) - \psi_{Y_1,+}(\beta,\theta)]$ is SE using the arguments in C.5 (for uncorrected estimating equations, except now for the local $(p+1)$-th polynomial estimator of the $(p+1)$-th derivative, instead of the local p-th polynomial estimator of the conditional mean). The the result follows from $h/b = O(1)$ and $\widehat{B}_+ \xrightarrow{p} B_+$ which does not depend on β or θ. Similarly we can show $c_n[\widehat{\varrho}_{Y_1,-}(\beta,\theta) - \varrho_{Y_1,-}(\beta,\theta)]$ is SE. Thus the first component of $v_n(\beta,\theta)$ is SE. Similarly we can show other components are SE.

D5. Verification of Assumptions M and M′. Note that $\sup_{(\beta,\theta),1\le i\le n}|\widehat{\psi}_{Y_1,+}(\beta,\theta;b)| = O_p((nb)^{-1})$ and the bound condition holds similarly for other $\widehat{\psi}$'s. Assumption M′ is satisfied since $\sup_{(\beta,\theta),1\le i\le n}|m_i(\beta,\theta)| \le O_p((nh)^{-1} + (nb)^{-1}h^{p+1}) = O_p((nh)^{-1} + (nh)^{-1}h^{p+2}b^{-1}) = O_p((nh)^{-1})$, under Assumption BW′.

17 Appendix E: Details in the inference of expected shortfall

For this example we write r_{t+H} and X_t as Y_i and X_i respectively, to be consistent with the generic framework above. The sample size is then $n = T - H$.

Let S be the $(p+1)$ by $(p+1)$ matrix with the (i,j)-th element $\int_{\mathcal{K}} u^{i+j-2}K(u)\mathrm{d}u$. Let S^* be the $(p+1)$ by $(p+1)$ matrix with the (i,j)-th element $\int_{\mathcal{K}} u^{i+j-2}K^2(u)\mathrm{d}u$. Assume S and S^* to be nonsingular.

E1. Asymptotic Variance. The matrices Ω and G in (9) take the form:

$$\Omega = \mu(x)^{-1}e_1'S^{-1}S^*S^{-1}e_1\begin{pmatrix}(1-\tau)\tau & \theta_0(1-\tau)\\ \theta_0(1-\tau) & \mathbb{E}[Y^2\mathbb{I}(Y<\beta_0)|X=x] - 2\tau\theta_0^2 + \tau^2\theta_0^2\end{pmatrix}$$

$$G = \begin{pmatrix}f(\beta_0|x) & 0\\ \beta_0 f(\beta_0|x) & -\tau\end{pmatrix}.$$

E2. Verification of Assumption UC(i). Let $g_0(\beta,\theta,X_i)=\mathbb{E}(g(Y_i,\beta,\theta)|X_i)$. We have

$$\begin{aligned}
& \sum_{i=1}^{n} W_p((X_i-x)/h)g(Y_i,\beta,\theta) \\
= \quad & e_1'[(nh)^{-1}\widehat{S}]^{-1}(nh)^{-1}\sum_{i=1}^{n}\varpi((X_i-x)/h)g(Y_i,\beta,\theta) \\
\overset{ULLN}{=} \quad & e_1'[(nh)^{-1}\widehat{S}]^{-1}\mathbb{E}[h^{-1}\varpi((X_i-x)/h)g(Y_i,\beta,\theta)]+o_p(1) \\
\overset{LIE}{=} \quad & e_1'[(nh)^{-1}\widehat{S}]^{-1}\mathbb{E}[h^{-1}\varpi((X_i-x)/h)g_0(\beta,\theta,X_i)]+o_p(1) \\
= \quad & e_1'[(nh)^{-1}\widehat{S}]^{-1}\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta,x+uh)\mu(x+uh)du+o_p(1) \\
\overset{(45)}{=} \quad & e_1'[(nh)^{-1}\widehat{S}]^{-1}\mu(x)g_0(\beta,\theta)\int_{\mathcal{K}}\varpi(u)du+o_p(1)=g_0(\beta,\theta)+o_p(1),
\end{aligned}$$

where all $o_p(1)$ terms are uniform in (β,θ), the second-to-last equality uses (45) below, and the last equality uses $(nh)^{-1}\widehat{S}\xrightarrow{p}\mu(x)S$ and $e_1'S^{-1}\int_{\mathcal{K}}\varpi(u)du=1$. The second equality uses ULLN (Andrews, 1987) for mixing sequences.

We have used

$$\sup_{(\beta,\theta)}\left|\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta,x+uh)\mu(x+uh)du-g_0(\beta,\theta)\mu(x)\int_{\mathcal{K}}\varpi(u)du\right|=o(1), \tag{45}$$

which follows from

$$\sup_{(\beta,\theta)}\left|\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta,x+uh)du-\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta)du\right|=o(1) \tag{46}$$

and

$$\sup_{(\beta,\theta)}\left|\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta,x+uh)[\mu(x+uh)-\mu(x)]du\right|=o(1). \tag{47}$$

Note that (47) holds by continuity of $\mu(\cdot)$ at x and bounded $g_0(\beta,\theta,\cdot)$ in a neighborhood of x uniformly in (β,θ) (Assumption ES (ii)), and (46) holds since

$$\begin{aligned}
& \sup_{(\beta,\theta)}\left|\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta,x+uh)du-\int_{\mathcal{K}}\varpi(u)g_0(\beta,\theta)du\right| \\
\leq \quad & \sup_{(\beta,\theta)}\int_{\mathcal{K}}|\varpi(u)||g_0(\beta,\theta,x+uh)-g_0(\beta,\theta)|du=o(1),
\end{aligned}$$

given that $\forall y$, $\overline{x}\mapsto f(y|\overline{x})$ is continuous at x and $\mathbb{E}(|Y||X=x)<C$. Thus (45) holds.

E3. Verification of Assumption UC (ii). It is satisfied by $V(\beta,\theta) = e_1' S^{-1} S^* S^{-1} e_1 \mathbb{E}[g_i(\beta,\theta) g_i(\beta,\theta)' | X = x]$. It can be shown similarly under Assumption ES.

