Quantile Spacings: A Simple Method for the Joint Estimation of Multiple Quantiles Without Crossing*

Lawrence D. W. Schmidt

Department of Economics, University of Chicago

Yinchu Zhu

Rady School of Management, University of California San Diego†

July 11, 2016

Abstract

We propose a simple but flexible parametric method for estimating multiple conditional quantiles. By construction, the estimated quantiles will satisfy the monotonicity requirement which must hold for any distribution, so, in contrast to many benchmark methods, they are not susceptible to the well-known quantile crossing problem. Rather than directly modeling the level of each individual quantile, we begin with a single quantile (usually the median), and then add or subtract sums of nonnegative functions (quantile spacings) to obtain the other quantiles. Our approach is thus a natural extension of the location-scale paradigm that also permits higher order moments (e.g., skewness and kurtosis) to vary. We propose two estimation methods and characterize the limiting behavior of each, establishing consistency, asymptotic normality, and the validity of bootstrap inference. The latter method, under an additional "linear index" assumption, respects monotonicity but preserves the computational tractability of standard linear quantile regression. We propose a simple interpolation method which generates a mapping from a finite number of quantiles to a probability density function. Simulation exercises demonstrate that the estimators perform well in finite samples. Finally, three applications demonstrate the utility of the method in time-series (forecasting), cross-sectional, and panel settings.

NOTE: Draft is still preliminary. All comments are very welcome!

*We are particularly grateful to Brendan Beare, Stéphane Bonhomme, Xiaohong Chen, James Hamilton, Ivana Komunjer, Andres Santos, Yixiao Sun, Allan Timmermann, Rossen Valkanov, and seminar participants at UCSD and the University of Chicago for many helpful suggestions and comments which helped to substantially improve the paper.

†Schmidt: ldwschmidt@uchicago.edu, Zhu: yinchu.zhu@rady.ucsd.edu

1 Introduction

While the vast majority of applied research is focused on estimating conditional expectations, as larger, high quality datasets continue to emerge, there is growing interest and potential for moving beyond means and variances towards studying conditional distributions.[1] One way to proceed is to develop a statistical model for a finite number of conditional quantiles, and, within that paradigm, the most common approach is to estimate a model in which the conditional quantile at each probability index is a linear function of observables.[2] This assumption is often motivated on the grounds of parsimony and computational efficiency.

Unfortunately, linear models often result in fitted quantile functions which do not satisfy the natural monotonicity requirement of a true quantile function, a phenomenon often referred to as the quantile crossing problem. For example, the fitted model might suggest that the conditional median is lower than the conditional tenth percentile, a clear logical inconsistency. Crossings can limit the utility of the multiple-quantile approach, since the distribution-approximation qualities discussed above break down at values of X for which these crossings occur. Comparisons across different quantiles, holding X fixed, convey little meaning in these regions of the support. Crossings are perhaps the most extreme manifestation of the potential misspecification of the constant marginal effect assumption implied by the additive separability of a linear model. Bondell et al. (2010) observe that "this problem is well-known, but no simple and general solution currently exists."[3]

We propose a simple but flexible parametric framework for joint estimation of multiple conditional quantiles. These quantiles will satisfy the monotonicity requirement by construction. The approach is quite intuitive, and it relates to a suggestion in Granger (2010). Rather than directly modeling the level of each individual quantile, we begin with a single quantile (usually the median), and then add or subtract sums of nonnegative functions (quantile spacings) to it to calculate other quantiles. Each nonnegative function parameterizes the distance between two adjacent quantiles. Our model is in many ways a natural extension of the location-scale paradigm that also permits higher order moments (e.g., skewness and kurtosis) to vary.[4] When

[1] Economists are fascinated by the evolution of the distributions of wealth and income over time, both unconditionally and conditional on observables. Researchers across many fields study sources of idiosyncratic risk faced by workers, entrepreneurs, investors, and firms, and how total income, wealth, and employment are reallocated in the cross section. Incorporating non-Gaussian shocks into both theoretical and forecasting models can lead to different implications, more robust decision-making and more reliable risk measurement. Policy interventions may have highly heterogeneous effects on distributions of potential outcomes. In all of these cases (and many others), distributions can be informative about underlying mechanisms and frictions.

[2] More precisely, the assumption is linearity in parameters; one could include nonlinear functions of X.

[3] We will discuss other proposed solutions to the crossing problem below.

[4] Our approach is analogous to methods for approximating intervals, where one models the midpoint and the range of the interval, rather than try to model the upper and lower bounds directly.

the model is correctly specified, the model parameters may be consistently estimated by minimizing a sum of "check" functions, a result which is closely related to White et al. (2015), hereafter WKM.[5] In addition to showing consistency and asymptotic normality, we establish the consistency of a weighted bootstrap inference procedure.

At first glance, an apparent drawback of enforcing monotonicity is that, since each spacing function is nonnegative, our model for conditional quantiles is nonlinear, which can complicate estimation in practice. Given some additional structure, this need not be the case. One can estimate the parameters of a "linear index"[6] model by recursively running a sequence of standard linear quantile regressions–which may be written as convex, linear programs–on transformations of the data, starting with the median and working outwards towards more extremal quantiles. In this case, the key computational advantages of standard linear quantile regression model are preserved. As above, we establish the consistency, asymptotic normality, and bootstrap validity, making implementation and inference straightforward.

We argue for a natural alternative to the linear model discussed above, in which each spacing is an exponentially affine function of X, which is a special case of the computationally-friendly linear index model discussed in the prior paragraph. Its parameters are easy-to-interpret semi-elasticities and it allows one to test a wide variety of relatively complicated hypotheses about distributions via simple Wald tests. This functional form mirrors a common practice in the literature in financial econometrics on volatility forecasting which, beginning with the exponential GARCH model proposed by Nelson (1991), models the log of conditional volatility as an affine function of observed variables (e.g. squared daily returns).[7]

The spacing approach and recursive estimation procedure also integrates nicely with other related econometric tools. Difference-in-difference methods, when combined with our quantile spacing model, naturally extend to estimate treatment effects on at different conditional quantiles.[8] Regression-discontinuity designs (RDD) can also easily be adapted to our quantile

[5] Angrist et al. (2006) address interpretation of the estimates in linear quantile models under misspecification and how that the resulting estimated quantile functions are the best approximation (in some sense) to the true conditional quantile function. Analogously, under misspecification, our method yields consistent estimates of a pseudo-true parameter value which has a similar interpretation.

[6] Each spacing is a known, positive, strictly increasing transformation of a linear combination of X's.

[7] One of Nelson's (1991) original motivations for the EGARCH specification was to flexibly allow for the presence of a negative correlation between returns and future volatility, which can only be done in restricted ways in standard GARCH specifications while still guaranteeing positivity of the conditional variance. The use of the exponential GARCH specifications tends to be most common in situations where additional observable regressors are included in the variance equation. Examples include seasonal indicators (Andersen and Bollerslev, 1997), macroeconomic variables (Engle et al. 2013), or multiple sources of high-frequency realized volatility measures (Hansen and Huang, 2015). It is also common to assume that log volatility follows an AR(1) in stochastic volatility models; see, e.g., Alizadeh et al. (2002). While our framework does not nest GARCH models, our results cover quantile versions of MIDAS specifications, such as those in Ghysels et al. (2006).

[8] The identifying assumptions associated with our linear index, exponential spacing approach have a similar

framework. This suggests a potential path for applying these methods to study causal effects of policy changes on conditional distributions of outcomes while controlling for observed characteristics. Our work can also extend existing quantile methods, such as instrumental variables quantile regression by Chernozhukov and Hansen (2005) and high-dimensional quantile regression by Belloni et al. (2011a), to account for multiple quantiles simultaneously.

In some applications, an estimate of a few individual quantiles is the object of interest. Robust measures of conditional skewness and kurtosis depend only on 3 and 4 estimated quantiles, respectively. In other cases, one may be interested in characterizing the entire conditional quantile function or its inverse–the conditional distribution function. We propose an interpolation method which generates a mapping from a finite number of quantiles to a probability density function. This flexible, parametric mapping allows us to perform simulation exercises and to move between quantiles and moments. Since the conditional quantiles are monotone by construction, this object is always well-defined.[9] We provide a method for constructing confidence intervals on functionals of the parameters, interpolated density and X, such average marginal effects and counterfactual decompositions similar to Machado and Mata (2005).

Next, we study the performance of the methodology in several simulation exercises. We find that the recursive estimation procedure yields consistent estimates in finite samples and our bootstrap inference procedures correctly control size. Estimates obtained by estimating a finite number of quantiles and interpolating, under certain distance metrics, yield more accurate estimates of the conditional quantile function relative to those generated following the method in Chernozhukov et al. (2010).

Finally, we conclude by demonstrating the utility of the method via three empirical examples, which apply the method in time series, cross-sectional, and panel applications. Our method provides a simple but powerful way to characterize how distributions evolve over time and/or conditional on characteristics. In all three cases, the spacing decomposition makes parameters quite interpretable; it naturally decomposes changes in cross-sectional distributions into factors which shift the entire distribution (between effects) from those which generate cross-sectional heterogeneity for agents with similar observables (within effects).

The first application involves using generating forecasts of conditional quantiles of the daily distribution of S&P 500 returns. Some variables primarily shift the location of the distribution, others can affect just the left and/or the right tail. We compare forecasts of a linear model with

flavor to the fully nonparametric "changes in changes" estimator of Athey and Imbens (2006).

[9] Under some regularity conditions, one could likely use such an interpolation method to construct a quasi-likelihood for the data. In this case, a QMLE estimator would choose the parameters of the model so to minimize the Kullback-Leibler divergence between the parametric model and the data. We do not formally explore this approach here, but we note that such an estimator would be a relatively standard case of QMLE.

those generated via the spacing method–a natural alternative benchmark which uses the same conditioning information, has the same computational complexity, and estimates the same number of parameters. Estimated quantiles generated using the spacing approach generate a higher in-sample fit and significantly outperform the linear model out of sample. Whereas the linear model has fitted quantiles that cross approximately 3% of the time, our spacing method does not suffer from this drawback.

Second, we present a very simple application of the method for conducting causal inference on the effect of a policy change on several conditional quantiles using a regression discontinuity approach. We revisit the question considered in Lalive (2008), who studies the effect of extensions of unemployment benefits on unemployment durations. The policy change differentially affected workers above the age of 50 relative to younger workers, so we are able to compare distributions of unemployment durations as a function of age for workers immediately above and below the cutoff. Interestingly, using our spacing approach, we find that the effects of the policy change are almost exclusively concentrated in the right tail of the distribution. This suggests that the average effect on durations which is attributed to a behavioral response to the policy is concentrated among a small fraction of workers ex-post.

