Analysis of the Forward Search using some new results for martingales and empirical processes

Søren Johansen[1,2] & Bent Nielsen[3]

20 February 2014
Preliminary version

Summary: The Forward Search is an iterative algorithm for avoiding outliers in a regression analysis suggested by Hadi and Simonoff (1993), see also Atkinson and Riani (2000). The algorithm constructs subsets of 'good' observations so that the size of the subsets increases as the algorithm progresses. It results in a sequence of regression estimators and forward residuals. Outliers are detected by monitoring the sequence of forward residuals. We show that the sequences of regression estimators and forward residuals converge to Gaussian processes. The proof involves a new iterated martingale inequality, a theory for a new class of weighted and marked empirical processes, the corresponding quantile process theory, and a fixed point argument to describe the iterative aspect of the procedure.

Keywords: Fixed point result, Forward Search, iterated exponential martingale inequality, quantile process, weighted and marked empirical process.

1 Introduction

1.1 The Forward Search algorithm

The Forward Search algorithm was suggested for the multivariate location model by Hadi (1992) and for multiple regression by Hadi and Simonoff (1993) and developed further by Atkinson and Riani (2000), see also Atkinson, Riani and Cereoli (2010). It is an algorithm for avoiding outliers in a regression analysis by recursively constructing subsets of 'good' observations. The algorithm starts with a robust estimate of the regression parameters based on all observations, and constructs the set of observations with the smallest m_0 absolute residuals. It continues by estimating the parameters by least squares based on the m_0 observations selected. From this estimate the absolute residuals of all observations are computed and ordered. The $(m_0+1)'$st largest absolute residual is the forward residual and it is used to monitor the algorithm. The set of m_0+1 observations with the smallest absolute residuals is the starting point for the next iteration. The results of the analysis are plots of the

[1]The first author is grateful to CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.
[2]Department of Economics, University of Copenhagen and CREATES, Department of Economics and Business, Aarhus University, DK-8000 Aarhus C. E-mail: soren.johansen@econ.ku.dk.
[3]Nuffield College & Department of Economics, University of Oxford & Programme on Economic Modelling, INET, Oxford. Address for correspondence: Nuffield College, Oxford OX1 1NF, UK. E-mail: bent.nielsen@nuffield.ox.ac.uk.

recursively estimated forward residuals and estimates. This paper provides an asymptotic theory for these forward plots when applied to multiple regression under the assumption of no outliers.

1.2 Purpose of paper and results.

In this paper the forward plots are analysed for a multiple regression model. The model for the 'good' observations has symmetric, zero mean errors with unknown scale, while the regressors can be stationary as well as stochastically and deterministically trending. The plots of forward residuals and estimators are embedded as stochastic processes in $D[0,1]$, and their asymptotic properties are derived using new results on empirical processes and martingales. The results can be applied to construct pointwise and simultaneous confidence bands for the forward plots.

The first result is that the process of forward residuals behaves asymptotically as if the parameters were known. That is, as the process of ordered absolute errors from an i.i.d. sample from the error distribution. Such empirical quantile processes are studied by analysing the empirical distribution function as an empirical process. In order to show that the estimation uncertainty is negligible we introduce a class of weighted and marked empirical processes, where the weights represent functions of the predictable regressors and the marks are functions of the regression error. A technical difficulty is, that because the empirical processes are constructed from estimated residuals, the argument of the empirical process is stochastically varying. We develop the theory of such processes, applying and generalizing the results of Koul and Ossiander (1994).

In the second result, the process of forward residuals is scaled by recursive estimates of the unknown standard error. The limiting process is Gaussian and the covariance function is found.

In the study of weighted and marked empirical processes the well known method of replacing the discontinuous processes by their smooth compensators is applied. The difference is a martingale. To justify this replacement some new iterated exponential martingale inequalities for the variation of the maximum of finitely many martingales are developed by an iterative application of an exponential inequality of Bercu and Touati (2008).

1.3 History and background

The forward search starts with a robust estimator. Examples of robust regression estimators are the least median squares estimator and the least trimmed squares estimator of Rousseeuw (1984). These estimators are known to have good breakdown properties, see Rousseeuw and Leroy (1987, §3.4), and an asymptotic theory for the least trimmed squares regression estimator is provided by Víšek (2006a,b,c). We will allow initial estimators $\hat{\beta}^{(m_0)}$ converging at a rate slower than the usual $n^{1/2}$-rate, for the stationary case, as for example the least median squares estimator, which is $n^{1/3}$-consistent in location-scale models.

Broadly speaking, we require three asymptotic tools. First, a theory for weighted and marked empirical processes to describe the least squares statistics. Secondly, an analysis of the corresponding quantile processes to describe the forward residuals. Thirdly, a fixed point result to describe the iteration involved.

In the empirical process theory the weights represent functions of the regressors and the marks are functions of the regression error. The results generalize those of Johansen and Nielsen (2009) who did not allow stochastic variation in the quantiles and those of Koul and Ossiander (1994) who did not allow marks. The proof combines a chaining argument with iterations of an exponential inequality for martingales by Bercu and Touati (2008).

The quantile process theory draws on the exposition of Csörgő (1983). It is found that in the case of a known variance, the forward residuals satisfy a Bahadur representation, so that, asymptotically, the forward residuals have the same distribution as the order statistics of the absolute regression errors. When the variance is estimated, an additional term appears in the asymptotic distribution.

The last ingredient is a fixed point result to describe the iterative result. A single step of the algorithm has been discussed for the location-scale case by Johansen and Nielsen (2010). Starting with Bickel (1975) there are a number of asymptotic results for one-step L- and M-estimators. These are predominantly concerned with objective functions that have continuous derivatives, thereby excluding the hard rejection as for the one-step Huber-skip function. The forward search gives a sequence of one-step estimators. Because the estimators are based on least squares in a sample selected by truncating the residuals, each estimator is a one-step Huber-skip estimator. Such estimators have been studied by Rupert and Carroll (1980) and Johansen and Nielsen (2009, 2013), Ronchetti and Welsh (2002).

There appears to be less work on iteration of one-step estimators. The case of smooth weights was considered by Dollinger and Staudte (1991), but the case of 0-1 weights does not appear to have been studied until recently. Cavaliere and Georgiev (2013) analysed a sequence of Huber-skip estimators for a first order autoregression with infinite variance errors, while Johansen and Nielsen (2013) analysed sequences of one-step Huber-skip estimators with a fixed critical value. Here we need a critical value which changes with m so we need a generalisation of the fixed point result of the latter paper.

Outline of the paper: The model and the Forward Search algorithm are defined in §2. The main asymptotic results are given in §3. The weighted and marked empirical process results are given in §4 while the iterated exponential martingale inequalities are presented in §5 with proofs following in Appendix A and Appendix B. The proofs of the main results follow in Appendix C.

2 Model and Forward Search algorithm

The multiple regression model is presented, and the Forward Search algorithm is defined including the forward residual and forward deletion residual.

2.1 Model

We assume that (y_i, x_i), $i = 1, \ldots, n$ satisfy the multiple regression equation with regressors of dimension $\dim x$

$$y_i = x_i'\beta + \varepsilon_i, i = 1, \ldots, n. \tag{2.1}$$

The errors, ε_i, are assumed independent and identically distributed with mean zero and variance σ^2, and ε_i/σ has known density f and distribution function $\mathsf{F}(c) = \mathsf{P}(\varepsilon_i \leq \sigma c)$. In practice, the distribution F will often be standard normal.

The forward search is an algorithm based on ordering absolute residuals and calculation of least squares estimators from the selected observations. Both these choices implicitly assume a symmetric density, because truncating the errors symmetrically gives in general an error distribution with mean different from zero and hence biased least squares estimators, at least for the location parameter, unless symmetry is assumed.

The distribution function of the absolute errors $|\varepsilon_i|/\sigma$ of a symmetric density is $\mathsf{G}(c) = P(|\varepsilon_1| \leq \sigma c) = 2\mathsf{F}(c) - 1$ with density $\mathsf{g}(c) = 2\mathsf{f}(c)$. We define the quantiles of the absolute errors as

$$c_\psi = \mathsf{G}^{-1}(\psi) = \mathsf{F}^{-1}\{(1+\psi)/2\}, \psi \in [0,1], \tag{2.2}$$

and the truncated moments

$$\tau_\psi = \int_{-c_\psi}^{c_\psi} u^2 \mathsf{f}(u) du \text{ and } \varkappa_\psi = \int_{-c_\psi}^{c_\psi} u^4 \mathsf{f}(u) du. \tag{2.3}$$

Then the conditional variance of ε_1/σ given $\{|\varepsilon_1| \leq \sigma c\}$ is

$$\varsigma_\psi^2 = \tau_\psi/\psi. \tag{2.4}$$

This will serve as a bias correction for the variance estimator based on the truncated sample. If $\mathsf{f} = \varphi$ is Gaussian then $\varsigma_\psi^2 = 1 - 2c_\psi\varphi(c_\psi)/\psi$.

2.2 Forward Search algorithm

The Forward Search algorithm is designed to avoid outliers in a linear multiple regression. The first step is given by the choice of a robust estimator, $\hat{\beta}^{(m_0)}$, of the regression parameter, and the choice of the size m_0 of the initial set of 'good' observations. The algorithm generates a sequence of sets of 'good' observations and least squares regression estimators based on these. The $(m+1)'$st step of the algorithm is given as follows.

Algorithm 2.1 *(Forward Search)*

1. *Given an estimator* $\hat{\beta}^{(m)}$ *compute absolute residuals* $\hat{\xi}_i^{(m)} = |y_i - x_i'\hat{\beta}^{(m)}|$.
2. *Find the* $(m+1)$*st smallest order statistics* $\hat{z}^{(m)} = \hat{\xi}_{(m+1)}^{(m)}$.
3. *Find set of* $(m+1)$ *observations with smallest residuals* $S^{(m+1)} = (i : \hat{\xi}_i^{(m)} \leq \hat{z}^{(m)})$.
4. *Compute the new least squares estimators on* $S^{(m+1)}$

$$\hat{\beta}^{(m+1)} = (\textstyle\sum_{i\in S^{(m+1)}} x_i x_i')^{-1}(\sum_{i\in S^{(m+1)}} x_i y_i), \tag{2.5}$$

$$(\hat{\sigma}^{(m)})^2 = \frac{1}{m}\textstyle\sum_{i\in S^{(m)}}(y_i - x_i'\hat{\beta}^{(m)})^2. \tag{2.6}$$

Introduce also the bias corrected variance estimator using $\varsigma_{m/n}^2$ from (2.4) so that

$$(\hat{\sigma}_{corr}^{(m)})^2 = \frac{(\hat{\sigma}^{(m)})^2}{\varsigma_{m/n}^2}. \tag{2.7}$$

Applying the algorithm for $m = m_0, \ldots, n-1$, results in sequences of order statistics $\hat{z}^{(m)} = \hat{\xi}^{(m)}_{(m+1)}$, least squares estimators $(\hat{\beta}^{(m)}, (\hat{\sigma}^{(m)})^2)$, along with the scaled forward residuals

$$\frac{\hat{z}^{(m)}}{\hat{\sigma}^{(m)}} = \frac{\hat{\xi}^{(m)}_{(m+1)}}{\hat{\sigma}^{(m)}}.$$

Atkinson and Riani (2000) propose to use the scaled minimum deletion residual

$$\frac{\hat{d}^{(m)}}{\hat{\sigma}^{(m)}} = \min_{i \notin S^{(m)}} \frac{\hat{\xi}^{(m)}_i}{\hat{\sigma}^{(m)}},$$

instead of the forward residuals. Thus the deletion residual is based on the smallest residual with respect to $\hat{\beta}^{(m)}$ among those observations that were not included in $S^{(m)}$ which in turn is based on $\hat{\beta}^{(m-1)}$, and the forward residual is the largest absolute residual in $S^{(m+1)}$ which is based on $\hat{\beta}^{(m)}$.

The plots of $\hat{\beta}^{(m)}$, $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$, and $\hat{d}^{(m)}/\hat{\sigma}^{(m)}$ against m are called forward plots, see Atkinson and Riani (2000, p.12-13). The primary objective of this paper is to derive the asymptotic distribution of these plots.

When the method was proposed by Hadi and Simonoff (1993), they also suggested scaling the residual by a leverage factor and replace the scaled residuals $\hat{\xi}^{(m)}_i/\hat{\sigma}^{(m)}$ above by

$$\frac{\xi^{(m)}_i}{\hat{\sigma}^{(m)}\sqrt{1-h^{(m)}_i}} \text{ for } i \in S^{(m)}, \quad \frac{\xi^{(m)}_i}{\hat{\sigma}^{(m)}\sqrt{1+h^{(m)}_i}} \text{ for } i \notin S^{(m)},$$

where $h^{(m)}_i = x'_i(\sum_{j \in S^{(m)}} x_j x'_j)^{-1} x_i$ is the leverage factor. Atkinson and Riani (2006) suggest a leverage factor $1 + h^{(m)}_i$ for all observations. Johansen and Nielsen (2009) prove that such a leverage factor does not change the asymptotic distribution for the one-step Huber skip estimator, and the methods presented there can be used to prove a similar result for the forward search.

3 The main results

Johansen and Nielsen (2010, Theorems 5.1–5.3) analysed a single step of the Forward Search applied in a location-scale setting. Those results show that the one-step version of the scaled residuals $\hat{z}^{(m)}/\hat{\sigma}^{(m)}$ has an asymptotic representation involving an empirical process and a term arising from the estimation error for the variance. The subsequent analysis shows how this result generalizes to a fully iterated Forward Search. This section first gives the assumptions, then the results, and finally presents some simulations. The derivatives of f are denoted $\dot{f}$ and $\ddot{f}$ and for more complicated expressions by d/dx.

3.1 Assumptions

In the following a series of sufficient assumptions are listed for the asymptotic theory of the Forward Search. When using the Forward Search, the density f is assumed known. The leading case is the normal density, φ, but the results are also discussed for the t-density.

Assumption 3.1 *Let $\mathcal{F}_i$ be an increasing sequence of σ fields so ε_{i-1} and x_i are $\mathcal{F}_{i-1}$-measurable and ε_i is independent of $\mathcal{F}_{i-1}$ with symmetric continuous, differentiable density* f *which is positive for $\mathsf{F}^{-1}(0) < c < \mathsf{F}^{-1}(1)$. For some $0 \leq \kappa < \eta \leq 1/4$ choose an $r \geq 2$ so $2^{r-1} \geq 1 + (1/4 + \kappa - \eta)(1 + \dim x)$. Let $q_0 = 1 + 2^{r+1}$. Suppose*

(i) *density satisfies*

- (a) *tail monotonicity: $c^q \mathsf{f}(c)$, $|c^{q-1}\dot{\mathsf{f}}(c)|$ are decreasing for large c and some $q > q_0$;*
- (b) *quantile process condition: $\gamma = \sup_{c>0} \mathsf{F}(c)\{1 - \mathsf{F}(c)\}|\dot{\mathsf{f}}(c)|/\{\mathsf{f}(c)\}^2 < \infty$;*
- (c) *unimodality: $\dot{\mathsf{f}}(c) \leq 0$ for $c > 0$ and $\lim_{c \to 0} \dot{\mathsf{f}}(c) < 0$;*
- (d) *logarithmic derivative: $\Delta(c) = [\frac{d}{dc}\{c\frac{d}{dc}\log \mathsf{f}(c)\}] < 0$ for $c > 0$;*
- (e) *strong quantile process condition: $\{1 - \mathsf{F}(c)\}/\{c\mathsf{f}(c)\} = \mathrm{O}(1)$ for $c \to \infty$;*

(ii) *regressors x_i are $\mathcal{F}_{i-1}$-measurable and a normalisation matrix N exists so that*

- (a) $\Sigma_n = N' \sum_{i=1}^n x_i x_i' N \overset{\mathsf{D}}{\to} \Sigma \overset{a.s.}{>} 0$;
- (b) $\max_{1 \leq i \leq n} |n^{1/2-\kappa} N' x_i| = \mathrm{O}_{\mathsf{P}}(1)$ *for some $\kappa < \eta$;*
- (c) $n^{-1}\mathsf{E}\sum_{i=1}^n |n^{1/2} N' x_i|^{q_0} = \mathrm{O}(1)$;

(iii) *initial estimator: $N^{-1}(\hat{\beta}^{(m_0)} - \beta) = \mathrm{O}_{\mathsf{P}}(n^{1/4-\eta})$ for some $\eta > 0$.*

Assumption 3.1(i) is satisfied for the normal distribution. For other distributions the regularity conditions involve a trade-off between four features: η, which indicates the rate of the initial estimator, κ, which indicates the order of magnitude of maximum of the normalised regressors, and $\dim x$, the dimension of the regressor. From these quantities a number r is defined, which controls the number of moments and the smoothness required for the density f. The number r is increasing in κ and $\dim x$ and decreasing in η. The number of required moments, $1 + 2^{r+1}$, is larger than 8 in order to control the estimation error for the variance.

Assumption 3.1 is satisfied in a range of situations. First some general comments. Condition (ia) is more severe than normally seen in empirical process theory due to the marks ε_i^p. Condition (ib) is used in Theorem D.1. Conditions (ic, id) are needed for controlling the iterative aspect of the Forward Search. Condition (id) to $\Delta(c)$ is also used in Rousseuw (1982) when discussing change-of-variance curves for M-estimators. It is satisfied for log concave densities. It is also the cross derivative of the log likelihood for location-scale families. Condition (ie) to Mill's ratio is milder than the condition employed for kernel density estimation by Csörgő (1983, p. 139). Condition (iia) is standard in regression analysis. Condition (iib) is discussed in Example 3.1 below.

As part of the proof, a class of weighted and marked empirical processes are analysed in §4 and at that point somewhat weaker assumptions are introduced, see Assumption 4.1.

Example 3.1 ***Assumption 3.1**(i) **for the reference distribution*** f.

(a) ***Standard normal distribution**, $\mathsf{f} = \varphi$. Condition (i) is satisfied: (ia) holds since $c^q\varphi(c) = -c^{q-1}\dot{\varphi}(c)$ is decreasing for large c for any q. (ib) holds with $\gamma = 1$, noting $\dot{\varphi}(c) = -c\varphi(c)$ and the Mill's ratio result $\{(4+c^2)^{1/2} - c\}/2 < \{1 - \Phi(c)\}/\varphi(c) < 1/c$, see Sampford (1953). (id) holds with $\Delta(c) = -2c$. (ie) holds since $\{1 - \Phi(c)\}/\{c\varphi(c)\} < 1/c^2 \to 0$ as $c \to \infty$.*

(b) ***Scaled distribution.** Consider a density $\mathsf{f}_\delta(c)$ that has variance δ^2 but otherwise satisfies condition (i). Then $\mathsf{f}(c) = \delta\mathsf{f}_\delta(c\delta)$ has unit variance, distribution function $\mathsf{F}(c) = \mathsf{F}_\delta(c\delta)$ and satisfies condition (i) with the same γ in part (b).*

(c) ***Scaled t-distribution.** The t-distribution with $d > 2^{r+1}$ degrees of freedom has density*

$\mathsf{f}_d(c) = C_d(1+c^2/d)^{-(d+1)/2}$ *with* $C_d = \Gamma\{(d+1)/2\}/\{(d\pi)^{1/2}\Gamma(d/2)\}$ *and variance* $\delta_d^2 = d/(d-2)$. *Thus, the t-distribution scaled by* δ_d *has density* $\mathsf{f}(c) = \mathsf{f}_d(c\delta_d)\delta_d$ *and distribution function* $\mathsf{F}(c) = \mathsf{F}_d(c\delta_d)$. *It suffices to check condition* (i) *for the density* f_δ. *Condition* (i) *is satisfied:* (ia) *for some constants* C, *it holds* $c^q\mathsf{f}_d(c) \sim Cc^{q-d-1}$ *and* $c^{q-1}|\dot{\mathsf{f}}_d(c)| = Cc^{q-1}\mathsf{f}_d(c)h(c)$, $h(c) = c/(1+c^2/d) \sim c^{-1}$, *so that* $c^{q-1}|\dot{\mathsf{f}}_d(c)| \sim Cc^{q-d-3}$. *Thus* $c^q\mathsf{f}_d(c)$ *and* $c^{q-1}|\dot{\mathsf{f}}_d(c)|$ *are both declining for large* c, *for* q *chosen so* $d+1 > q > q_0$. (ib) *holds with the stated* γ *since* $1 - c^{-2}d/(d+2) < h(c)\{1-\mathsf{F}_d(c)\}/\mathsf{f}_d(c) < 1$, *see Soms (1976, equation 3.2).* (ic) *is well-known to hold.* (id) *holds with* $\Delta(c) = -2\gamma\{h(c)\}^2/c < 0$. (ie) *holds since* $\{1-\mathsf{F}_d(c)\}/\{c\mathsf{f}_d(c)\} < 1/\{ch(c)\} \to 1/d$ *as* $c \to \infty$.

Example 3.2 ***Assumption 3.1***(ii) ***for the regressors*** x_i.
(a) ***Stationary regressors.*** *Let* $N = n^{-1/2}I_{\dim x}$. *To ensure* (iic) *it is necessary that* $\mathsf{E}|x_i|^{q_0} < \infty$. *By Boole's inequality and the triangle inequality then* $n^{1/2-\kappa}\max_{1\leq i\leq n}|N'x_i| = \mathrm{O}_{\mathsf{P}}(n^{1-\kappa q_0})$ *so* (iib) *holds for all* $\eta > \kappa = q_0^{-1}$.
(b) ***Deterministic regressors*** *such as* $x_i = (1,i)'$. *Let* $N = \mathrm{diag}(n^{-1/2}, n^{-3/2})$. *Then* $n^{1/2}N'x_i = (1, i/n)'$. *Thus condition* (ii) *follows with* $\kappa = 0$.
(c) ***Random walk regressors*** *such as* $x_i = \sum_{s=1}^{i-1}\varepsilon_s$. *Let* $N = n^{-1}$. *Then* $n^{-1/2}x_{\mathrm{int}(n\psi)}$ *converges to a Brownian motion by Donsker's invariance principle, see Billingsley (1968). Condition* (iia, iib) *follows from the continuous mapping theorem with* $\kappa = 0$. *As* x_i *is defined in terms of* ε_i *which has moments of order* q_0, *so has* x_i *and* (iic) *follows.*

3.2 The results

The forward plot of for instance $\hat{z}^{(m)}$ is a process on $m = m_0, \ldots, n-1$. It is useful to embed it in the space $D[0,1]$ of right continuous process on $[0,1]$ with limits from the left, endowed with the uniform norm since all limiting processes will be continuous. Thus, define

$$\hat{z}_\psi = \begin{cases} \hat{z}^{(m)} & \text{for } m = \mathrm{int}(n\psi) \text{ and } m_0/n \leq \psi \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Embed in a similar way $\hat{\beta}^{(m)}$, $\hat{\sigma}^{(m)}$ as $\hat{\beta}_\psi$, $\hat{\sigma}_\psi$.

The main results are described in terms of three processes

$$\mathbb{G}_n(c_\psi) = n^{-1/2}\textstyle\sum_{i=1}^n\{1_{(|\varepsilon_i/\sigma|\leq c_\psi)} - \psi\}, \tag{3.2}$$
$$\mathbb{L}_n(c_\psi) = \tau_\psi^{-1}n^{-1/2}\textstyle\sum_{i=1}^n[\{(\varepsilon_i/\sigma)^2 - c_\psi^2\}1_{(|\varepsilon_i/\sigma|\leq c_\psi)} - (\tau_\psi - c_\psi^2\psi)], \tag{3.3}$$
$$\mathbb{K}_n(c_\psi) = n^{1/2}\textstyle\sum_{i=1}^n N'x_i(\varepsilon_i/\sigma)1_{(|\varepsilon_i/\sigma|\leq c_\psi)}, \tag{3.4}$$

The first two are asymptotically Gaussian processes and the same holds for the third if the regressors are stationary, see Theorem 3.6.

The main results give asymptotic representations of the forward residuals $\hat{z}_\psi/\sigma$ scaled with known scale, of the bias corrected variance, and of the forward residuals $\hat{z}_\psi/\hat{\sigma}_{\psi,corr}$ scaled with the bias corrected variance estimator. Next, it is shown that the forward residuals and the deletion residuals have the same asymptotic representation after an initial burn-in period. Finally, an asymptotic representation is given for the forward plot of regression estimators. The proof of these results are given in Appendix D.

Theorem 3.1 *Suppose Assumption 3.1 holds. Let $\psi_0 > 0$. Then*

$$\sup_{\psi_0 \leq \psi \leq n/(n+1)} |2\mathsf{f}(c_\psi) n^{1/2}(\sigma^{-1}\hat{z}_\psi - c_\psi) + \mathbb{G}_n(c_\psi)| \overset{\mathsf{P}}{\to} 0. \tag{3.5}$$

Moreover, if $\hat{c}_{m/n}$ are the order statistics of $\xi_i/\sigma = |\varepsilon_i|/\sigma$, then

$$\sup_{\psi_0 \leq \psi \leq n/(n+1)} |\mathsf{f}(c_\psi) n^{1/2}(\sigma^{-1}\hat{z}_\psi - \hat{c}_\psi)| \overset{\mathsf{P}}{\to} 0. \tag{3.6}$$

If β and σ were known, the residuals are the errors, ε_i, and the ordering of the absolute residuals $\xi_i = |y_i - \beta' x_i| = |\varepsilon_i|$ can be done once, so that $\sigma^{-1}\hat{z}_m = \sigma^{-1}\xi_{(m+1)} = \hat{c}_{(m+1)/n}$, and the left hand side of (3.6) is trivially zero. In this situation (3.5) reduces to the Bahadur (1966) representations for the order statistics of the errors ξ_i, see also Theorem D.1 in the Appendix. Theorem 3.1 therefore has the interpretation that in the forward search the process $\sigma^{-1}\hat{z}_\psi$ behaves asymptotically as if the parameters were known.