E4. Verification of Assumption SE. We want to show $(\beta,\theta) \mapsto v_n(\beta,\theta)$ is SE. We have the decomposition

$$\begin{aligned} v_n(\beta,\theta) &= \sqrt{nh}[\widehat{m}(\beta,\theta) - g_0(\beta,\theta)] \\ &= \underbrace{e_1'[(nh)^{-1}\widehat{S}]^{-1} n^{-1/2} \sum_{i=1}^{n-H} [h^{-1/2}\varpi((X_i - x)/h) g(Y_i,\beta,\theta) - \mathbb{E} h^{-1/2}\varpi((X_i - x)/h) g(Y_i,\beta,\theta)]}_{:=T_1} \\ &\quad \underbrace{+(nh)^{-1/2} \sum e_1'[(nh)^{-1}\widehat{S}]^{-1} \mathbb{E}\varpi((X_i - x)/h) g(Y_i,\beta,\theta) - \sqrt{nh} g_0(\beta,\theta)}_{:=T_2}. \end{aligned}$$

As above, SE of $v_n(\beta,\theta)$ follows by SE of T_1 and negligibility of T_2, as we will show below.

We now show that $T_2 = o_p(1)$ uniformly in β and θ, which will be done by examining the two elements of T_2. The (1,1)-element of $(nh)^{-1/2} \sum e_1'[(nh)^{-1}\widehat{S}]^{-1} \mathbb{E}\varpi((X_i - x)/h) g(Y_i,\beta,\theta)$ is,

$$\begin{aligned} &\overset{LIE}{=} && (nh)^{-1/2} n e_1'[(nh)^{-1}\widehat{S}]^{-1} \mathbb{E}\varpi((X_i - x)/h)(F(\beta|X_i) - \tau) \\ &= && (nh)^{-1/2} n e_1'[(nh)^{-1}\widehat{S}]^{-1} \int_{\mathcal{X}} \varpi((u - x)/h)(F(\beta|u) - \tau)\mu(u) du \\ &\overset{z=(u-x)/h}{=} && (nh)^{1/2} \mu(x) e_1'[(nh)^{-1}\widehat{S}]^{-1} \int_{\mathcal{K}} \varpi(z)(F(\beta|x + zh) - \tau) dz + o_p(1) \\ &= && (nh)^{1/2} \mu(x) e_1'[(nh)^{-1}\widehat{S}]^{-1} \int_{\mathcal{K}} \varpi(z)[z^{p+1} h^{p+1} F^{(p+1)}(\beta|\dot{x}) + F(\beta|x) - \tau] dz + o_p(1) \\ &= && n^{1/2} h^{(2p+3)/2} e_1' S^{-1} \int_{\mathcal{K}} z^{p+1} \varpi(z) F^{(p+1)}(\beta|x) dz + (nh)^{1/2}(F(\beta|x) - \tau) + o_p(1), \end{aligned}$$

where $\dot{x}$ is a point between x and $x + zh$, and the last two equalities use $e_1' S^{-1} \int_{\mathcal{K}} z^k \varpi(z) dz = \mathbb{I}_{(k=0)}$ for $k = 0, 1, ..., p$ (Fan and Gijbels, 1996, (3.17)). Thus the first element of T_2 is $o_p(1)$ by the undersmoothing condition Assumption BW. By Assumption ES (v) and the bounded $F^{(p+1)}(\cdot|x)$ in a small neighborhood of β_0, the $o_p(1)$ terms above are uniform in β.

The (2,1)-element of T_2 follows similarly, except that Assumption ES (ii) is used in arguing for uniformity.

Consider the term T_1. We will show that the class of functions $\{h^{-1/2}\varpi((X_i - x)/h) g(Y_i,\beta,\theta) : (\beta,\theta) \in int(B \times \Theta)\}$ satisfies the L^2-continuity condition Andrews (1993,1994). The arguments for

its first element follows similarly, and we focus on its second element. We have

$$\begin{aligned}
& \mathbb{E} \sup_{|(\beta',\theta')-(\beta,\theta)|<\delta} |h^{-1/2}\varpi((X_t-x)/h)[Y_{t+H}\mathbb{I}_{(Y_{t+H}<\beta')} - \tau\theta' - Y_{t+H}\mathbb{I}_{(Y_{t+H}<\beta)} + \tau\theta]|^2 \\
= \; & h^{-1}\mathbb{E}\varpi^2((X_t-x)/h) \sup_{|(\beta',\theta')-(\beta,\theta)|<\delta} |Y_{t+H}[\mathbb{I}_{(Y_{t+H}<\beta')} - \mathbb{I}_{(Y_{t+H}<\beta)}] - \tau(\theta'-\theta)|^2 \\
\leq \; & h^{-1}\mathbb{E}\varpi^2((X_t-x)/h) \sup_{|(\beta',\theta')-(\beta,\theta)|<\delta} [2Y_{t+H}^2|\mathbb{I}_{(Y_{t+H}<\beta')} - \mathbb{I}_{(Y_{t+H}<\beta)}| + 2\tau^2(\theta'-\theta)^2] \\
\leq \; & h^{-1}\mathbb{E}\varpi^2((X_t-x)/h)[2Y_{t+H}^2\mathbb{I}_{(\beta-\delta<Y_{t+H}<\beta+\delta)} + 2\tau^2\delta^2] \leq O(\delta+\delta^2),
\end{aligned}$$

provided that $\mathbb{E}(Y_{t+H}^2|X_t = x) < \infty$. So by Andrews (1993, (4.7)), the L^2-bracketing number $N_2^B(\varepsilon) \leq C(1/\varepsilon)^2$. Then using Andrews (1993, Result 3(c), p. 200), T_1 is SE under Assumption ES (i) on the serial dependence, Assumption ES (ii), and the property (a) in the verification of Assumption SE for Example 1.

E5. Verification of Assumption AN. It follows from the asymptotic normality for the local polynomial estimator (Fan and Gijbels, 1996).