In a panel context, our preferred specification provides a natural way for studying the evolution of a cross sectional distribution over time. To demonstrate this, we highlight several interesting results from recent work by Schmidt et al. (2016), hereafter STW, who use our econometric method to study a run on money market mutual funds that developed during September 2008 after the failure of Lehman Brothers. STW find that the model for multiple quantiles reveals a number of very interesting insights about the forces which combine to generate runs. We highlight several insights which emerge via using our spacing method to study these data.

Related Literature: Beginning with Koenker and Bassett (1978), well-understood methods exist for estimating a single conditional quantile, for both linear and nonlinear models. These models are semi-parametric in nature, rely on few distributional assumptions, have some desirable robustness properties, and are valid under relatively weak conditions. We refer the reader to the monograph byKoenker (2005), which provides a thorough treatment of quantile regression methods and notable applications.

A number of previous studies consider potential remedies to the quantile crossing problem, including Mammen (1991), Dette and Volgushev (2008), Chernozhukov et al. (2009), Chernozhukov et al. (2010) and Qu and Yoon (2015). Most assume that a linear model is correctly specified (making crossings a finite sample problem only), and proposes statistical solutions to eliminate crossings in finite sample estimates.[10] While many elegant solutions have been

[10]For example, Chernozhukov et al. (2010) address the problem by monotonically rearranging the estimated

proposed to the finite-sample problem, the observed non-monotonicities sometimes call into question the validity of the functional form (usually, linearity) assumed and can result in parameters which are difficult to interpret for high-dimensional X. Two alternatives are to change the functional form or use fully nonparametric methods.[11] Parameters can sometimes be difficult to interpret and/or estimation/inference can be challenging. In contrast, monotonicity is easily satisfied by the spacing approach, which is similar in spirit to the location-scale model of He (1997). These spacings have useful interpretations and are often direct objects of interest.

Conditional quantiles have been the object of interest in many studies in economics and finance. Important examples include the studies on wage structure and dynamics, such as Buchinsky (1994), Gosling et al. (2000), Abadie et al. (2002) and Machado and Mata (2005). One of the most widely used quantile models in finance is the value at risk (VaR) model and has been applied to measure the market risk (e.g. Engle and Manganelli (2004) and firm cash flow risk (e.g. Adrian and Brunnermeier (2011)).

A number of recent papers in macro and finance have estimated at multiple quantiles as a parsimonious approach to study conditional distributions evolving over time. Applications to stock returns include Kim and White (2003), White Jr et al. (2008), Cenesizoglu and Timmermann (2008), White et al. (2015) and Ghysels et al. (2016a). Covas et al. (2014) investigate bank-level capital shortfalls conditional on macroeconomic variables. Guvenen et al. (2014) examine conditional quantiles of the distribution of idiosyncratic income risk over the business cycle. Kelly and Jiang (2014), Herskovic et al. (2015), Kehrig (2015), and Salgado et al. (2015) study the relation between aggregate risk factors and changes in the shape of distributions of firm stock returns, profitability, and productivity measures. Decker et al. (2016) study the asymmetry in the firm growth rate distribution.

The organization of the remainder of the paper is as follows. In Section 2, we present our model for conditional quantiles along with some examples and potential applications illustrating the model. In Section 3, we propose two distinct procedures for estimation and inference and establish the theoretical properties for both. In Section 4, we discuss via several examples how our method can be combined with commonly used econometric models. We study the finite sample properties of our procedure through Monte Carlo simulations in Section 5. Section 6 provides three empirical applications. Technical details are contained in Appendix A.

quantile curve, and provides limit theory for the rearranged curve. Belloni et al. (2011b) extend this result to allow for a large number of regressors, in combination with model selection techniques, in order to get a sieve approximation of the conditional quantile function.

[11] Gouriéroux and Jasiak (2008) propose an alternative model, exploiting the fact that a mixture of quantile functions is itself a quantile function. Using this property, it is straightforward to guarantee monotonicity of a conditional quantile function which is a mixture of time-invariant quantile functions with weights which depend on the conditioning variables. Like us, Qu and Yoon (2015) propose to approximate a distribution by interpolating between a finite number of quantiles, which are estimated using local linear methods.

2 A Flexible Model for Conditional Quantiles

In this section, we outline the basic motivation and structure of our spacing approach and highlight a number of useful features and potential applications of a natural alternative to the linear model which preserves monotonicity.

2.1 Motivation and basic setup

Before introducing any notation, it may be useful to clarify the objects of interest. Our goal is to estimate a finite number (p) of quantiles of a scalar-valued real random variable y_i, where these quantiles will be allowed to vary as a function of a finite dimensional vector of conditioning variables, x_i.

Remark 2.1. Throughout, we will restrict our attention to the case where the support of y_i is $\mathbb{R}$. It is straightforward to modify this setup for situations in which y_i has a bounded support. One need only to specify a link function from the real line to the interval of interest. We will abstract away from these cases for ease of exposition.

We specify a simple but flexible parametric functional form which guarantees that the conditional quantiles will satisfy the basic monotonicity restrictions. Much of the motivation for our method was succinctly described by Granger (2010) (emphasis added):

> A single quantile series can appear to be very much like any other economic series and standard methods of analysis can be used, such as building AR(l) models. However, if one moves to the series coming from a pair of quantiles Q1 and Q2, the situation changes, as there will necessarily be an inequality to hold, such as $Q2 > Q1$, and this is difficult to maintain in a standard linear model. This problem is even worse with a group of quantiles, all of which fall into a ranking with many inequalities, and this is not possible to achieve with the standard VAR model, for example.
>
> *An easier way to proceed is to start with the median, say, and then adding the 'spacing' S to then obtain the 'next' quantile Q. Note that $S > 0$. One can proceed to any quantile above the median by starting with the median and then adding several spacings, all of which are positive. Compared to ordered series, such as quantiles, it is much easier to model positive processes, such as spacings, plus the median.*

Following Granger (2010)'s suggestion, we will model the level of a single quantile, where a natural choice is the conditional median. All other quantiles are defined by adding/subtracting a series of nonnegative spacing functions to that quantile.

Let $0 < \alpha_1 < \cdots < \alpha_p < 1$. For $j = 1, \ldots, p$, the α_j^{th} quantile of y_i conditional on x_i satisfies

$$q_j(x_i) \equiv \inf\{y \in \mathbb{R} \mid P(y_i \leq y \mid x_i) \geq \alpha_j\}. \tag{1}$$

We parametrize

$$q_j\left(x_i;\theta\right) = \begin{cases} q_{L,j}(x_i;\theta_j,\theta_{j+1},\cdots,\theta_{j^*}) & if\ j < j^* \\ q_*\left(x_i;\theta_{j^*}\right) & if\ j = j^* \\ q_{U,j}(x_i;\theta_{j^*},\cdots,\theta_j) & if\ j > j^* \end{cases} \tag{2}$$

where

$$q_{L,j}(x_i;\theta_j,\theta_{j+1},\cdots,\theta_{j^*}) = f_{j^*}\left(x_i;\theta_{j^*}\right) - \sum_{k=j}^{j^*-1} g_k\left(f_k(x_i;\theta_k)\right) \tag{3}$$

$$q_{U,j}(x_i;\theta_{j^*},\cdots,\theta_j) = f_{j^*}\left(x_i;\theta_{j^*}\right) + \sum_{k=j^*+1}^{j-j^*} g_k\left(f_k(x_i;\theta_k)\right). \tag{4}$$

We stack the parameter $\theta = (\theta_1', \cdots, \theta_p')'$. Here, $g_k(\cdot)$'s are invertible nonnegative functions and $f_k(\cdot;\theta_k)$'s are transformation functions. $g_k(\cdot)$ is known and $f_k(\cdot;\theta_k)$ is known up to the parameter θ_k.

Under mild regularity conditions, a decomposition as in (2) is essentially without loss of generality. We can always choose to think about multiple quantiles in terms of a single "level" quantile and spacings between adjacent quantiles. The key advantage of this way of breaking up the distribution is that the nonnegativity restriction on the spacings is sufficient to guarantee that the fitted quantiles are properly ordered; a necessary condition implied by any correctly-specified model.

Another advantage of this framework is that it has a recursive structure. As we will demonstrate below, in many cases one can estimate quantiles sequentially rather than jointly, beginning with the j^{*th} quantile and working outwards toward each tail. The presence of the transformation functions $g_k(\cdot)$ make our model nonlinear. As such, having the ability to do sequential estimation has immense practical advantages, especially under the additional linear index assumption described in the next section.

2.2 A useful special case: the linear index model

Thus far, we have placed little structure on the function $f_k(x_i;\theta_k)$. A natural special case is when $f_k(\cdot)$ is linear in parameters:

$$f_k(x_i;\theta_k) = \tilde{f}_k(x_i)'\theta_k. \tag{5}$$

Note that equation (5) does require the model to be linear in x_i, so we can include nonlinear basis functions of x_i within the model.[12]

In addition to being simple and parsimonious, the linear index structure greatly facilitates estimation. Given this restriction, one can also view the model (2) as an iteratively transformed linear quantile model. The main idea is to notice that, once we transform the residuals layer by layer, we actually obtain linear quantile models. Suppose that we know the value of $(\theta_{j^*}, \theta_{j^*+1}, \cdots, \theta_j)$ for some $j > j^*$. This means that $P(y_i \leq q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j) \mid x_i) = \alpha_j$ and $P(y_i \leq q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j) + g_{j+1}(f_{j+1}(x_i;\theta_{j+1})) \mid x_i) = \alpha_{j+1}$. Thus, $P(y_i - q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j) \leq g_{j+1}(f_{j+1}(x_i;\theta_{j+1})) \mid \{y_i - q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j) > 0\}$ and $x_i) = (\alpha_{j+1} - \alpha_j)/(1-\alpha_j)$. Since $g_{j+1}(\cdot)$ is invertible, we have

$$P\left(g_{j+1}^{-1}(y_i - q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j)) \leq f_{j+1}(x_i;\theta_{j+1}) \mid x_i \text{ and } \{y_i - q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j) > 0\}\right) = \frac{\alpha_{j+1} - \alpha_j}{1-\alpha_j}. \tag{6}$$

Thus, if we are given $(\theta_{j^*}, \cdots, \theta_j)$, then we can use it to construct the transformed residual $g_{j+1}^{-1}(y_i - q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j))$ and the next layer of the model becomes a simple quantile model of the positive transformed residuals on $f_{j+1}(x_i;\theta_{j+1})$, where the only unknown parameter is θ_{j+1}. An analogous argument applies to the lower quantiles. As a result, our model can be viewed a collection of simple models for transformed residuals. When $f_j(\cdot;\theta_j)$ is a linear function of θ_j, each layer becomes a linear quantile regression, which may be estimated by solving a convex, linear program. In this special case, due to the transformation, our model ensures monotonicity while specifying linear models on the "untransformed" residuals does not. In Section 3, we will exploit the linearity and develop a simple estimation procedure under which the nonlinear nature of the transformation does not complicate the computation.