Theorem 3.2 *Suppose Assumption 3.1 holds. Let $\psi_0 > 0$. Then*

$$\sup_{\psi_0 \leq \psi \leq n/(n+1)} |n^{1/2}(\sigma^{-2}\hat{\sigma}^2_{\psi,corr} - 1) - \mathbb{L}_n(c_\psi)| = \mathrm{o}_{\mathsf{P}}(1).$$

Remark 3.1 *In Theorems 3.1 and 3.2 the supremum is taken over a smaller interval for ψ than the unit interval. A left end point larger than 0 is needed to ensure consistency. The results potentially hold with a right end point equal to 1. Proving this would, however, add significantly to the length of the proof without practical benefit since the last forward residual is based on the set $S^{(n-1)}$ with $n-1$ selected observations.*

Remark 3.2 *The least squares estimator for the variance is $\hat{\sigma}^2_{1,corr} = \hat{\sigma}^2_1$, noting that $\tau_1 = 1$ and $\varsigma_1 = 1$. Least squares theory shows that $n^{1/2}(\hat{\sigma}^2_1/\sigma^2 - 1) = n^{-1}\sum_{i=1}^n(\varepsilon_i^2/\sigma^2 - 1) + \mathrm{o}_{\mathsf{P}}(1)$. To see that Theorem 3.2 matches this result, note that the leading term of the least squares approximation is $\lim_{\psi\to 1}\tau_\psi^{-1} n^{-1/2}\sum_{i=1}^n\{(\varepsilon_i/\sigma)^2 1_{(|\varepsilon_i/\sigma|\leq c_\psi)} - \tau_\psi\}$. It is therefore necessary that the other term in $\mathbb{L}_n(\psi)$ satisfies $\lim_{\psi\to 1}\tau_\psi^{-1} c_\psi^2 n^{-1/2}\sum_{i=1}^n\{1_{(|\varepsilon_i/\sigma|\leq c_\psi)} - \psi\} = \lim_{\psi\to 1} c_\psi^2 \mathbb{G}_n(\psi) = \mathrm{o}_{\mathsf{P}}(1)$. Since ε_i has more than 8 moments then $c_\psi^2 = \mathrm{o}\{(1-\psi)^{-1/4}\}$, see also item 5 of the proof of Lemma D.14. Combine this with Theorems D.2(a), D.3 to see that $\lim_{\psi\to 1} c_\psi^2 \mathbb{G}_n(c_\psi) = \mathrm{o}_{\mathsf{P}}(1)$.*

Combining Theorem 3.1 and 3.2 gives an asymptotic representation of the forward residuals with a biased corrected scale.

Theorem 3.3 *Suppose Assumption 3.1 holds. Let $c_\psi = \mathsf{G}^{-1}(\psi)$ and $\psi_0 > 0$. Then the bias corrected scaled forward residuals has the expansion*

$$\sup_{\psi_0 \leq \psi \leq n/(n+1)} |2\mathsf{f}(c_\psi) n^{1/2}(\frac{\hat{z}_\psi}{\hat{\sigma}_{\psi,corr}} - c_\psi) + \mathbb{G}_n(c_\psi) + c_\psi \mathsf{f}(c_\psi)\mathbb{L}_n(c_\psi)| \overset{\mathsf{P}}{\to} 0.$$

The above results generalize those of Johansen and Nielsen (2010, Theorems 5.1, 5.3) which hold for a single forward step for location-scale models. It is interesting to note that the results do not depend on the type of regressors for the model. In particular, the results do not depend on whether the regressors include an intercept or not, which sets the results aside from empirical processes of residuals, compare for instance Engler and Nielsen (2009, Theorem 2.1) and Lee and Wei (1999, Theorem 3.2).

In finite samples the forward residuals and the deletion residuals can be different, see for instance Johansen and Nielsen (2010, §2.2). The next result implies that $\widehat{d}^{(m)}$ and $\widehat{z}^{(m)}$ have the same asymptotic distribution.

Theorem 3.4 *Suppose Assumption 3.1 holds. Let* $m_0 = \text{int}(n\psi_0)$ *where* $\psi_0 > 0$. *Then for all* ψ_1 *so* $\psi_0 < \psi_1 < 1$ *it holds*

$$\sup_{\psi_1 \leq \psi \leq n/(n+1)} |\mathsf{f}(c_\psi) n^{1/2} (\widehat{z}^{(m)} - \widehat{d}^{(m)})| \xrightarrow{\mathsf{P}} 0.$$

The last result is for the forward plot of the estimator error $N^{-1}(\widehat{\beta}^{(m)} - \beta)$, which can be analysed in two stages. First it is established that $N^{-1}(\widehat{\beta}^{(m)} - \beta)$ satisfies a recursion of the form

$$N^{-1}(\widehat{\beta}^{(m+1)} - \beta) = \rho_{m/n} N^{-1}(\widehat{\beta}^{(m)} - \beta) + (\psi\Sigma_n)^{-1}\mathbb{K}_n(c_\psi) + e_{m/n}\{N^{-1}(\widehat{\beta}^{(m)} - \beta)\},$$

where $\rho_\psi = 2c_\psi \mathsf{f}(c_\psi)/\psi$ is an 'autoregressive coefficient' and e_ψ is a vanishing remainder term. By iterating this relation one can prove the result in the next theorem.

Theorem 3.5 *Suppose Assumption 3.1 holds and that* $\psi > 2c_\psi \mathsf{f}(c_\psi)$. *Let* $m_0 = \text{int}(n\psi_0)$ *where* $\psi_0 > 0$. *Then, for all* ψ_1 *so* $\psi_0 < \psi_1 < 1$, *the forward plot of the estimator has the expansion*

$$\sup_{\psi_1 \leq \psi \leq 1} |N^{-1}(\widehat{\beta}_\psi - \beta) - \frac{1}{\psi - 2c_\psi \mathsf{f}(c_\psi)} \Sigma_n^{-1} \mathbb{K}_n(c_\psi)| = \mathrm{o}_{\mathsf{P}}(1).$$

This result generalizes that of Johansen and Nielsen (2010, Theorem 5.2) who considered as single forward step for the location model. The condition $\psi > 2c_\psi \mathsf{f}(c_\psi)$ implies that $0 < \rho_\psi < 1$, see Lemma D.10, so the recursion is a contraction. Johansen and Nielsen (2013) established a similar result for the iterated one-step Huber-skip estimator for a fixed ψ.

3.3 Applications of the result for the forward residuals

The statements of Theorems 3.1, 3.3, 3.4 for the forward residuals and Theorem 3.2 do not depend on the type of regressor. Thus, to apply these theorems it suffices to analyse the asymptotically Gaussian processes $\mathbb{G}_n$, $\mathbb{L}_n$ and $\mathbb{K}_n$ for the chosen reference distribution.

Theorem 3.6 *Suppose Assumptions 4.1 holds. Then* $\mathbb{G}_n$ *and* $\mathbb{L}_n$ *converge to zero mean Gaussian processes with variances given by*

$$\mathsf{Var}\{\mathbb{G}_n(c_\psi)\} = 2\psi(1-\psi), \tag{3.7}$$

$$\mathsf{Var}\{\mathbb{L}_n(c_\psi)\} = \frac{1}{\tau_\psi^2}\{\varkappa_\psi - \tau_\psi^2 + c_\psi^2(1-\psi)(c_\psi^2\psi - 2\tau_\psi)\}, \tag{3.8}$$

where the truncated moments τ_ψ and $\varkappa_\psi$ are given in (2.3). The covariance of the processes $\mathbb{G}_n$ and $\mathbb{L}_n$ satisfies

$$\mathsf{Cov}\{\mathbb{G}_n(c_\psi), \mathbb{L}_n(c_\psi)\} = \frac{1}{\tau_\psi}(\tau_\psi - c_\psi^2\psi)(1-\psi) < 0. \tag{3.9}$$

The following pointwise results arise for $\psi_0 \leq \psi \leq \psi_1$, for some $\psi_0 > 0$ and $\psi_1 < 1$,

$$n^{1/2}(\frac{\hat{z}_\psi}{\hat{\sigma}_\psi} - \frac{c_\psi}{\varsigma_\psi}) = n^{1/2}\frac{\hat{z}_\psi\zeta_\psi - \hat{\sigma}_\psi c_\psi}{\hat{\sigma}_\psi\varsigma_\psi} \overset{\mathsf{D}}{\rightarrow} \mathbb{N}(0, \omega_\psi), \tag{3.10}$$

where ω_ψ has contributions from $\hat{z}_\psi$, from $\hat{\sigma}_\psi$, and from their covariance so that

$$\omega_\psi = \frac{1}{2\mathsf{f}(c_\psi)}\left[\mathsf{Var}\{\mathbb{G}_n(c_\psi)\} + 2c_\psi\mathsf{f}(c_\psi)\mathsf{Cov}\{\mathbb{G}_n(c_\psi), \mathbb{L}_n(c_\psi)\} + c_\psi^2\mathsf{f}^2(c_\psi)\mathsf{Var}\{\mathbb{L}_n(c_\psi)\}\right].$$

Using l'Hôpital's rule it is seen that $c_0/\varsigma_0 = \sqrt{3}$.

The above results shed light on some previously suggested distribution approximations for the deletion residuals. The approximation of Atkinson and Riani (2006, Theorem 2) has an asymptotic variance that matches that of the process $\mathbb{G}_n$, while omitting the estimation error for the scale. Riani and Atkinson (2007) presented an approximation to the distribution of the deletion residuals that evolves around order statistics of certain t-distributed variables. Due to Theorem E.1 in Appendix E that approximation also has an asymptotic variance matching that of the process $\mathbb{G}_n$.

Example 3.3 ***Some particular reference distributions.***
(a) ***Standard normal distribution.*** *If $\mathsf{f} = \varphi$ then $c_\psi = \Phi^{-1}\{(1+\psi)/2\}$ and*

$$\tau_\psi = 2\int_0^{c_\psi} x^2\varphi(x)dx = 2\{\Phi(x) - x\varphi(x)\}\Big|_0^{c_\psi} = \psi - 2c_\psi\varphi(c_\psi),$$
$$\varkappa_\psi = 2\int_0^{c_\psi} x^4\varphi(x)dx = 2\{3\Phi(x) - (x^3+3x)\varphi(x)\}\Big|_0^{c_\psi} = 3\psi - 2(c_\psi^3 + 3c_\psi)\varphi(c_\psi).$$

(b) ***Scaled t-distribution*** *with d degrees of freedom of Example 3.1(c) has density $\mathsf{f}(c) = s_d\mathsf{f}_d(cs_d)$ where f_d is the t-density with d degrees of freedom variance $s_d^2 = d/(d-2)$. Then $c_\psi = s_d^{-1}\mathsf{F}_d^{-1}\{(1+\psi)/2\}$ and $\psi = 2\mathsf{F}_d(c_\psi s_d) - 1$ It holds*

$$\tau_\psi = (d-1)\{2\mathsf{F}_{d-2}(c_\psi) - 1\} - (d-2)\{2\mathsf{F}_d(c_\psi s_d) - 1\}$$
$$\varkappa_\psi = (d-2)^2\left[\frac{(d-1)(d-3)}{(d-2)(d-4)}\left\{2\mathsf{F}_{d-4}\left(\frac{c_\psi}{s_{d-2}}\right) - 1\right\} - 2\frac{d-1}{d-2}\{2\mathsf{F}_{d-2}(c_\psi) - 1\} + \{2\mathsf{F}_d(c_\psi s_d) - 1\}\right].$$

Note that for $c_\psi \to \infty$ then the distribution functions approach unity so that

$$\tau_\psi \to 1, \qquad \varkappa_\psi \to 3\frac{d-2}{d-4}.$$

which are the variance and the kurtosis of the scaled t distribution.

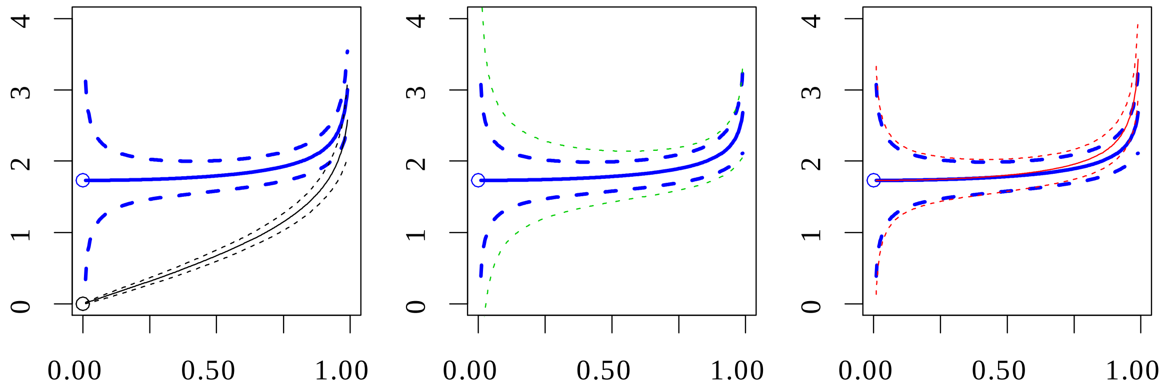

Figure 1: The plots show c_ψ/ς_ψ (thick line through $(0, c_0/\varsigma_0) = (0, \sqrt{3})$) and the limits $(c_\psi \pm 2(\omega_\psi/n)^{1/2})/\varsigma_\psi$ for $\hat{z}_\psi/\hat{\sigma}_\psi$ and $\mathsf{f} = \varphi$, given as thick dashed lines, for $n = 128$. Panel a compares this to $c_\psi \pm 2(\omega_\psi/n)^{1/2}$ for $\hat{z}_\psi/\hat{\sigma}_{\psi,corr}$ through $(0, c_0) = (0, 0)$, panel b compares it to $c_\psi \pm 2(\psi(1-\psi)/n)^{1/2}/\{2\varphi(c_\psi)\}$ for $\hat{z}_\psi/\sigma$ (dotted lines), and panel c compares it to $(c_\psi \pm 2(\omega_\psi/n)^{1/2})/\varsigma_\psi$ calculated for $\mathsf{f} = \mathsf{t}_5$.

Figure 1 illustrates the asymptotic results (3.10) when f is standard normal or t_5 and $n = 128$. For the standard normal case of Example 3.3(a) the asymptotic results of the forward residuals $\hat{z}_\psi/\hat{\sigma}_\psi$ based on the biased estimator $\hat{\sigma}_\psi$ are shown in all three panels. The bold, solid line is the asymptotic mean c_ψ/ς_ψ, noting that $c_0/\varsigma_0 = \sqrt{3}$. The bold, dashed lines are the 5% and 95% quantiles $\{c_\psi \pm 2(\omega_\psi/n)^{1/2}\}/\varsigma_\psi$, which contain 90% of the scaled forward residuals with $n = 128$, chosen for comparability with the data example in Riani and Atkinson (2007, Figure 1). The 5% and 95% quantiles fan out for small ψ due to the biased estimate of the variance $\hat{\sigma}^{(m)}$. For ψ close to 1 the quantiles and therefore the limits diverge.

In panel a these results are compared to the results for the bias-corrected forward residuals $\hat{z}_\psi/\hat{\sigma}_{\psi,cor}$ which passes through $(0, 0)$ because the curves for $\hat{z}_\psi/\hat{\sigma}_\psi$ are multiplied by ς_ψ, and $\varsigma_0 = 0$.

In panel b the result for estimated variance are compared to the results for known variance, which are actually wider. This phenomenon is also seen for empirical processes of estimated residuals, see Engler and Nielsen (2009, equation 2.10).

Finally panel c compares the result for $f = \phi$ with the results for $f = t_5$. With 5 degrees of freedom Assumption 3.1 is not met. For higher degrees of freedom the results will be in between the t_5 and the normal results.[4]

3.4 Application of the result for the forward estimators

In an application of Theorem 3.5 for the forward estimators, the distribution of the kernel $\Sigma_n^{-1}\mathbb{K}_n(c_\psi)$ depends on the type of regressors. Building on the analysis in Johansen and Nielsen (2009, section 1.4,1.5, 2013) we present a result for the stationary case. For situations with deterministic trends or unit roots see those papers. In the case of stationary and

[4]Graphics were done using R 2.13, see R Development Core Team (2011).

autoregressive regressors we take $N = n^{-1/2}$, and the normalised matrix of squared regressors, $\Sigma_n = n^{-1}\sum_{i=1}^n x_i x_i'$, described in Assumption 3.1(iia) has a deterministic limit

Theorem 3.7 *Suppose Assumptions 4.1 holds and that x_i is stationary and autoregressive with finite variance. Then $\Sigma_n \overset{\mathsf{P}}{\to} \Sigma > 0$ and $\mathbb{K}_n$ conveges to a zero mean Gaussian process with variance given as*

$$\mathsf{Var}\{\mathbb{K}_n(c_\psi)\} = \tau_\psi \sigma^2 \Sigma. \tag{3.11}$$

Theorem 3.7 then implies that

$$n^{1/2}(\widehat{\beta}_\psi - \beta) \overset{\mathsf{D}}{\to} \mathsf{N}[0, \frac{\tau_\psi \sigma^2}{\{\psi - 2c_\psi \mathsf{f}(c_\psi)\}^2}\Sigma^{-1}],$$

which matches Johansen and Nielsen (2010, Corollary 5.3).

4 A class of auxiliary weighted and marked empirical processes

It is useful to consider an auxiliary class of weighted and marked empirical distribution functions for errors ε_i as opposed to absolute errors $|\varepsilon_i|$. The analysis of this class generalizes that of Koul and Ossiander (1994) in two respects. First, the standardized estimation error b is permitted to diverge at a rate of $n^{1/4-\eta}$ rather than being bounded. Secondly, non-bounded marks of the type ε_i^p are allowed. These results are therefore of independent interest. This class of weighted and marked empirical distribution functions is defined for $b \in \mathbb{R}^{\dim x}$ and $c \in \mathbb{R}$ by

$$\widehat{\mathsf{F}}_n^{g,p}(b,c) = \frac{1}{n}\sum\nolimits_{i=1}^n g_{in}\varepsilon_i^p 1_{(\varepsilon_i \le \sigma c + x_{in}' b)}, \tag{4.1}$$

with $(\varepsilon_{i-1},\dots,\varepsilon_1,x_i,\dots,x_1)$-measurable weights g_{in} and marks ε_i^p. By proving results that hold uniformly in b, suitably bounded, we can handle the Forward Search. This allows an analysis of the order statistics of the residuals at a given step m of the Forward Search, since the order statistics depend on the previous estimation error $\hat{b} = N^{-1}(\widehat{\beta}^{(m)} - \beta)$, but are scale invariant. In turn, we can apply the results for the estimation errors $N^{-1}(\hat{\beta}^{(m+1)} - \beta)$ and $n^{1/2}(\hat{\sigma}_{corr}^{(m+1)} - \sigma)$.

4.1 Assumptions

We will keep track of the assumptions in a more explicit way than done above. In the analysis of the one-sided empirical processes the density f is not necessarily symmetric.

Assumption 4.1 *Let $\mathcal{F}_i$ be an increasing sequence of σ fields so $\varepsilon_{i-1}, x_i, g_{in}$ are $\mathcal{F}_{i-1}$-measurable and ε_i is independent of $\mathcal{F}_{i-1}$ with continuous, differentiable density f which is positive for $\mathsf{F}^{-1}(0) < c < \mathsf{F}^{-1}(1)$. Let p, r, η, κ, ν be given so $p, r \in \mathbb{N}_0$, $0 \le \kappa < \eta \le 1/4$ and $\nu \le 1$. Suppose*

(i) *density satisfies:*

(a) *moments:* $\int_{-\infty}^{\infty} |\varepsilon|^{2^r p/\nu} \mathsf{f}(u) du < \infty$;

(*b*) *boundedness:* $\sup_{c\in\mathbb{R}}\{(1+|c|^{2^r p-1})\mathsf{f}(c)+(1+|c|^{2^r p})|\dot{\mathsf{f}}(c)|\}<\infty$;
(*c*) *smoothness: a* $C_{\mathsf{H}}\in\mathbb{N}$ *exist so that for all* $a>0$

$$\frac{\sup_{c\geq a}(1+c^{2^r p})\mathsf{f}(c)}{\inf_{0\leq c\leq a}(1+c^{2^r p})\mathsf{f}(c)}\leq C_{\mathsf{H}},\qquad \frac{\sup_{c\leq -a}(1+|c|^{2^r p})\mathsf{f}(c)}{\inf_{-a\leq u\leq 0}(1+|c|^{2^r p})\mathsf{f}(c)}\leq C_{\mathsf{H}}.$$

(*ii*) *regressors* x_i *satisfy* $\max_{1\leq i\leq n}|n^{1/2-\kappa}N'x_i|=\mathsf{O}_\mathsf{P}(1)$ *for some normalisation matrix* N;
(*iii*) *weights* g_{in} *are matrix valued and satisfy*
(*a*) $n^{-1}\mathsf{E}\sum_{i=1}^n|g_{in}|^{2^r}(1+|n^{1/2}N'x_i|)=\mathsf{O}(1)$;
(*b*) $n^{-1}\sum_{i=1}^n|g_{in}|(1+|n^{1/2}N'x_i|^2)=\mathsf{O}_\mathsf{P}(1)$.

Remark 4.1 *Some discussion of Assumption 4.1 is given*
(*a*) ***The case of no marks*** $p=0$. *This is the situation discussed in Koul and Ossiander (1994). The primary role of* r *is to control the tail behaviour of the density. When* $p=0$ *then* $2^r p=0$ *for all* $r\in\mathbb{N}_0$, *so* r *can be chosen as* $r=0$ *and the assumption simplifies considerably.*
(*b*) ***The tail condition in Assumption 4.1***(*ia*) *is used for some* $\nu<1$ *for the tightness result in Theorem 4.4. Otherwise* $\nu=1$ *suffices.*
(*c*) ***The smoothness of density in Assumption 4.1***(*ic*) *is satisfied if* $\mathsf{h}_r(c)=(1+\epsilon^{2^r p})\mathsf{f}(\epsilon)$ *is monotone for* $|c|>d_1$ *for some* $d_1\geq 0$. *Indeed, choose* $d_2\geq d_1$ *so that* $\sup_{c\geq d_2}\mathsf{h}_r(c)=\inf_{0\leq c\leq d_2}\mathsf{h}_r(c)=\mathsf{h}_r(d_2)$. *Then choose* C_H *larger than* $\sup_{0\leq c\leq d_2}\mathsf{h}_r(c)/\inf_{0\leq c\leq d_2}\mathsf{h}_r(c)$. *A similar argument applies for* $c<0$. *Note, that the smoothness condition implies that the density has connected support.*
(*d*) ***Sufficient condition for Assumption 4.1***(*i*). *If* f *is symmetric and differentiable with* $c^q\mathsf{f}(c)$, $c^{q-1}|\dot{\mathsf{f}}(c)|$ *both decreasing for large* c *for some* $q>1+2^r p$, *then Assumption 4.1(i) holds. Indeed, (ia) holds, since when* $c^q\mathsf{f}(c)$ *is decreasing, then* $c^{2^r p/\nu}\mathsf{f}(c)$ *is integrable for some* $\nu<1$. *Further, (ib) holds, since, first, the continuity and decreasingness of* $c^q\mathsf{f}(c)$ *and hence of* $\mathsf{f}(c)$ *implies* $(1+|c|^{1+2^r p})\mathsf{f}(c)$ *is bounded, and, secondly, since* $\dot{\mathsf{f}}(c)<0$ *so that* $|c^{q-1}\dot{\mathsf{f}}(c)|$ *decreases then* $(1+|c|^{2^r p})|\dot{\mathsf{f}}(c)|$ *is bounded. Finally, (ic) holds due to the remark (c) above.*

4.2 The empirical process results

The weighted and marked empirical distribution function $\widehat{\mathsf{F}}_n^{g,p}(b,c)$ defined in (4.1) is analysed through martingale arguments. Thus, introduce the sum of conditional expectations

$$\overline{\mathsf{F}}_n^{g,p}(b,c)=\frac{1}{n}\sum_{i=1}^n g_{in}\mathsf{E}_{i-1}\{\varepsilon_i^p 1_{(\varepsilon_i\leq\sigma c+x_{in}'b)}\},\tag{4.2}$$

and the weighted and marked empirical process

$$\mathbb{F}_n^{g,p}(b,c)=n^{1/2}\{\widehat{\mathsf{F}}_n^{g,p}(b,c)-\overline{\mathsf{F}}_n^{g,p}(b,c)\}.\tag{4.3}$$

Three results follows. These are proved in the subsequent Appendix C. The first result shows that the dependence of $\mathbb{F}_n^{g,p}$ on the estimation error b is negligible.

Theorem 4.1 *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 4.1(i, ii, iiia) holds with $\nu = 1$, some $\eta > 0$ and an r so $2^{r-1} \geq 1 + (1/4 + \kappa - \eta)(1 + \dim x)$. Then, for any $B > 0$ and $n \to \infty$, it holds that*

$$\sup_{0\leq\psi\leq 1} \sup_{|b|\leq n^{1/4-\eta}B} |\mathbb{F}_n^{g,p}(b, c_\psi) - \mathbb{F}_n^{g,p}(0, c_\psi)| = \mathrm{o}_{\mathsf{P}}(1).$$

For the standard empirical process with weights $g_{in} = 1$ and marks $\varepsilon_i^p = 1$ the order of the remainder term can be improved as follows. In terms of the Assumption 4.1 note that when $p = 0$ then r will be irrelevant except for the condition on the regressors in part $(iiia)$.

Theorem 4.2 *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 4.1(i, ii, iiia) holds with $\nu = 1$, $p = 0$, $r = 2$ and some $\eta > 0$. Then, for any $B > 0$, any $\omega < \eta - \kappa \leq 1/4$ and $n \to \infty$, it holds that*

$$\sup_{0\leq\psi\leq 1} \sup_{|b|,|d|\leq n^{1/4-\eta}B} |\mathbb{F}_n^{1,0}(b, c_\psi + n^{\kappa-1/2}d) - \mathbb{F}_n^{1,0}(0, c_\psi)| = \mathrm{o}_{\mathsf{P}}(n^{-\omega}).$$

The next results presents a linearization of $\overline{\mathsf{F}}_n^{g,p}(b, c)$.

Theorem 4.3 *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 4.1(ib, iiib) holds with $r = 0$ and some $\eta > 0$. Then, for all $B > 0$ and $n \to \infty$, it holds that*

$$\sup_{0\leq\psi\leq 1} \sup_{|b|\leq n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{F}}_n^{g,p}(b, c_\psi) - \overline{\mathsf{F}}_n^{g,p}(0, c_\psi)\} - \sigma^{p-1}c_\psi^p \mathsf{f}(c_\psi) n^{-1}\textstyle\sum_{i=1}^n g_{in} n^{1/2} x_{in}' b| = \mathrm{O}_{\mathsf{P}}(n^{-2\eta}).$$

Finally, the weighted and marked empirical process $\mathbb{F}_n^{g,p}(0, c_\psi)$ in (4.3) is tight. It holds by construction that $\mathbb{F}_n^{g,p}(0, 0) = 0$. Following Billingsley (1968, Theorem 15.5) tightness in the space $D[0, 1]$ endowed with the uniform metric, then follows from the next result.

Theorem 4.4 *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. Suppose Assumption 4.1(ia, iiia) holds with $r = 2$ and some $\nu < 1$. Then, for all $\epsilon > 0$, it holds*

$$\lim_{\phi\downarrow 0} \limsup_{n\to\infty} \mathsf{P}\{\sup_{0\leq\psi\leq\psi^\dagger\leq 1:\psi^\dagger-\psi\leq\phi} |\mathbb{F}_n^{g,p}(0, c_{\psi^\dagger}) - \mathbb{F}_n^{g,p}(0, c_\psi)| > \epsilon\} \to 0.$$

The proofs of these results are given in Section C, but first we establish some martingale results and discuss a metric on $\mathbb{R}$ which is applied in the chaining argument needed in the proofs.

5 Iterated exponential martingale inequalities

Chaining arguments will be used to handle tightness properties of the empirical processes. This reduces the problem to a problem of finding the tail probability for the maximum of a certain family of martingales. We first give a general result on a bound of a finite number of martingales, which we prove by iterating a martingale inequality by Bercu and Touati (2008). Subsequently, two special cases are analysed where the number of elements in the martingale family is increasing and where it is fixed.