E6. Verification of Assumptions M and M′: For Assumption M, we need

$$\sup_{(\beta,\theta),1\leq i\leq n} |\varpi((X_i-x)/h)g_i(\beta,\theta)| = o_p((nh)^{1/2}). \tag{48}$$

Note that $\mathbb{P}\left(\sup_{(\beta,\theta),1\leq i\leq n}(nh)^{-1/a}|\varpi((X_i-x)/h)g_i(\beta,\theta)| > C\right) = \mathbb{P}(\sup_{1\leq i\leq n}(nh)^{-1}|\sup_{(\beta,\theta)}\varpi((X_i-x)/h)g_i(\beta,\theta)|^a > C^a) \leq \mathbb{P}(n^{-1}\sum_{i=1}^n h^{-1}|\sup_{(\beta,\theta)}\varpi((X_i-x)/h)g_i(\beta,\theta)|^a > C^a) \overset{Markov}{\leq} h^{-1}\mathbb{E}|\sup_{(\beta,\theta)}\varpi((X_i-x)/h)g_i(\beta,\theta)|^a/C^a$. Assumption M is thus satisfied if $\mathbb{E}(|Y_i|^a|X_i = x) < \infty$ for $a > 2$ (Assumption ES(ii)). Assumption M′ is satisfied under Assumption ES (ii').

References

Agarwal, V. and N.Y. Naik (2004): "Risks and Portfolio Decisions Involving Hedge Funds," *Review of Financial Studies*, 17, 63-98.

Anatolyev, S. (2005): "GMM, GEL, Serial Correlation, and Asymptotic Bias," *Econometrica*, 73, 983–1002.

Andrews, D.W.K. (1987): "Consistency in Nonlinear Econometric Models: A Generic Uniform Law of Large Numbers," *Econometrica*, 55, 1465-1472.

Andrews, D.W.K. (1993): "An Introduction to Econometric Applications of Empirical Process Theory for Dependent Random Variables," *Econometric Reviews,* 12, 183-216.

Andrews, D.W.K. (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. McFadden. New York: North-Holland, 2247-2294.

Andrews, D.W.K., and X. Cheng (2012): "Estimation and Inference with Weak, Semi-strong, and Strong Identification," *Econometrica*, 80, 2153-2211.

Angrist, J., D. Lang, and P. Oreopoulos (2009): "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 1, 136–63.

Antoine, B., H. Bonnal, and E. Renault (2007): "On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood," *Journal of Econometrics*, 138, 461–87.

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999): "Coherent Measures of Risk," *Mathematical Finance* 9, 203-228.

Calonico, S., M. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, forthcoming.

Card, D., D. Lee, Z. Pei, and A. Weber (2012): "Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in a Generalized Regression Kink Design," NBER Working Paper #18564.

Carroll, R., D. Ruppert and A. Welsh (1998): "Local Estimating Equations," *Journal of the American Statistical Association*, 93, 214-227.

Chen, X., H. Hong, and M. Shum (2007): "Nonparametric Likelihood Ratio Model Selection Tests Between Parametric Likelihood and Moment Condition Models," *Journal of Econometrics*, 141, 109-140.

Chernozhukov, V., and H. Hong (2003): "An MCMC Approach to Classical Estimation," *Journal of Econometrics*, 115, 293-346.

Chib, S. (2001): "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, Vol. 5., ed. by J. J. Heckman and E. Leamer. North-Holland, Amestradam, 3564–3634.

Curto, V.E., and R.G. Fryer (2015): "The Potential of Urban Boarding Schools for the Poor: Evidence from SEED," *Journal of Labor Economics*, Forthcoming.

Donald, S., G. Imbens, and W. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions," *Journal of Econometrics*, 117, 55-93.

Fan, J., and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall.

Fan, J., and Q. Yao (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Fan, Y., M. Gentry, and T. Li (2011): "A New Class of Asymptotically Efficient Estimators for Moment Condition Models," *Journal of Econometrics*, 162, 268–77.

Fan, Y., and R. Liu (2014): "A Direct Approach to Inference in Nonparametric and Semiparametric Quantile Models," Working paper, University of Washington.

Frandsen, B., M. Frolich, and B. Melly (2012): "Quantile Treatment Effects in the Regression Discontinuity Design," *Journal of Econometrics,* 168, 382-395.

Gagliardini, P, C. Gourieroux, and E. Renault (2011): "Efficient Derivative Pricing by the Extended Method of Moments," *Econometrica*, 79, 1181-1232.

Gan, L., and J. Jiang (1999): "A Test for Global Maximum," *Journal of the American Statistical Association*, 94, 847-854.

Gozalo, P., and O. Linton (2000): "Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression," *Journal of Econometrics*, 99, 63-106.

Guggenberger, P., and R. J. Smith (2005): "Generalized Empirical Likelihood Estimators and Tests under Partial, Weak and Strong Identification," *Econometric Theory,* 21, 667-709.

Hahn, J., Todd, P., and W. van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design," *Econometrica*, 69, 201-209.

Hall, A.R., and A. Inoue (2003): "The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models," *Journal of Econometrics*, 114, 361-394.

Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments estimators," *Econometrica*, 50, 1029–1054.

Hong, H., A. Mahajan, and D. Nekipelov (2012): "Extremum estimation and numerical derivatives," Working paper, UC-Berkeley.

Imbens, G.W., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G.W., and T. Lemieux (2008): "Regression Discontinuity Designs: a Guide to Practice," *Journal of Econometrics*, 142, 615-635.

Kai, B., R. Li, and H. Zou (2010): "Local Composite Quantile Regression Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression," *Journal of the Royal Statistical Society (Series B)*, 72, 49–69.

Kitamura, Y. (1997): "Empirical Likelihood Methods with Weakly Dependent Processes," *Annals of Statistics*, 25, 2084–2102.

Kitamura, Y. (2001): "Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 69, 1661–1672.

Kitamura, Y. (2006): "Empirical Likelihood Methods in Econometrics: Theory and Practice," in *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress*, ed. by R. Blundell, P. Torsten and W. K. Newey. Cambridge University Press, Cambridge.