[12] One can also potentially include a low-dimensional parameter vector in $\tilde{f}_k(\cdot)$, and still benefit from many of the computational gains associated with the linear index specification. In the recursive estimation procedure described below, the optimization over θ is globally convex, so we can effectively "concentrate out" these parameters before searching for this lower dimensional parameter vector.

2.3 Interpreting parameters

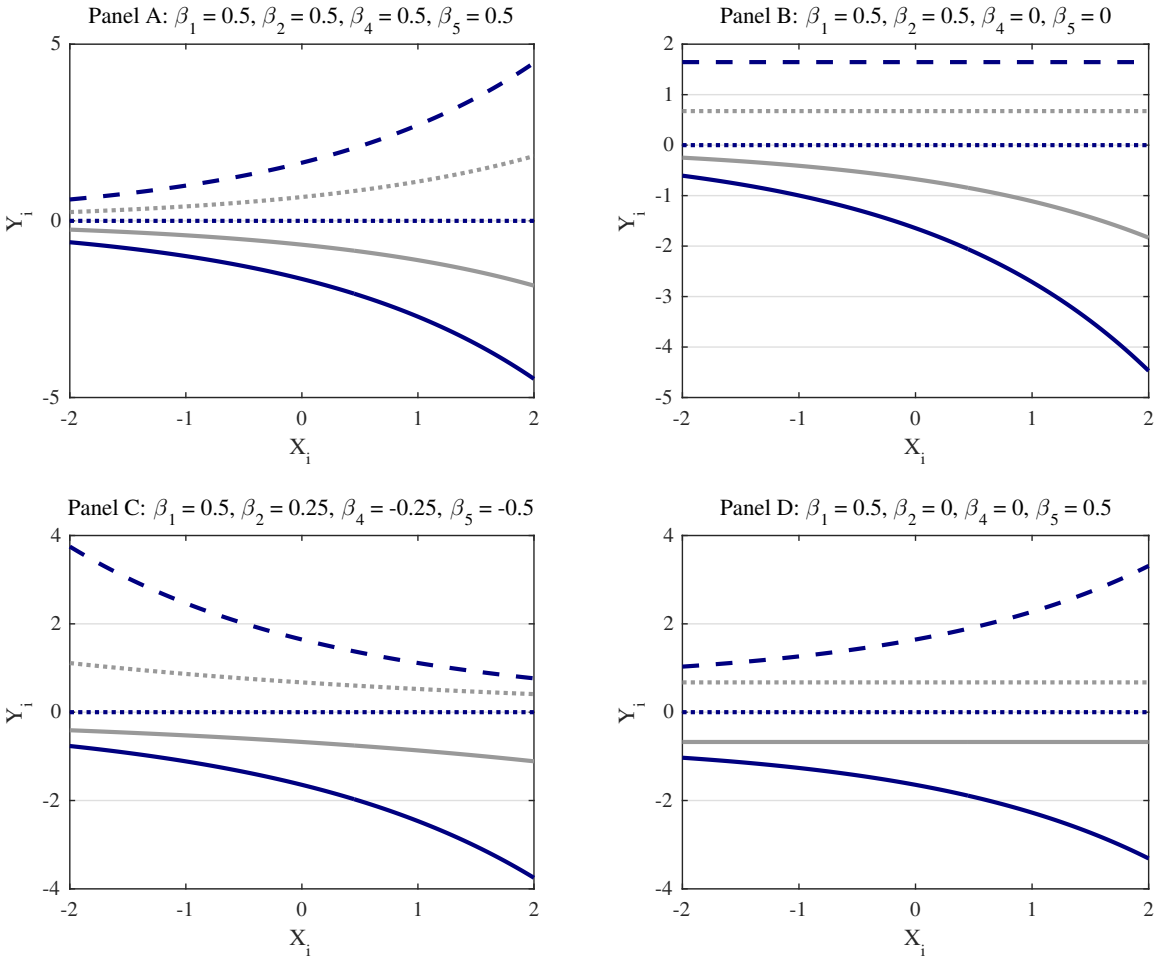
Next we discuss the role of the link function which maps $f_k(\cdot)$ onto the positive real line, and its effect on the associated parameters. Just as a linear specification is a natural choice for $f_k(\cdot)$, we argue that the exponential function is a natural choice for $g_k(\cdot)$. When the two are coupled together, the interpretation of parameters is quite straightforward. Suppose that $\tilde{f}_k(x_i) = (1, x_i')$, then the slope coefficient on x_i for spacing $j+1$ equals $\frac{\partial(q_{j+1}(x_i)-q_j(x_i))}{\partial x_i}\frac{1}{(q_{j+1}(x_i)-q_j(x_i))}$, i.e., the semi-elasticity of the distance between these two quantiles, or the percentage change in the difference between the α_{j+1}^{th} and α_j^{th} quantiles in response to a 1 unit change in x. Since the components of x are multiplicatively separable, the semi-elasticity interpretation is the same independent of other parameters.

Let's briefly contrast this interpretation in the linear index, exponential spacing model with the standard one from a linear quantile regression model. For concreteness, suppose that a binary indicator for whether an individual receives an experimental intervention d, an element of x, is assumed to enter the conditional quantile function linearly, as is common to assume in practice. Thus, we are assuming that the effect of the intervention on the difference between two quantiles is constant, equal to $\frac{\partial}{\partial z}(q_k(x) - q_j(x)) = (\beta_{k,z} - \beta_{j,z})$. If d is binary and there are no controls, a linear model is fully saturated and essentially without loss of generality.

However, things change as soon as we add controls to the regression. Suppose that, restricting attention to the control group, I identify a variable w (again assume it is binary) such that the $Std(y|w=1,d=0) = 2 \cdot Std(y|w=0,d=0)$. Perhaps the outcome y is earnings and w is an indicator for working in finance, an industry in which people tend to have more volatile earnings. Thus, if we ran a univariate quantile regression of y on w, we would generally expect that $\frac{\partial}{\partial w}(q_k(x) - q_j(x)) > 0$. Suppose that we were to run a linear quantile regression estimation of y on w and d and find that $(\hat{\beta}_{k,z} - \hat{\beta}_{j,z})$ and $(\hat{\beta}_{k,w} - \hat{\beta}_{j,w})$ are both positive. Additive separability implies that the effect of the intervention on the difference between quantiles is smaller in proportional terms (i.e., relative to $Std(y|w)$ in the control group) for people working in finance relative to people working in other industries. This may or may not be reasonable depending on the nature of the intervention, but it does suggest that assuming a constant marginal effect may not always be natural in a multivariate setting.

While crossings are perhaps the most extreme manifestation of potential misspecification of a linear model, the difficulty of interpreting constant marginal effects points to a deeper conceptual challenge. To be clear, there many ways to address this issue maintaining linearity in parameters. One could estimate a different model for the effect of the treatment on each subgroup, or, equivalently, one could interact d with w (estimate a nonlinear model that is

Figure 1: Conditional quantiles of linear index-exponential spacing model



This figure plots the $(0.05, 0.25, 0.5, 0.75, 0.95)$-quantiles of $Y_i|X_i$ for linear index model with exponential spacings for various choices of slope coefficients. Constant terms were chosen to match the quantiles of a standard normal distribution when $X_i = 0$. The median is fixed at zero throughout.

linear in parameters). Both approaches are quite feasible when the dimension of x is low, but the curse of dimensionality can quickly kick in when the dimension of x is high. When some functional form restrictions for dimension reduction are required, the marginal effects on spacings from a multiplicatively separable model could be more reasonable in many cases.

Figure 1 plots the conditional quantile functions associated with the linear index-exponential spacing model for a variety of parameter values. We set $\alpha = (0.05, 0.25, 0.5, 0.75, 0.95)'$ and choose the constant terms to match the quantiles of a standard normal random variable when

$x_i = 0$. The horizontal axis shows the value of the conditioning variable X and the vertical axis plots the values of the α_1 through α_p quantiles of $Y|X$. By varying slope coefficients inside the exponential function, the spacing approach can easily generate rich variation in variance, skewness, and kurtosis with 5 quantiles. We do not label the curves, since the bottom line always corresponds with the α_1 quantile, the next line up corresponds with α_2, etc. Graphs of this form are likely to provide a powerful data visualization tool, particularly in situations where $dim(X) \gg 2$, since they allow us to look at "slices" of $Y|X$ which cannot necessarily be observed using scatter plots alone. Graphically, the $g_j(\cdot)$ functions provide the vertical distance between adjacent level curves in the graph.

Regardless of the specific functional form of $g_k(\cdot)$, since the link function is always positive and strictly increasing, positive coefficients in the linear index specification have an unambiguous interpretation; they indicate that a particular segment of the distribution becomes more spread out as x increases. If we model five quantiles which are symmetric about the median, the four slope coefficients on the spacings characterize the effect of a one unit change in x on the width of the "left tail", "left shoulder", "right shoulder", and "right tail" of the distribution, respectively.

2.4 Interpolating between quantiles

Next, we discuss our method to construct a mapping from p quantiles to a conditional quantile function. Given that we have guaranteed that the conditional quantiles satisfy the necessary monotonicity constraints, this essentially reduces to an interpolation exercise. In this section, we propose two alternative approaches for interpolating between adjacent quantiles.