Theorem 5.1 *For ℓ so $1 \le \ell \le L$ let $z_{\ell,i}$ be $\mathcal{F}_i$-adapted so $\mathsf{E} z_{\ell,i}^{2^{\overline{r}}} < \infty$ for some $\overline{r} \in \mathbb{N}$. Let $D_r = \max_{1\le\ell\le L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^{2^r}$ for $1 \le r \le \overline{r}$. Then, for all $\kappa_0, \kappa_1, \ldots, \kappa_{\overline{r}} > 0$, it holds*

$$\mathsf{P}\{\max_{1\le\ell\le L} |\textstyle\sum_{i=1}^n (z_{\ell,i} - \mathsf{E}_{i-1} z_{\ell,i})| > \kappa_0\} \le L\frac{\mathsf{E} D_{\overline{r}}}{\kappa_{\overline{r}}} + \sum_{r=1}^{\overline{r}} \frac{\mathsf{E} D_r}{\kappa_r} + 2L \sum_{r=0}^{\overline{r}-1} \exp(-\frac{\kappa_r^2}{14\kappa_{r+1}}).$$

The proof is given in Appendix A.

Theorem 5.2 *For ℓ so $1 \le \ell \le L$ let $z_{\ell,i}$ be $\mathcal{F}_i$-adapted so $\mathsf{E} z_{\ell,i}^{2^{\overline{r}}} < \infty$ for some $\overline{r} \in \mathbb{N}$. Let $D_r = \max_{1\le\ell\le L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^{2^r}$ for $1 \le r \le \overline{r}$. Suppose, for some $\varsigma \ge 0$, $\lambda > 0$, that $L = \mathrm{O}(n^\lambda)$ and $\mathsf{E} D_r = \mathrm{O}(n^\varsigma)$ for $r \le \overline{r}$. Then, if $\nu > 0$ is chosen such that*
(i) $\varsigma < 2\nu$
(ii) $\varsigma + \lambda < \nu 2^{\overline{r}}$,
it holds that for all $\kappa > 0$ and $n \to \infty$ that

$$\lim_{n\to\infty} \mathsf{P}\{\max_{1\le\ell\le L} |\textstyle\sum_{i=1}^n (z_{\ell,i} - \mathsf{E}_{i-1} z_{\ell,i})| > \kappa n^\nu\} = 0.$$

Proof of Theorem 5.2. Apply Lemma 5.1 with $\kappa_q = (\kappa n^\upsilon)^{2^q} (28\lambda \log n)^{1-2^q}$ for any $\kappa > 0$ so that $\kappa_0 = \kappa n^\upsilon$ and $\kappa_q^2/\kappa_{q+1} = 28\lambda \log n$ and exploit conditions (i, ii) to see that the probability of interest satisfies

$$\mathcal{P}_n = \mathrm{O}\{n^\lambda \frac{n^\varsigma (\log n)^{2^{\overline{r}}-1}}{n^{\nu 2^{\overline{r}}}} + \sum_{r=1}^{\overline{r}} \frac{n^\varsigma (\log n)^{2^r-1}}{n^{\nu 2^r}} + 2n^\lambda \overline{r} n^{-2\lambda}\} = \mathrm{o}(1),$$

as desired since $\varsigma + \lambda < \nu 2^{\overline{r}}$ and $\varsigma < 2\nu \le \nu 2^r$ for $r \ge 1$. ■

Theorem 5.3 *For ℓ so $1 \le \ell \le L$ let $z_{\ell,i}$ be $\mathcal{F}_i$-adapted so $\mathsf{E} z_{\ell,i}^4 < \infty$. Suppose $\mathsf{E} \max_{1\le\ell\le L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^{2^q} \le Cn$ for $q = 1, 2$ and some $C > 0$. Then it holds, for all $\theta > 0$,*

$$\mathsf{P}\{\max_{1\le\ell\le L} |\textstyle\sum_{i=1}^n (z_{\ell,i} - \mathsf{E}_{t-1} z_{\ell,i})| > \kappa n^{1/2}\} \le \frac{(L+1)\theta^3 C}{\kappa n} + \frac{\theta C}{\kappa} + 4L \exp(-\frac{\kappa\theta}{14}).$$

Proof of Theorem 5.3. Apply Lemma 5.1 with $\kappa_q = \kappa n^{2^{q-1}} \theta^{1-2^q}$ for any $\kappa, \theta > 0$ so that $\kappa_0 = \kappa n^{1/2}$ and $\kappa_q^2/\kappa_{q+1} = \kappa\theta$ to get the bound

$$\mathcal{P} \le \frac{(L+1)\theta^3}{\kappa n^2} \mathsf{E} \max_{1\le\ell\le L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^4 + \frac{\theta}{\kappa n} \mathsf{E} \max_{1\le\ell\le L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^2 + 4L \exp(-\frac{\kappa\theta}{14}).$$

Exploit the moment conditions to get the desired result. ■

6 Conclusion

The intention of the Forward Search is to determine the number of outliers by looking at the forward plot of the forward residuals. The main results for the Forward Search given in §3 describe the asymptotic distribution of that process in a situation where there are no outliers. We can therefore add pointwise confidence bands to the forward plot, using Theorem 3.3.

These give an impression of the pointwise variation we would expect for the forward plot if there were in fact no outliers. In practice we would want to make a simultaneous decision based on the entire graph, but this is for future research.

We suspect that the iterated martingale inequalities will be useful in a variety of situations. For instance, in ongoing research, we are finding that the inequalities are helpful in establishing consistency and asymptotic distribution results for the Huber-skip regression estimator.

A Proofs of martingale inequalities

Proof of Theorem 5.1. 1. *Notation.* For $0 \leq r \leq \overline{r}$ define $A_{\ell,r} = \sum_{i=1}^{n}(z_{\ell,i}^{2^r} - \mathsf{E}_{i-1} z_{\ell,i}^{2^r})$ and

$$\mathcal{P}_r(\kappa_r) = \mathsf{P}(\max_{1\leq\ell\leq L} A_{\ell,r} > \kappa_r), \qquad \mathcal{Q}_r(\kappa_r) = \mathsf{P}(\max_{1\leq\ell\leq L} |A_{\ell,r}| > \kappa_r),$$

where $\mathcal{Q}_0(\kappa_0)$ is the probability of interest, while $\mathcal{P}_r(\kappa_r) \leq \mathcal{Q}_r(\kappa_r)$.

2. *The terms $\mathcal{Q}_r(\kappa_r)$ for $0 \leq r < \overline{r}$.* We first prove that for any $\kappa_r, \kappa_{r+1} > 0$ then

$$\mathcal{Q}_r(\kappa_r) \leq 2L\exp(-\frac{\kappa_r^2}{14\kappa_{r+1}}) + \mathcal{P}_{r+1}(\kappa_{r+1}) + \frac{\mathsf{E}D_{r+1}}{\kappa_{r+1}}. \tag{A.1}$$

The idea is now to apply the following inequality, for sets $\mathcal{A}, \mathcal{B}$,

$$\mathsf{P}(\mathcal{A}) = \mathsf{P}(\mathcal{A} \cap \mathcal{B}) + \mathsf{P}(\mathcal{A} \cap \mathcal{B}^c) \leq \mathsf{P}(\mathcal{A} \cap \mathcal{B}) + \mathsf{P}(\mathcal{B}^c).$$

In the first term, $\mathcal{A}$ relates to the tails of a martingale and $\mathcal{B}$ to the central part of the distribution of the quadratic variation. Thus the first term can be controlled by a martingale inequality. In the second term, $\mathcal{B}^c$ relates to the tail of the quadratic variation. The sum of the predictable and the total quadratic variation of $A_{\ell,r}$ is $B_{\ell,r} = \sum_{i=1}^{n} B_{\ell,r,i}$ where $B_{\ell,r,i} = (z_{\ell,i}^{2^r} - \mathsf{E}_{i-1} z_{\ell,i}^{2^r})^2 + \mathsf{E}_{i-1}(z_{\ell,i}^{2^r} - \mathsf{E}_{i-1} z_{\ell,i}^{2^r})^2$. It holds

$$\mathcal{Q}_r(\kappa_r) \leq \mathsf{P}\{(\max_{1\leq\ell\leq L} |A_{\ell,r}| > \kappa_r) \cap (\max_{1\leq\ell\leq L} B_{\ell,r} < 7\kappa_{r+1})\} + \mathsf{P}(\max_{1\leq\ell\leq L} B_{\ell,r} \geq 7\kappa_{r+1}). \tag{A.2}$$

Consider the first term in (A.2), $\mathcal{S}_{1,r}$ say. By Boole's inequality this satisfies

$$\mathcal{S}_{1,r} \leq \sum_{\ell=1}^{L} \mathsf{P}\{(|A_{\ell,r}| > \kappa_r) \cap (\max_{1\leq\ell\leq L} B_{\ell,r} < 7\kappa_{r+1})\}.$$

Noting that $(\max_{1\leq\ell\leq L} B_{\ell,r} \leq 7\kappa_{r+1}) \subset (B_{\ell,r} \leq 7\kappa_{r+1})$ gives the further bound

$$\mathcal{S}_{1,r} \leq \sum_{\ell=1}^{L} \mathsf{P}\{(|A_{\ell,r}| > \kappa_r) \cap (B_{\ell,r} < 7\kappa_{r+1})\}.$$

Since $A_{\ell,r}$ is a martingale the exponential inequality of Bercu and Touati (2008) shows

$$\mathsf{P}\{(|A_{\ell,r}| > \kappa_r) \cap (B_{\ell,r} < 7\kappa_{r+1})\} \leq 2\exp\{-\kappa_r^2/(14\kappa_{r+1})\}.$$

Taken L times, this gives the first term in (A.1).

Consider the second term in (A.2), $\mathcal{S}_{2,r}$ say. Ignore the indices on $B_{\ell,r,i}, z_{\ell,i}^{2^r}$ and apply the inequality $(z - \mathsf{E}z)^2 \leq 2(z^2 + \mathsf{E}^2 z)$ along with $\mathsf{E}^2 z \leq \mathsf{E}z^2$ and $\mathsf{E}(z - \mathsf{E}z)^2 \leq \mathsf{E}z^2$ to get that $B = (z - \mathsf{E}z)^2 + \mathsf{E}(z - \mathsf{E}z)^2 \leq 2z^2 + 3\mathsf{E}z^2 = 2(z^2 - \mathsf{E}z^2) + 5\mathsf{E}z^2$. Thus,

$$\mathcal{S}_{2,r} \leq \mathsf{P}\{\max_{1\leq\ell\leq L} \sum_{i=1}^{n}(z_{\ell,i}^{2^{r+1}} - \mathsf{E}_{i-1} z_{\ell,i}^{2^{r+1}}) \geq \kappa_{r+1}\} + \mathsf{P}(\max_{1\leq\ell\leq L} \sum_{i=1}^{n} \mathsf{E}_{i-1} z_{\ell,i}^{2^{r+1}} \geq \kappa_{r+1}).$$

Use the notation from above and then the Markov inequality to get

$$\mathcal{S}_{2,r} \leq \mathcal{P}_{r+1}(\kappa_{r+1}) + \mathsf{P}(D_{r+1} \geq \kappa_{r+1}) \leq \mathcal{P}_{r+1}(\kappa_{r+1}) + \frac{1}{\kappa_{r+1}}\mathsf{E}D_{r+1},$$

which are the last terms of (A.1).

3. *The term* $\mathcal{P}_{\bar r}(\kappa_{\bar r})$. Apply the inequality $|z| - \mathsf{E}_{i-1}|z| \leq |z|$ and then Boole's and Markov's inequalities to get

$$\mathcal{P}_{\bar r}(\kappa_{\bar r}) \leq \mathsf{P}(\max_{1\leq \ell \leq L} \sum_{i=1}^n z_{\ell,i}^{2^{\bar r}} > \kappa_{\bar r}) \leq L \max_{1\leq \ell\leq L} \mathsf{P}(\sum_{i=1}^n z_{\ell,i}^{2^{\bar r}} > \kappa_{\bar r}) \leq \frac{L}{\kappa_{\bar r}} \max_{1\leq \ell \leq L} \mathsf{E}\sum_{i=1}^n z_{\ell,i}^{2^{\bar r}}.$$

Apply iterated expectations and interchange maximum and expectation to get

$$\mathcal{P}_{\bar r}(\kappa_{\bar r}) \leq \frac{L}{\kappa_{\bar r}} \max_{1\leq \ell\leq L} \mathsf{E}\sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^{2^{\bar r}} \leq \frac{L}{\kappa_{\bar r}} \mathsf{E} \max_{1\leq \ell \leq L} \sum_{i=1}^n \mathsf{E}_{i-1} z_{\ell,i}^{2^{\bar r}} = \frac{L}{\kappa_{\bar r}} \mathsf{E} D_{\bar r}.$$

4. *Combine expressions.* Since $\mathcal{P}_{r+1}(\kappa_{r+1}) \leq \mathcal{Q}_{r+1}(\kappa_{r+1})$ then write (A.1) as

$$\mathcal{Q}_r(\kappa_r) \leq 2L\exp(-\frac{\kappa_r^2}{14\kappa_{r+1}}) + \mathcal{Q}_{r+1}(\kappa_{r+1}) + \frac{\mathsf{E}D_{r+1}}{\kappa_{r+1}} \qquad \text{for } r = 0,\ldots,\bar r - 2,$$
$$\mathcal{Q}_r(\kappa_r) \leq 2L\exp(-\frac{\kappa_r^2}{14\kappa_{r+1}}) + \mathcal{P}_{r+1}(\kappa_{r+1}) + \frac{\mathsf{E}D_{r+1}}{\kappa_{r+1}} \qquad \text{for } r = \bar r - 1.$$

Then sum from $r = 0$ to $\bar r - 1$ and insert the bound $\mathcal{P}_{\bar r}(\kappa_{\bar r}) \leq \kappa_{\bar r}^{-1} L \mathsf{E} D_{\bar r}$. ■

B A metric on $\mathbb{R}$ and some inequalities

A metric is set up that will be used for the chaining argument. Then a number of inequalities are shown, mostly related to this metric.

Introduce the function

$$J_{i,p}(x,y) = (\varepsilon_i/\sigma)^p \{1_{(\varepsilon_i \leq \sigma y)} - 1_{(\varepsilon_i \leq \sigma x)}\}, \tag{B.1}$$

where $p \in \mathbb{N}_0$ and ε_i/σ has density f. We will be interested in powers of $J_{i,p}(x,y)$ of order 2^r where $r \in \mathbb{N}$ was chosen in Assumption 4.1(i). Note that $2^r p$ is even for $p \in \mathbb{N}_0$ and $r \in \mathbb{N}$ so that $\varepsilon_i^{2^r p}$ is non-negative. Thus, define the increasing function

$$\mathsf{H}_r(x) = \int_{-\infty}^x (1 + \varepsilon^{2^r p}) \mathsf{f}(\varepsilon) d\varepsilon,$$

with derivative $\dot{\mathsf{H}}_r(x) = (1 + x^{2^r p})\mathsf{f}(x)$, along with the constant

$$H_r = \mathsf{H}_r(\infty) = \int_{-\infty}^{\infty} (1 + \varepsilon^{2^r p}) \mathsf{f}(\varepsilon) d\varepsilon < \infty.$$

It follows that, for $x \leq y$ and $0 \leq s \leq r$ then

$$0 \leq |\mathsf{E}\{J_{i,p}(x,y)\}^{2^s}| \leq \mathsf{E}\{|J_{i,p}(x,y)|^{2^s}\} < \mathsf{H}_r(y) - \mathsf{H}_r(x), \tag{B.2}$$

noting that, for $q \geq p \geq 0$ and $\varepsilon \in \mathbb{R}$, then $|\varepsilon^p| < 1 + |\varepsilon|^q$.

For the chaining, partition the range of $\mathsf{H}_r(c)$ into K intervals of equal size H_r/K. That is, partition the support into K intervals defined by the endpoints

$$-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty, \tag{B.3}$$

and for $1 \leq k \leq K$,

$$\mathsf{E}[\{J_{i,p}(c_{k-1}, c_k)\}^{2^r}] = \mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-1}) = \frac{H_r}{K}.$$

Let $c_{-k} = c_0$ for $k \in \mathbb{N}$.

The number of intervals K will be chosen so large that c_-, c_+ exists which are (weakly) separated from zero by grid points in the sense that $c_{k_- -1} \leq c_- \leq c_{k_-} \leq 0$ and $0 \leq c_{k_+ -1} \leq c_+ \leq c_{k_+}$ and so that

$$\dot{\mathsf{H}}_r(c_-) = \dot{\mathsf{H}}_r(c_+) = H_r/(C_{\mathsf{H}} K^{1/2}). \tag{B.4}$$

This can be done for sufficiently large K since f is continuous and since the function $\dot{\mathsf{H}}_r(c) = (1 + c^{2^r p})\mathsf{f}(c)$ is integrable by Assumption 4.1(ia).

The first inequality concerns the H_r-distance of certain pertubations of the $]c_{k-1}, c_k]$ intervals.

Lemma B.1 *Suppose Assumption 4.1(i) holds with $\nu = 1$ only. Then a constant $C > 0$ exists so that for all K satisfying (B.4) then*

$$\sup_{1 \leq k \leq K} \sup_{|d| \leq K^{-1/2}} \{\mathsf{H}_r(c_k + d) - \mathsf{H}_r(c_{k-1} + d)\} \leq C H_r/K.$$

Proof of Lemma B.1. 1. *Definitions.* Consider positive c_k only with a similar argument for negative c_k. Let $\mathcal{H} = \mathsf{H}_r(c_k + d) - \mathsf{H}_r(c_{k-1} + d)$. Let $\dot{\mathsf{H}}_r(c) = (1 + c^{2^r p})\mathsf{f}(c)$ and

$$\underline{\dot{\mathsf{H}}}_r(c) = \inf_{0 \leq d \leq c} \dot{\mathsf{H}}_r(d), \qquad \overline{\dot{\mathsf{H}}}_r(c) = \sup_{d \geq c} \dot{\mathsf{H}}_r(d),$$

which are decreasing in c. Assumption 4.1(ic) then implies

$$C_{\mathsf{H}}^{-1} \overline{\dot{\mathsf{H}}}_r(c) \leq \underline{\dot{\mathsf{H}}}_r(c) \leq \dot{\mathsf{H}}_r(c) \leq \overline{\dot{\mathsf{H}}}_r(c) \leq C_{\mathsf{H}} \underline{\dot{\mathsf{H}}}_r(c). \tag{B.5}$$

Since $\ddot{\mathsf{H}}_r(c) = 2^r p c^{2^r p - 1}\mathsf{f}(c) + (1 + c^{2^r p})\dot{\mathsf{f}}(c)$ then Assumption 4.1(ib) gives

$$\sup_{c \in \mathbb{R}} |\ddot{\mathsf{H}}_r(c)| < \infty. \tag{B.6}$$

2. *Apply the mean-value theorem* to get, for some c_ℓ^* so $c_{\ell-1} \leq c_\ell^* \leq c_\ell$, that

$$H_r/K = \mathsf{H}_r(c_\ell) - \mathsf{H}_r(c_{\ell-1}) = (c_\ell - c_{\ell-1})\dot{\mathsf{H}}_r(c_\ell^*). \tag{B.7}$$

Two inequalities for $\dot{\mathsf{H}}_r(c)$ arise from (B.5) and condition (B.4). These are

$$\dot{\mathsf{H}}_r(c) \leq \overline{\dot{\mathsf{H}}}_r(c) \leq \overline{\dot{\mathsf{H}}}_r(c_+) \leq C_{\mathsf{H}} \dot{\mathsf{H}}_r(c_+) = H_r/K^{1/2} \text{for} c \geq c_+, \tag{B.8}$$

$$\dot{\mathsf{H}}_r(c) \geq \underline{\dot{\mathsf{H}}}_r(c) \geq \underline{\dot{\mathsf{H}}}_r(c_+) \geq \overline{\dot{\mathsf{H}}}_r(c_+)/C_{\mathsf{H}} \geq \dot{\mathsf{H}}_r(c_+)/C_{\mathsf{H}} = H_r/(C_{\mathsf{H}}^2 K^{1/2}) \text{for} 0 \leq c \leq c_+. \tag{B.9}$$

In parallel to (B.9), which is derived for positive c, it holds for negative c that

$$\dot{\mathsf{H}}_r(c) \geq H_r/(C_{\mathsf{H}}^2 K^{1/2}) \qquad \text{for } 0 \geq c \geq c_-. \tag{B.10}$$

3. *Small arguments* $c_- \leq c_k^* \leq c_+$. Combine (B.7), (B.9) and (B.10) to get

$$c_k - c_{k-1} = H_r/\{K\dot{\mathsf{H}}_r(c_k^*)\} \leq C_{\mathsf{H}}^2/K^{1/2}. \tag{B.11}$$

Two second order Taylor expansions give

$$\mathsf{H}_r(c_k + d) - \mathsf{H}_r(c_k) = d\dot{\mathsf{H}}_r(c_k) + (d^2/2)\ddot{\mathsf{H}}_r(c_k^{**}),$$
$$\mathsf{H}_r(c_{k-1} + d) - \mathsf{H}_r(c_{k-1}) = d\dot{\mathsf{H}}_r(c_{k-1}) + (d^2/2)\ddot{\mathsf{H}}_r(c_{k-1}^{**}),$$

where c_k^{**}, c_{k-1}^{**} satisfy $\max(|c_k^* - c_k|, |c_{k-1}^* - c_{k-1}|) \leq |d| \leq K^{-1/2}$. The difference is, when recalling the definition of $\mathcal{H}$ in item 1,

$$\mathcal{H} - \{\mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-1})\} = d\{\dot{\mathsf{H}}_r(c_k) - \dot{\mathsf{H}}_r(c_{k-1})\} + (d^2/2)\{\ddot{\mathsf{H}}_r(c_k^{**}) - \ddot{\mathsf{H}}_r(c_{k-1}^{**})\}.$$

It holds $\mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-1}) = \mathsf{H}_r/K$. The mean-value theorem gives that for a $\tilde{c}_k$ so $c_{k-1} \leq \tilde{c}_k \leq c_k$ then $\dot{\mathsf{H}}_r(c_k) - \mathsf{H}_r'(c_{k-1}) = (c_k - c_{k-1})\ddot{\mathsf{H}}_r(\tilde{c}_k)$. Insert this and rearrange to get

$$0 \leq \mathcal{H} = \frac{H_r}{K} + d(c_k - c_{k-1})\ddot{\mathsf{H}}_r(\tilde{c}_k) + \frac{d^2}{2}\{\ddot{\mathsf{H}}_r(c_k^{**}) - \ddot{\mathsf{H}}_r(c_{k-1}^{**})\}.$$

Using the bound to $c_k - c_{k-1} \leq C_{\mathsf{H}}^2/K^{1/2}$ from (B.11) and the bound $|d| \leq K^{-1/2}$ it follows that $0 \leq \mathcal{H} \leq C/K$ where $C = H_r + (C_{\mathsf{H}}^2 + 1)\sup_{c\in\mathbb{R}} |\ddot{\mathsf{H}}_r(c)|$ does not depend on K.

4. *Large arguments* $c_k^* \geq c_+$ *so* $k \geq k_+ + 2$. Expansion (B.7) and inequality (B.8) imply

$$c_k - c_{k-1} = H_r/\{K\dot{\mathsf{H}}_r(c_k^*)\} \geq K^{-1/2} \geq |d|.$$

The same holds for $c_{k+1} - c_k$ and $c_{k-1} - c_{k-2}$. Therefore

$$c_k + d \leq c_k + |d| \leq c_k + c_{k+1} - c_k = c_{k+1},$$
$$c_{k-1} + d \geq c_{k-1} - |d| \geq c_{k-1} - (c_{k-1} - c_{k-2}) = c_{k-2}.$$

It then holds that $0 \leq \mathcal{H} \leq \mathsf{H}_r(c_{k+1}) - \mathsf{H}_r(c_{k-2}) = C/K$, where $C = 3H_r$ does not depend on K.

5. *Intermediate arguments* $c_k^* \geq c_+$ *and* $k_+ \leq k \leq k_+ + 1$. In this case $c_{k_+-1} \leq c_+ \leq c_k^* \leq c_k \leq c_{k_++1}$. Consider the length of the interval $]c_{k_+-1}, c_+]$. The mean-value theorem shows

$$\mathsf{H}_r/K = \mathsf{H}_r(c_{k_+}) - \mathsf{H}_r(c_{k_+-1}) \geq \mathsf{H}_r(c_+) - \mathsf{H}_r(c_{k_+-1}) = (c_+ - c_{k_+-1})\dot{\mathsf{H}}_r(c_{k_+}^*),$$

for $0 \leq c_{k_+-1} \leq c_{k_+}^* \leq c_+$. Insert the inequality (B.9) and rearrange to get

$$c_+ - c_{k_+-1} \leq C_{\mathsf{H}}^2 K^{-1/2}. \tag{B.12}$$

Now, rewrite $0 \leq \mathcal{H} = \mathcal{H}_1 - \mathcal{H}_2$ where

$$\mathcal{H}_1 = \mathsf{H}_r(c_k + d) - \mathsf{H}_r(c_+), \qquad \mathcal{H}_2 = \mathsf{H}_r(c_{k-1} + d) - \mathsf{H}_r(c_+).$$

For the term $\mathcal{H}_1$, note that following the argument in item 5, then $c_k - c_{k-1}$ and $c_{k+1} - c_k$ are greater than $|d|$, so that $c_{k_+-1} \leq c_{k-1} \leq c_k + d \leq c_{k+1} \leq c_{k_++2}$. Since $c_{k_+-1} \leq c_+ \leq c_{k_+}$ it holds $|\mathcal{H}_1| \leq \mathsf{H}_r(c_{k_++2}) - \mathsf{H}_r(c_{k_+-1}) \leq 3H_r/K$.

For the term $\mathcal{H}_2$ use the mean value theorem to get $\mathcal{H}_2 = \delta_{k,d}\dot{\mathsf{H}}_r(c_+) + (\delta_{k,d}^2/2)\ddot{\mathsf{H}}_r(c^{**})$, where $\delta_{k,d} = c_{k-1} + d - c_+$ while c^{**} satisfies $|c^{**} - c_+| \leq |\delta_{k,d}|$. For the linear term note that (B.12) and the bound $|d| \leq K^{-1/2}$ imply $|\delta_{k,d}| \leq (C_{\mathsf{H}}^2 + 1)K^{-1/2}$, whereas (B.4) shows $\dot{\mathsf{H}}_r(c_+) = H_r/(C_{\mathsf{H}}K^{1/2})$. For the quadratic term note that $\delta_{k,d}^2 \leq (C_{\mathsf{H}}^2+1)^2K^{-1}$, while $\ddot{\mathsf{H}}_r(c^{**})$ is bounded by (B.6). Therefore $|\mathcal{H}_2| \leq K^{-1}\{(C_{\mathsf{H}}^2 + 1)H_r/C_{\mathsf{H}} + (C_{\mathsf{H}}^2 + 1)^2 \sup_{c\in\mathbb{R}} |\ddot{\mathsf{H}}_r(c)|/2\}$.