Kitamura, Y., G. Tripathi, and H. Ahn (2004): "Empirical Likelihood-Based Inference in Conditional Moment Restriction Models," *Econometrica,* 72, 1667-1714.

Kleibergen, F. (2005): "Testing Parameters in GMM Without Assuming That They Are Identified," *Econometrica*, 73, 1103–1123.

Lazar, N. (2003): "Bayesian Empirical Likelihood," *Biometrika,* 90, 319-326.

Lee, D.S. (2008): "Randomized Experiments from Non-random Selection in U.S. House Elections," *Journal of Econometrics*, 142, 675–697.

Lewbel, A. (2007): "A Local Generalized Method of Moments Estimator," *Economics Letters*, 94, 124-128.

Lindo, J., N. Sanders, and P. Oreopoulos (2010): "Ability, Gender, and Performance Standards: Evidence from Academic Probation," *American Economic Journal: Applied Economics*, 2, 95-117.

Linton, O., and Z. Xiao (2013): "Estimation of and Inference about the Expected Shortfall for Time Series with Infinite Variance," *Econometric Theory,* 29, 771-807.

Marmer, V., D. Feir, and T. Lemieux (2014): "Weak Identification in Fuzzy Regression Discontinuity," working paper, UBC.

Molanes Lopez, E., I. van Keilegom, and N. Veraverbeke (2009): "Empirical Likelihood for Non-Smooth Criterion Functions," *Scandinavian Journal Statistics*, 36, 413-432.

NEWEY, W.K. (1984): "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters,* 14, 201-206.

NEWEY, W.K. (1985): "Generalized Method-of-Moments Specification Testing," *Journal of Econometrics,* 29, 229-256.

NEWEY, W.K., AND D.L. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, ed. by R. F. Engle and D. McFadden. New York: North-Holland, 2111-2245.

NEWEY, W.K., AND R.J. SMITH (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219-255.

OTSU, T. (2006): "Generalized Empirical Likelihood Inference for Nonlinear and Time Series Models under Weak Identification," *Econometric Theory*, 22, 513-527.

OTSU, T. (2008): "Conditional Empirical Likelihood Estimation and Inference for Quantile Regression Models," *Journal of Econometrics*, 142, 508-538.

OTSU, T., K.-L. XU, AND Y. MATSUSHITA (2013): "Estimation and Inference of Discontinuity in Density," *Journal of Business and Economic Statistics*, 31, 507-524.

OTSU, T., K.-L. XU, AND Y. MATSUSHITA (2014): "Empirical Likelihood for Regression Discontinuity Design," *Journal of Econometrics*, forthcoming.

OWEN, A. (2001) *Empirical Likelihood*. Chapman and Hall/CRC.

PARENTE, P., AND R.J. SMITH (2011): "GEL Methods for Non-Smooth Moment Indicators," *Econometric Theory,* 27, 74-113.

POLLARD, D. (1991): "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory,* 7, 186-199.

PORTER, J. (2003): "Estimation in the Regression Discontinuity Model," *Mimeo*, Department of Economics, University of Wisconsin.

QIN, J., AND J. LAWLESS (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300-325.

SCHENNACH, S.M. (2005): "Bayesian Exponentially Tilted Empirical Likelihood," *Biometrika*, 92, 31-46.

SMITH, R.J. (2011): "GEL Criteria for Moment Condition Models," *Econometric Theory*, 27, 1192–235.

STOCK, J.H., AND J.H. WRIGHT (2000): "GMM With Weak Instruments," *Econometrica*, 68, 1055–1096.

VAN DER VAART, A.W. (2000): *Asymptotic Statistics*. Cambridge University Press.

XU, K.-L. (2013): "Nonparametric Inference for Conditional Quantiles of Time Series," *Econometric Theory*, 29, 673-698.

XUE, L., AND L. ZHU (2007): "Empirical Likelihood for a Varying Coefficient Model with Longitudinal Data," *Journal of the American Statistical Association*, 102, 642-654.

YANG, Y., AND X. HE (2012): "Bayesian Empirical Likelihood for Quantile Regression," *Annal of Statistics*, 40, 1102–1131.

ZHAO, Z., AND Z. XIAO (2014): "Efficient Regressions via Optimally Combinning Quantile Information," *Econometric Theory*, forthcoming.

ZHU, S., AND M. FUKUSHIMA (2009): "Worst-Case Conditional Value-at-Risk with Application to Robust Portfolio Management," *Operations Research* 57,1155-1168.

DGP 1:								
	Point Estimator			Bias Corrected			MCMC Acceptance Rate	
C_{bw}	Bias	SE	RMSE	Bias	SE	RMSE	Mean	STD
1	0.0162	0.0838	0.0854	0.0121	0.1047	0.1054	0.381	0.112
2	0.0288	0.0618	0.0682	0.0191	0.0757	0.0780	0.311	0.096
3	0.0441	0.0512	0.0668	0.0332	0.0631	0.0713	0.263	0.083
3.5	0.0491	0.0461	0.0674	0.0400	0.0564	0.0691	0.247	0.080
DGP 2:								
	Point Estimator			Bias Corrected			MCMC Acceptance Rate	
C_{bw}	Bias	SE	RMSE	Bias	SE	RMSE	Mean	STD
1	0.0072	0.0796	0.0799	0.0044	0.1039	0.1040	0.381	0.118
2	0.0241	0.0561	0.0611	-0.0026	0.0648	0.0649	0.293	0.091
3	0.0605	0.0466	0.0764	0.0211	0.0517	0.0558	0.246	0.075
3.5	0.0828	0.0429	0.0933	0.0318	0.0494	0.0587	0.232	0.074

Table 1: Other (secondary) information in simulations. The last two columns contain the average and standard deviation of acceptance rates (over replications) in generating Markov chains (which are used with bias-corrected local estimating equations, as in Figures 3 and 4).