Our first approach exploits the monotonicity of well-defined quantile functions. We use this approach in simulation exercises below:

Algorithm 2.1. *Let $\Psi(\cdot)$ be a baseline quantile function, say the inverse function of the c.d.f of $N(0,1)$. We require $Q(\alpha, x_i; \theta)$, the interpolated quantile function, to have the form*

$$Q(\alpha, x_i; \theta) = \begin{cases} a_1(x_i;\theta) + b_1(x_i;\theta)\Psi(\alpha) & \forall \alpha \in (0, \alpha_1] \bigcup [\alpha_p, 1) \\ a_j(x_i;\theta) + b_j(x_i;\theta)\Psi(\alpha) & \forall \alpha \in (\alpha_{j-1}, \alpha_j] \text{ for } j \in \{2, \cdots, p\} \end{cases}$$

For example, let $Z \sim N(0,1)$. Then the above interpolation simply means that on $(\alpha_{j-1}, \alpha_j]$, the quantile function $Q(\cdot, x_i; \theta)$ coincides with the quantile function of $a_j + b_j Z$. Now we determine $\{(a_j, b_j)\}_{j=1}^p$ by requiring that $\forall j \in \{1, \cdots, p\}$, $q_j(x_i;\theta) = Q(\alpha_j, x_i; \theta)$. It is also straightforward to recover the distribution and density functions.

Since we are building the quantile function from a well-defined baseline quantile function, the interpolated quantile function is automatically an increasing function. One can also use different baseline quantile functions for different j. This method is computationally straightforward because it can be easily shown that $\forall 1 \leq j \leq p$, $a_j(x_i;\theta) = q_j(x_i;\theta) - b_j(x_i;\theta)\Psi(\alpha_j)$ and

$$b_j(x_i;\theta) \quad = \quad \begin{cases} \left[q_p(x_i;\theta) - q_1(x_i;\theta)\right] / \left[\Psi(\alpha_p) - \Psi(\alpha_1)\right] & \text{if } j = 1 \\ \left[q_j(x_i;\theta) - q_{j-1}(x_i;\theta)\right] / \left[\Psi(\alpha_j) - \Psi(\alpha_{j-1})\right] & \text{if } 2 \leq j \leq p \end{cases}$$

Hence, $Q(\alpha, x_i;\theta)$ is a linear combination of $q_j(x_i;\theta)$'s, where the linear combination only depends on α. If $q_j(x_i;\theta)$'s are continuously differentiable in θ, so is the interpolated $Q(\alpha, x_i;\theta)$. Hence, one can expect that the mapping $\theta \mapsto Q(\cdot,\cdot;\theta)$ is Hadamard differentiable in some appropriate space and, by the functional delta method, any valid bootstrap procedure for inference of θ translates an inference procedure for the conditional quantile function.

The left panel of Figure 2 shows the interpolated quantile functions that are obtained for the same specifications as in Figure 1 by using Algorithm 2.1. Different lines correspond with various values of X. For the majority of the specifications considered, the method yields relatively smooth, continuous, strictly increasing CDFs. While this method guarantees monotonicity of the interpolated quantile functions and differentiability with respect to the parameters, the fitted density functions are not guaranteed to be continuous in y. Close inspection of Panels C1 and D1 reveals that the conditional quantile functions have kinks, which implies that the interpolated densities are discontinuous in y. Our alternative approach, which we describe momentarily, imposes more smoothness restrictions.

When $\Psi(\cdot)$ is taken to be the quantile function of $N(0,1)$, the conditional mean of y_i is given by a closed-end formula and thus it is easy to compare the estimated conditional mean with estimates obtained from other methods. Let $F(y, x_i;\theta)$ be the cumulative distribution function corresponding to the quantile function $Q(\alpha, x_i;\theta)$ defined above with $\Psi = \Phi^{-1}$, where $\Phi(\cdot)$ is the c.d.f of $N(0,1)$. Then, we can obtain, by straight-forward computations, that

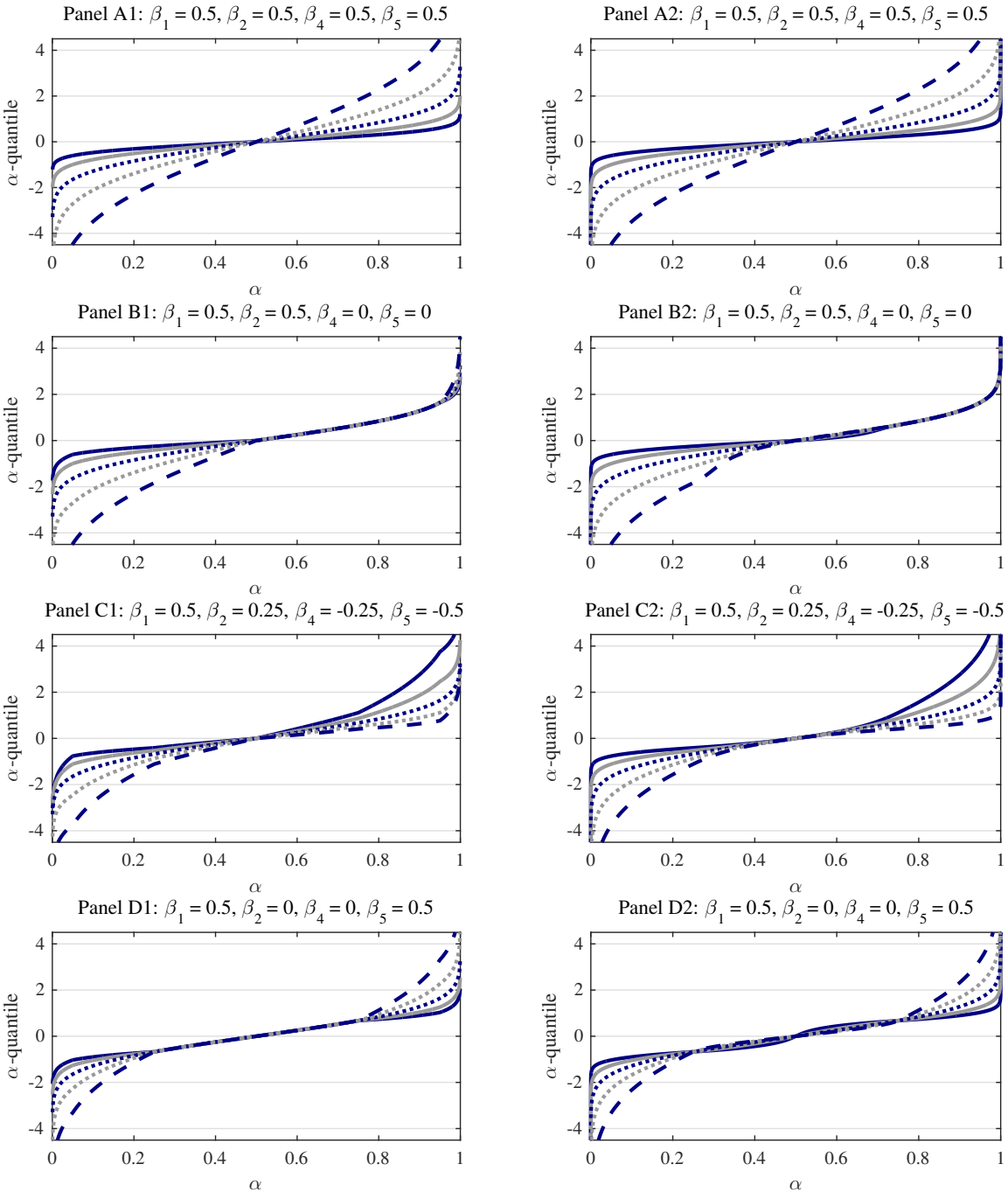
$$\int_{-\infty}^{\infty} y dF(y, x_i;\theta) = \sum_{j=1}^{p+1} D\left[a_j(x_i;\theta), b_j(x_i;\theta), q_{j-1}(x_i;\theta), q_j(x_i;\theta)\right], \tag{7}$$

where $q_0(x_i;\theta) = -\infty$, $q_{p+1}(x_i;\theta) = \infty$, $a_{p+1}(x_i;\theta) = a_1(x_i;\theta)$, $b_{p+1}(x_i;\theta) = b_1(x_i;\theta)$ and

$$D\left[a, b, q_L, q_U\right] = \frac{b}{\sqrt{2\pi}}\left[\exp\left(-\frac{(q_L - a)^2}{2b^2}\right) - \exp\left(-\frac{(q_U - a)^2}{2b^2}\right)\right] + a\left[\Phi\left(\frac{a - q_L}{b}\right) - \Phi\left(\frac{a - q_U}{b}\right)\right].$$

A logarithm transformation is commonly used when the quantity of interest can only take

Figure 2: Interpolated quantile functions of linear index-exponential spacing model



This figure plots the $(0.05, 0.25, 0.5, 0.75, 0.95)$-quantiles of $Y_i|X_i$ for linear index model with exponential spacings for various choices of slope coefficients. Different lines correspond with different $X_i \in \{-2, -1, 0, 1, 2\}$. Left and right panels use Algorithms 2.1 and 2.2, respectively, to interpolate between quantiles. Constant terms were chosen to match the quantiles of a standard normal distribution when $X_i = 0$. The median is fixed at zero throughout.

nonnegative values. In these situations, the response variable y is the log of the quantity of interest and thus we would typically the conditional mean of $\exp(y)$. When $\Psi = \Phi^{-1}$, we also have a closed-end solution for the conditional mean of $\exp(y_i)$ whose conditional quantile function is $Q(\cdot, x_i; \theta)$:

$$\int_{-\infty}^{\infty} e^y dF(y, x_i; \theta) = \sum_{j=1}^{p+1} B\left[a_j(x_i;\theta), b_j(x_i;\theta), q_{j-1}(x_i;\theta), q_j(x_i;\theta)\right], \tag{8}$$

where

$$B\left[a, b, q_L, q_U\right] = \exp\left(a + \frac{1}{2}b^2\right)\left[\Phi\left(\frac{-q_L + a + b^2}{b}\right) - \Phi\left(\frac{-q_U + a + b^2}{b}\right)\right].$$

Our second algorithm generates smooth, continuous densities and is computationally straightforward. However, via this method, it is not straight-forward to impose monotonicity of the interpolated quantile function. We use a parametric, location-scale model to match the tails and interpolate between interior quantiles using a quadratic spline.