Combine to get $\mathcal{H} \leq |\mathcal{H}_1| + |\mathcal{H}_2| \leq C/K$ for some constant C not depending on K. ∎

The next lemma shows how small fluctuations in the arguments of the function $J_{i,p}$ can be controlled in terms of $J_{i,p}$ functions defined on the grid points. The proof uses Lemma B.1.

Lemma B.2 *Suppose Assumption 4.1(i) holds with $\nu = 1$ only. For any c let c_k be a right grid point for $k < K$ that is $c_{k-1} < c \leq c_k$ and let c_k is a left grid point for $c > c_{K-1}$ so $k = K - 1$. Then an integer $k_J > 0$ exists so that for all K satisfying (B.4) and all $c, d, d_m \in \mathbb{R}$ so $|d| \leq K^{-1/2}$ and $|d - d_m| \leq K^{-1}$, then integers $k^\dagger, k^\ddagger$ exists so*

$$|J_{i,p}(c, c+d) - J_{i,p}(c_k, c_k + d_m)| \leq |J_{i,p}(c_{k-k_J}, c_k)| + |J_{i,p}(c_{k^\dagger - k_J}, c_{k^\dagger})| + |J_{i,p}(c_{k^\ddagger - k_J}, c_{k^\ddagger})|.$$

Proof of Lemma B.2. 1. *Decomposition.* Only the case $k < K$ is proved. The proof for $k = K$ is similar. Let $\sigma = 1$ for notational simplicity. Write

$$\mathcal{J} = J_{i,p}(c, c+d) - J_{i,p}(c_k, c_k + d_m) = \varepsilon_i^p(\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3),$$

in terms of indicator functions $\mathcal{I}_1 = 1_{(c<\varepsilon_i\leq c_k)}$, $\mathcal{I}_2 = 1_{(\varepsilon_i\leq c_k+d)} - 1_{(\varepsilon_i\leq c_k+d_m)}$ and $\mathcal{I}_3 = 1_{(c+d<\varepsilon_i\leq c_k+d)}$. It follows that $|\mathcal{J}| \leq |\varepsilon_i^p|(\mathcal{I}_1 + |\mathcal{I}_2| + \mathcal{I}_3)$.

2. *Bound for $\mathcal{I}_1$.* Since $c_{k-1} < c \leq c_k$ then $0 \leq \mathcal{I}_1 = 1_{(c<\varepsilon_i\leq c_k)} \leq 1_{(c_{k-1}<\varepsilon_i\leq c_k)}$.

3. *Bound for $\mathcal{I}_2$.* Write $d = d_m + (d - d_m)$ where $|d - d_m| \leq K^{-1}$. Let $c^\dagger = c_k + d_m$. Then it holds $|\mathcal{I}_2| \leq 1_{(c^\dagger - K^{-1}<\varepsilon_i\leq c^\dagger + K^{-1})}$. Using first this inequality and then the mean value theorem it holds

$$\mathcal{E}_2 = \mathsf{E}(|\varepsilon_i^p\mathcal{I}_2|) \leq \mathsf{H}_r(c^\dagger + K^{-1}) - \mathsf{H}_r(c^\dagger - K^{-1}) \leq 2H_r^{-1}\sup_{c\in\mathbb{R}} \dot{\mathsf{H}}_r(c)H_r/K.$$

Therefore, a $k^\dagger$ exists so $|\mathcal{I}_2| \leq 1_{(c_{k^\dagger - k_J}<\varepsilon_i\leq c_{k^\dagger})}$ where $k_J \leq 2H_r^{-1}\sup_{c\in\mathbb{R}} \dot{\mathsf{H}}_r(c) + 2$.

4. *Bound for $\mathcal{I}_3$.* Since $c_{k-1} < c \leq c_k$ then $\mathcal{I}_3 \leq 1_{(c_{k-1}+d<\varepsilon_i\leq c_k+d)}$. Using first this inequality and then Lemma B.1 noting that $|d| \leq K^{-1/2}$ it holds

$$\mathcal{E}_3 = \mathsf{E}(|\varepsilon_i^p|\mathcal{I}_3) \leq \mathsf{H}_r(c_k + d) - \mathsf{H}_r(c_{k-1} + d) \leq CH_r/K.$$

Therefore, a $k^\ddagger$ exists so $|\mathcal{I}_3| \leq 1_{(c_{k^\ddagger - k_J}<\varepsilon_i\leq c_{k^\ddagger})}$ where $k_J \leq C + 1$. ∎

The next inequality gives a tightness type result for the function H_r.

Lemma B.3 *Let $c_\psi = \mathsf{F}^{-1}(\psi)$. For all densities satisfying Assumption 4.1(ia) for some $\nu < 1$, then a $C_\nu > 0$ exists so that for all $0 \leq \phi \leq 1$ it holds*

$$\max_{0\leq\psi\leq 1-\phi}\{\mathsf{H}_r(c_{\psi+\phi}) - \mathsf{H}_r(c_\psi)\} \leq C_\nu\phi^{1-\nu}.$$

Proof of Lemma B.3. Let $\psi_0 = \mathsf{F}(0)$. Note that $2^r p$ is even for $r \in \mathbb{N}$, $p \in \mathbb{N}_0$.

1. *Let* $\psi \geq \psi_0$. Then $\mathsf{H}_r(c_{\psi+\phi}) - \mathsf{H}_r(c_\psi)$ is increasing in ψ since

$$\frac{d}{d\psi}\{\mathsf{H}_r(c_{\psi+\phi}) - \mathsf{H}_r(c_\psi)\} = \frac{\dot{\mathsf{H}}_r(c_{\psi+\phi})}{\mathsf{f}(c_{\psi+\phi})} - \frac{\dot{\mathsf{H}}_r(c_\psi)}{\mathsf{f}(c_\psi)} = c_{\psi+\phi}^{p2^r} - c_\psi^{p2^r} > 0.$$

Thus, $\max_{\psi_0 \leq \psi \leq 1-\phi}\{\mathsf{H}_r(c_{\psi+\phi}) - \mathsf{H}_r(c_\psi)\} \leq \mathsf{H}_r(\infty) - \mathsf{H}_r(c_{1-\phi})$. This bound satisfies

$$\mathsf{H}_r(\infty) - \mathsf{H}_r(c_{1-\phi}) = \int_{c_{1-\phi}}^{\infty} (1 + \epsilon^{p2^r})\mathsf{f}(\epsilon)d\epsilon = \phi + \int_{c_{1-\phi}}^{\infty} \epsilon^{p2^r}\mathsf{f}(\epsilon)d\epsilon.$$

Assumption 4.1(ia) shows $\mathsf{E}\varepsilon^{p2^r/\nu} \leq C$ for some $C > 0$ so $1 - \mathsf{F}(\epsilon) \leq C\epsilon^{-p2^r/\nu}$ by the Chebychev inequality. Hence, $\epsilon^{p2^r} \leq C^\nu\{1 - \mathsf{F}(\epsilon)\}^{-\nu}$, so that

$$\mathsf{H}_r(\infty) - \mathsf{H}_r(c_{1-\phi}) \leq \phi + C^\nu \int_{c_{1-\phi}}^{\infty} \{1 - \mathsf{F}(\epsilon)\}^{-\nu}\mathsf{f}(\epsilon)d\epsilon$$

Substituting $\psi = \mathsf{F}(\epsilon)$ so $d\psi = \mathsf{f}(\epsilon)d\epsilon$ gives

$$\mathsf{H}_r(\infty) - \mathsf{H}_r(c_{1-\phi}) \leq \phi + C^\nu \int_{1-\phi}^{1} (1 - x)^{-\nu}dx = \phi + \frac{C^\nu}{1-\nu}\phi^{1-\nu}.$$

2. *Let* $\psi \leq \psi_0 - \phi$. Apply a similar argument as in item 1, to show that $\mathsf{H}_r(c_{\psi+\phi}) - \mathsf{H}_r(c_\psi)$ is decreasing because $c_\psi < c_{\psi+\phi} \leq 0$. Thus, $\mathsf{H}_r(c_\phi) - \mathsf{H}_r(-\infty)$ satisfies the same bound.

3. *Let* $\psi_0 - \phi \leq \psi \leq \psi_0$. Then

$$\mathcal{H} = \max_{\psi_0 - \phi \leq \psi \leq \psi_0}\{\mathsf{H}_r(c_{\psi+\phi}) - \mathsf{H}_r(c_\psi)\} \leq \mathsf{H}_r(c_{\psi_0+\phi}) - \mathsf{H}_r(c_{\psi_0-\phi}).$$

Using the mean value theorem then, for some ψ^* so $\psi_0 - \phi \leq \psi^* \leq \psi_0 + \phi$,

$$\mathcal{H} \leq \mathsf{H}_r\{\mathsf{F}^{-1}(\psi_0 + \phi)\} - \mathsf{H}_r\{\mathsf{F}^{-1}(\psi_0 - \phi)\} = \mathsf{H}_r\{\mathsf{F}^{-1}(\psi^*)\}2\phi \leq 2H_r\phi.$$

4. *Combine results.* Note that $\phi \leq \phi^{1-\nu}$. Let $C_\nu = \max\{2H_r, 1 + C^\nu/(1-\nu)\}$. ■

C Proofs of auxiliary Theorems 4.1–4.4

Proof of Theorem 4.1. Without loss of generality let $\sigma = 1$. Let $\tilde{R}(b, c_\psi) = \mathbb{F}_n^{g,p}(b, c_\psi) - \mathbb{F}_n^{g,p}(0, c_\psi)$ and $\mathcal{R}_n = \sup_{0 \leq \psi \leq 1} \sup_{|b| \leq n^{1/4-\eta}B} |\mathbb{F}_n^{g,p}(b, c_\psi) - \mathbb{F}_n^{g,p}(0, c_\psi)|$.

1. *Partition the support.* For $\delta, n > 0$ partition axis as laid out in (B.3) with $K = \mathrm{int}(H_r n^{1/2}/\delta)$ using Assumption 4.1(ia) with $\nu = 1$ only.

2. *Assign c_ψ to the partitioned support.* Consider $1/2 \leq \psi \leq 1$ only, noting that a similar argument can be made for $0 \leq \psi \leq 1/2$. Thus, for each c_ψ there exists c_{k-1}, c_k so $c_{k-1} < c_\psi \leq c_k$.

3. *Construct b-balls.* For a $\zeta > \kappa$ cover the set $|b| \leq n^{1/4-\eta}B$ with $M = \mathrm{O}\{n^{(1/4-\eta+\zeta)\dim x}\}$ balls of radius $n^{-\zeta}$ with centers b_m. Thus, for any b there exists a b_m so $|b - b_m| < n^{-\zeta}$.

4. *Apply chaining.* For $k < K$ so $c_\psi \leq c_{K-1}$ then relate c_ψ to the nearest right grid point so $\tilde{R}(b, c_\psi) = \tilde{R}(b_m, c_k) + \{\tilde{R}(b, c_\psi) - \tilde{R}(b_m, c_k)\}$, whereas for $k = K$ so $c_\psi > c_{K-1}$ related

c to nearest left grid point so $\tilde{R}(b,c_\psi)=\tilde{R}(b_m,c_{K-1})+\{\tilde{R}(b,c_\psi)-\tilde{R}(b_m,c_{K-1})\}$. Therefore $\mathcal{R}_n\leq\sum_{j=1}^3\mathcal{R}_{n,j}$, where

$$\begin{aligned}\mathcal{R}_{n,1}&=\max_{1\leq k<K}\max_{1\leq m\leq M}|\tilde{R}(b_m,c_k)|,\\ \mathcal{R}_{n,2}&=\max_{1\leq k<K}\max_{1\leq m\leq M}\sup_{c_{k-1}<c_\psi\leq c_k}\sup_{|b-b_m|<n^{-\zeta}}|\tilde{R}(b,c_\psi)-\tilde{R}(b_m,c_k)|\\ &\quad+\max_{1\leq m\leq M}\sup_{c_{K-1}<c_\psi}\sup_{|b-b_m|<n^{-\zeta}}|\tilde{R}(b,c_\psi)-\tilde{R}(b_m,c_{K-1})|.\end{aligned}$$

Thus, it suffices to show that $\mathsf{P}(\mathcal{R}_{n,j}>\gamma)$ vanishes for $j=1,2$.

5. *The term* $\mathcal{R}_{n,1}$. Use Lemma 5.2 to see that $\mathcal{R}_{n,1}=\mathrm{o}_\mathsf{P}(1)$. To see this let $\upsilon=1/2$ and let g_{in} have coordinates g_{in}^*. Then write $\tilde{R}(b_m,c_k)$ as $n^{-1/2}\sum_{i=1}^n(z_{\ell i}-\mathsf{E}_{i-1}z_{\ell i})$ with $z_{\ell i}=g_{in}^*J_{i,p}(c_k,c_k+\sigma^{-1}x_{in}'b_m)$, see definition in (B.1), and where ℓ represents the indices k,m. The conditions of Lemma 5.2 need to be demonstrated.

The parameter λ. The set of indices ℓ has size $L=\mathrm{O}(n^\lambda)$ where $\lambda=1/2+(1/4-\eta+\zeta)\dim x$ since $K=\mathrm{O}(n^{1/2})$ and $M=\mathrm{O}\{n^{(1/4-\eta+\zeta)\dim b}\}$.

The parameter ς. Since $|1_{(\varepsilon_i\leq c_k+x_{in}'b_m)}-1_{(\varepsilon_i\leq c_k)}|\leq 1_{(c_k-|x_{in}||b_m|<\varepsilon_i\leq c_k+|x_{in}||b_m|)}$ then, for $1\leq q\leq r$,

$$\mathsf{E}_{i-1}(J_{i,p})^{2^q}\leq\mathsf{H}_r(c_k+|x_{in}||b_m|)-\mathsf{H}_r(c_k-|x_{in}||b_m|)\leq 2|x_{in}||b_m|\sup_{v\in\mathbb{R}}\mathsf{H}_r'(v),$$

when using the mean-value theorem. Since $|b_m|\leq n^{1/4-\eta}B$ while $\sup_{v\in\mathbb{R}}\mathsf{H}_r'(v)<\infty$ by Assumption 4.1(ib) then

$$D_q=\max_{1\leq\ell\leq L}\textstyle\sum_{i=1}^n\mathsf{E}_{i-1}(z_{\ell i})^{2^q}\leq C_1(n^{-1}\sum_{i=1}^n|g_{in}^*|^{2^q}|n^{1/2}x_{in}|)n^{3/4-\eta}. \tag{C.1}$$

Thus, $\mathsf{E}D_q=\mathrm{O}(n^\varsigma)$ where $\varsigma=3/4-\eta$ by Assumption 4.1($iiia$).

Condition (i) is that $\varsigma<2\upsilon$. This holds since $0<\eta$ so that $\varsigma=3/4-\eta<1=2\upsilon$.

Condition (ii) is that $\varsigma+\lambda<2\upsilon$. If $\zeta>\kappa$ is chosen sufficiently small then

$$\varsigma+\lambda=1+(1/4+\kappa-\eta)(1+\dim x)+(\zeta-\kappa)\dim x-\kappa<\upsilon 2^r=2^{r-1},$$

provided r is chosen so $2^{r-1}\geq 1+(1/4+\kappa-\eta)(1+\dim x)$.

6. *Decompose* $\mathcal{R}_{n,2}$. It will be argued that $\mathcal{R}_{n,2}\leq 3(\widetilde{\mathcal{R}}_{n,2}+2\overline{\mathcal{R}}_{n,2})+\mathrm{o}_\mathsf{P}(1)$, where

$$\widetilde{\mathcal{R}}_{n,2}=\max_{1\leq k\leq K}n^{-1/2}\textstyle\sum_{i=1}^n|g_{in}|\{|J_{i,p}(c_{k-k_J},c_k)|-\mathsf{E}_{i-1}|J_{i,p}(c_{k-k_J},c_k)|\}, \tag{C.2}$$

$$\overline{\mathcal{R}}_{n,2}=\max_{1\leq k\leq K}n^{-1/2}\textstyle\sum_{i=1}^n|g_{in}|\mathsf{E}_{i-1}|J_{i,p}(c_{k-k_J},c_k)|. \tag{C.3}$$

To see this, let c_k denote nearest right grid point for $c_\psi\leq c_{K-1}$ while $c_k=c_{K-1}$ for $c_\psi>c_{K-1}$. Note first that $\tilde{R}_\mathsf{F}^p(b,c_\psi)-\tilde{R}_\mathsf{F}^p(b_m,c_k)$ involves the functions

$$\mathcal{J}_i=J_{i,p}(c_\psi,c_\psi+x_{in}'b)-J_{i,p}(c_k,c_k+x_{in}'b_m).$$

Assumption 4.1(ii) gives that $\max_{1\leq i\leq n}|x_{in}|=\mathrm{O}_\mathsf{P}(n^{\kappa-1/2})$. Thus, for all $\epsilon>0$ an $C_x>0$ exists so that the set $(\max_{1\leq i\leq n}|x_{in}|\leq n^{\kappa-1/2}C_x)$ has probability of at least $1-\epsilon$. On that set and with $d = x_{in}'b$ and $d = x_{in}'b$ then $|d| = \mathrm{O}(n^{-1/4+\kappa-\eta}) = \mathrm{o}(K^{-1/2})$ for $\eta - \kappa > 0$ and $|d - d_m| = \mathrm{O}(n^{-1/2+\kappa-\zeta}) = \mathrm{o}(K^{-1})$ for $\zeta - \kappa > 0$. Thus, for sufficiently large n then $|d| < K^{-1/2}$ and $|d - d_m| < K^{-1}$. Lemma B.2 using Assumption 4.1(i) then shows that a k_J exists so that for all c, d, d_m there exist $k^\dagger, k^\ddagger$ so

$$|\mathcal{J}_i| \leq |J_{i,p}(c_{k-k_J}, c_k)| + |J_{i,p}(c_{k^\dagger-k_J}, c_{k^\dagger})| + |J_{i,p}(c_{k^\ddagger-k_J}, c_{k^\ddagger})|, \tag{C.4}$$

As a consequence it holds, as desired, $\mathcal{R}_{n,2} \leq 3(\widetilde{\mathcal{R}}_{n,2} + 2\overline{\mathcal{R}}_{n,2}) + \mathrm{o}_{\mathsf{P}}(1)$.

7. *The term* $\widetilde{\mathcal{R}}_{n,2}$ is $\mathrm{o}_{\mathsf{P}}(1)$ by Lemma 5.2. To see this note that $\widetilde{\mathcal{R}}_{n,2}$ is the maximum of a family of martingale of the required form with $\ell = k$ so $L = K$ and $z_{\ell i} = |g_{in}||J_{i,p}(c_{k-k_J}, c_k)|$ and it suffices to set $\bar{r} = 2$.

Condition (i) holds with $\lambda = 1/2$ since $K = \mathrm{int}(H_r n^{1/2}/\delta)$.

Condition (ii) holds with $\varsigma = 1/2$ since $\mathsf{E}_{i-1}(J_{i,p})^{2^{\bar{r}}} \leq \mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-k_J}) = k_J H_r/K$ so that $\sum_{i=1}^n \mathsf{E}_{i-1}(J_{i,p})^{2^{\bar{r}}} = \mathrm{O}(n^{1-1/2})$, uniformly in ℓ, i.

It holds that $\lambda + \varsigma = 1$ which is less than $2^{\bar{r}} = 4$.

8. *Bounding* $\bar{\mathcal{R}}_{n,2}$. Note $\mathsf{E}_{i-1}|J_{i,p}(c_{k-k_J}, c_k)| \leq 2k_J\delta n^{-1/2}$ uniformly in i, k by the same argument as in item 7. It follows that $\bar{\mathcal{R}}_{n,2} \leq 2k_J\delta n^{-1}\sum_{i=1}^n |g_{in}|$. Here $n^{-1}\sum_{i=1}^n |g_{in}| = \mathrm{O}_{\mathsf{P}}(1)$ by Markov's inequality and Assumption 4.1($iiia$), so that $\bar{\mathcal{R}}_{n,2} = \mathrm{O}_{\mathsf{P}}(\delta)$. Thus, choosing δ sufficiently small then $\bar{\mathcal{R}}_{n,2}$ is small in probability. ∎

Proof of Theorem 4.2. It suffices to show, for all $\omega < \eta - \kappa$ where $\eta - \kappa \leq 1/4$, that

$$\mathcal{S}_1 = \sup_{0\leq\psi\leq 1} \sup_{|b|\leq n^{1/4-\eta}B} \sup_{d\in\mathbb{R}} |\mathbb{F}_n^{1,0}(b, c_\psi + n^{\kappa-1/2}d) - \mathbb{F}_n^{1,0}(0, c_\psi + n^{\kappa-1/2}d)| = \mathrm{o}_{\mathsf{P}}(n^{-\omega}),$$

$$\mathcal{S}_2 = \sup_{0\leq\psi\leq 1} \sup_{|d|\leq n^{1/4-\eta}B} |\mathbb{F}_n^{1,0}(0, c_\psi + n^{\kappa-1/2}d) - \mathbb{F}_n^{1,0}(0, c_\psi)| = \mathrm{o}_{\mathsf{P}}(n^{-\omega}).$$

For each term the proof of Theorem 4.1 is used with minor modifications. Since $p = 0$ then $2^r p = 0$ for all r, which simplifies the assumptions.

A. The term $\mathcal{S}_1$. The steps of the proof of Theorem 4.1 are modified as follows.

1. Choose $K = \mathrm{int}(H_r n^{1/2+\omega}/\delta)$ where $\omega < \eta - \kappa \leq 1/4$.
2. For each $c_\psi + n^{\kappa-1/2}d$ there exists c_{k-1}, c_k depending on n so $c_{k-1} < c_\psi + n^{\kappa-1/2}d \leq c_k$.
3. Choose $\zeta \geq \eta$ which implies $\zeta > \kappa$ since $\kappa < \eta$. The b-set is now $|b| \leq n^{1/4+\kappa-\eta}B$ so that the number of b-balls is $M = \mathrm{O}\{n^{(1/4+\kappa-\eta+\zeta)\dim x}\}$.
4. Note that in the chaining argument c_ψ is replaced by $c_\psi + n^{\kappa-1/2}d$. This only affects $\mathcal{R}_{n,2}$.
5. The term $\mathcal{R}_{n,1}$. Use Lemma 5.2 to see that $\mathcal{R}_{n,1} = \mathrm{o}_{\mathsf{P}}(n^{-\omega})$, now using $\upsilon = 1/2 - \omega > 1/2 + \kappa - \eta$. Define $z_{\ell i}$ as before. Since $p = 0$, $g_{in} = 0$ then $|J_{i,p}(x, y)|^{2^r} = |J_{i,p}(x, y)|$ and $|z_{\ell i}^{2^r}| = |z_{\ell i}|$ for any $r \in \mathbb{N}_0$. The inequality (C.1) for D_q holds as before, uniformly in $q \in \mathbb{N}$ so $\varsigma = 3/4 - \eta$, but $\lambda = 1/2 + \omega + (1/4 + \kappa - \eta + \zeta)\dim x$. *Condition* ($i$) holds since $\eta \leq 1/4$ and $\kappa \geq 0$ so $\varsigma = 3/4 - \eta \leq 1 + \kappa - 2\eta < 2\upsilon$. *Condition* ($ii$) holds since $\varsigma + \lambda < \infty$ while $\upsilon > 0$. Thus, for any ζ and sufficiently large $\bar{r}$ then $\varsigma + \lambda < \upsilon 2^{\bar{r}}$.
6. Lemma B.2 is an analytic result holding in finite samples. So the argument is not affect the dependence of c_k on n through $c_\psi + n^{\kappa-1/2}d$. In particular, (C.4) holds as stated and therefore the decomposition of $\mathcal{R}_{n,2}$ holds, noting that K is now chosen differently.
7. Apply Lemma 5.2 with $\bar{r} = 2$, but with λ, ς chosen differently. *Condition* (i) holds with $\lambda = 1/2 + \omega$ holds since $K = \mathrm{int}(H_r n^{1/2+\omega}/\delta)$. *Condition* ($ii$) holds with $\varsigma = 1/2 - \omega$

since $\mathsf{E}_{i-1}(J_{i,p})^4 = \mathsf{E}_{i-1}(J_{i,p}) \leq \mathsf{H}_r(c_k) - \mathsf{H}_r(c_{k-k_J}) = k_J H_r/K$ so that $\sum_{i=1}^n \mathsf{E}_{i-1}(J_{i,p})^4 = \mathrm{O}(n^{1-1/2-\omega})$, uniformly in ℓ, i. It holds that $\lambda + \varsigma = 1$ which is less than $2^2(1/2-\omega)$ for all $\omega < 1/4$. Lemma 5.2 then shows $\tilde{\mathcal{R}}_{n,2} = \mathrm{o}_{\mathsf{P}}(n^{-\omega})$ for all $\omega < 1/4$.

8. Note $\mathsf{E}_{i-1}|J_{i,p}(c_{k-k_J}, c_k)| \leq 2k_J\delta n^{-\omega-1/2}$ uniformly in i, k by the same argument as in item 7. It follows that $\bar{\mathcal{R}}_{n,2} \leq (n^{-1}\sum_{i=1}^n |g_{in}|^{2^r})n^{-\omega} = \mathrm{O}_{\mathsf{P}}(n^{-\omega})$.