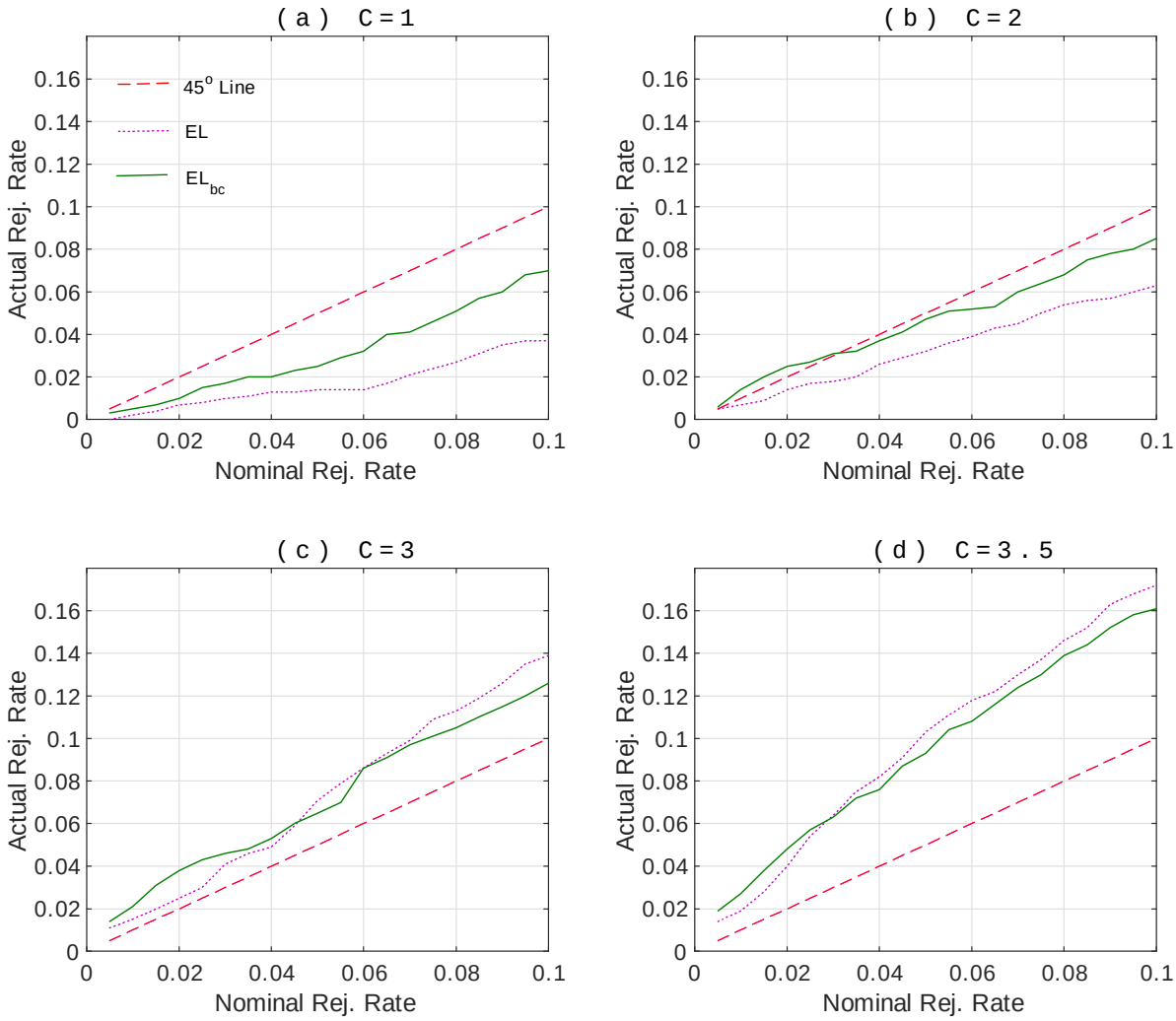

Figure 1: Null rejection rates (bias-corrected EL, ($\mathrm{EL}_{\mathrm{bc}}$) vs. uncorrected EL): DGP 1. The bandwidth used : $h = Cn^{-1/5}$, where $C \in \{1, 2, 3, 3.5\}$