Algorithm 2.2. *The quadratic spline method proceeds as follows:*

1. *Choose the parameters of a distribution which is known up to location and scale to match α_1 and α_2. Proceed analogously for upper quantiles, choosing different parameters for α_{p-1} and α_p.*
2. *Interpolate the density between remaining interior interval $q_j(x_i;\theta) - q_j(x_i;\theta)$ using a quadratic spline. The spline parameters are identified by the following conditions:*
   (a) *The integral of the polynomial between the α_j and α_{j+1} quantiles equals $\alpha_{j+1} - \alpha_j$*
   (b) *The spline is continuous and differentiable in its interior*
   (c) *The density is continuous at its endpoints: i.e., the polynomial connects smoothly with the parametric density in both tails.*

Solving for the parameters of the density, distribution, and quantile functions is fast and straightforward using Algorithm 2.2. One can solve for the relevant parameters by inverting a linear system of equations. Due to the non-crossing property, this system will always have a well-behaved solution. The only downside, however, one cannot guarantee that the interpolated density is nonnegative. However, in our experiments thus far, we have encountered this issue extremely infrequently. The right panel of Figure 2 shows the interpolated quantile functions obtained via this approach. While there are subtle differences between the two methods,

results are generally pretty similar. In the simulation results below, we work with Algorithm 2.1.

Remark 2.2. An alternative to the method proposed in Algorithm 2.2 is to use a spline to characterize the *log* of the density, rather than the density itself. This has the advantage of guaranteeing nonnegativity of the density, but solving for the spline coefficients requires solving a nonlinear system of equations.

3 Estimation and inference

In this section, we assume that we observe an i.i.d sample of $\{(y_i, x_i)\}_{i=1}^n$. Our method can also be used for time series data but establishing the theoretical properties in the time series setup requires additional complications in technicality. We will use the check function $\rho_\tau(u) = u(\tau - \mathbf{1}\{u < 0\})$ and the function $\psi_\tau(u) = \tau - \mathbf{1}\{u < 0\}$.

3.1 A general method

For quantile regressions, we use

$$\hat{\theta}_{gen} = \arg\min_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{p} \rho_{\alpha_j} \left(y_i - q_j(x_i; \theta) \right). \tag{9}$$

Similar estimators are considered by White et al. (2015), but here we do not assume that the model is correctly specified. As a result, the population counterpart of $\hat{\theta}_{general}$ should be intepreted as the pseudo-true parameter defined by

$$\theta_0 = \arg\min_{\theta} \sum_{j=1}^{p} E\rho_{\alpha_j} \left(y_i - q_j(x_i; \theta) \right).$$

One can show consistency and asymptotic normality of $\hat{\theta}_{general}$ by applying the usual theory on M-estimators based on empirical processes.

Assumption 1. *The following conditions hold:*
(i) The parameter space Θ *is a compact subset of a Euclidean space.*
(ii) θ_0 *lies in the interior of* Θ *and* $\forall \varepsilon > 0$, $\inf_{\|\theta - \theta_0\| > \varepsilon, \theta \in \Theta} H(\theta) > H(\theta_0)$, *where* $H(\theta) = \sum_{j=1}^{p} E\rho_{\alpha_j} \left(y_i - q_j(x_i; \theta) \right)$.
(iii) $H(\cdot)$ *is twice continuously differentiable and* $\nabla_\theta^2 H(\theta_0)$ *is nonsingular.*

(iv) In a neighborhood of zero, $y_i - q_j(x_i;\theta_0)$ has bounded p.d.f for $1 \le j \le p$.
(v) For some $\delta > 0$, $E\sup_\theta \|\nabla_\theta q_j(x_i;\theta)\|^{2+\delta} < \infty$ for $1 \le j \le p$.

Remark 3.1. Assumption 1 is mild and similar to commonly imposed conditions in the literature. Assumption 1(i) imposes compactness, which is commonly for M-estimators with non-convex objection functions. Assumption 1(ii) ensures that the model parameter is well identified and rules out non-standard asymptotics that arises due to the parameter lying on the boundary of the parameter space. Assumption 1(iii) is needed to ensure that the estimator is $\sqrt{n}$-consistent. Conditions on the p.d.f similar to Assumption 1(iv) are also very common in the literature of quantile regressions. It rules out point probability mass at zero. Conditions A1 and A2 in Section 4.2 of Koenker (2005) are similar to Assumption 1(i)-(iv). We also impose moment conditions on the derivative of the objective function in Assumption 1(v). This condition, when the parameter space is compact, is typically guaranteed by certain moment conditions of x_i or compact support of x_i.

Theorem 1. *Under Assumption 1, we have*

$$\sqrt{n}(\hat{\theta}_{general} - \theta_0) \to^d N(0, V_*)$$

and $V_* = \left[\nabla^2_\theta H(\theta_0)\right]^{-1} Var\left[v_i(\theta_0)v_i(\theta_0)'\right]\left[\nabla^2_\theta H(\theta_0)\right]^{-1}$, *where* $v_i(\theta) = -\sum_{j=1}^p \nabla_\theta q_j(x_i;\theta)\psi_{\alpha_j}(y_i - q_j(x_i;\theta))$.

Remark 3.2. Since the objective function in (9) is nonsmooth and possibly nonconvex, we recommend computing $\hat{\theta}_{general}$ using the MCMC method proposed by Chernozhukov and Hong (2003). Another advantage of this method is that, by Theorem 4 of Chernozhukov and Hong (2003), the MCMC method automatically yields a consistent estimator for $\nabla^2_\theta H(\theta_0)$. We can estimate $Var\left[v_i(\theta_0)v_i(\theta_0)'\right]$ by the sample covariance of $v_i(\hat{\theta}_{general})$.

3.2 An iterative method

In this section, we develop computationally fast procedures for estimation and inference of the parameters in (2) and establish their asymptotic validity. The strategy is to exploit (6), which is a simple quantile regression model after we transform the residuals. We first define the pseudo-true parameter values.

Definition 1. Let $\theta_* = (\theta_{1,*}, \cdots, \theta_{p,*})$ be defined recursively as follows.

1. Define

$$\theta_{j^*,*} = \arg\min_{\theta_{j^*} \in \Theta_{j^*}} E\rho_{\alpha_{j^*}}\left(y_i - m(x_i;\theta_{j^*})\right).$$

2. For $j \in \{j^*, \cdots, p-1\}$, define

$$\theta_{j+1,*} = \underset{\theta_{j+1} \in \Theta_{j+1}}{\arg\min}\, E\left[\mathbf{1}_{\{\varepsilon_{j,i}>0\}} \rho_{\tau_{j+1}^U}\left(g_{j+1}^{-1}(\varepsilon_{j,i}) - f_{j+1}(x_i;\theta_{j+1})\right)\right],$$

where $\varepsilon_{j,i} = y_i - q_{U,j}(x_i;\theta_{j^*,*},\cdots,\theta_{j,*})$ and $\tau_{j+1}^U = (\alpha_{j+1}-\alpha_j)/(1-\alpha_j)$.

3. For $j \in \{2, \cdots, j^*\}$, define

$$\theta_{j-1,*} = \underset{\theta_{j-1} \in \Theta_{j-1}}{\arg\min}\, E\left[\mathbf{1}_{\{\varepsilon_{j,i}<0\}} \rho_{\tau_{j-1}^L}\left(g_{j-1}^{-1}(-\varepsilon_{j,i}) - f_{j-1}(x_i;\theta_{j-1})\right)\right],$$

where $\varepsilon_{j,i} = y_i - q_{L,j}(x_i;\theta_{j,*},\cdots,\theta_{j^*,*})$ and $\tau_{j-1}^L = (\alpha_j - \alpha_{j-1})/\alpha_j$.

In practice, we will use the following algorithm instead in order to avoid potential problem with $g_j^{-1}(\cdot)$ near zero. We choose a sequence $a_n \downarrow 0$ for the truncation, instead of zero.

Algorithm 3.1. *Implement the following steps:*

1. *Estimate the central quantile:*

$$\hat{\theta}_{j^*} = \underset{\theta_{j^*}}{\arg\min} n^{-1} \sum_{i=1}^n \rho_{\alpha_{j^*}}(y_i - q_*(x_i;\theta_{j^*})).$$

2. *Estimate the upper quantiles sequentially: for each $j \in \{j^*, \cdots, p-1\}$, compute*

$$\begin{aligned}
\hat{\varepsilon}_{j,i}^U &= y_i - q_{U,j}(x_i;\hat{\theta}_{j^*},\cdots,\hat{\theta}_j) \\
\hat{\theta}_{j+1} &= \underset{\theta_{j+1}}{\arg\min} n^{-1} \sum_{i=1}^n \mathbf{1}_{\left\{\hat{\varepsilon}_{j,i}^U > a_n\right\}} \rho_{\tau_{j+1}^U}\left(g_{j+1}^{-1}(\hat{\varepsilon}_{j,i}^U) - f_{j+1}(x_i;\theta_{j+1})\right).
\end{aligned}$$

3. *Estimate the lower quantiles sequentially: for each $j \in \{2, \cdots, j^*\}$, compute*

$$\begin{aligned}
\hat{\varepsilon}_{j,i}^L &= y_i - q_{L,j}(x_i;\hat{\theta}_j,\cdots,\hat{\theta}_{j^*}) \\
\hat{\theta}_{j-1} &= \underset{\theta_{j-1}}{\arg\min} n^{-1} \sum_{i=1}^n \mathbf{1}_{\left\{\hat{\varepsilon}_{j,i}^L < -a_n\right\}} \rho_{\tau_{j-1}^L}\left(g_{j-1}^{-1}\left(-\hat{\varepsilon}_{j,i}^L\right) - f_{j-1}(x_i;\theta_{j-1})\right).
\end{aligned}$$

Notice that in each step, we only estimate one component of θ instead of the entire vector θ. As a result, although the optimization in general has non-convex and non-smooth objective functions, the computational burden is typically manageable since we only optimize on a small number of parameters, usually one or two. In the case of linear $f_j(\cdot;\theta_j)$ in θ_j, each step reduces to the linear quantile regression, which admits very fast algorithms. We also introduce the

following weighted bootstrap procedure discussed in Ma and Kosorok (2005) and Chapter 21 of Kosorok (2007).