B. The term $\mathcal{S}_2$. Rewrite

$$\mathcal{S}_2 = \sup_{0\leq\psi\leq 1} \sup_{|d|\leq n^{1/4-\eta}B} |\mathbb{F}_n^{1,0}(0, c_\psi + n^{\kappa-1/2}d) - \mathbb{F}_n^{1,0}(0, c_\psi)|.$$

Choosing the regressor as $x_{in}^* = n^{\kappa-1/2}$, then $\mathbb{F}_n^{1,0}(0, c_\psi + n^{\kappa-1/2}d) = \mathbb{F}_n^{1,0}(d, c_\psi)$. Apply the argument of part A. ∎

Proof of Theorem 4.3. The expression of interest is

$$R(b, c_\psi) = n^{1/2}\{\overline{\mathsf{F}}_n^{g,p}(b, c_\psi) - \overline{\mathsf{F}}_n^{g,p}(0, c_\psi)\} - \sigma^{p-1}c_\psi^p \mathsf{f}(c_\psi)n^{-1}\textstyle\sum_{i=1}^n g_{in}n^{1/2}x_{in}'b.$$

Recalling the definition of $\overline{\mathsf{F}}_n^{g,p}$ from (4.2) this satisfies $R(b, c_\psi) = n^{-1/2}\sum_{i=1}^n g_{in}\mathcal{S}_i(b, c_\psi)$ where

$$\mathcal{S}_i(b, c_\psi) = \mathsf{E}_{i-1}[\varepsilon_i^p\{1_{(\varepsilon_i\leq\sigma c_\psi + b'x_{in})} - 1_{(\varepsilon_i\leq\sigma c_\psi)}\}] - \sigma^{p-1}x_{in}'bc_\psi^p\mathsf{f}(c_\psi).$$

A bound is needed for $\mathcal{S}_i(b, c_\psi)$. Let $h_{in} = \sigma^{-1}x_{in}'b$ and $\mathsf{g}(c) = c^p\mathsf{f}(c)$. Write $\mathcal{S}_i(b, c_\psi)$ as an integral and Taylor expand to second order to get

$$\mathcal{S}_i(b, c_\psi) = \int_{c_\psi}^{c_\psi+h_{in}} g(c)dc - h_{in}g(c_\psi) = \frac{1}{2}h_{in}^2\dot{\mathsf{g}}(c^*),$$

for an intermediate point so $|c^* - c_\psi| \leq |h_{in}|$. Exploit the bound $|b| \leq n^{1/4-\eta}B$ to get

$$|\mathcal{S}_i(b, c_\psi)| \leq \frac{1}{2}\sigma^{-2}|b|^2|x_{in}|^2 \sup_{c\in\mathbb{R}}|\dot{\mathsf{g}}(c^*)| = |x_{in}|^2 \sup_{c\in\mathbb{R}}|\dot{\mathsf{g}}(c)|\mathrm{O}(n^{1/2-2\eta}).$$

Thus, by the triangular inequality then

$$|R(b, c_\psi)| \leq n^{-1/2}\textstyle\sum_{i=1}^n|g_{in}||S_i(b, c_\psi)| \leq \mathrm{O}(n^{-2\eta})n^{-1}\sum_{i=1}^n|g_{in}||n^{1/2}x_{in}|^2 \sup_{c\in\mathbb{R}}|\dot{\mathsf{g}}(c)|.$$

Due to Assumption 4.1($ib, iiib$), this expression is of order $\mathrm{O}_{\mathsf{P}}(n^{-2\eta})$ uniformly in ψ, b. ∎

Proof of Theorem 4.4. 1. *Coefficients* $\sigma, \epsilon, \phi, r$. Without loss of generality let $\sigma = 1$ and $0 < \phi < 1$ and $\epsilon < 1$. Since $\psi^\dagger - \psi \leq \phi$ then Lemma B.3 with Assumption 4.1(ia) shows that $0 < \nu < 1$ and $C_1 > 0$ exist so $\mathsf{H}_r(c_{\psi^\dagger}) - \mathsf{H}_r(c_\psi) \leq C_1\phi^{1-\nu}$. Now, take $0 < \epsilon$ and n as well as $0 < \phi^{(1-\nu)/4} \leq \epsilon^2$ as given. Throughout, constants $C_j > 0$ for $j = 1, 2\ldots$ will be found not depending on ϕ, n, ϵ. Let $r = 2$.

2. *Fine grid.* Given ϵ, ϕ, n let $\overline{m}$ satisfy $2^{-\overline{m}} \leq n^{-1/2}\epsilon\phi^{(1-\nu)/4} \leq 2^{1-\overline{m}}$.

3. *Coarse grid.* Let $\underline{m}$ satisfy $2^{-\underline{m}-1}H_r < C_1\phi^{1-\nu} \leq 2^{-\underline{m}}H_r$. For large n then $\overline{m} > \underline{m}$.

4. *Partition support.* For each of $m = \underline{m}, \ldots, \overline{m}$ partition axis as laid out in (B.3) with $K_m = 2^m$ points. For each m, points $c_{k_m,m}$ and $c_{k^\dagger,m}$ exist so $\underline{c}_m = c_{k_m-1,m} < c_\psi \leq c_{k_m,m} = \overline{c}_m$ and $\underline{c}_m^\dagger = c_{k_m^\dagger-1,m} < c_{\psi^\dagger} \leq c_{k_m^\dagger,m} = \overline{c}_m^\dagger$. It holds that $\overline{c}_{m-1} = c_{k_{m-1},m-1}$ equals either

$\overline{c}_m = c_{k_m,m}$ or $c_{k_m+1,m}$ so that $\overline{c}_{m-1} \geq \overline{c}_m$ and $\mathsf{H}(\overline{c}_{m-1}) - \mathsf{H}(\overline{c}_m)$ is either zero or $2^{-m}H_r$. There is at most one $\underline{m}$-grid point in the interval $c_\psi, c_{\psi^\dagger}$.

5. *Decompose* $J_{i,p}(c_\psi, c_{\psi^\dagger})$, see definition in (B.1). Split the $c_\psi, c_{\psi^\dagger}$ interval into three intervals where the partitioning points are $\overline{c}_{\overline{m}}$ and $\underline{c}^\dagger_{\overline{m}}$ which are the fine grid points to the right of c_ψ and to the left of $c_{\psi^\dagger}$, respectively. Note, that if $c_\psi, c_{\psi^\dagger}$ are in the same $\overline{m}$-interval then $\overline{c}_{\overline{m}} > \underline{c}^\dagger_{\overline{m}}$ and if they are in neighbouring $\overline{m}$-interval then $\overline{c}_{\overline{m}} = \underline{c}^\dagger_{\overline{m}}$. Thus,

$$J_{i,p}(c_\psi, c_{\psi^\dagger}) = J_{i,p}(c_\psi, \overline{c}_{\overline{m}}) + J_{i,p}(\underline{c}^\dagger_{\overline{m}}, c_{\psi^\dagger}) - 1_{(\overline{c}_{\overline{m}} > \underline{c}^\dagger_{\overline{m}})} J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}}) + 1_{(\overline{c}_{\overline{m}} < \underline{c}^\dagger_{\overline{m}})} J_{i,p}(\overline{c}_{\overline{m}}, \underline{c}^\dagger_{\overline{m}}).$$

Consider the fourth term. An iterative argument can be made. Since $\overline{c}_{\overline{m}} < \underline{c}^\dagger_{\overline{m}}$ then the coarser $(\overline{m}-1)$-grid satisfies $\overline{c}_{\overline{m}} \leq \overline{c}_{\overline{m}-1} \leq \underline{c}^\dagger_{\overline{m}-1} \leq \underline{c}^\dagger_{\overline{m}}$, so that

$$J_{i,p}(\overline{c}_{\overline{m}}, \underline{c}^\dagger_{\overline{m}}) = J_{i,p}(\overline{c}_{\overline{m}}, \overline{c}_{\overline{m}-1}) + J_{i,p}(\overline{c}_{\overline{m}-1}, \underline{c}^\dagger_{\overline{m}-1}) + J_{i,p}(\underline{c}^\dagger_{\overline{m}-1}, \underline{c}^\dagger_{\overline{m}}).$$

If $\overline{c}_{\overline{m}-1} = \underline{c}^\dagger_{\overline{m}-1}$ then $J_{i,p}(\overline{c}_{\overline{m}-1}, \underline{c}^\dagger_{\overline{m}-1}) = 0$ and the iteration stops noting that for $m < \overline{m}-1$ then m-grid points cross over so $\overline{c}_m \geq \overline{c}_{\overline{m}-1} = \underline{c}^\dagger_{\overline{m}-1} \geq \underline{c}^\dagger_m$. If $\overline{c}_{\overline{m}-1} < \underline{c}^\dagger_{\overline{m}-1}$ then the argument can be made again for $J_{i,p}(\overline{c}_{\overline{m}-1}, \underline{c}^\dagger_{\overline{m}-1})$. In the m-th step the iteration continues if $\overline{c}_m < \underline{c}^\dagger_m$, so that if there are no other m-grid points between $\overline{c}_{\overline{m}}, \underline{c}^\dagger_{\overline{m}}$ the contribution from the $(m-1)$-step is zero. Since there is at most one $\underline{m}$-point in the interval $c_\psi, c_{\psi^\dagger}$, then the $\underline{m}$-step will either give a zero contribution or the grid points will have crossed over at an earlier stage. Therefore the fourth term satisfies

$$1_{(\overline{c}_{\overline{m}} < \underline{c}^\dagger_{\overline{m}})} J_{i,p}(\overline{c}_{\overline{m}}, \underline{c}^\dagger_{\overline{m}}) = \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^\dagger_m)} \{J_{i,p}(\overline{c}_m, \overline{c}_{m-1}) + J_{i,p}(\underline{c}^\dagger_{m-1}, \underline{c}^\dagger_m)\}.$$

6. *Decompose* $\mathcal{S} = n^{1/2}\{\widetilde{\mathsf{F}}(0,0,c_{\psi^\dagger}) - \widetilde{\mathsf{F}}(0,0,c_\psi)\}$. Due to the decomposition of $J_{i,p}(c_\psi, c_{\psi^\dagger})$ in item 5 then $|\mathcal{S}| \leq |Z_1| + |Z_2| + |Z_3| + |Z_4| + |Z_5|$, where

$$\begin{aligned}
Z_1 &= \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n g_{in} [J_{i,p}(c_\psi, \overline{c}_{\overline{m}}) - \mathsf{E}_{i-1}\{J_{i,p}(c_\psi, \overline{c}_{\overline{m}})\}], \\
Z_2 &= \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n g_{in} [J_{i,p}(\underline{c}^\dagger_{\overline{m}}, c_{\psi^\dagger}) - \mathsf{E}_{i-1}\{J_{i,p}(\underline{c}^\dagger_{\overline{m}}, c_{\psi^\dagger})\}], \\
Z_3 &= 1_{(\overline{c}_{\overline{m}} > \underline{c}^\dagger_{\overline{m}})} \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n g_{in} [J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}}) - \mathsf{E}_{i-1}\{J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})\}], \\
Z_4 &= \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^\dagger_m)} \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n g_{in} [J_{i,p}(\overline{c}_m, \overline{c}_{m-1}) - \mathsf{E}_{i-1}\{J_{i,p}(\overline{c}_m, \overline{c}_{m-1})\}], \\
Z_5 &= \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} 1_{(\overline{c}_m < \underline{c}^\dagger_m)} \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n g_{in} [J_{i,p}(\underline{c}^\dagger_{m-1}, \underline{c}^\dagger_m) - \mathsf{E}_{i-1}\{J_{i,p}(\underline{c}^\dagger_{m-1}, \underline{c}^\dagger_m)\}].
\end{aligned}$$

7. *The term* Z_1. Since $|J_{i,p}(c_\psi, \overline{c}_{\overline{m}})| \leq |J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})|$ it holds

$$|Z_1| \leq \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n |g_{in}| [|J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})| + \mathsf{E}_{i-1}\{|J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})|\}].$$

Since $\mathsf{E}_{i-1}\{|J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})|\} \leq \mathsf{H}_r(\overline{c}_{\overline{m}}) - \mathsf{H}_r(\underline{c}_{\overline{m}}) = 2^{-\overline{m}}H_r$ then Assumption 4.1(ia, iii) shows, for some $C_2 > 0$, that

$$\mathsf{E} \textstyle\sum_{i=1}^n |g_{in}| \mathsf{E}_{i-1}\{|J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})|\} \leq nC_2 2^{-\overline{m}} H_r. \tag{C.5}$$

Noting that $2^{-\overline{m}} \le n^{-1/2}\epsilon\phi^{(1-\nu)/4}$ and using the Markov inequality then

$$\mathsf{P}(|Z_1| > \epsilon) \le \frac{2}{\epsilon\sqrt{n}}\mathsf{E}\textstyle\sum_{i=1}^n |g_{in}|\mathsf{E}_{i-1}\{|J_{i,p}(\underline{c}_{\overline{m}}, \overline{c}_{\overline{m}})|\} \le 2C_2 H_r \phi^{(1-\nu)/4}.$$

8. *The terms Z_2 and Z_3.* Apply the same argument as in item 7.
9. *The term Z_4: finding martingale.* Introduce martingales

$$M_{\ell,m,n} = \frac{1}{\sqrt{n}}\textstyle\sum_{i=1}^n g_{in}[J_{i,p}(c_{\ell,m}, c_{\ell+1,m}) - \mathsf{E}_{i-1}\{J_{i,p}(c_{\ell,m}, c_{\ell+1,m})\}].$$

Recall that for instance $\overline{c}_m = c_{k_m,m}$ while $\overline{c}_{m-1}$ either equals $c_{k_m,m}$ or $c_{k_m+1,m}$ so that $\overline{c}_m, \overline{c}_{m-1}$ are at most 1 step apart in the m-grid. It then holds that

$$|Z_4| \le \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} |M_{k_m,m,n}|.$$

The point $c_{k_m,m}$ satisfies $\underline{c}_{\underline{m}} < c_\psi < c_{k_m,m} \le \overline{c}_{\underline{m}}$. Decompose the interval $\underline{c}_{\underline{m}}, \overline{c}_{\underline{m}}$ of length $2^{-\underline{m}}H_r$ into $2^{m-\underline{m}}$ intervals of length $2^{-m}H_r$ with left endpoint $c_{k_{\underline{m}}+\ell}$ for $0 \le \ell < 2^{m-\underline{m}}$, say. The interval $c_{k_m,m}, c_{k_m+1,m}$ is one of those. This gives rise to a further bound

$$|Z_4| \le \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} \max_{k_{\underline{m}} \le \ell < k_{\underline{m}} + 2^{m-\underline{m}}} |M_{\ell,m,n}|.$$

Note that $\sum_{m=\underline{m}+1}^{\overline{m}} 2^{(\underline{m}-m)/4} \le \sum_{j=1}^{\infty} 2^{-j/4} = (2^{1/4}-1)^{-1} < 6$. It therefore holds

$$\mathsf{P}(|Z_4| > \epsilon) \le \mathsf{P}\textstyle\bigcup_{m=\underline{m}+1}^{\overline{m}} \{\max_{k_{\underline{m}} \le \ell < k_{\underline{m}} + 2^{m-\underline{m}}} |M_{\ell,m,n}| > \frac{2^{(\underline{m}-m)/4}\epsilon}{6}\}$$

Using Boole's inequality then

$$\mathsf{P}(|Z_4| > \epsilon) \le \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} \mathsf{P}\{\max_{k_{\underline{m}} \le \ell < k_{\underline{m}} + 2^{m-\underline{m}}} |M_{\ell,m,n}| > \frac{2^{(\underline{m}-m)/4}\epsilon}{6}\}.$$

10. *The term Z_4: apply Lemma 5.3* with $z_{\ell,i} = g_{in}^* J_{i,p}(c_{\ell-1,m}, c_{\ell,m})$ where g_{in}^* is a coordinated of g_{in} and with $L = 2^{m-\underline{m}}$ while $\kappa = 2^{(\underline{m}-m)/4}\epsilon/6$. Noting that for $q = 1, 2$ then $\mathsf{E}_{i-1}|J_{i,p}(c_{\ell-1,m}, c_{\ell,m})|^{2^q} \le 2^{-m}H_r$. Therefore the moment condition holds with $C = 2^{-m}H_r C_2$ since

$$\mathsf{E}\max_{1\le\ell\le L}\textstyle\sum_{i=1}^n \mathsf{E}_{i-1}|z_{\ell,i}|^q \le \mathsf{E}\max_{1\le\ell\le L}\sum_{i=1}^n |g_{in}|^q 2^{-m}H_r = n2^{-m}H_r(n^{-1}\mathsf{E}\sum_{i=1}^n |g_{in}|^q) \le n2^{-m}H_r C_2,$$

as in the argument leading to (C.5) under Assumption 4.1(ia, iii). The Lemma then shows that for all $\theta_m > 0$ then

$$\mathsf{P}(|Z_4| > \epsilon) \le \textstyle\sum_{m=\underline{m}+1}^{\overline{m}} (\mathcal{A}_m\mathcal{B}_m + \mathcal{C}_m),$$

where

$$\mathcal{A}_m = \frac{C\theta_m}{\kappa}2^{-m/2}, \qquad \mathcal{B}_m = 2^{-m/2}\{1 + (L+1)\frac{\theta_m^2}{n}\}, \qquad \mathcal{C}_m = 4L\exp(-\frac{\kappa\theta_m}{14}).$$

Choose $\theta_m = 14\kappa^{-1}\{\log(4^{m-\underline{m}}) + \log\phi^{-1}\}$. If $\mathcal{A}_m, \mathcal{B}_m$ are exponentially decreasing in $m - \underline{m}$ and proportional to ϕ^α for some $\alpha > 0$ while $\mathcal{C}_m$ is bounded then $\mathsf{P}(|Z_4| > \epsilon) < C\phi^\alpha$ for some constant $C > 0$.

11. *The term $\mathcal{A}_m$.* Use that $\kappa = C2^{(\underline{m}-m)/4}\epsilon$, the definition of θ_m to get

$$\mathcal{A}_m = C\{\log(4^{m-\underline{m}}) + \log\phi^{-1}\}2^{-(m-\underline{m})/2}\epsilon^{-2}2^{-\underline{m}/2}.$$

Use the bounds $\epsilon^{-2} \le \phi^{-(1-\nu)/4}$ and $2^{-\underline{m}/2} < C\phi^{(1-\nu)/2}$ to get

$$\mathcal{A}_m < C\{(m-\underline{m})\log 4 + \log\phi^{-1}\}2^{-(m-\underline{m})/2}\phi^{(1-\nu)/4}.$$

Since $\phi^{(1-\nu)/8}\log\phi^{-1}$ is bounded and $\phi^{(1-\nu)/8} < 1$ then

$$\mathcal{A}_m < C\{(m-\underline{m}) + 1\}2^{-(m-\underline{m})/2}\phi^{(1-\nu)/8},$$

which is exponentially decreasing in $m - \underline{m}$ and proportional to $\phi^{(1-\nu)/8}$.

12. *The term $\mathcal{B}_m$.* Use that $2^{-m/2} \le 1$ and $(L+1) \le 2L$. Note that $\theta_m^2 = 2^m\kappa^2\mathcal{A}_m^2$ where $\mathcal{A}_m$ is bounded due to item 11. Therefore

$$\mathcal{B}_m \le 1 + C2^{-m/2}2^{m-\underline{m}}n^{-1}2^m\kappa^2 = 1 + C2^{m/2}2^{m-\underline{m}}n^{-1}\kappa^2.$$

Use that $n^{-1} \le C2^{-\overline{m}}\epsilon^{-2}\phi^{-(1-\nu)/2}$ and $\kappa^2 = C2^{(\underline{m}-m)/2}\epsilon^2$ to get

$$\mathcal{B}_m - 1 \le C2^{m/2}2^{m-\underline{m}}2^{-\overline{m}}\epsilon^{-2}\phi^{-(1-\nu)/2}2^{-(m-\underline{m})/2}\epsilon^2 = C2^{-(\overline{m}-m)}2^{-\underline{m}/2}\phi^{-(1-\nu)/2} \le C,$$

since $2^{-(\overline{m}-m)} \le 2$ and $\phi^{-(1-\nu)/2} < C2^{\underline{m}/2}$.

13. *The term $\mathcal{C}_m$.* Insert expression for θ_m to get

$$\mathcal{C}_m = C2^{m-\underline{m}}\exp\{-\log(4^{m-\underline{m}}) - \log\phi^{-1}\} = C2^{-(m-\underline{m})}\phi,$$

which is exponentially decreasing in $m - \underline{m}$ and proportional to $\phi \le \phi^{(1-\nu)/8}$.

14. *The terms Z_5.* Apply the same argument as for Z_4.

15. *Combine* the bounds from items 7,8,10,14 to get

$$\mathsf{P}(|\mathcal{S}| > \epsilon) \le \textstyle\sum_{j=1}^5 \mathsf{P}(|Z_j| > \epsilon) \le 3(2C_1H_r\phi^{(1-\nu)/4}) + 2C\phi^{(1-\nu)/8}.$$

uniformly in $\underline{m}, \overline{m}, n$. For a given $\epsilon > 0$ the only constraint to ϕ is that $0 < \phi^{(1-\nu)/4} \le \epsilon^2$. Thus, the probability vanishes as $\phi \downarrow 0$. ■

D Proofs of main Theorems 3.1-3.5

The main results for the forward search are proved in a series of steps. Theorem 3.1 shows that asymptotically the forward residuals behaves like the quantile of the absolute errors $|\varepsilon_i|$. It is therefore useful to start by reviewing some known results from the theory of quantile processes. Secondly, the forward search problem is reformulated in terms of a weighted and marked absolute empirical distribution function $\widehat{\mathsf{G}}_n$. At this point we work with absolute errors and it is natural to move from the general densities of Assumption 4.1 to the symmetric densities of Assumption 3.1. Thirdly, this empirical distribution function is analysed using the results from Section 4. Fourthly, the corresponding quantile processes are analysed. Fifthly, a single step of the Forward Search is analysed using these results. Sixthly, the iteration of the Forward Search is analysed.

D.1 Some known results from the theory of quantile processes

Introduce the empirical distribution function of the absolute errors, $|\varepsilon_i|/\sigma$, that is

$$\widehat{\mathsf{G}}_n(c) = \frac{1}{n}\sum_{i=1}^{n} 1_{(|\varepsilon_i| \le \sigma c)}. \tag{D.1}$$

The empirical quantiles of the absolute errors, $|\varepsilon_i|/\sigma$, are defined as

$$\hat{c}_\psi = \widehat{\mathsf{G}}_n^{-1}(\psi) = \inf\{c : \widehat{\mathsf{G}}_n(c) \ge \psi\}. \tag{D.2}$$

Empirical quantiles and empirical distribution functions are linked as follows.

Theorem D.1 *Csörgő (1983, Corollaries 6.2.1, 6.2.2). Suppose that* f *is symmetric, differentiable, positive for* $\mathsf{F}^{-1}(0) < c < \mathsf{F}^{-1}(1)$, *satisfying* $\gamma = \sup_{c>0} \mathsf{F}(c)\{1-\mathsf{F}(c)\}|\dot{\mathsf{f}}(c)|/\{\mathsf{f}(c)\}^2 < \infty$, *and decreasing for large* c. *Then, for all* $\zeta > 0$, *it holds*
(a) $\sup_{0\le\psi\le1} |2\mathsf{f}(c_\psi)n^{1/2}(\hat{c}_\psi - c_\psi) + n^{1/2}\{\widehat{\mathsf{G}}_n(c_\psi) - \psi\}| = \mathrm{o}_\mathsf{P}(n^{\zeta-1/4})$;
(b) $\sup_{0\le\psi\le1} |2\mathsf{f}(c_\psi)n^{1/2}(\hat{c}_\psi - c_\psi) - n^{1/2}\{\mathsf{G}(\hat{c}_\psi) - \psi\}| = \mathrm{o}_\mathsf{P}(n^{\zeta-1/2})$;
(c) $\sup_{0\le\psi\le1} |n^{1/2}\{\mathsf{G}(\hat{c}_\psi) - \psi\} + n^{1/2}\{\widehat{\mathsf{G}}_n(c_\psi) - \psi\}| = \mathrm{o}_\mathsf{P}(n^{\zeta-1/4})$.

The result in Theorem D.1(a) shows that the empirical quantile $\hat{c}_\psi$ satisfies, for $0 < \psi < 1$,

$$n^{1/2}(\hat{c}_\psi - c_\psi) = \frac{1}{2\mathsf{f}(c_\psi)} n^{1/2}\{\psi - \widehat{\mathsf{G}}_n(c_\psi)\} + \mathrm{o}_\mathsf{P}(1).$$

This is known as the Bahadur (1966) representation. The results in parts (b, c) combine to that of (a) and were studied by Kiefer (1967). More detail can be found in Csörgő (1983) who also gives almost sure, logarithmic rates.

Some weighted versions of the above results are also needed.

Theorem D.2 *(Shorack 1979, Csörgő, 1983, Theorem 5.1.1). Let the function* q_ψ *be symmetric about* $\psi = 1/2$ *(it suffices if* q_ψ *is bounded below by such a function), such that on* $0 \le \psi \le 1/2$ *then* q_ψ *is increasing and continuous, and satisfies* $q_\psi = \{\psi \log\log(1/\psi)\}^{1/2} g_\psi$ *for a function* g_ψ *so* $\lim_{\psi\to0} g_\psi = \infty$. *Then, a probability space exists on which one can define a Brownian bridge* $\mathbb{B}_n$ *for each* n, *so that*
(a) $\sup_{0\le\psi\le1} |\{\mathbb{G}_n(c_\psi) - \mathbb{B}_n(\psi)\}/q_\psi| = \mathrm{o}_\mathsf{P}(1)$;
(b) $\sup_{1/(n+1)\le\psi\le n/(n+1)} |\{\mathsf{f}(c_\psi)n^{1/2}(\hat{c}_\psi - c_\psi) - \mathbb{B}_n(\psi)\}/q_\psi| = \mathrm{o}_\mathsf{P}(1)$ *provided the assumptions of Theorem D.1 hold.*

In Theorem D.2 a possible choice of q_w is $\{\psi(1-\psi)\}^\alpha$ for $\alpha < 1/2$, which will be used in the proof of the main Theorem. Finally, a continuity property of the Brownian bridge is needed.

Theorem D.3 *(Revuz and Yor, 1998, Theorem I.2.2) A Brownian motion* $\mathbb{W}$ *is locally Hölder continuous of order* α *for all* $\alpha < 1/2$. *That is*

$$\sup_{0\le\psi<\psi^\dagger\le1} \frac{|\mathbb{W}(\psi^\dagger) - \mathbb{W}(\psi)|}{(\psi^\dagger - \psi)^\alpha} \overset{a.s.}{<} \infty.$$

Thus, for a Brownian bridge $\mathbb{B}$ *then* $\lim_{\psi\to0} \mathbb{B}(\psi)/\psi^\alpha = 0$ *a.s.*

D.2 Absolute empirical process representation

Normalizations are needed for estimators and regressors. Depending on the stochastic properties of the regressor x_i choose a normalization matrix N and define

$$\hat{b} = N^{-1}(\hat{\beta} - \beta), \qquad x_{in} = N'x_i,$$

so that $\sum_{i=1}^n x_{in}x_{in}'$ converges, $n^{-1/2}\sum_{i=1}^n |x_{in}|$ is bounded, and $x_i'(\hat{\beta} - \beta) = x_{in}'b$. If, for example, (y_i, x_i) is stationary then $N = n^{-1/2}I_{\dim x}$ so that $b = n^{1/2}(\hat{\beta}-\beta)$ and $x_{in} = n^{-1/2}x_i$. If x_i is a random walk then $N = n^{-1}$.

Introduce matrix-valued weights g_{in} of the form 1, $n^{1/2}Nx_i$ or $nNx_ix_i'N$, so that $n^{-1}\sum_{i=1}^n |g_{in}|$ is bounded. In the stationary case g_{in} will be 1, x_i or x_ix_i'. When x_i is a random walk g_{in} is 1, $n^{-1/2}x_i$ or $n^{-1}x_ix_i'$.

Define the *weighted and marked absolute empirical distribution functions*

$$\widehat{\mathsf{G}}_n^{g,p}(b, c) = \frac{1}{n}\sum_{i=1}^n g_{in}\varepsilon_i^p 1_{(|\varepsilon_i - x_{in}'b| \leq \sigma c)}, \tag{D.3}$$

for $b \in \mathbb{R}^{\dim x}$ and $c \geq 0$ and with weights g_{in} and marks ε_i^p. Four combinations of weights and marks are of interest in the analysis of the Forward Search. The deletion residuals involve $g_{in} = 1$, $p = 0$. The least squares estimator involves $g_{in} = n^{1/2}N'x_i$, $p = 1$ and $g_{in} = nN'x_ix_i'N$, $p = 0$. The variance estimator involves the mentioned terms as well as $g_{in} = 1$, $p = 2$. When $p = 0$ the marks are $\varepsilon_i^0 = 1$ so that $\widehat{\mathsf{G}}_n^{g,0}$ is a weighted absolute empirical distribution function similar to those studied by Koul and Ossiander (1994). When also $b = 0$ then $\widehat{\mathsf{G}}_n^{1,0}$ equals the empirical distribution function $\widehat{\mathsf{G}}_n$ of (D.1).