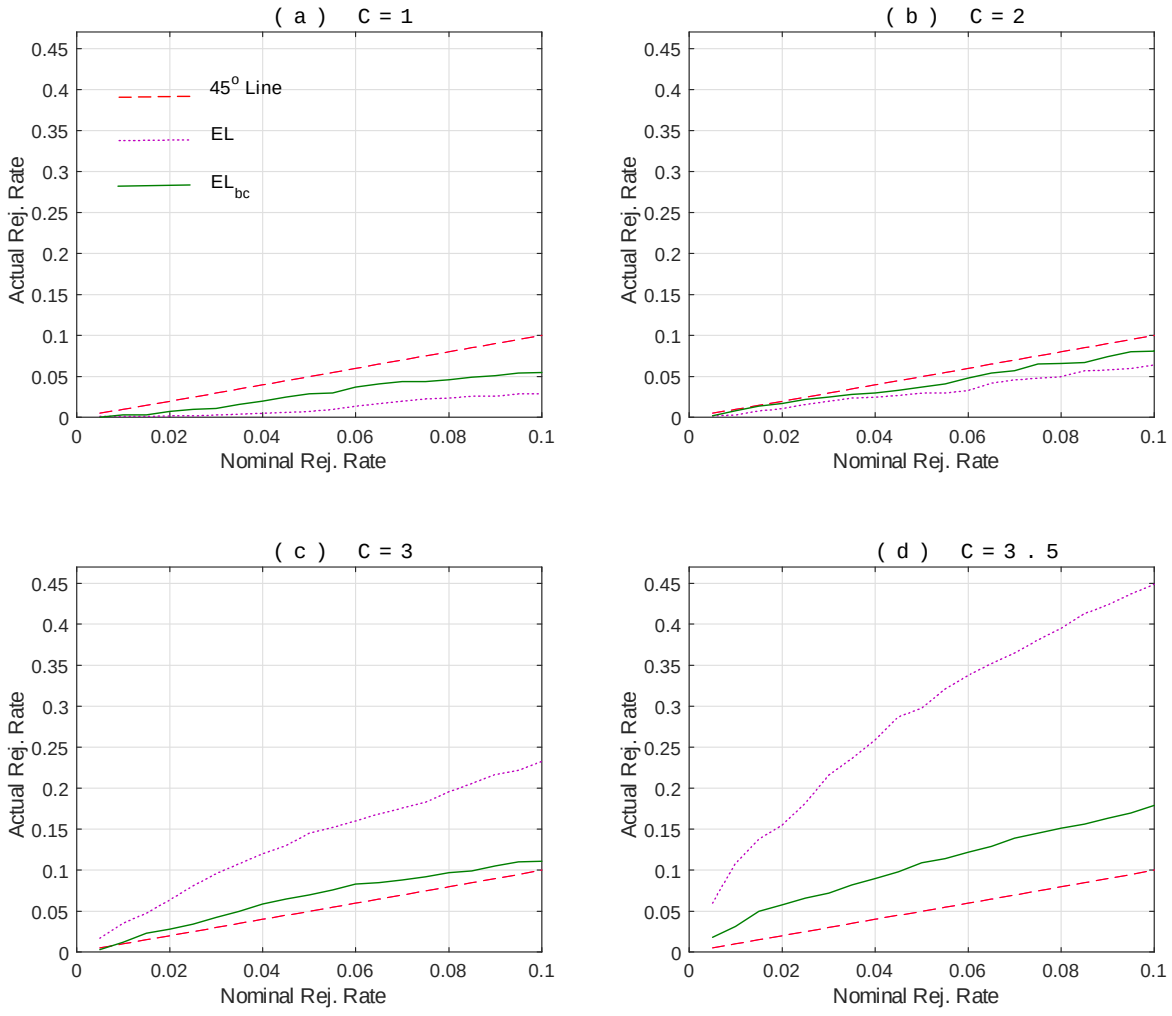

Figure 2: Null rejection rates (bias-corrected EL, ($\mathrm{EL}_{\mathrm{bc}}$) vs. uncorrected EL): DGP 2. The bandwidth used : $h = Cn^{-1/5}$, where $C \in \{1, 2, 3, 3.5\}$

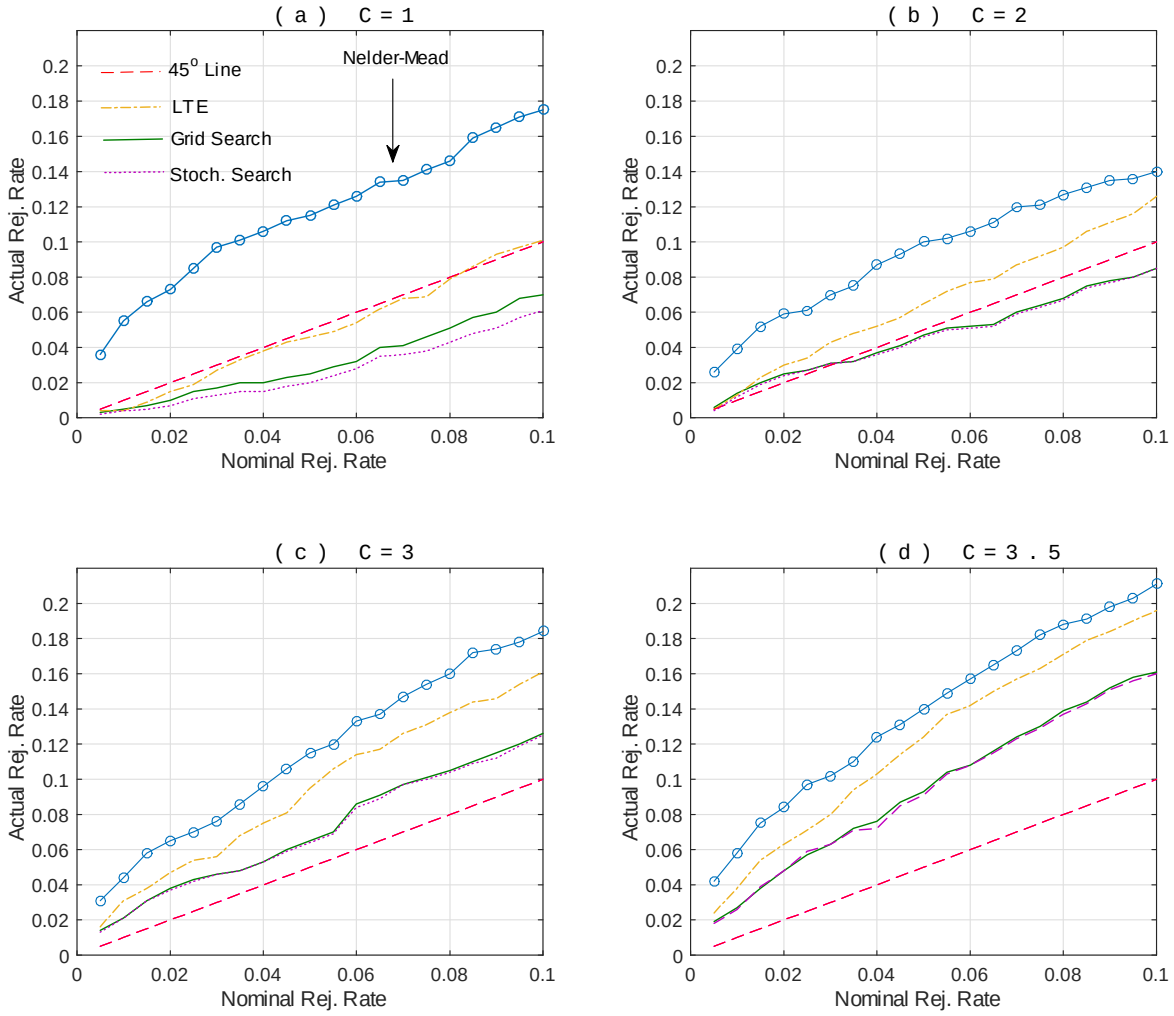

Figure 3: Null rejection rates (bias-corrected EL with various ways of eliminating the nuisance parameter): DGP 1. The bandwidth used : $h = Cn^{-1/5}$, where $C \in \{1, 2, 3, 3.5\}$