Algorithm 3.2. *Implement the following steps:*

1. *We generate i.i.d random variable ξ_i with $E\xi_i = 1$.*

2. *Estimate the central quantile:*

$$\hat{\theta}^*_{j^*} = \arg\min_{\theta_{j^*}} n^{-1} \sum_{i=1}^{n} \xi_i \rho_{\alpha_{j^*}}(y_i - q_*(x_i; \theta_{j^*})).$$

3. *Estimate the upper quantiles sequentially: for each $j \in \{j^*, \cdots, p-1\}$, compute*

$$\begin{aligned}
\hat{\varepsilon}^{U,*}_{j,i} &= y_i - q_{U,j}(x_i; \hat{\theta}^*_{j^*}, \cdots, \hat{\theta}^*_j) \\
\tau^U_{j+1} &= (\alpha_{j+1} - \alpha_j)/(1 - \alpha_j) \\
\hat{\theta}_{j+1} &= \arg\min_{\theta_{j+1}} n^{-1} \sum_{i=1}^{n} \xi_i \mathbf{1}_{\left\{\hat{\varepsilon}^{U,*}_{j,i} > a_n\right\}} \rho_{\tau^U_{j+1}} \left(g^{-1}_{j+1}(\hat{\varepsilon}^{U,*}_{j,i}) - f_{j+1}(x_i; \theta_{j+1}) \right).
\end{aligned}$$

4. *Estimate the lower quantiles sequentially: for each $j \in \{2, \cdots, j^*\}$, compute*

$$\begin{aligned}
\hat{\varepsilon}^{L,*}_{j,i} &= y_i - q_{L,j}(x_i; \hat{\theta}^*_j, \cdots, \hat{\theta}^*_{j^*}) \\
\tau^L_{j-1} &= (\alpha_j - \alpha_{j-1})/\alpha_j \\
\hat{\theta}^*_{j-1} &= \arg\min_{\theta_{j-1}} n^{-1} \sum_{i=1}^{n} \xi_i \mathbf{1}_{\left\{\hat{\varepsilon}^{L,*}_{j,i} < -a_n\right\}} \rho_{\tau^L_{j-1}} \left(g^{-1}_{j-1}\left(-\hat{\varepsilon}^{L,*}_{j,i}\right) - f_{j-1}(x_i; \theta_{j-1}) \right).
\end{aligned}$$

Remark 3.3. This procedure is easy to implement in that the optimization problem can be recast as linear programs similar to the usual linear quantile regression. Another reason for using the weighted bootstrap instead of the nonparametric bootstrap is theoretical. As pointed out in Ma and Kosorok (2005) and Kosorok (2007), the validity of nonparametric bootstrap probably holds but is much more difficult to establish. On the other hand, the validity of the weighted bootstrap almost automatically follows by the argument used to derive the asymptotic normality.

Remark 3.4. In the Monte Carlo simulations and empirical applications, we draw the weights from the exponential distribution with parameter one. The reader is referred to Barbe and Bertail (2012) for rigorous discussions regarding how to choose the weights.

We denote $\hat{\theta} = (\hat{\theta}'_1, \cdots, \hat{\theta}'_p)'$, $\theta_* = (\theta'_{1,*}, \cdots, \theta'_{p,*})'$ and $\hat{\theta}^* = (\hat{\theta}^{*\prime}_1, \cdots, \hat{\theta}^{*\prime}_p)'$. The asymptotic properties are derived under Assumption 2 in Appendix A. In addition to Assumption

1, Assumption 2 imposes more restrictions because the iterative estimator contains $g_j^{-1}(\cdot)$, which tends to infinity around zero. Roughly speaking, for each value of θ, we view $\varepsilon_{j,t}$ viewed as a function of (y_i, x_i) and θ and impose moment conditions on $g_{j+1}^{-1}(|\varepsilon_{j,i}|)$ and $\varepsilon_{j,i} - g_j(f_{j+1}(x_i; \theta_{j+1}))$. To simplify the proof, we also impose compact support for the variables. One can easily verify that, under compact support and parameter space, Assumption 2 is satisfied by $g_j(x) = \exp(x)$.

Theorem 2. *Let Assumption 2 in Appendix A hold. Suppose that $a_n = O(n^{-c})$ for some $c \in (0, \infty)$ and $\sup_{x \geq a_n} |dg^{-1}(x)/dx| = O(n^a)$ for some $a < 1$. Then $n^{1/2}(\hat{\theta} - \theta_*) \to^d N(0, M_*)$ for some matrix M_*. Moreover, if, in addition, $E\xi_i = 1$, then the distribution of $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ conditional on the data converges to $N(0, M_*)$ in probability.*

Remark 3.5. The expression for M_* is extremely complicated and a plug-in estimation approach is not practical. However, one can simply use the bootstrap procedure to estimate M_*.

Remark 3.6. Notice that one can also use the methods introduced in Chernozhukov and Hong (2003) for estimation and inference for the model specified in (2). However, their methods require a version of the information equality, which holds under correct specification. Under misspecification, their methods require estimates for the variance of the derivative of the objective function. On the other hand, our bootstrap procedure does not require such external inputs and thus may be easier to implement in practice.

Our bootstrap inference also makes it convenient to conduct inference on certain functions of the model parameter. For example, if one might be interested in building a confidence interval for the quantile function evaluated at certain values of x_i, then this can be rephrased as inference on a function of θ. Let $F_X(\cdot)$ be the cumulative distribution function of x_i and $\hat{F}_X(\cdot)$ the empirical distribution function. Let $\mathcal{F}$ be a space of distribution functions. Suppose that $\phi(\cdot, \cdot)$ is a mapping from $\Theta \times \mathcal{F}$ to a Euclidean space. We establish the following result.

Corollary 1. *Suppose that ϕ is Hadamard-differentiable with derivative ϕ'. Then under the conditions of Theorem 2,*

$$\sqrt{n}\left(\phi(\hat{\theta}, \hat{F}_X) - \phi(\theta_*, F_X)\right) \to^d \phi'(W),$$

where W is a zero-mean Gaussian process. Moreover, the distribution of $\sqrt{n}\left(f(\hat{\theta}^*, \hat{F}_X^*) - f(\hat{\theta}, \hat{F}_X)\right)$ conditional on the data converges to $\phi'(W)$ in probability, where $\hat{F}_X^*(x) = n^{-1}\sum_{i=1}^n \xi_i \mathbf{1}\{x_i \leq x\}$ and ξ_i is defined in Algorithm 3.2.

Corollary 1 applies the functional Delta method for the inference problem for $\phi(\theta, F)$. This result is quite powerful, as it states that we can conduct inference on functions of θ and the

empirical distribution of x_i. If we want to interpolate between quantiles so as to approximate a density for a given x_i, the theorem tells us how to perform valid inference on the density or distribution forecast. It also allows us to include smooth functions of the empirical distribution of x_i, such as the average quantile treatment effect: $E\left[\partial q_j(x_i;\theta)/\partial x_i'\right]$, which is a function of θ and the distribution of x_i. We could also, for example, calculate similar effects where we re-weight the data using a subset of x_i (e.g., the distribution of individuals who take up a training program).

4 Potential applications and extensions

Before talking about estimation and inference, we wish to provide a number of examples in order to highlight the flexibility of our specification. Sections 4.1 and 4.2 discuss assumptions under which cross-sectional identification techniques, differences-in-differences and regression discontinuity methods, respectively, have natural extensions to the case of multiple quantiles. Section 4.3 briefly discusses the extension of instrumental variables quantile regression methods to the case of the spacing approach. Section 4.6 describes an extension to allow for high-dimensional parameters. Section 4.4 suggests some simple ways to test for symmetry, location/scale restrictions, Granger causality, and nonlinear dependence within our framework. Section 4.5 discusses how to impose some restrictions across spacings while maintaining the computational advantages of the recursive method. Section 4.7 talks about how our method can provide interesting new ways to model (and test for) nonlinear dependence, with some examples motivated by the empirical finance literature.

4.1 Differences-in-differences / event studies

In this section and the one that follows it, suppose that we want to evaluate the effect of a binary treatment, D_{it}, on an outcome Y_{it}. Groups are indexed by i and time periods are indexed by t. We observe the outcome for a large number of individuals in each group and for each time period, and we wish to characterize the effect of the treatment on the distribution of Y_{it}. It is common to estimate the effect of the treatment by running the following regression,

$$E_t[Y_{it}|A_i, Z_{it}, D_{it}] = \alpha_0 + \rho_0 D_{it} + \lambda_{0t} + A_i'\gamma_0 + Z_{it}'\beta_0, \tag{10}$$

where $A_i'\gamma_0$ is a group-specific, constant unobservable and λ_{0t} is a time-specific unobservable mean shifter, often a group-specific dummy variable. The parameter $\rho_0 = E_t[Y_{it}|A_i, Z_{it}, D_{it} =$

$1]-E_t[Y_{it}|A_i, Z_{it}, D_{it}=0] \equiv E_t[Y_{it}^1-Y_{it}^0|A_i, Z_{it}]$ is identified via the parallel trends assumption:

$$E_t[Y_{i,t}^0|A_i, Z_{it}]-E_{t-1}[Y_{i,t-1}^0|A_i, Z_{i,t-1}] = \lambda_{0t}-\lambda_{0,t-1} = E_t[Y_{j,t}^0|A_j, Z_{jt}]-E_{t-1}[Y_{j,t-1}^0|A_j, Z_{j,t-1}]. \tag{11}$$

One could make an analogous assumption about the conditional median by simply replacing conditional expectations of Y_{it} with the alternative condition that

$$P\left[Y_{it} - \alpha_0 + \rho_0 D_{it} + \lambda_{0t} + A_i'\gamma_0 + z_{it}'\beta_0 \leq 0 \mid A_i, Z_{it}, D_{it}\right] = \tfrac{1}{2}. \tag{12}$$

Ex-ante, it is not obvious why restriction (12) on the conditional median would be more or less reasonable than our earlier assumption in (10) on the conditional mean. The restriction in (12) imposes that the counterfactual location shift in Y_{it}^0 for the treated group equals the observed location shift in the control group, which is analogous to (11).

In principle, one could estimate a linear specification like (12) for other quantiles as well, but such a linear-in-parameters specification is only valid under a very strong additive separability assumption. As emphasized in Athey and Imbens (2006), one cannot allow for changes in the scale of the distribution of the non-treated group between periods, and requires assuming that "the underlying distribution of unobservables must be identical in all subpopulations, eliminating an important potential source of intrinsic heterogeneity." As an alternative, Athey and Imbens (2006) argue for a more flexible "changes-in-changes" estimator which relies on weaker assumptions and may be estimated by nonparametric methods.

The basic idea of the Athey and Imbens (2006) approach is to construct an estimate for the counterfactual distribution of Y_{it}^0 for treated individuals by *shifting and rescaling* the observed distribution of pre-treatment outcomes in a manner consistent with the observed change in the control group. A simple parameterization of our linear index, exponential spacing model has a very similar flavor, except that we put more parametric structure on the procedure. In particular, we impose a parallel trends assumption on the log of the distance between adjacent quantiles, which amounts to calculating a counterfactual by rescaling different segments of the conditional quantile function. This structure may be advantageous in situations with many right hand side variables. Suppose we parameterize the model so that the j^{*th} quantile satisfies a restriction like (12) and the spacings satisfy

$$q_{j+1}(x_{i,t};\theta) - q_j(x_{i,t};\theta) = \exp(\alpha_j + \rho_j D_{it} + \lambda_{jt} + A_i'\gamma_j + z_{it}'\beta_j), \tag{13}$$

where x_{it} is a vector of time dummies, group dummies, and a treatment indicator.