The Forward Search Algorithm 2.1 can now be cast as follows. Step $(m + 1)$ results in an order statistic

$$\hat{z}^{(m)} = \sigma \inf\{c : \widehat{\mathsf{G}}_n^{1,0}(\hat{b}^{(m)}, c) \geq \frac{m+1}{n}\}, \tag{D.4}$$

where $g_{in} = 1$, $p = 0$, so that

$$\frac{m+1}{n} = \widehat{\mathsf{G}}_n^{1,0}(\hat{b}^{(m)}, \frac{\hat{z}^{(m)}}{\sigma}) = \frac{1}{n}\sum_{i=1}^n 1_{(|\varepsilon_i - x_{in}'\hat{b}^{(m)}| \leq \hat{z}^{(m)})} = \frac{1}{n}\sum_{i\in S^{(m+1)}} 1. \tag{D.5}$$

The least squares estimator has estimation error

$$\hat{b}^{(m+1)} = N^{-1}(\hat{\beta}^{(m)} - \beta) = \{\widehat{\mathsf{G}}_n^{xx,0}(\hat{b}^{(m)}, \frac{\hat{z}^{(m)}}{\sigma})\}^{-1}\{n^{1/2}\widehat{\mathsf{G}}_n^{x,1}(\hat{b}^{(m)}, \frac{\hat{z}^{(m)}}{\sigma})\}, \tag{D.6}$$

while the bias corrected least squares variance estimator satisfies

$$n^{1/2}\{(\hat{\sigma}_{cor}^{(m+1)})^2 - \sigma^2\} = \frac{n^{1/2}}{\tau_{m/n}}[\widehat{\mathsf{G}}_n^{1,2}(\hat{b}^{(m)}, \frac{\hat{z}^{(m)}}{\sigma}) - \{\hat{b}^{(m+1)}\}'\widehat{\mathsf{G}}_n^{xx,0}(\hat{b}^{(m)}, \frac{\hat{z}^{(m)}}{\sigma})\{\hat{b}^{(m+1)}\}]. \tag{D.7}$$

D.3 The absolute empirical distribution

The process $\widehat{\mathsf{G}}_n$ is now analysed using the auxillary Theorems 4.1-4.4 for the process $\widehat{\mathsf{F}}_n$. Only the four combinations of g_{in}, p are now considered as outlined in Section D.2. When checking

Assumption 4.1 it suffices to check the conditions for the hybrid case where $g_{in} = nN'x_ix_i'N$ *and* $p = 2$. The process $\widehat{\mathsf{G}}_n$ can be expressed in terms of $\widehat{\mathsf{F}}_n$ quite easily by

$$\widehat{\mathsf{G}}_n^{g,p}(b,c) = \widehat{\mathsf{F}}_n^{g,p}(b,c) - \lim_{c^+\downarrow c} \widehat{\mathsf{F}}_n^{g,p}(b,-c^+). \tag{D.8}$$

The asymptotic arguments are made on the probability scale $\psi = \mathsf{G}(c_\psi)$. When f is symmetric then the probability scales of G and F are related in a simple linear fashion, see (2.2), so that (D.8) translates into

$$\widehat{\mathsf{G}}_n^{g,p}\{b,\mathsf{G}^{-1}(\psi)\} = \widehat{\mathsf{F}}_n^{g,p}\{b,\mathsf{F}^{-1}(\frac{1+\psi}{2})\} - \lim_{\psi^+\downarrow\psi} \widehat{\mathsf{F}}_n^{g,p}\{b,\mathsf{F}^{-1}(\frac{1-\psi^+}{2})\}. \tag{D.9}$$

Therefore, results for $\widehat{\mathsf{F}}_n$ transfer to $\widehat{\mathsf{G}}_n$. The corresponding conditional mean process is

$$\overline{\mathsf{G}}_n^{g,p}(b,c) = \frac{1}{n}\sum_{i=1}^n g_{in}\mathsf{E}_{i-1}\{\varepsilon_i^p 1_{(|\varepsilon_i - x_{in}'b|\le\sigma c)}\}, \qquad p = 0,1,2. \tag{D.10}$$

Form also the empirical process

$$\mathbb{G}_n^{g,p}(b,c) = n^{1/2}\{\widehat{\mathsf{G}}_n^{g,p}(b,c) - \overline{\mathsf{G}}_n^{g,p}(b,c)\}. \tag{D.11}$$

For later use note that $\mathsf{E}_{i-1}\{\varepsilon_i^p 1_{(|\varepsilon_i|\le\sigma c)}\} = 0$ for odd p since f is symmetric and $b = 0$. Errors in estimating the quantile are denoted $d = n^{1/2}(c_\psi^b - c_\psi)$. Estimation errors represented by b, d vanish uniformly as shown in the next result. Due to the two-sidedness of the absolute residuals and symmetry of f, only one of the error terms $x_{in}'b$ and $n^{-1/2}d$ enters the asymptotic expansion depending on the choice of p.

Lemma D.4 *For each* ψ *let* $c_\psi = \mathsf{G}^{-1}(\psi)$. *Suppose Assumption 3.1*(ia, iib, iic) *holds for some* $0 \le \kappa < \eta \le 1/4$. *Then, for all* $B, \epsilon > 0$ *and all* $\omega < \eta - \kappa \le 1/4$, *it holds*
(a) $\sup_{0\le\psi\le1}\sup_{|b|,|d|\le n^{1/4-\eta}B} |n^{1/2}\{\overline{\mathsf{G}}_n^{g,p}(b, c_\psi + n^{\kappa-1/2}d) - \overline{\mathsf{G}}_n^{g,p}(0,c_\psi)\}$
$-\sigma^{p-1}c_\psi^p\mathsf{f}(c_\psi)n^{-1/2}\sum_{i=1}^n g_{in}\{1_{(p\ odd)}x_{in}'b + 1_{(p\ even)}n^{\kappa-1/2}\sigma d\}|| = \mathsf{O}_\mathsf{P}\{n^{2(\kappa-\eta)}\}$;
(b) $\sup_{0\le\psi\le1}\sup_{|b|,|d|\le n^{1/4-\eta}B} |\mathbb{G}_n^{g,p}(b, c_\psi + n^{\kappa-1/2}d) - \mathbb{G}_n^{g,p}(0,c_\psi)| = \mathsf{o}_\mathsf{P}(1)$;
(b') $\sup_{0\le\psi\le1}\sup_{|b|,|d|\le n^{1/4-\eta}B} |\mathbb{G}_n^{1,0}(b, c_\psi + n^{\kappa-1/2}d) - \mathbb{G}_n^{1,0}(0,c_\psi)| = \mathsf{o}_\mathsf{P}(n^{-\omega})$;
(c) $\lim_{\phi\downarrow0}\limsup_{n\to\infty}\mathsf{P}\{\sup_{0\le\psi\le\psi^\dagger\le1:\psi^\dagger-\psi\le\phi} |\mathbb{G}_n^{g,p}(0,c_{\psi^\dagger}) - \mathbb{G}_n^{g,p}(0,c_\psi)| > \epsilon\} \to 0$.

Proof of Lemma D.4. (a) Assumption 3.1(ia, iic) implies Assumption 4.1$(ib, iiib)$ with $r = 0$, $p \le 2$ and $g_{in} = 1, n^{1/2}x_{in}$ or $nx_{in}x_{in}'$, and hence the assumptions of Theorem 4.3. First, we want to apply this result to $\overline{\mathsf{F}}_n^{g,p}(b, c_\psi + n^{\kappa-1/2}d)$. Thus, rewrite

$$\overline{\mathsf{F}}_n^{g,p}(b, c_\psi + n^{\kappa-1/2}d) = n^{-1}\sum_{i=1}^n g_{in}\mathsf{E}_{i-1}\varepsilon_i^p 1_{\{\varepsilon_i - x_{in}'b \le \sigma(c_\psi + n^{\kappa-1/2}d)\}} = n^{-1}\sum_{i=1}^n g_{in}\mathsf{E}_{i-1}\varepsilon_i^p 1_{(\varepsilon_i - \overline{x}_{in}'\overline{b}\le\sigma c_\psi)},$$

for $\overline{b} = (b', n^\kappa d)'$ and $\overline{x}_{in} = (x_{in}', n^{-1/2}\sigma)'$, where $|\overline{b}| \le 2n^{1/4+\kappa-\eta}B$ while $\overline{x}_{in}$ satisfies Assumption 4.1$(iiib)$ because $|\overline{x}_{in}|^2 = |x_{in}|^2 + n^{-1}\sigma^2$. Therefore we find, using that $\overline{\mathsf{G}}_n^{g,p}$ can be expressed in terms of $\overline{\mathsf{F}}_n^{g,p}$ as in (D.8), that $n^{1/2}\{\overline{\mathsf{G}}_n^{g,p}(b, c_\psi + n^{\kappa-1/2}d) - \overline{\mathsf{G}}_n^{g,p}(0,c_\psi)\}$ has correction term

$$\sigma^{p-1}c_\psi^p\mathsf{f}(c_\psi)n^{-1}\sum_{i=1}^n g_{in}n^{1/2}(x_{in}'b + n^{\kappa-1/2}\sigma d) - \sigma^{p-1}(-c_\psi)^p\mathsf{f}(-c_\psi)n^{-1}\sum_{i=1}^n g_{in}n^{1/2}(x_{in}'b - n^{\kappa-1/2}\sigma d)$$
$$= \sigma^{p-1}c_\psi^p\mathsf{f}(c_\psi)n^{-1/2}\sum_{i=1}^n g_{in}[\{1-(-1)^p\}x_{in}'b + \{1+(-1)^p\}n^{\kappa-1/2}\sigma d],$$

due to the $1_{(p\ odd)}x_{in}'b$symmetry of f. This reduces as desired.

(b) Let $c_\psi^\dagger = c_\psi + n^{\kappa-1/2}d$. Rewrite $\mathcal{G} = \mathbb{G}_n^{g,p}(b, c_\psi^\dagger) - \mathbb{G}_n^{g,p}(0, c_\psi)$ as $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$ where

$$\mathcal{G}_1 = \mathbb{G}_n^{g,p}(b, c_\psi^\dagger) - \mathbb{G}_n^{g,p}(0, c_\psi^\dagger), \qquad \mathcal{G}_2 = \mathbb{G}_n^{g,p}(0, c_\psi^\dagger) - \mathbb{G}_n^{g,p}(0, c_\psi).$$

The term $\mathcal{G}_1$ is $\mathrm{o}_{\mathsf{P}}(1)$ uniformly in $|b| \le n^{1/4-\eta}B$, $0 \le \psi \le 1$. To see this, expand $\mathbb{G}_n^{g,p}$ in a similar fashion to (D.8). Apply Theorem 4.1, noting that Assumption 3.1(ia, iib, iic) implies Assumption 4.1($i, ii, iiia$) with $p \le 2$, $g_{in} = 1, n^{1/2}x_{in}$ or $nx_{in}x_{in}'$ and the chosen r.

The term $\mathcal{G}_2$. Apply Theorem 4.4 noting that Assumption 3.1(ia, iic) implies Assumption 4.1($ia, iiia$) with $r = 2$ and some $\nu < 1$.

(b') Similar to (b), but using Theorem 4.2.

(c) Assumption 3.1(ia, iic) implies Assumption 4.1($ia, iiia$) using the Cauchy-Schwarz inequality. Expand $\mathbb{G}_n^{g,p}$ and apply Theorem 4.4. ■

D.4 Proof of Theorems 3.6 and 3.7.

Proof of Theorems 3.6 and 3.7. Tightness follows from Lemma D.4(c) and convergence of finite dimensional distributions follows from the central limit theorem for martingale differences, see Helland (1982, Theorem 3.2b) using Assumption 3.1(iic). ■

D.5 A first analysis of the order statistics

The Forward Search evolves around order statistics $\hat{z}^{(m)}$ defined in (D.4). A process version gives quantiles

$$\hat{c}_\psi^b = \inf\{c : \widehat{\mathsf{G}}_n^{1,0}(b, c) \ge \psi\}. \tag{D.12}$$

Setting $b = 0$ gives $\hat{c}_\psi^0 = \widehat{\mathsf{G}}_n^{-1}(\psi)$ as defined in (D.2) and studied in Theorem D.1. The first result gives an algebraic bound to the distance between $\hat{c}_\psi^b$ and $\hat{c}_\psi^0$. Probabilistic bounds follow.

Lemma D.5 *It holds, for all b, ψ, that* $\sigma|\hat{c}_\psi^b - \hat{c}_\psi^0| < 2|b| \max_{1\le i\le n} |x_{in}|$.

Proof of Lemma D.5. 1. *A property of* $\widehat{\mathsf{G}}_n$. The quantile $\sigma\hat{c}_\psi^0$ is the left-continuous inverse of the right-continuous function $\widehat{\mathsf{G}}_n^{1,0}(0, c) = \widehat{\mathsf{G}}_n(c)$ in (D.2). Thus,

$$\widehat{\mathsf{G}}_n(y) \le \widehat{\mathsf{G}}_n(\hat{c}_\psi^0) \le \widehat{\mathsf{G}}_n(z) \qquad \Rightarrow \qquad y \le \hat{c}_\psi^0 \le z. \tag{D.13}$$

2. *A lower bound.* Let $x_{\max} = \max_{1\le i\le n} |x_{in}|$. It holds that

$$\mathcal{S}_i = [-\sigma\hat{c}_\psi^b + x_{in}'b, \sigma\hat{c}_\psi^b + x_{in}'b] \subset [-\sigma\hat{c}_\psi^b - x_{\max}|b|, \sigma\hat{c}_\psi^b + x_{\max}|b|] = \mathcal{S},$$

so that for all $0 \le \psi \le 1$ then, with $z = \hat{c}_\psi^b + \sigma^{-1}x_{\max}|b|$ it holds

$$\widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) \le \frac{1}{n}\textstyle\sum_{i=1}^n 1_{(|\varepsilon_i|\le\sigma z)} = \widehat{\mathsf{G}}_n^{1,0}(0, z) = \widehat{\mathsf{G}}_n(z).$$

Since, $\widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) = n^{-1}\text{int}(\psi n)$ for all b, ψ then

$$0 = \widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) - \widehat{\mathsf{G}}_n^{1,0}(0, \hat{c}_\psi^0) \leq \widehat{\mathsf{G}}_n(z) - \widehat{\mathsf{G}}_n(\hat{c}_\psi^0),$$

which implies that $\sigma z = \sigma \hat{c}_\psi^b + x_{\max}|b| \geq \sigma \hat{c}_\psi^0$ by inequality (D.13).

3. *An upper bound for $\psi < 1$.* It holds, for $y = \hat{c}_\psi^b - \sigma^{-1} 2 x_{\max}|b|$ that

$$\mathcal{S}_i = [-\sigma \hat{c}_\psi^b + x_{in}' b, \sigma \hat{c}_\psi^b + x_{in}' b] \supset [-\sigma y, \sigma y] = \mathcal{S},$$

noting that the smaller set is empty if $y < 0$. It will therefore hold that

$$\widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) \geq \frac{1}{n} \sum_{i=1}^n 1_{(|\varepsilon_i| \leq \sigma y)} = \widehat{\mathsf{G}}_n(y).$$

Actually, this inequality must be strict. Indeed, at least one $i^\dagger$ exists so that $\sigma \hat{c}_\psi^b = |\varepsilon_{i^\dagger} - x_{i^\dagger n}' b|$. For this (these) $i^\dagger$ it holds that $\varepsilon_{i^\dagger} \in \mathcal{S}_i$ but $\varepsilon_{i^\dagger} \notin \mathcal{S}$. Thus, it holds $\widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) > \widehat{\mathsf{G}}_n(y)$. Proceed as before to see that

$$0 = \widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) - \widehat{\mathsf{G}}_n^{1,0}(0, \hat{c}_\psi^0) > \widehat{\mathsf{G}}_n(y) - \widehat{\mathsf{G}}_n(\hat{c}_\psi^0), \tag{D.14}$$

which implies that $y = \hat{c}_\psi^b - \sigma^{-1} 2 x_{\max}|b| < \hat{c}_\psi^0$ by inequality (D.13). ∎

The next result introduces a convergence rate for $\hat{c}_\psi^b - \hat{c}_\psi^0$.

Lemma D.6 *Suppose Assumptions 3.1(ia, iib, iic) holds. Then, for all $\omega < \eta - \kappa$,*

$$\sup_{0 \leq \psi \leq 1} \sup_{|b| \leq n^{1/4-\eta} B} n^{1/2} |\mathsf{f}(\hat{c}_\psi^0)(\hat{c}_\psi^b - \hat{c}_\psi^0)| = \mathrm{o}_{\mathsf{P}}(n^{-\omega}).$$

Proof of Lemma D.6. By definition $\widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^b) = \widehat{\mathsf{G}}_n^{1,0}(0, \hat{c}_\psi^0) = n^{-1}\text{int}(n\psi)$. Combine the inequality of Lemma D.5 with Assumption 3.1 (iib) showing $\max_{1 \leq i \leq n} |x_{in}| = \mathrm{O}_{\mathsf{P}}(n^{\kappa - 1/2})$ to get that $\hat{c}_\psi^b - \hat{c}_\psi^0 = \mathrm{O}_{\mathsf{P}}(n^{-1/4+\kappa-\eta})$ for $|b| \leq n^{1/4-\eta} B$. Thus, for any $\epsilon > 0$ a $C > 0$ exists so that the set $\mathcal{C}_n = \{|n^{1/2-\kappa}(\hat{c}_\psi^b - \hat{c}_\psi^0)| \leq n^{1/4-\eta} C\}$ has probability $\mathsf{P}(\mathcal{C}_n) > 1 - \epsilon$. On this set it holds, with $d = n^{1/2-\kappa}(\hat{c}_\psi^b - \hat{c}_\psi^0)$, that

$$0 = \widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^0 + n^{\kappa - 1/2} d) - \widehat{\mathsf{G}}_n^{1,0}(0, \hat{c}_\psi^0).$$

Lemma D.4(a) using Assumption 3.1(ia, iic) shows that

$$n^{1/2}\{\overline{\mathsf{G}}_n^{1,0}(b, c_\psi + n^{\kappa-1/2} d) - \overline{\mathsf{G}}_n^{1,0}(0, c_\psi)\} - 2\sigma^{-1} \mathsf{f}(c_\psi) n^\kappa d = \mathrm{O}_{\mathsf{P}}(n^{2\kappa - 2\eta}) = \mathrm{o}_{\mathsf{P}}(n^{-\omega}),$$

uniformly in $0 \leq \psi \leq 1$ and $|b|, |d| \leq n^{1/4+\kappa-\eta} B$, for all $\omega < \eta - \kappa < 2(\eta - \kappa)$. Lemma D.4($b'$) using Assumption 3.1($ia, iib, iic$) shows that, uniformly in $0 \leq \psi \leq 1$ and $|b|, |d| \leq n^{1/4-\eta} B$,

$$\mathbb{G}_n^{1,0}(b, c_\psi + n^{\kappa-1/2} d) - \mathbb{G}_n^{1,0}(0, c_\psi) = \mathrm{o}_{\mathsf{P}}(n^{-\omega}),$$

for all $\omega < \eta - \kappa$. Using the definition $\mathbb{G}_n^{1,0} = n^{1/2}(\widehat{\mathsf{G}}_n^{1,0} - \overline{\mathsf{G}}_n^{1,0})$ then

$$0 = n^{1/2}\{\widehat{\mathsf{G}}_n^{1,0}(b, \hat{c}_\psi^0 + n^{\kappa-1/2} d) - \widehat{\mathsf{G}}_n^{1,0}(0, \hat{c}_\psi^0)\} = 2\sigma^{-1} \mathsf{f}(\hat{c}_\psi^0) n^\kappa d + \mathrm{o}_{\mathsf{P}}(n^{-\omega}).$$

Inserting$d = n^{1/2-\kappa}(\hat{c}_\psi^b - \hat{c}_\psi^0)$ we get the desired result. ∎

The next result provides a modification of Csörgő (1983, equation 2.8).

Lemma D.7 *Let $c_\psi = \mathsf{G}^{-1}(\psi)$. Suppose* f *is symmetric and decreasing for large c and that Assumption 3.1(ib) holds. Then, for all ψ^* so $|\psi^* - \psi| \leq |\mathsf{G}(\hat{c}^0_\psi) - \psi|$, it holds*
(a) $\sup_{0\leq\psi\leq c_n} |1 - \mathsf{f}(c_\psi)/\mathsf{f}(c_{\psi^*})| = \mathrm{o}_\mathsf{P}(1)$, *for any sequence $c_n \to 0$ so $nc_n \to \infty$;*
(b) $\sup_{0\leq\psi\leq n/(n+1)} |1 - \mathsf{f}(c_\psi)/\mathsf{f}(c_{\psi^*})| = \mathrm{O}_\mathsf{P}(1)$.

Proof of Lemma D.7. (a) By (2.2) then $\mathsf{G}^{-1}(\psi) = \mathsf{F}^{-1}(y)$ for $y = (1+\psi)/2$ varying in $1/2 \leq y \leq 1-(2n+2)^{-1}$. Let $\gamma = \sup_{c\in\mathbb{R}} \mathsf{F}(c)\{1-\mathsf{F}(c)\}|\dot{\mathsf{f}}(c)|/\{\mathsf{f}(c)\}^2$ which is finite by Assumption 3.1(ib). It is first argued that for all $\epsilon > 0$ and $0 < c < 1$ and all n then

$$\mathsf{P}\{\sup_{1/2+c\leq y\leq 1-c} |\frac{\mathsf{f}\{\mathsf{F}^{-1}(y)\}}{\mathsf{f}\{\mathsf{F}^{-1}(y^*)\}} - 1| > \epsilon\} \leq 4\{1+\mathrm{int}(\gamma)\}\{\exp(-nch_1) + \exp(-nch_2)\}, \quad \text{(D.15)}$$

where, with $h(\lambda) = \lambda + \log(1/\lambda) - 1$ then $h_1 = h[(1+\epsilon)^{\{1+\mathrm{int}(\gamma)\}/2}]$ and $h_2 = h[1/(1+\epsilon)^{\{1+\mathrm{int}(\gamma)\}/2}]$. This is nearly the statement of Theorem 1.5.1 of Csörgő (1983), which, however, has the denominator $\mathsf{f}(\theta_{y,n})$ instead of $\mathsf{f}\{\widehat{\mathsf{F}}_n^{-1}(y^*)\}$ where $\theta_{y,n}$ is a particular intermediate point between $\widehat{\mathsf{F}}_n^{-1}(y)$ and $\mathsf{F}^{-1}(y)$ rather than any intermediate point. Csörgő states that the proof of this Theorem is similar to that of his Theorem 1.4.3. Equation (1.4.18.2) of that proof uses a bound only depending on $\widehat{\mathsf{F}}_n^{-1}(y)$ and $\mathsf{F}^{-1}(y)$ and not on the particular intermediate point $\theta_{y,n}$. This proves (D.15).

The inequality (D.15) implies that for any sequence $c_n \to 0$ so $nc_n \to \infty$ then

$$\mathsf{P}\{\sup_{1/2+c_n\leq y\leq 1-c_n} |\frac{\mathsf{f}\{\mathsf{F}^{-1}(y)\}}{\mathsf{f}\{\widehat{\mathsf{F}}_n^{-1}(y^*)\}} - 1| > \epsilon\} \to 0.$$

The reason is that $h(\lambda) > 0$ for all $\lambda > 0$ so $\lambda \neq 1$. Consider the tails.

Left hand tail. Use that c_n vanishes, that $\mathsf{G}(\hat{c}^0_\psi) - \psi = \mathrm{O}_\mathsf{P}(n^{-1/2})$ by Theorem 3.6, and that f is uniformly continuous in a neighbourhood of zero because f is bounded, positive and continuous.

(b) *Right hand tail.* It suffices to argue that

$$\lim_{\epsilon\to\infty}\limsup_{n\to\infty}\mathsf{P}\{\sup_{1-c_n\leq y\leq 1-(2n+2)^{-1}} |\frac{\mathsf{f}\{\mathsf{F}^{-1}(y)\}}{\mathsf{f}\{\widehat{\mathsf{F}}_n^{-1}(y^*)\}} - 1| > \epsilon\} = 0. \quad \text{(D.16)}$$

Apply the inequality (D.15) with $c = (2n+2)^{-1}$ so that $nc \sim 1/2$. Then use that $h_1, h_2 \to \infty$ for $\epsilon \to \infty$ since $h(\lambda) \to \infty$ for $\lambda \to \infty$. ■

The next result relates $\hat{c}^0_\psi$ to c_ψ.

Lemma D.8 *Suppose Assumptions 3.1(ia, ib) holds with $q = 1$ only. Then*

$$\sup_{0\leq\psi\leq 1} |(\hat{c}^0_\psi)^k \mathsf{f}(\hat{c}^0_\psi) - (c_\psi)^k \mathsf{f}(c_\psi)| = \mathrm{o}_\mathsf{P}(1) \qquad \textit{for } k = 0, 1.$$

Proof of Lemma D.8. 1. *Consider ψ so $0 \leq \psi \leq 1 - 1/z_n$ for any sequence $0 < z_n <$* $\mathrm{o}(n^{1/2})$. Rewrite the process of interest as

$$(\hat{c}^0_\psi)^k \mathsf{f}(\hat{c}^0_\psi) - (c_\psi)^k \mathsf{f}(c_\psi) = \{(\hat{c}^0_\psi)^k - (c_\psi)^k\}\mathsf{f}(c_\psi) + (\hat{c}^0_\psi)^k \mathsf{f}(\hat{c}^0_\psi)\{1 - \frac{\mathsf{f}(c_\psi)}{\mathsf{f}(\hat{c}^0_\psi)}\}. \quad \text{(D.17)}$$

The first term is zero for $k = 0$ and $\mathrm{O}_\mathsf{P}(n^{-1/2})$ for $k = 1$ due to Lemmas 3.6, D.1(a) using Assumption 3.1(ib). For the second term, note that $(\hat{c}^0_\psi)^k \mathsf{f}(\hat{c}^0_\psi)$ is bounded uniformly in $0 \leq \psi \leq 1$ due to Assumption 3.1(ia) with $q = 1$, while $1 - \mathsf{f}(c_\psi)/\mathsf{f}(\hat{c}^0_\psi)$ vanishes by Lemma D.7(a) using Assumption 3.1(ib).