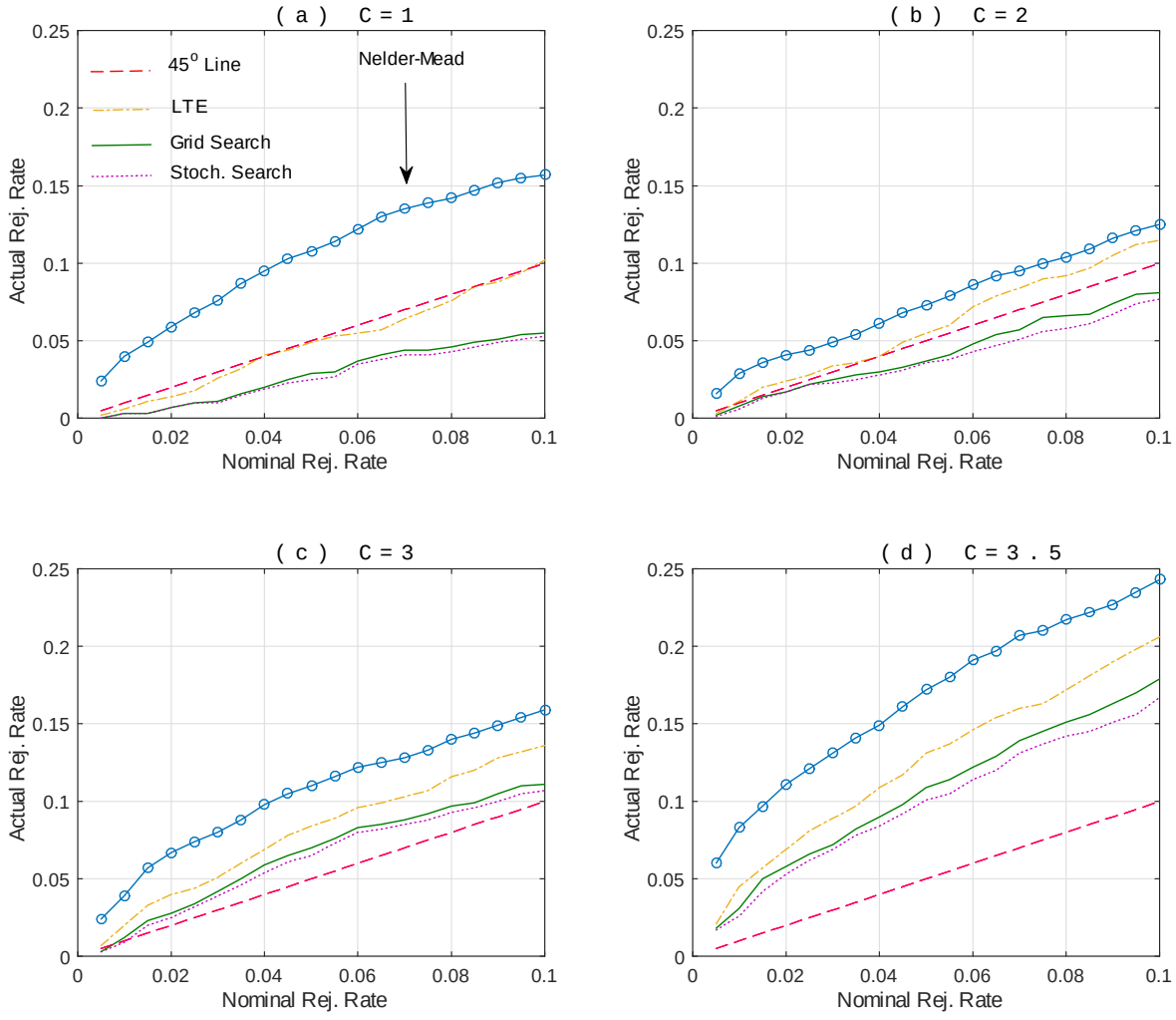

Figure 4: Null rejection rates (bias-corrected EL with various ways of eliminating the nuisance parameter): DGP 2. The bandwidth used : $h = Cn^{-1/5}$, where $C \in \{1, 2, 3, 3.5\}$

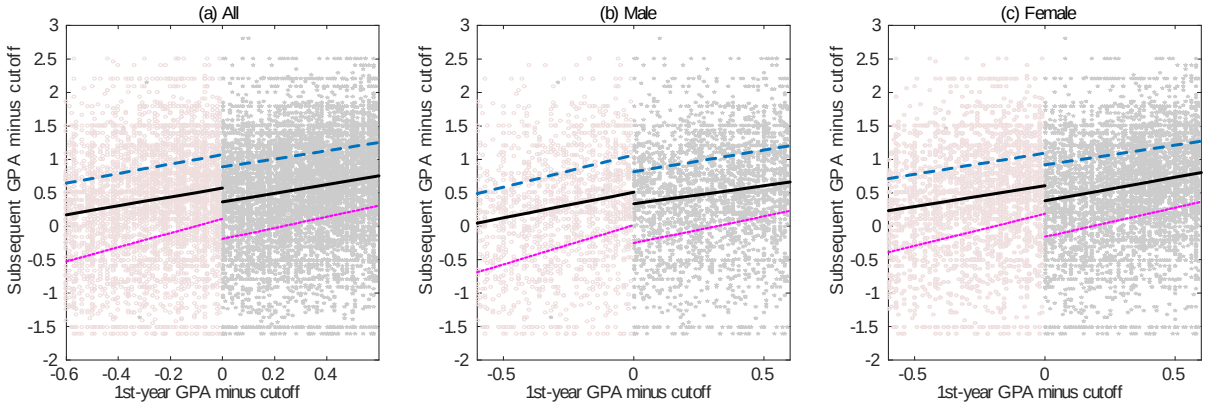

Figure 5: Observations in the $h = 0.6$ neighborhood, and three linear quartile regression lines (using observations in right and left neighborhoods), i.e. $\tau \in \{0.25, 0.5, 0.75\}$.