Let's discuss each of these ingredients in turn. The factor $\alpha_j + A_i'\gamma_j$ soaks up sources of time-

invariant, group-specific heterogeneity in the distance between quantiles. $z_{it}'\beta_j$ allows other observable, time-varying characteristics to scale up or down the distances between quantiles. We need to assume that this scaling factor is invariant to the treatment, which is analogous to the interpretation on the controls in the standard OLS specifications. The factor λ_{jt} captures common sources of time series variation in this distance. Here, the parallel trends assumption kicks in. Under the null hypothesis that $\rho_j = 0$, we are assuming that the distance between quantiles of the counterfactual distribution of Y_{it}^0 scales up or down by $\lambda_{jt}\%$. Any remaining changes in $q_{j+1}(x_{i,t};\theta) - q_j(x_{i,t};\theta)$ are attributed to the treatment, and the coefficient ρ_j has the semi-elasticity interpretation discussed above.

Often, we have data for many periods, and we may be interested in tracing out the impact of a treatment (or any sequence of past shocks more generally) on a variable of interest, a different form of a differences-in-differences estimator which is sometimes called an event study. For example, in the literature on job displacement, it is common to compare the earnings of individuals involved in mass layoff events with a control group of otherwise similar workers who were not displaced. Given a similar parallel trends assumption to the one made above, one can compare the distributions of the outcome (e.g., earnings) between the groups prior to and after treatment by including leads and lags of D_{it} in the regressions as well. Inspection of coefficients on the leads can help to assess the validity of the parallel trends assumption.

4.2 Regression discontinuity

Our spacing approach integrates nicely with standard regression-discontinuity methods. Suppose that I know that a person is treated if and only if some continuous variable $z_i \geq \bar{z}$. If the density of z_i is continuous in the neighborhood of $\bar{z}$, then I can trace out the causal impact of the treatment by comparing the conditional quantiles of Y_i for individuals with z_i immediately above and below the cutoff. The difference between the two conditional quantile functions is a local estimate of the treatment on the distributions of potential outcomes. These differences are estimable using quantile regression methods.

Since regression discontinuity methods are inherently local, there isn't a clear reason to prefer our spacing method to a linear one in the absence of covariates. However, once covariates are involved, one needs to take a stand on how the distributions change these observable variables. In this case, many of the same practical issues from the previous section become relevant. When the covariates have a nontrivial effect on the spread of the outcome distribution, the additive separability assumption implicit in the standard linear quantile model is perhaps a bit less natural than the multiplicative separability of our exponential spacing approach.

In addition to being potentially useful for producing estimates of treatment effects, our approach is also easily applicable to testing the validity of the identifying assumption, namely that all relevant factors other than the treatment are continuous in the neighborhood of the cutoff.

4.3 IV Quantile regression without crossing

We can adapt the model in (2) to IV quantile models. Let $q_j(x_i;\theta)$ be parametrized as in (2). If x_i is endogeneous and instrument variables z_i are available, we can identify θ using the following moment condition:

$$P\left(y_i - q_j(x_i;\theta) < 0 \mid z_i\right) = \alpha_j. \tag{14}$$

IV quantile models for one single quantile have been considered in the literature, such as Chernozhukov and Hansen (2005). In order to model multiple quantiles simultaneously, one still needs to impose the monotonicity just as in the case without endogeneity. Since the structure of $q_j(x_i;\theta)$ in (2) automatically guarantees this basic requirement, the same specification can be applied to IV quantile models using the moment condition in 14. Moreover, the layer-by-layer interpretation in (6) is still valid.

Thus, one can apply the iterative estimation principle to IV quantile models. To see how, suppose that I want to calculate the parameters of the $(j+1)^{th}$ quantile, with $j \geq j^*$. Then, simple rearrangement of (14) yields an alternative testable restriction:

$$P\left[\left(g_{j+1}^{-1}(\varepsilon_{j,i}^U) - f_{j+1}(x_i;\theta_{j+1})\right) \mid z_i, \varepsilon_{j,i}^U \equiv y_i - q_{U,j}(x_i;\hat{\theta}_{j^*},\cdots,\hat{\theta}_j) > 0\right] = \tau_{j+1}^U = \frac{\alpha_{j+1}-\alpha_j}{1-\alpha_j},$$

where an analogous condition holds for quantiles to the left of the j^{*th} one. If we make the linear index assumption, then one can iteratively apply the "inverse quantile regression" procedure proposed in Chernozhukov and Hansen (2005) to the transformed residuals. Estimating the linear index spacing model with the iterative method does not add any additional computational complexity relative to the linear in parameters model. As estimation is typically much harder in IV quantile models even for linear specifications, the iterative estimation scheme is extremely useful in this context.

4.4 Specification Testing

In the previous section, we demonstrated how we could generate a wide variety of conditional distributions through the use of simple parametric restrictions. If the quantile spacings are defined as in the linear index-exponential spacing model it is extremely straightforward to test necessary conditions for a variety of restrictions on the data generating process. For concreteness, let's remain in the setting from the previous setting, where $f_j(X_t) = (1, X_t)$. We discuss how to perform these tests using a series of numbered remarks.

Remark 4.1 (Testing for Linear Dependence). Perhaps a theory suggests that Y_t and X_t are related to one another only through a location shift. In other words, $Y_t|X_t$ does not exhibit nonlinear dependence. To test this assumption, one can estimate an unrestricted model, then test whether the slope coefficients on X_t in the spacings $j = 1, \ldots, p$, are significantly different from zero using a Wald statistic.

Remark 4.2 (Testing for Location/Scale). Suppose that one wants to test whether Y_t and X_t are related to one another only through location and scale shifts. To test this assumption, one can estimate an unrestricted model, then test the restrictions that $\beta_{1j} = \beta_{1k}$ for $j \neq k$ using a Wald test.

Remark 4.3 (Testing for Symmetry). To test whether $Y_t|X_t$ has a symmetric distribution, choose p to be odd, let $j^* = (p+1)/2$, and $\alpha_j = 1-\alpha_{p-j+1}$ for $j = 1, \ldots, j^*-1$. Then, estimate an unrestricted model, then test the restrictions that $\beta_{1j} = \beta_{1,p-j+1}$ and $\beta_{0j} = \beta_{0,p-j+1}$ for $j = 1, \ldots, j^* - 1$ using a Wald test.

Remark 4.4 (Testing for Granger Causality). To test whether X_t does not Granger cause Y_t, we simply need to include lagged values (or functions of lagged values) of X_t and Y_t in W_t, then test whether the slope coefficients on the lagged values of X_t are equal to zero.

4.5 Imposing cross-equation restrictions on spacings during estimation

In some cases, the researcher may have a prior reason to believe that a particular variable or subset of variables has a symmetric effect on the conditional quantiles of Y_i (i.e., it has the same proportional effect on both tails). Whereas the previous section discussed how to test these parametric restrictions, here we briefly discuss how in order to incorporate them in estimation, which can potentially yield efficiency gains. As in the previous section, we restrict attention to the linear index-exponential spacing model, in which identical slope coefficients yield these proportional increases in scale.

In this case, a slight modification of the recursive estimation procedure is applicable when the quantiles of interest are symmetrically located about the origin, so $\alpha_{j^*} - \alpha_{j^*-k} = \alpha_{j^*+k} - \alpha_{j^*}$

for all positive integer k and $\alpha_{j^*} = 1/2$. For simplicity, let's assume that there are three quantiles. Partition the vector $X_i = (W_i', Z_i')'$, where we want to impose the restriction that $\beta_{z1} = \beta_{z3} = \gamma$ for all elements of Z_i. In this case, one can estimate γ imposing this restriction via running the following second-stage quantile regression:

$$\log |Y_i - X_i'\hat{\beta}_2| = \beta_{w1}' 1\{Y_i - X_i'\hat{\beta}_2 < 0\} W_i + \beta_{w3}' 1\{Y_i - X_i'\hat{\beta}_2 > 0\} W_i + \gamma' Z_i + u_{it}, \qquad (15)$$

where the identifying assumption is $P[u_{it} < 0 = 2(\alpha_1 - 1/2)]$.[13] Again, one can work iteratively out towards the tails, imposing these equality restrictions, in a similar fashion as described in Algorithm 3.1.

When more than 3 quantiles are of interest, it is also easy to restrict the equality of certain slope coefficients across central and more extreme spacings. For example, suppose that $\alpha = \{0.05, 0.25, 0.5, 0.75, 0.95\}$, with $j^* = 3$. Efficiency considerations might motivate a restriction that the slope coefficient on Z_i is the same in both spacings to the left of the median ($\gamma = \beta_{z1} = \beta_{z2}$). Given the recursive structure of our setup, one can obtain a consistent estimate of $\hat{\gamma}$ from the estimation of $\hat{\beta}_{z2}$, and impose this restriction when estimating β_1 by running the quantile regression model,

$$\log |Y_i - X_i'\hat{\beta}_3 - \exp(X_i'\hat{\beta}_2)| - \hat{\beta}_{z2}' Z_i = \beta_{w1}' W_i + u_{it}, \qquad P[u_{it} < 0 = \tfrac{\alpha_2 - \alpha_1}{\alpha_2}], \qquad (16)$$

on the negative residuals. While these cases are technically not covered by our existing asymptotic results in Theorem 2, the extension is extremely straightforward.

4.6 High-dimensional quantile regression

Belloni et al. (2011a) studied the estimation problem for single quantile linear models with sparse high-dimensional parameters. If we set $f_j(\cdot)$'s in the model (2) to be linear in $\theta_j \in \mathbb{R}^{d_j}$, then our model can be used to simulaneously model several quantiles with high-dimensional parameters in the sense that $d_j \gg n$. One might expect the quantile crossing problem to be more common simply due to the large number of covariates. For example, $d_j = O(\exp(n^\alpha))$ for some $\alpha \in (0, 1)$. Our model not only automatically guarantees the monotonicity of quantiles but also provides a feasible approach for estimation. To our best knowledge, there is no general theory regarding the estimation of nonlinear quantile models and the computational burden for minimizing a regularized version of (9) is formidable: an optimization problem of a non-smooth and non-convex function over a high-dimensional space. However, our specification in

[13] This procedure is quite similar to He (1997), who proposes an iterative quantile regression procedure for estimating location scale models.