2. *Consider ψ so $\psi_n \leq \psi \leq 1$ for any sequence $\psi_n \to 1$.* Assumption 3.1(ia) and the continuity of f implies that $(c_\psi)^k \mathsf{f}(c_\psi)$ is continuous and convergent for $\psi \to 1$. Rewrite

$$\hat{c}^0_{\psi_n} = \mathsf{G}^{-1}\{\mathsf{G}(\hat{c}^0_{\psi_n})\} = \mathsf{G}^{-1}[\psi_n + \{\mathsf{G}(\hat{c}^0_{\psi_n}) - \psi_n\}] \geq \mathsf{G}^{-1}(\psi_n - g_n),$$

where $g_n = \sup_{0 \leq \psi \leq 1}\{\mathsf{G}(\hat{c}^0_\psi) - \psi\} = \mathrm{O}_\mathsf{P}(n^{-1/2})$ due to Lemmas 3.6, D.1(c) using Assumption 3.1(ib). By the continuity of G^{-1} then $\hat{c}^0_{\psi_n} \to \mathsf{G}^{-1}(1)$ in probability and therefore $(\hat{c}^0_\psi)^k \mathsf{f}(\hat{c}^0_\psi) - (c_\psi)^k \mathsf{f}(c_\psi)$ vanishes in probability. ∎

D.6 A one-step result for the least squares estimator

A one-step result for the least squares estimator now follows. Equation (D.6) represents the one-step least squares estimator $\hat{\beta}^{(m+1)}$ in terms of $\widehat{\mathsf{G}}^{g,p}_n$. That expression has the random quantities $\hat{b}^{(m)}$ and $\sigma^{-1}\hat{z}^{(m)}$ as arguments. Replacing these by a deterministic quantity b and the residual $\hat{c}^b_\psi$ defined in (D.12) gives the following asymptotic uniform linearization result for the one-step least squares estimator if we insert the initial estimator $b = \hat{b}^{(m)}$.

Lemma D.9 *Let $c_\psi = \mathsf{G}^{-1}(\psi)$ and*

$$\rho_\psi = 2c_\psi \mathsf{f}(c_\psi)/\psi \qquad \text{(D.18)}$$

Suppose Assumption 3.1($ia - ib, ii$) hold for some $0 \leq \kappa < \eta \leq 1/4$. Then, for all $\psi_0 > 0$ it holds
(a) $\sup_{0\leq\psi\leq1} \sup_{|b|\leq n^{1/4-\eta}B} |n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, \hat{c}^b_\psi) - \mathbb{G}^{x,1}_n(0, c_\psi) - 2c_\psi \mathsf{f}(c_\psi)\Sigma_n b| = \mathrm{o}_\mathsf{P}(1)$;
(b) $\sup_{0\leq\psi\leq1} \sup_{|b|\leq n^{1/4-\eta}B} |\widehat{\mathsf{G}}^{xx,0}_n(b, \hat{c}^b_\psi) - \Sigma_n\psi| = \mathrm{O}_\mathsf{P}(n^{\kappa-\eta-1/4})$;
(c) $\sup_{\psi_0\leq\psi\leq1} \sup_{|b|\leq n^{1/4-\eta}B} |\{\widehat{\mathsf{G}}^{xx,0}_n(b, \hat{c}^b_\psi)\}^{-1} n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, \hat{c}^b_\psi) - (\psi\Sigma_n)^{-1}\mathbb{G}^{x,1}_n(0, c_\psi) - \rho_\psi b| = \mathrm{o}_\mathsf{P}(1)$.

Proof of Lemma D.9. (a) The inequality of Lemma D.5 implies that $\hat{c}^b_\psi - \hat{c}^0_\psi = \mathrm{O}_\mathsf{P}(n^{\kappa-\eta-1/4})$ uniformly in $0 \leq \psi \leq 1$ and $|b| \leq n^{1/4-\eta}B$ since $\max_{1\leq i\leq n} |x_{in}| = \mathrm{O}_\mathsf{P}(n^{\kappa-1/2})$ by Assumption 3.1 (iib). Start by expanding $\widehat{\mathsf{G}}^{x,1}_n$. By definition

$$n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, c_\psi + n^{\kappa-1/2}d) = \mathbb{G}^{x,1}_n(b, c_\psi + n^{\kappa-1/2}d) + n^{1/2}\overline{\mathsf{G}}^{x,1}_n(b, c_\psi + n^{\kappa-1/2}d).$$

Lemma D.4(a, b), using Assumption 3.1(ia, iib, iic) along with the definitions $g_{in} = n^{1/2}x_{in}$ and $\Sigma_n = \sum_{i=1}^n x_{in}x'_{in}$ gives, uniformly in $|b|, |d| \leq n^{1/4-\eta}B$ and $0 \leq \psi \leq 1$,

$$n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, c_\psi + n^{\kappa-1/2}d) = \mathbb{G}^{x,1}_n(0, c_\psi) + n^{1/2}\overline{\mathsf{G}}^{x,1}_n(0, c_\psi) + E_n + \mathrm{o}_\mathsf{P}(1).$$

Note that $\overline{\mathsf{G}}^{x,1}_n(0, c_\psi) = 0$ due to the symmetry of f. Replace c_ψ by $\hat{c}^0_\psi$ and d by $n^{1/2-\kappa}(\hat{c}^b_\psi - \hat{c}^0_\psi)$, which is $\mathrm{O}_\mathsf{P}(n^{1/4-\eta})$. Thus it holds

$$n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, \hat{c}^b_\psi) = \mathbb{G}^{x,1}_n(0, \hat{c}^0_\psi) + 2\hat{c}^0_\psi \mathsf{f}(\hat{c}^0_\psi)\Sigma_n b + \mathrm{o}_\mathsf{P}(1), \qquad \text{(D.19)}$$

uniformly in $|b| \leq n^{1/4-\eta}B$ and $0 \leq \psi \leq 1$. The two terms are analysed in turn.

First term. Theorem 3.6 shows $a_\psi = n^{1/2}\{\mathsf{G}(\hat{c}^0_\psi) - \psi\}$ is tight. Expand

$$\hat{c}^0_\psi = \mathsf{G}^{-1}\{\mathsf{G}(\hat{c}^0_\psi)\} = c_{\mathsf{G}(\hat{c}^0_\psi)} = c_{\psi + n^{-1/2}a_\psi}. \tag{D.20}$$

Lemma D.4(c) using Assumption 3.1(ia, iib, iic) shows $\mathbb{G}^{x,1}_n(0, \hat{c}^0_\psi) = \mathbb{G}^{x,1}_n(0, c_\psi) + \mathrm{o}_\mathsf{P}(1)$.

Second term. Use that $\hat{c}^0_\psi \mathsf{f}(\hat{c}^0_\psi) = c_\psi \mathsf{f}(c_\psi) + \mathrm{o}_\mathsf{P}(1)$ uniformly in ψ by Lemma D.8 using Assumptions 3.1(ia, ib).

(b) An expansion as in (D.19) gives

$$\widehat{\mathsf{G}}^{xx,0}_n(b, \hat{c}^b_\psi) = n^{-1/2}\mathbb{G}^{xx,0}_n(0, \hat{c}^0_\psi) + \overline{\mathsf{G}}^{xx,0}_n(0, \hat{c}^0_\psi) + 2\sigma^{-1}\mathsf{f}(\hat{c}^0_\psi)\Sigma_n(\hat{c}^b_\psi - \hat{c}^0_\psi) + \mathrm{o}_\mathsf{P}(n^{-1/2}),$$

uniformly in b, ψ. The three terms are analysed in turn.

First term. This is $n^{-1/2}\mathbb{G}^{xx,0}_n(0, \hat{c}^0_\psi) = n^{-1/2}\mathbb{G}^{xx,0}_n(0, c_\psi) + \mathrm{o}_\mathsf{P}(n^{-1/2})$ by an argument as for the first term of (D.19).

Second term. Use the definition of Σ_n and Theorem 3.6, D.1(c) using Assumption 3.1(ib) showing $\mathsf{G}(\hat{c}^0_\psi) = \psi + \mathrm{O}_\mathsf{P}(n^{-1/2})$ uniformly in ψ along with the tightness of Σ_n by Assumption 3.1(iia) to see that

$$\overline{\mathsf{G}}^{xx,0}_n(0, \hat{c}^0_\psi) = \frac{1}{n}\sum_{i=1}^n n x_{in}x'_{in}\mathsf{E}_{i-1}1_{(|\varepsilon_i| \le \sigma\hat{c}^0_\psi)} = \sum_{i=1}^n x_{in}x'_{in}\mathsf{G}(\hat{c}^0_\psi) = \Sigma_n\psi + \mathrm{O}_\mathsf{P}(n^{-1/2}).$$

Third term. This is $\mathrm{O}_\mathsf{P}(n^{\kappa-\eta-1/4})$ since $\mathsf{f}(\hat{c}^0_\psi) = \mathsf{f}(c_\psi) + \mathrm{o}_\mathsf{P}(1)$ uniformly in $0 \le \psi \le 1$ by Lemma D.8 using Assumptions 3.1(ia, ib), while $\hat{c}^b_\psi - \hat{c}^0_\psi = \mathrm{O}(n^{\kappa-\eta-1/4})$ and Σ_n is tight by Assumption 3.1(iia).

(c) Combine (a), (b). The denominator from (b) satisfies

$$\widehat{\mathsf{G}}^{xx,0}_n(b, \hat{c}^b_\psi) = \psi\Sigma_n\{1 + \mathrm{o}_\mathsf{P}(1)\},$$

for $\psi \ge \psi_0 > 0$ and since $\Sigma_n \to \Sigma$ in distribution where $\Sigma > 0$ *a.s.* by Assumption 3.1(iia). Combine with the expression for the numerator in (a). ■

D.7 The forward plot of least squares estimators

The Forward Plot of least squares estimators is now considered. The one-step result in Lemma D.9 implies that the Forward Search iteration can be viewed as a fixed point problem. Indeed, the one-step result in Lemma D.9 implies an autoregressive relation between the one-step updated estimation error $\hat{b}^{(m+1)}$ and the previous estimation error $\hat{b}^{(m)}$. It holds that

$$\hat{b}^{(m+1)} = \rho_\psi\hat{b}^{(m)} + (\psi\Sigma_n)^{-1}\mathbb{G}^{x,1}_n(0, c_\psi) + e_\psi(\hat{b}^{(m)}), \tag{D.21}$$

for $\psi = m/n + \mathrm{o}(1)$, an "autoregressive coefficient" ρ_ψ defined in (D.18) and a vanishing remainder term e_ψ. This autoregressive representation generalizes Theorem 5.2 of Johansen and Nielsen (2010) which was concerned with a location-scale model, a fixed $\psi \sim m/n$, and convergent initial estimators, $\hat{b}^{(m)} = \mathrm{O}(1)$.

It is first established that ρ_ψ has nice properties for unimodal densities f.

Lemma D.10 *Suppose Assumption 3.1(ic) holds. Then $0 < \rho_\psi < 1$ for $0 < \psi < 1$ while $\lim_{\psi\to 0}\rho_\psi = 1$ and $\lim_{\psi\to 1}\rho_\psi = 0$.*

Proof of Lemma D.10. For $c > 0$ then $\mathsf{f}(x)1_{(|x|\leq c)} \geq \mathsf{f}(c)1_{(|x|\leq c)}$ because f is symmetric and non-increasing by Assumption 3.1(ic). Integration gives

$$\psi = 2\int_0^{c_\psi} \mathsf{f}(x)dx \geq 2c_\psi \mathsf{f}(c_\psi) = \rho_\psi \psi,$$

where equality holds for $\mathsf{f}(x) = \mathsf{f}(c)$ for $|x| \leq c$, by continuity of f. This is, however, ruled out by assuming $\lim_{c\to 0} \dot{\mathsf{f}}(c) < 0$. It holds $\lim_{c\to 0}(2c)^{-1}2\int_0^c \mathsf{f}(x)dx = \mathsf{f}(0)$ while $\rho_\psi\psi/(2c_\psi) = \mathsf{f}(c_\psi)$ so $\lim_{\psi\to 0}\rho_\psi = 1$. Similarly, $2\int_0^\infty \mathsf{f}(x)dx = 1$ and $\lim_{\psi\to 1} c\mathsf{f}(c) \to 0$ so $\lim_{\psi\to 1}\rho_\psi = 0$. ∎

Lemma D.11 *Suppose Assumption 3.1(id) holds. Then ρ_ψ is strictly decreasing.*

Proof of Lemma D.11. Let $\tau_k = 2\int_0^c x^k \mathsf{f}(x)dx$ for $k \in \mathbb{N}_0$. It holds $\lim_{c\to 0}\tau_k = 0$ and $\tau_k > 0$ for $c > 0$. The derivatives with respect to c are

$$\dot{\tau}_k = 2c^k \mathsf{f}, \qquad \ddot{\tau}_k = \dot{\tau}_{k-1}(k + c\frac{\dot{\mathsf{f}}}{\mathsf{f}}).$$

Consider the ratio $R_k = \dot{\tau}_{k+1}/\tau_k$, noting that $R_0 = \xi_1^\psi/\psi$. l'Hôpital's rule gives

$$\lim_{c\to 0} R_k = \lim_{c\to 0}\frac{\ddot{\tau}_{k+1}}{\dot{\tau}_k} = k+1.$$

Moreover, R_k has derivative

$$\dot{R}_k = \frac{\ddot{\tau}_{k+1}\tau_k - \dot{\tau}_{k+1}\dot{\tau}_k}{\tau_k^2} = \frac{\dot{\tau}_k}{\tau_k^2}M_k.$$

where $M_k = \{k + 1 + c\dot{\mathsf{f}}/\mathsf{f}\}\tau_k - \dot{\tau}_{k+1}$. It has to be argued that $\dot{R}_k < 0$ for $c > 0$. Since $\dot{\tau}_k, \tau_k > 0$ then $\dot{R}_k < 0$ if and only if $M_k < 0$. Now, $\lim_{c\to 0} M_k = 0$ so a sufficient condition is that $\dot{M}_k < 0$. But

$$\dot{M}_k = \tau_k\{\frac{d}{dc}\log \mathsf{f}(c) + c\frac{d^2}{dc^2}\log \mathsf{f}(c)\} = \tau_k\frac{d}{dc}(c\frac{d}{dc}\log \mathsf{f}(c))$$

which is negative if and only if $\Delta(c) = \frac{d}{dc}(c\frac{d}{dc}\log \mathsf{f}(c)) < 0$. ∎

The next result investigates the forward estimator $\hat{\beta}^{(m+1)}$. There are two results: first, the forward search preserves the order of the initial estimator, and, secondly, by infinite iteration a slowly converging initial estimator can be improved to consistency at a standard rate. The proof of this result is related to that of Johansen and Nielsen (2011, Theorem 3.3).

Lemma D.12 *Suppose Assumption 3.1$(ia - id, ii, iii)$ holds. Then, for all $\psi_1 > \psi_0 > 0$ so $m_0/n = \psi_0 + \mathrm{o}(1)$, it holds*
(a) $\sup_{\psi_0\leq\psi\leq 1}|N^{-1}(\hat{\beta}_\psi - \beta)| = \mathrm{O}_\mathsf{P}(n^{1/4-\eta})$;
(b) $\sup_{\psi_1\leq\psi\leq 1}|N^{-1}(\hat{\beta}_\psi - \beta)| = \mathrm{O}_\mathsf{P}(1)$.

Proof of Lemma D.12. Due to the embedding (3.1) it suffices to evaluate $N^{-1}(\hat\beta_\psi - \beta)$ at the grid points $\psi = m/n$. Introduce notation $K_\psi^n = \Sigma_n^{-1}\mathbb{G}_n^{x,1}(0, c_\psi)$.

(a) Solve the autoregressive equation (D.21) recursively to get

$$\hat b^{(m+1)} = \sum_{k=m_0}^m (\prod_{\ell=k}^{m-1} \rho_{\ell/n})\{\frac{n}{k} K_{k/n}^n + e_{k/n}(\hat b^{(k)})\} + (\prod_{k=m_0}^m \rho_{k/n})\hat b^{(m_0)}.$$

with the convention that an empty product equals unity. Lemmas D.10, D.11 using Assumption 3.1(ic, id) show that the coefficient ρ_ψ is strictly decreasing and less than unity. For $m \geq \psi_0 n$ then $\rho_{m/n} \leq \rho_{m_0/n} = \rho_0$ for some $\rho_0 < 1$ giving the bound

$$|\hat b^{(m+1)}| \leq (\sum_{k=m_0}^m \rho_0^{m-k})\{\sup_{\psi_0\leq\psi\leq 1} |\psi^{-1}K_\psi^n| + \max_{m_0\leq k<m} |e_{k/n}(\hat b^{(k)})|\} + \rho_0^{m-m_0+1}|\hat b^{(m_0)}|. \quad \text{(D.22)}$$

For $\psi \geq \psi_0 > 0$ then $\psi^{-1}K_\psi^n$ is tight by Lemma D.4(c) using Assumption 3.1(ia, iib, iic). The bound $\sum_{k=m_0}^m \rho_0^{m-k} \leq \sum_{k=0}^\infty \rho_0^k = C$ is finite, while $\hat b^{(m_0)} = \mathrm{O}(n^{1/4-\eta})$ by Assumption 3.1(iii). Moreover, $\sup_{\psi_0\leq\psi\leq 1}\sup_{|b|\leq 3n^{1/4-\eta}B} |e_\psi(b)| = \mathrm{o}_\mathsf{P}(1)$ for any $B > 0$ by Lemma D.9 using Assumption 3.1(ia, ib, ii). Thus, for all $\epsilon, \zeta > 0$ constants $B, n_0 > 0$ exist so that for $n \geq n_0$, the set

$$\mathcal{A}_n = (C|\hat b^{(m_0)}| \leq n^{1/4-\eta}B) \cap (C \sup_{0\leq\psi\leq 1} |K_\psi^n| \leq B) \cap (C \sup_{\psi_0\leq\psi\leq 1} \sup_{|b|\leq 3n^{1/4-\eta}B} |e_\psi(b)| \leq \zeta/2)$$

has probability larger than $1 - \epsilon$. An induction over m is now used to prove that

$$\max_{m_0\leq k\leq m} |\hat b^{(k)}| \leq 3n^{1/4-\eta}B \qquad \text{for } m = m_0, \ldots, n,$$

on the set $\mathcal{A}_n$, which implies the desired result. As induction start, for $m + 1 = m_0$, then $|\hat b^{(m_0)}| \leq n^{1/4-\eta}B$ on the set $\mathcal{A}_n$. Suppose the result holds for some m. This implies that

$$C \sup_{\psi_0\leq\psi\leq 1} \max_{m_0\leq k<m} |e_\psi(\hat b^{(k)})|\} \leq \zeta/2 \quad \text{(D.23)}$$

on the set $\mathcal{A}_n$. Thus, the bound (D.22) becomes $|\hat b^{(m)}| \leq 2n^{1/4-\eta}B + \zeta/2 \leq 3n^{1/4-\eta}B$. Thus, the result holds for $m + 1$.

(b) Consider (D.22). Here $\sum_{k=0}^n \rho_0^k$ is finite, $\sup_{\psi_0\leq\psi\leq 1} |\psi^{-1}K_\psi^n| = \mathrm{O}_\mathsf{P}(1)$ due to tightness and $\sup_{0\leq\psi\leq 1} \max_{m_0\leq k<n} |e_\psi(\hat b^{(k)})|\} = \mathrm{o}_\mathsf{P}(1)$ due to (D.23). Let $m \geq \psi_1 n$ and $m_0 = \psi_0 n + \mathrm{O}(1)$ for some $\psi_1 > \psi_0 > 0$. Since $\rho_0^{m-m_0}$ declines exponentially then $\rho_0^{m-m_0} < n^{-1/4}$ for large n so that $\max_{m\geq m_1} \rho_0^{m-m_0}|\hat b^{(m_0)}| = \mathrm{o}_\mathsf{P}(1)$. ■

D.8 Proof of Theorem 3.1

Lemmas D.1, D.5 are now combined to show that the forward residuals scaled with a known variance, $\sigma^{-1}\hat z_\psi$, have the same Bahadur representation as the quantile process for the innovations $\sigma^{-1}\varepsilon_i$. This is the main Theorem stated with slightly weaker conditions.

Lemma D.13 *Suppose Assumption 3.1($ia - id, ii, iii$) holds. Let $\psi_0 > 0$. Then*

$$\sup_{\psi_0\leq\psi\leq n/(n+1)} |2\mathsf{f}(c_\psi)n^{1/2}(\sigma^{-1}\hat z_\psi - c_\psi) + \mathbb{G}_n^{1,0}(c_\psi)| = \mathrm{o}_\mathsf{P}(1).$$

Proof of Lemma D.13. Due to the embedding (3.1) it suffices to evaluate the forward residuals at the grid points $\psi = m/n$. It is first argued that the forward plot of the estimators is bounded in the sense that for all $\epsilon > 0$ a $B > 0$ exists so that the set $\mathcal{C}_n = (\sup_{\psi_0 \le \psi \le 1} |N^{-1}(\hat{\beta}_\psi - \beta)| \le n^{1/4-\eta}B)$ has $\mathsf{P}(\mathcal{C}_n) \ge 1 - \epsilon$. This follows from Lemma D.12 using Assumption 3.1$(ia - id, ii, iii)$. Now, on $\mathcal{C}_n$ it holds that $\sigma^{-1}\hat{z}_\psi = \hat{c}^b_\psi$, see (D.4), for some $|b| \le n^{1/4-\eta}B$. Thus it suffices to show that

$$\sup_{\psi_0 \le \psi \le n/(n+1)} \sup_{|b| \le n^{1/4-\eta}B} |\mathbb{C}^b_\psi| = \mathrm{o}_\mathsf{P}(1) \qquad \text{for} \qquad \mathbb{C}^b_\psi = 2\mathsf{f}(c_\psi)n^{1/2}(\hat{c}^b_\psi - c_\psi) + \mathbb{G}^{1,0}_n(c_\psi).$$

Now, write $(\hat{c}^b_\psi - c_\psi) = (\hat{c}^0_\psi - c_\psi) + (\hat{c}^b_\psi - \hat{c}^0_\psi)$, so that

$$\mathbb{C}^b_\psi = \{2\mathsf{f}(c_\psi)n^{1/2}(\hat{c}^0_\psi - c_\psi) + \mathbb{G}^{1,0}_n(c_\psi)\} + 2\frac{\mathsf{f}(c_\psi)}{\mathsf{f}(\hat{c}^0_\psi)}n^{1/2}\mathsf{f}(\hat{c}^0_\psi)(\hat{c}^b_\psi - \hat{c}^0_\psi).$$

The first term is $\mathrm{o}_\mathsf{P}(n^{\zeta-1/4})$ for all $\zeta > 0$ uniformly in $0 \le \psi \le 1$ by Theorem D.1(a) using Assumption 3.1(ib). In the second term the ratio $\mathsf{f}(c_\psi)/\mathsf{f}(\hat{c}^0_\psi)$ is $\mathrm{O}_\mathsf{P}(1)$ uniformly in $0 \le \psi \le n/(n+1)$ by Lemma D.7 using Assumption 3.1(ia, ib), while $n^{1/2}\mathsf{f}(\hat{c}^0_\psi)(\hat{c}^b_\psi - \hat{c}^0_\psi) =\mathrm{o}_\mathsf{P}(1)$ uniformly in $0 \le \psi \le 1$ by Lemma D.6 using Assumption 3.1(ia, iib, iic) ■

D.9 Proofs of Theorems 3.2 and 3.3

The above theory for $\sigma^{-1}\hat{z}_\psi$ involves the population variance σ^2. The next results gives an asymptotic expansion for $\hat{\sigma}^2_{\psi,cor}$, recalling, from (2.7) that

$$\hat{\sigma}^2_{\psi,cor} - \sigma^2 = \frac{1}{\tau_\psi}[\widehat{\mathsf{G}}^{1,2}_n(\hat{b}, \hat{c}^{\hat{b}}_\psi) - \{\widehat{\mathsf{G}}^{x,1}_n(\hat{b}, \hat{c}^{\hat{b}}_\psi)\}'\{\widehat{\mathsf{G}}^{xx,0}_n(\hat{b}, \hat{c}^{\hat{b}}_\psi)\}^{-1}\{\widehat{\mathsf{G}}^{x,1}_n(\hat{b}, \hat{c}^{\hat{b}}_\psi)\}].$$

Compare also the definitions in (3.2), (3.3) with (D.11) to see

$$\mathbb{G}_n(c_\psi) = \mathbb{G}^{1,0}_n(0, c_\psi), \qquad \mathbb{L}_n(c_\psi) = \sigma^{-2}\mathbb{G}^{1,2}_n(0, c_\psi) - c^2_\psi\mathbb{G}^{1,0}_n(0, c_\psi). \tag{D.24}$$

The main Theorem 3.2 then follows immediately from the next result.

Lemma D.14 *Suppose Assumption 3.1(ia, ib, ie, ii) holds. Then*

$$\sup_{\psi_0 \le \psi \le n/(n+1)} \sup_{|b| \le n^{1/4-\eta}B} |n^{1/2}(\hat{\sigma}^2_{\psi,cor} - \sigma^2) - \sigma^2\tau^{-1}_\psi\mathbb{L}_n(c_\psi)| = \mathrm{o}_\mathsf{P}(1).$$

Proof of Lemma D.14. 1. *Regression correction term.* Lemma D.9(a, b) using Assumption 3.1(ia, ib, ii) shows that

$$n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, \hat{c}^b_\psi) = \mathbb{G}^{x,1}_n(0, c_\psi) + 2c_\psi\mathsf{f}(c_\psi)\Sigma_n b + \mathrm{o}_\mathsf{P}(1), \tag{D.25}$$

$$\widehat{\mathsf{G}}^{xx,0}_n(b, \hat{c}^b_\psi) = \Sigma_n\psi + \mathrm{o}_\mathsf{P}(1), \tag{D.26}$$

uniformly in $|b| \le n^{1/4-\eta}B$, $\psi_0 \le \psi \le 1$. Evaluate the terms of the expansion for $n^{1/2}\widehat{\mathsf{G}}^{x,1}_n(b, \hat{c}^b_\psi)$. The first term is $\mathbb{G}^{1,0}_n(0, c_\psi) = \mathrm{O}_\mathsf{P}(1)$ since $\mathbb{G}^{1,0}_n$ is tight by Lemma D.4(c) using Assumption 3.1(ia, iib, iic). The second term is $\mathrm{O}_\mathsf{P}(n^{1/4-\eta})$ since b is of that order. Therefore

$\widehat{\mathsf{G}}_n^{x,1}(b,\hat{c}_\psi^b) = \mathrm{O}_\mathsf{P}(n^{-1/4-\eta})$. Note that $\Sigma_n \to \Sigma$ in distribution where $\Sigma > 0$ *a.s.* by Assumption 3.1(iia). Due to Lemma D.12(a) using Assumption 3.1($ia - id, ii, iii$) it follows that

$$\{\widehat{\mathsf{G}}_n^{x,1}(\hat{b},\hat{c}_\psi^{\hat{b}})\}'\{\widehat{\mathsf{G}}_n^{xx,0}(\hat{b},\hat{c}_\psi^{\hat{b}})\}^{-1}\{\widehat{\mathsf{G}}_n^{x,1}(\hat{b},\hat{c}_\psi^{\hat{b}})\} = \mathrm{O}_\mathsf{P}(n^{-1/2-2\eta}).$$

This vanishes, even when scaled by $n^{1/2}$.