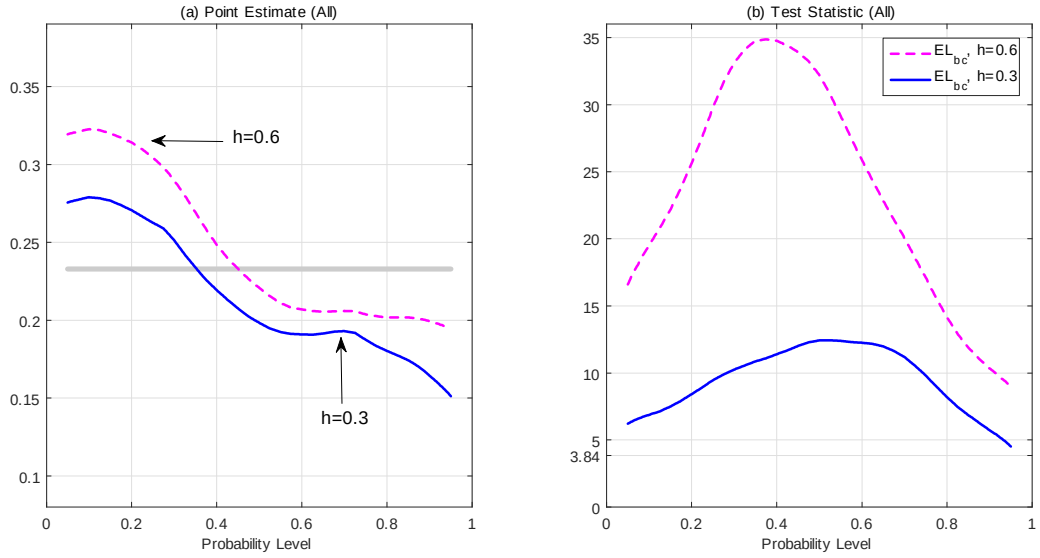

Figure 6: The entire sample: (a) The local linear estimate of the quantile treatment effect (the grey bar stands for the estimated ATE, as reported in Lindo et al., 2010); (b) The concentrated EL test statistic (with bias correction) of significance.

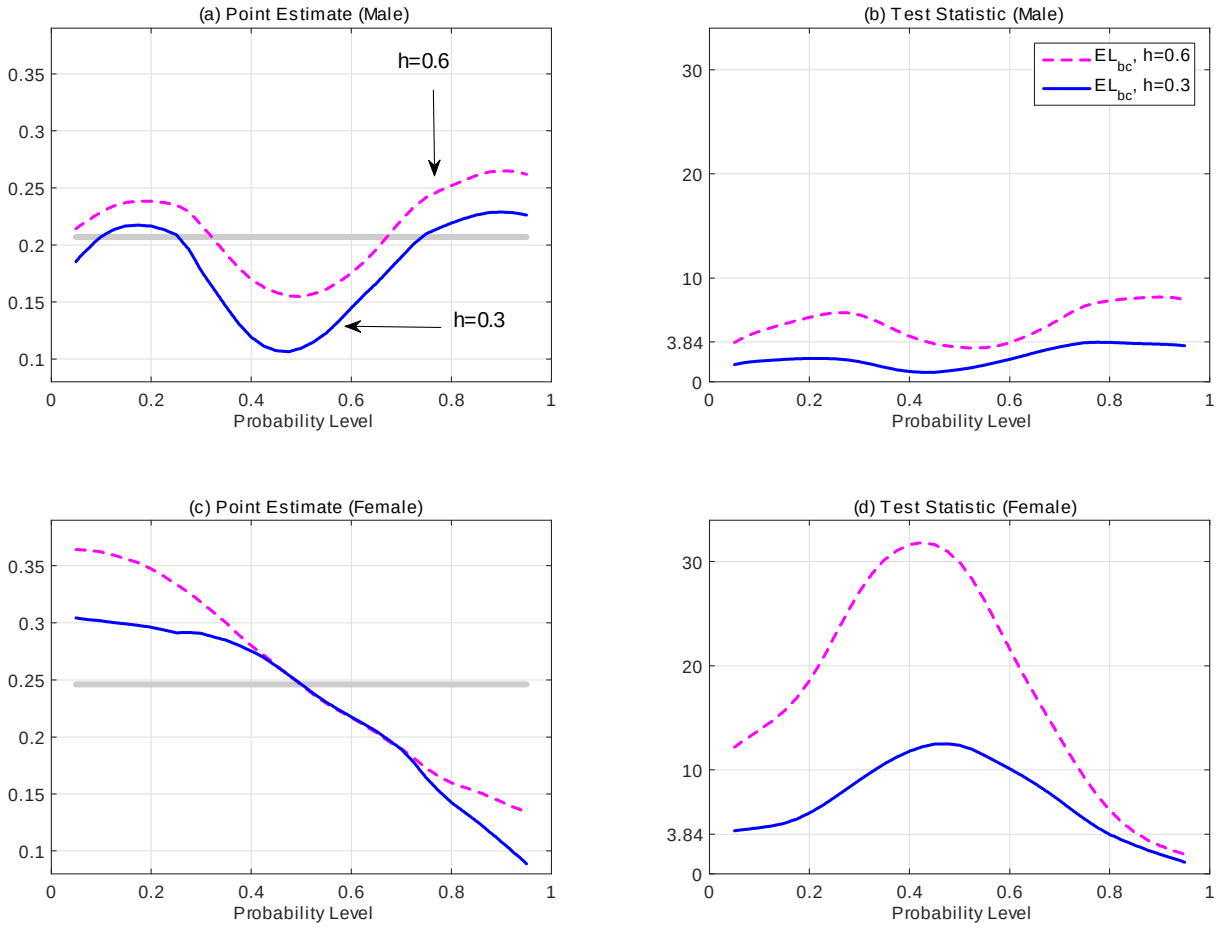

Figure 7: The subsamples of male and female students: (a) & (c) The local linear estimate of the quantile treatment effect (the grey bar stands for the estimated ATE, as reported in Lindo et al., 2010); (b) & (d) The concentrated EL test statistic (with bias correction) of significance.