2. *The leading term.* It remains to argue that

$$n^{1/2}\{\widehat{\mathsf{G}}_n^{1,2}(b,\hat{c}_\psi^b) - \tau_\psi\sigma^2\} - \sigma^2\{\sigma^{-2}\mathbb{G}_n^{1,2}(0,c_\psi) - c_\psi^2\mathbb{G}_n^{1,0}(c_\psi)\} = \mathrm{o}_\mathsf{P}(1),$$

uniformly in $|b| \le n^{1/4-\eta}B$, $\psi_0 \le \psi \le n/(n+1)$. Lemma D.5 shows that $\sigma|\hat{c}_\psi^b - \hat{c}_\psi^0| < 2|b|\max_{1\le i\le n}|x_{in}|$ for all b,ψ. Since $|b| \le n^{1/4-\eta}B$ while $\max_{1\le i\le n}|x_{in}| = \mathrm{O}_\mathsf{P}(n^{\kappa-1/2})$ by Assumption 3.1 (iib) then $\hat{c}_\psi^b - \hat{c}_\psi^0 = \mathrm{O}_\mathsf{P}(n^{\kappa-\eta-1/4})$ uniformly in $0 \le \psi \le 1$ and $|b| \le n^{1/4-\eta}B$. Thus, we start by expanding $n^{1/2}\widehat{\mathsf{G}}_n^{1,2}(b,\hat{c}_\psi^0 + n^{\kappa-1/2}d)$. By definition

$$n^{1/2}\widehat{\mathsf{G}}_n^{1,2}(b,c_\psi + n^{\kappa-1/2}d) = \mathbb{G}_n^{1,2}(b,c_\psi + n^{\kappa-1/2}d) + n^{1/2}\overline{\mathsf{G}}_n^{1,2}(b,c_\psi + n^{\kappa-1/2}d).$$

Apply Lemma D.4(a,b) using Assumption 3.1(ia, iib, iic) to get

$$n^{1/2}\widehat{\mathsf{G}}_n^{1,2}(b,c_\psi + n^{\kappa-1/2}d) = \mathbb{G}_n^{1,2}(0,c_\psi) + n^{1/2}\overline{\mathsf{G}}_n^{1,2}(0,c_\psi) + 2\sigma(c_\psi)^2\mathsf{f}(c_\psi)n^\kappa d + \mathrm{o}_\mathsf{P}(1),$$

uniformly in $|b|,|d| \le n^{1/4-\eta}B$, $0 \le \psi \le 1$. By definition it holds

$$n^{1/2}\widehat{\mathsf{G}}_n^{1,2}(b,c_\psi + n^{\kappa-1/2}d) = n^{1/2}\widehat{\mathsf{G}}_n^{1,2}(0,c_\psi) + 2\sigma(c_\psi)^2\mathsf{f}(c_\psi)n^\kappa d + \mathrm{o}_\mathsf{P}(1).$$

We can replace c_ψ by $\hat{c}_\psi^0$. Moreover, since $\hat{c}_\psi^b - \hat{c}_\psi^0 = \mathrm{O}_\mathsf{P}(n^{\kappa-\eta-1/4})$ we can replace $n^\kappa d$ by $n^{1/2}(\hat{c}_\psi^b - \hat{c}_\psi^0)$ on a set with large probability. When also subtracting $n^{1/2}\tau_\psi\sigma^2$ on both sides and adding and subtracting $n^{1/2}\tau_{\mathsf{G}(\hat{c}_\psi^0)}\sigma^2$ on the right hand side we get

$$n^{1/2}\{\widehat{\mathsf{G}}_n^{1,2}(b,\hat{c}_\psi^b) - \tau_\psi\sigma^2\} = n^{1/2}\{\widehat{\mathsf{G}}_n^{1,2}(0,\hat{c}_\psi^0) - \tau_{\mathsf{G}(\hat{c}_\psi^0)}\sigma^2\} + 2\sigma(\hat{c}_\psi^0)^2\mathsf{f}(\hat{c}_\psi^0)n^{1/2}(\hat{c}_\psi^b - \hat{c}_\psi^0) + \sigma^2 n^{1/2}\{\tau_{\mathsf{G}(\hat{c}_\psi^0)} - \tau_\psi\} + \mathrm{o}_\mathsf{P}(1), \quad \text{(D.27)}$$

uniformly in $|b| \le n^{1/4-\eta}B$, $0 \le \psi \le 1$. The three terms are analysed in turn.

3. *First term of* (D.27). Since $\overline{\mathsf{G}}_n^{1,2}(0,c) = \sigma^2\tau_{\mathsf{G}(c)}$ the first term equals $\mathbb{G}_n^{1,2}(0,\hat{c}_\psi^0)$. Theorem 3.6 shows that $\hat{c}_\psi^0 = c_{\psi+n^{-1/2}\phi}$ where $\phi = n^{1/2}\{\mathsf{G}(\hat{c}_\psi^0) - \psi\}$ is tight. The tightness of $\mathbb{G}_n^{1,2}$ established in Lemma D.4(c) using Assumption 3.1(ia, iib, iic) then implies that the first term equals $\mathbb{G}_n^{1,2}(0,c_\psi) + \mathrm{o}_\mathsf{P}(1)$ uniformly in $0 \le \psi \le 1$.

4. *The order of* $\hat{c}_\psi^0$ is $\mathrm{o}_\mathsf{P}(n^{1/8})$. The reason is that $\hat{c}_\psi^0 \le \max_{i\le n}|\varepsilon_i|$, that $\mathsf{E}|\varepsilon_i|^q < \infty$ for some $q > 8$ by Assumption 3.1(ia), and that Boole's and Markov's inequalities imply that $\mathsf{P}(\max_i|\varepsilon_i| > Cn^\upsilon) \le \sum_{i=1}^n \mathsf{P}(|\varepsilon_i| > Cn^{1/8}) \le n(Cn^{1/8})^{-q}\mathsf{E}|\varepsilon_i|^q$ vanishes.

5. *The order of* c_ψ^2 is $\mathrm{o}\{(1-\psi)^{-1/4}\}$. The reason is that $\mathsf{E}|\varepsilon_i|^q < \infty$ for some $q > 8$ by Assumption 3.1(ia) and that $1 - \mathsf{F}(c_\psi) = \mathsf{P}(|\varepsilon_i| > \sigma\psi_\psi)$ is bounded by $c_\psi^{-q}\mathsf{E}(|\varepsilon_i/\sigma|^q)$ by the Markov inequality. Thus, $c_\psi^2 = \mathrm{O}\{(1-\psi)^{-2/q}\}$. In particular, for $\psi \le 1 - n^{-1}$ then $c_\psi^2 = \mathrm{O}(n^{2/q}) = \mathrm{o}(n^{1/4})$.

6. *Second term of* (D.27). It holds that $\mathsf{f}(\hat{c}^0_\psi)n^{1/2}(\hat{c}^b_\psi - \hat{c}^0_\psi) = \mathrm{o}_\mathsf{P}(n^{-\omega})$ for all $\omega < \eta - \kappa$ uniformly in $0 \le \psi \le 1$, $b \le n^{1/4-\eta}B$ by Lemma D.6 using Assumption 3.1(ia, iib, iic). By item 4 then $(\hat{c}^0_\psi)^2 = \mathrm{o}_\mathsf{P}(n^{2\upsilon})$ for all $\upsilon > \nu(\eta-\kappa)/2$ for some $\nu < 1$. Thus, the second term vanishes.

7. *Third term of* (D.27). Argue that this equals $-\sigma^2 c^2_\psi \mathbb{G}^{1,0}_n(c_\psi) + \mathrm{o}_\mathsf{P}(1)$. Recall the definition $\tau_\psi = 2\int_0^{c_\psi}\epsilon^2\mathsf{f}(\epsilon)d\epsilon$ and expand

$$\mathcal{S}_3 = n^{1/2}(\tau_{\psi+n^{-1/2}\phi} - \tau_\psi) - c^2_\psi\phi = n^{1/2}\int_{c_\psi}^{c_{\psi+n^{-1/2}\phi}}(\epsilon^2 - c^2_\psi)2\mathsf{f}(\epsilon)d\epsilon.$$

If $\mathcal{S}_3$ can be proved to vanish when $\phi = n^{1/2}\{\mathsf{G}(\hat{c}^0_\psi) - \psi\}$ then the desired expression follows, noting that $\phi = -\mathbb{G}^{1,0}_n(c_\psi) + \mathrm{o}_\mathsf{P}(1)$ by Theorem D.1(c). Thus, consider $\mathcal{S}_3$. Changing variable $y = \mathsf{G}(\epsilon)$, $dy = 2\mathsf{f}(\epsilon)d\epsilon$, and Taylor expanding gives

$$\mathcal{S}_3 = n^{1/2}\int_\psi^{\psi+n^{-1/2}\phi}(c^2_y - c^2_\psi)dy = \phi(c^2_{\psi^*} - c^2_\psi),$$

for some ψ^* so $|\psi^* - \psi| \le \phi$. Rewrite this, for some $\upsilon > 0$,

$$\mathcal{S}_3 = \{\psi(1-\psi)\}^{-2\upsilon}\{\frac{\psi(1-\psi)}{\mathsf{f}(c_\psi)}\}(c_{\psi^*} + c_\psi)[\frac{\phi}{\{\psi(1-\psi)\}^{1/2-\upsilon}}][\frac{\mathsf{f}(c_\psi)n^{1/2}(c_{\psi^*} - c_\psi)}{\{\psi(1-\psi)\}^{1/2-\upsilon}}]n^{-1/2}.$$

Insert $\phi = n^{1/2}\{\mathsf{G}(\hat{c}^0_\psi) - \psi\}$. Consider the five components of $\mathcal{S}_3$ individually. The first component is $\mathrm{O}(n^{2\upsilon})$ for $\psi_0 \le \psi \le n/(n+1)$. The second component is $\mathrm{O}(n^{1/8})$ for $\psi_0 \le \psi \le 1 - n^{-1}$. The reason is that $\psi(1-\psi)/\mathsf{f}(c_\psi) = \mathrm{O}(c_\psi) = \mathrm{O}(n^{1/8})$ by Assumption 3.1(ie) and item 5. The third component is $\mathrm{o}_\mathsf{P}(n^{1/4})$ due to items 4,5. The fourth component equals $n^{1/2}\{\mathsf{G}(\hat{c}^0_\psi) - \psi\}/\{\psi(1-\psi)\}^{1/2-\upsilon}$, which is $\mathrm{o}_\mathsf{P}(1)$ uniformly in $1/(n+1) \le \psi \le n/(n+1)$, see Theorem D.2($a$). The fifth component is seen to be $\mathrm{o}_\mathsf{P}(1)$ by first bounding $|c_{\psi^*} - c_\psi| \le |c_{\psi+n^{-1/2}\phi} - c_\psi| = |c_{\hat{\psi}} - c_\psi|$ where $\hat{\psi} = \mathsf{G}(\hat{c}^0_\psi)$ and then combining the result for the fourth component with the result that $\{\mathsf{f}(c_\psi)n^{1/2}(c_{\hat{\psi}} - c_\psi) - \phi\}/\{\psi(1-\psi)\}^{1/2-\upsilon}$ is $\mathrm{o}_\mathsf{P}(1)$ uniformly in $1/(n+1) \le \psi \le n/(n+1)$, see Theorem D.2($b$) using Assumption 3.1($ib$). The sixth component is $n^{-1/2}$. Overall it holds that $\mathcal{S}_3$ is of order $\mathrm{o}_\mathsf{P}(n^{2\upsilon+1/8+1/4+0+0-1/2}) = \mathrm{o}_\mathsf{P}(n^{2\upsilon-1/8}) = \mathrm{o}_\mathsf{P}(1)$ uniformly in $\psi_0 < \psi < 1 - n^{-1}$ since υ can be chosen sufficiently small. ■

Proof of Theorem 3.3. Note, first the identity

$$\frac{\hat{z}_\psi}{\hat{\sigma}_{\psi,cor}} - c_\psi = \frac{\hat{z}_\psi/\sigma - c_\psi}{\hat{\sigma}_{\psi,cor}/\sigma} - c_\psi\frac{\hat{\sigma}^2_{\psi,cor} - \sigma^2}{\hat{\sigma}_{\psi,cor}(\hat{\sigma}_{\psi,cor} + \sigma)}.$$

Multiply this by $2\mathsf{f}(c_\psi)n^{1/2}$. Use that $n^{1/2}(\hat{\sigma}^2_{\psi,cor}/\sigma^2 - 1)$ and $n^{1/2}(\hat{z}_\psi/\sigma - c_\psi)$ have the leading terms $\tau_\psi^{-1}\mathbb{L}_n(c_\psi)$ and $\mathbb{G}_n(c_\psi)$, respectively, due to Theorems 3.1, 3.2. In particular $\hat{\sigma}_{\psi,cor}$ is consistent for σ. ■

D.10 Proof of Theorem 3.5

Proof of Theorem 3.5. 1. *Expand* using (D.25), (D.26) in item 1 of the proof of Lemma D.14 to get

$$\begin{aligned} b &= \{\widehat{\mathsf{G}}_n^{xx,0}(b, \hat{c}_\psi^b)\}^{-1}\{n^{1/2}\widehat{\mathsf{G}}_n^{x,1}(b, \hat{c}_\psi^b)\} \\ &= \{\Sigma_n\psi + \mathrm{o_P}(1)\}^{-1}\{\mathbb{G}_n^{1,0}(0, c_\psi) + 2c_\psi \mathsf{f}(c_\psi)\Sigma_n b + \mathrm{o_P}(1)\}, \end{aligned}$$

uniformly in $|b| \leq n^{1/4-\eta}B$, $\psi_0 \leq \psi \leq 1$ using Assumptions 3.1(ia, ib, ii). In particular, the results hold for $|b| \leq B$. Recall the notation $\rho_\psi = 2c_\psi \mathsf{f}(c_\psi)/\psi$ defined in (D.18). Note that $\rho_\psi < 1$ for $\psi_0 \leq \psi$ as established in Lemma D.10 using Assumptions 3.1(ic), and that the process $\mathbb{G}_n^{1,0}$ is tight as established in Lemma D.4(c) using Assumptions 3.1(ia, iib, iic). It then follows that

$$b = (\Sigma_n\psi)^{-1}\mathbb{G}_n^{1,0}(0, c_\psi) + \rho_\psi b + \mathrm{o_P}(1),$$

noting that $|b| \leq B$, or equivalently that

$$(1 - \rho_\psi)b = (\Sigma_n\psi)^{-1}\mathbb{G}_n^{1,0}(0, c_\psi) + \mathrm{o_P}(1).$$

Since $\rho_\psi < 1$ then $1 - \rho_\psi$ is bounded away from zero so we can divide by $1 - \rho_\psi$. Due to Lemma D.12(b) using Assumption 3.1$(ia - id, ii, iii)$ then $\widehat{b}_\psi = \mathrm{O_P}(1)$ for $\psi_1 \leq \psi \leq 1$ so that b can be replaced by $\widehat{b}$. ■

D.11 Proof of Theorems 3.4

Proof of Theorem 3.4. We first find a lower bound $\widehat{d}^{(m)} \leq \widehat{z}^{(m)}$ and then an upper bound for $\widehat{z}^{(m)}$, and finally show that the difference is small.

1. *Lower bound problem.* It holds $\widehat{d}^{(m)} \leq \widehat{z}^{(m)}$. Indeed, if $S^{(m)}$ is the ranks of $\widehat{\xi}_1^{(m)}, \dots, \widehat{\xi}_m^{(m)}$ then $\widehat{d}^{(m)} = \widehat{z}^{(m)}$. If $S^{(m)}$ does not have this form then its complement must include one of the ranks of $\widehat{\xi}_1^{(m)}, \dots, \widehat{\xi}_m^{(m)}$, for instance that of $i^\dagger$. In that situation $\widehat{d}^{(m)} \leq \widehat{\xi}_{i^\dagger}^{(m)} \leq \widehat{\xi}_{(m)}^{(m)} \leq \widehat{\xi}_{(m+1)}^{(m)} = \widehat{z}^{(m)}$.

2. *Relation between the ranks of $S^{(m)}$ and $\widehat{\xi}_{(m)}^{(m-1)}$.* The set $S^{(m)}$ is the ranks of $\widehat{\xi}_{(1)}^{(m-1)}, \dots, \widehat{\xi}_{(m)}^{(m-1)}$. It follows that for all $i \notin S^{(m)}$ then $\widehat{\xi}_i^{(m-1)} \geq \widehat{\xi}_{(m+1)}^{(m-1)} \geq \widehat{\xi}_{(m)}^{(m-1)} = \widehat{z}^{(m-1)}$.

3. *Inequality for deletion residual.* The absolute residual for observation i based on the set $S^{(m)}$, $\xi_i^{(m-1)}$, satisfies

$$\xi_i^{(m-1)} = |y_i - x_i'\widehat{\beta}^{(m-1)}| \leq |y_i - x_i'\widehat{\beta}^{(m)}| + |x_i'(\widehat{\beta}^{(m)} - \widehat{\beta}^{(m-1)})| \leq \xi_i^{(m)} + \max_{1\leq i\leq n}|N'x_i||N^{-1}(\widehat{\beta}^{(m)} - \widehat{\beta}^{(m-1)})|.$$

For $i \notin S^{(m)}$ we have $\xi_i^{(m-1)} \geq \xi_m^{(m-1)} = \hat{z}^{(m-1)}$ and $\widehat{d}^{(m)} = \min_{i \notin S^{(m)}} \widehat{\xi}_i^{(m)}$ giving

$$\hat{z}^{(m-1)} \leq \widehat{d}^{(m)} + \max_{1\leq i\leq n}|N'x_i||N^{-1}(\widehat{\beta}^{(m)} - \widehat{\beta}^{(m-1)})|$$

and therefore, using $\widehat{d}^{(m)} \leq \widehat{z}^{(m)}$ we find

$$0 \leq \widehat{z}^{(m)} - \widehat{d}^{(m)} \leq \widehat{z}^{(m)} - \widehat{z}^{(m-1)} + |N^{-1}(\widehat{\beta}^{(m)} - \widehat{\beta}^{(m-1)})| \max_i |Nx_i|.$$

4. *Embed in the interval* $[0,1]$ *using* $\psi = m/n$. The limiting processes for $\widehat{z}^{(m)}$ and $\widehat{\beta}^{(m)}$ are tight and continuous due to Theorems 3.1, D.9. Therefore

$$\sup_{\psi_0 \le \psi \le n/(n+1)} |2\mathsf{f}(c_\psi)(\widehat{z}_\psi - \widehat{z}_{\psi-1/n})| = \mathrm{o}_\mathsf{P}(n^{-1/2}),$$
$$\sup_{\psi_0 \le \psi \le n/(n+1)} |N^{-1}(\widehat{\beta}^{(m)} - \widehat{\beta}^{(m-1)})| = \mathrm{o}_\mathsf{P}(n^{-1/2}).$$

Moreover, $\max_i |Nx_i|$ is $\mathrm{O}_\mathsf{P}(n^{\kappa-1/2})$ for some $\kappa < \eta \le 1/4$ by Assumption 3.1(*iib*). Combine to get the desired result

$$0 \le 2\mathsf{f}(c_{m/n})(\widehat{z}^{(m)} - \widehat{d}^{(m)}) \le 2\mathsf{f}(c_{m/n})|\widehat{z}^{(m)} - \widehat{z}^{(m-1)}| + 2\mathsf{f}(c_{m/n})|N^{-1}(\widehat{\beta}^{(m)} - \widehat{\beta}^{(m-1)})| \max_i |Nx_i|$$
$$\le \mathrm{o}_\mathsf{P}(n^{-1/2}) + \mathrm{o}_\mathsf{P}(n^{-1/2}n^{\kappa-1/2}) = \mathrm{o}_\mathsf{P}(n^{-1/2}).$$

■

E A result on order statistics of t-distributed variables

x x

Theorem E.1 *Let* $v_1, \dots, v_n$ *be independent absolute* $\mathsf{t}_{m-\dim x}$ *distributed. Consider the* $m+$ x 1 *smallest order statistic* $\hat{v}^{(m)}$. *Suppose* $\dim x$ *is fixed while* $m \sim \psi n$ *for some* $0 < \psi < 1$. *Let* φ *be the standard normal density. Then as* $n \to \infty$ *it holds*

$$2\varphi(c_{m/n})n^{1/2}(\hat{v}^{(m)} - c_{m/n}) \overset{\mathsf{D}}{\to} \mathsf{N}\{0, \psi(1-\psi)\}.$$

Sketch of the proof of Theorem E.1. Let $\hat{v}^{(m)}$ be the $(m+1)'$st quantile of a sample of n scaled, absolute $\mathsf{t}_{m-\dim x}$ variables. To get a handle on the asymptotic distribution of $\hat{v}^{(m)}$ consider first the $(m+1)'$st smallest order statistic, $\hat{w}^{(m)}$ say, from n draws of absolute standard normal variables. This satisfies

$$2\varphi(c_{m/n})n^{1/2}(\hat{w}^{(m)} - c_{m/n}) \overset{\mathsf{D}}{\to} \mathsf{N}\{0, \psi(1-\psi)\},$$

for $m \sim \psi n$ and $c_\psi = \mathsf{G}^{-1}(\psi)$ due to Theorems 3.6 and D.1(a). The absolute standard normal variables have distribution function $2\Phi(y) - 1$. For the $\mathsf{t}_{m-\dim x}$-order statistic $\hat{v}^{(m)}$ it is useful to Edgeworth expand $\mathsf{P}(\mathsf{t}_{m-\dim x} \le y) = 2\{\Phi(y) + \mathrm{O}(n^{-1})\} - 1$, for $m \sim \psi n$, which indicates that the same asymptotic distribution arises as in the normal case. A more formal argument will keep track of the remainder terms. The starting point could be the expression for $\mathsf{P}(\hat{v}^{(m)} \le y)$ in terms of the distribution of an F variate as given in Guenther (1977, equation 3). This can be expanded using the approximation to the $\log \mathsf{F}$ distribution by Aroian (1941, Section 15). These considerations lead to the following result. ■

References

Aroian, L.A. (1941) A study of R.A. Fisher's z distribution and the related F distribution. *Annals of Mathematical Statistics,* **12**, 429–448.

Atkinson, A.C. (1994) Fast very robust methods for detection of multiple outliers. *Journal of the American Statistical Association,* **89**, 1329–1339.

Atkinson, A.C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis.* New York: Springer.

Atkinson, A.C. and Riani. M. (2006) Distribution Theory and Simulations for Tests of Outliers in Regression. *Journal of Computational and Graphical Statistics*, **15**, 460–476.

Atkinson, A.C., Riani, M. and Cerioli, A. (2010a) The forward search: Theory and data analysis (with discussion). *Journal of the Korean Statistical Society,* **39**, 117–134.

Atkinson, A.C., Riani, M. and Cerioli, A. (2010b) Rejoinder: The forward search: Theory and data analysis. *Journal of the Korean Statistical Society,* **39**, 161–163.

Bahadur, R.R. (1966) A note on quantiles in large samples. *Annals of Mathematical Statistics,* **37**, 577–580.

Bercu, B. and Touati, A. (2008) Exponential inequalities for self-normalized martingales with applications. *Annals of Applied Probability,***18**, 1848–1869.

Bickel, P.J. (1975) One-step Huber estimates in the linear model. *Journal of the American Statistical Association,* **70**, 428–434.

Billingsley, P. (1968) *Convergence of Probability Measures.* New York: Wiley.

Csörgő, M. (1983) *Quantile Processes with Statistical Applications.* CBMS-NFS Regional Conference Series in Applied Mathematics 42, Society for Industrial and Applied Mathematics.

Cavaliere, G. and Georgiev, I. (2013) Exploiting infinite variance through dummy variables in nonstationary autoregressions. *Econometric Theory,* **29**, 1162–1195.

Dollinger, M.B. and Staudte, R.G. (1991) Influence functions of iteratively reweighted least squares estimators. *Journal of the American Statistical Association,* **86**, 709–716.

Engler, E. and Nielsen, B. (2009) The empirical process of autoregressive residuals. *Econometrics Journal,* **12**, 367–381.

Guenther, W.C. (1977) An easy method for obtaining percentage points of order statistics. *Technometrics,* **19**, 319–321

Hadi, A.S. (1992) Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society B*, **54**, 761–771.

Hadi, A.S. and Simonoff, J.S. (1993) Procedures for the Identification of Multiple Outliers in Linear Models *Journal of the American Statistical Association,* **88**, 1264-1272.

Helland, I. (1982) Central limit theorems for martingales with discrete or continuous time. *Scandinavian Journal of Statistics,* **9**, 79–94.

Johansen, S. and Nielsen, B. (2009) An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, J.L. and Shephard, N. (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, pp. 1–36. Oxford: Oxford University Press.

Johansen, S. and Nielsen, B. (2010) Discussion: The forward search: Theory and data analysis. *Journal of the Korean Statistical Society,* **39**, 137–145.

Johansen, S. and Nielsen, B. (2013) Outlier detection in regression using an iterated one-step approximation to the Huber-skip estimator. *Econometrics,* **1**, 53–70.

Kiefer, J. (1967) On Bahadur's representation of sample quantiles. *Annals of Mathematical Statistics,* **38**, 1323–1342.

Koul, H.L. and Ossiander, M. (1994) Weak convergence of randomly weighted dependent residual empiricals with applications to autoregression. *Annals of Statistics,* **22**, 540–582.

Lee, S. and Wei, C.Z. (1999) On residual empirical processes of stochastic regression models with applications to time series. *Annals of Statistics,* **27**, 237–61.

R Development Core Team (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Revuz, D. and Yor, M. (1998) *Continuous Martingales and Brownian motion.* 3rd edition. Berlin: Springer.

Riani, M. and Atkinson, A.C. (2007) Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in Data Analysis and Classification,* **1**, 123–141.

Riani, M., Atkinson, A.C. and Cerioli, A. (2009) Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society B*, **71,** 447–466.

Ronchetti, E and Welsh, A.H. (2002) A journey in single steps: robust one-step M-estimation in linear regression *Journal of Statistical Planning and Inference,* **103**, 287–310.

Rousseeuw, P.J. (1982) Most robust M-estimators in the infinitesimal sense. *Zeitschrift für Warhscheinlichkeitstheorie und verwandte Gebiete,* **61**, 541–551.

Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of the American Statistical Association,* **79**, 871–880.

Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection.* New York: Wiley.

Ruppert, D. and Carroll, R.J. (1980) Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association,* **75**, 828–838.

Sampford, M.R. (1953) Some inequalities on Mill's ratio and related functions. *Annals of Mathematical Statistics,* **24**, 130–132.

Shorack, G.R. (1979) Weak convergence of empirical and quantile processes in sup-norm metrics via KMT-constructions. *Stochastic Processes and their Applications,* **9**, 95–98.

Soms, A.P. (1976) An asymptotic expansion for the tail area of the t-distribution. *Journal of the American Statistical Association,* **71**, 728-730.

Víšek, J.Á. (2006a) The least trimmed squares. Part I: Consistency. *Kybernetika,* **42**, 1–36.

Víšek, J.Á. (2006b) The least trimmed squares. Part II: $\sqrt{n}$-consistency. *Kybernetika,* **42**, 181–202.

Víšek, J.Á. (2006c) The least trimmed squares. Part III: Asymptotic normality. *Kybernetika,* **42**, 203–224.