(2) still admits the layer-by-layer interpretation in (6) and thus the estimation can proceed via an iterative scheme in which each iteration involves only the linear high-dimensional quantile regression studied by Belloni et al. (2011a). The computational burden is minimal since the optimization reduces to linear programs. We believe that the asymptotic theory can be developed by adapting the arguments in Belloni et al. (2011a). Since this is outside the scope of the current paper, we leave this possibility to future research.

4.7 Interacting the approach with copula methods

Copulas provide a very flexible way of allowing for nonlinear dependence between random variables. However, they have two main shortcomings: 1) one must be able to specify the marginal distributions of both variables, and 2) extensions to higher dimensions are challenging. Moreover, obtaining the distribution of $Y_t|X_t$ often involves numerical integration.

In our framework, one obtains a wide variety of ways to model the dependence between Y_t and a vector of conditioning variables X_t. This is quite useful in situations where we care more about the distribution of $Y_t|X_t$, rather than the joint distribution of the two. A simple example from finance would be a setting in which Y_t is the excess return on an individual stock and X_t is a vector of returns for a set of priced risk factors. Often, we would prefer to leave the distribution of the factors unspecified. This framework makes it easy to do so. In addition, “factor loadings" would be directly comparable across firms.

Our approach can also be complementary with copula models, as one obtains a variety of new ways to specify the marginal distributions which are inputs to a copula. In the example above, our model could provide the marginal distributions of two individual assets conditional on the factor returns X_t and/or their own lags. We could then combine these marginals with a copula in order to characterize their joint distribution.

5 Monte Carlo simulations

5.1 Inference of model parameter

We specify $p = 3$ quantile functions with $(\alpha_1, \alpha_2, \alpha_3) = (0.25, 0.5, 0.75)$ with $j^* = 2$. We generate $x_i \in \mathbb{R}^3$ from $N(0, I_3)$. For $j \in \{1, 2, 3\}$, we generate $\theta_{\alpha_j} \in \mathbb{R}^3$ from $N(0, I_3)$. The response variable is generated by $y_i = Q(x_i, u_i; \theta)$, where u_i is drawn from the uniform distribution on the interval $(0, 1)$. The quantile function $Q(x, u, \theta)$ is defined using Algorithm

Table 1: Coverage probabilites

	95% confidence interval						90% confidence interval					
	Weighted bootstrap			i.i.d bootstrap			Weighted bootstrap			i.i.d bootstrap		
$n = 100$	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3	α_1	α_2	α_3
$\theta_{\alpha_j,1}$	0.972	0.938	0.985	0.973	0.939	0.981	0.934	0.898	0.958	0.936	0.898	0.953
$\theta_{\alpha_j,2}$	0.980	0.967	0.996	0.983	0.967	0.995	0.956	0.925	0.984	0.956	0.930	0.983
$\theta_{\alpha_j,3}$	0.982	0.969	0.995	0.985	0.970	0.995	0.957	0.936	0.982	0.962	0.936	0.976
$n = 500$												
$\theta_{\alpha_j,1}$	0.949	0.948	0.961	0.951	0.949	0.960	0.906	0.903	0.917	0.906	0.903	0.913
$\theta_{\alpha_j,2}$	0.957	0.963	0.976	0.958	0.965	0.976	0.917	0.924	0.941	0.919	0.922	0.942
$\theta_{\alpha_j,3}$	0.963	0.964	0.976	0.964	0.963	0.977	0.922	0.922	0.943	0.923	0.923	0.945
$n = 2000$												
$\theta_{\alpha_j,1}$	0.943	0.950	0.941	0.942	0.949	0.943	0.894	0.905	0.888	0.895	0.903	0.889
$\theta_{\alpha_j,2}$	0.949	0.961	0.961	0.950	0.961	0.960	0.898	0.914	0.915	0.899	0.914	0.917
$\theta_{\alpha_j,3}$	0.950	0.961	0.963	0.950	0.957	0.964	0.900	0.920	0.921	0.901	0.919	0.919

The jth component of θ_{α_k} is denoted by $\theta_{\alpha_k,j}$. We construct confidence intervals for each component of the parameter $\theta = (\theta_{\alpha_1}, \theta_{\alpha_2}, \theta_{\alpha_3}) \in \mathbb{R}^9$.

2.1:

$$Q(x,u;\theta) = \begin{cases} \frac{q_j(x;\theta)[\Phi(\alpha)-\Phi(\alpha_{j-1})]-q_{j-1}(x;\theta)[\Phi(\alpha)-\Phi(\alpha_j)]}{\Phi(\alpha_j)-\Phi(\alpha_{j-1})} & u \in (\alpha_{j-1}, \alpha_j] \\ \frac{q_5(x;\theta)[\Phi(\alpha)-\Phi(\alpha_1)]-q_1(x;\theta)[\Phi(\alpha)-\Phi(\alpha_5)]}{\Phi(\alpha_5)-\Phi(\alpha_1)} & u \in (0,\alpha_1] \bigcup (\alpha_p, 1), \end{cases}$$

where $\Phi(\cdot)$ is the quantile function of $N(0,1)$ and

$$\begin{cases} q_1(x;\theta) = q_2(x;\theta) - \exp(x'\theta_{\alpha_1}) \\ q_2(x;\theta) = x'\theta_{\alpha_2} \\ q_3(x;\theta) = q_2(x;\theta) + \exp(x'\theta_{\alpha_3}) \end{cases}$$

The coverage probabilities of confidence intervals for each of the 9 entries of $\theta = (\theta_{\alpha_1}, \theta_{\alpha_2}, \theta_{\alpha_3})$ are calculated using 5000 repetitions. The result is reported in Table 1.

As shown in Table 1, even in small samples, our weighted bootstrap delivers valid confidence intervals for the parameters. Interestingly, although it is hard to prove the validity of the i.i.d bootstrap (nonparametric bootstrap) for θ, we still find, in simulations, that our weighted bootstrap delivers results that are close to those from the i.i.d bootstrap.[14]

We also consider the following difference-in-difference DGP. In time period t in state s, the response variable for individual i is modeled by $y_{i,s,t} = Q(s,t,d_{i,t},u_{i,s,t};\theta)$, where $d_{i,t}$ is the

[14]Notice that, since i.i.d bootstrap is valid for θ_{α_2}, which is the regression coefficient of a simple quantile model, the two bootstrap schemes are comparable for inference on θ_{α_2}.

treatment, $1 \leq t \leq T$, $1 \leq s \leq S$ and $u_{i,s,t}$ is a random variable from $U(0,1)$ independent of $(t, s, d_{i,t})$. The quantile function is defined using Algorithm 2.1:

$$Q(s,t,d,u;\theta) = \begin{cases} \frac{q_j(s,t,d;\theta)[\Phi(\alpha)-\Phi(\alpha_{j-1})]-q_{j-1}(s,t,d;\theta)[\Phi(\alpha)-\Phi(\alpha_j)]}{\Phi(\alpha_j)-\Phi(\alpha_{j-1})} & u \in (\alpha_{j-1}, \alpha_j] \\ \frac{q_5(s,t,d;\theta)[\Phi(\alpha)-\Phi(\alpha_1)]-q_1(s,t,d;\theta)[\Phi(\alpha)-\Phi(\alpha_5)]}{\Phi(\alpha_5)-\Phi(\alpha_1)} & u \in (0, \alpha_1] \bigcup (\alpha_p, 1), \end{cases} \tag{17}$$

where $\Phi(\cdot)$ is the quantile function of $N(0,1)$, $(\alpha_1, \cdots, \alpha_5) = (0.1, 0.25, 0.5, 0.75, 0.9)$, $\theta = \{\beta_1, \cdots, \beta_5\} \bigcup \{\alpha_{j,s} \mid 1 \leq j \leq 5, 1 \leq s \leq S\} \bigcup \{\lambda_{j,t} \mid 1 \leq j \leq 5, 1 \leq t \leq T\}$ and

$$\begin{cases} q_1(s,t,d;\theta) = q_2(s,t,d;\theta) - \exp(d\beta_1 + a_{1,s} + \lambda_{1,t}) \\ q_2(s,t,d;\theta) = q_3(s,t,d;\theta) - \exp(d\beta_2 + a_{2,s} + \lambda_{2,t}) \\ q_3(s,t,d;\theta) = d\beta_3 + a_{3,s} + \lambda_{3,t} \\ q_4(s,t,d;\theta) = q_3(s,t,d;\theta) + \exp(d\beta_4 + a_{4,s} + \lambda_{4,t}) \\ q_5(s,t,d;\theta) = q_4(s,t,d;\theta) + \exp(d\beta_5 + a_{5,s} + \lambda_{5,t}). \end{cases} \tag{18}$$

Notice that, under the above specification, for $1 \leq j \leq 5$, we have $P(y_{i,s,t} \leq q_j(s,t,d_{i,t}) \mid s,t,d_{i,t}) = q_j(s,t,d_{i,t};\theta)$.

The parameters used in simulations are chosen as follows. We set $\beta_1 = \cdots = \beta_5 = 0$ and other components of θ are randomly generated from $N(0,1)$. We use $S = 50$ and $T = 4$. The treatment is binary $d_{i,t} \in \{0,1\}$. For each (s,t), we generate 100 individuals, 50 of which are treated ($d_{i,t} = 1$). In each random sample, there are 20000 observations and the dimension of θ is 270 (after deleting certain indicators to avoid a singular regressor matrix). In each random sample, 200 bootstrap samples are generated to construct confidence intervals for the treatment effect $(\beta_1, \cdots, \beta_5)'$. The performance of the procedure is evaluated in terms of coverage probability using 400 random samples. We consider the weighted bootstrap procedure whose theoretical properties are established in Section 3, as well as the nonparametric bootstrap. The results are reported in Table 2.

As we can see from Table 2, our method provides decent coverage for the treatment effects. Deviations from the nominal coverage probability 95% are due to the "small" sample. Although the sample size is large $n = 20000$, the number of parameters in θ is close to 300. Therefore, we should not expect our procedure to have the same performance in this case as in a case where samples with 20000 observations are used to estimate 5 parameters.

Table 2: Coverage probability of treatment effect of 95%-confidence intervals

	β_1	β_2	β_3	β_4	β_5
Weighted bootstrap	0.975	0.950	0.958	0.970	0.975
Nonparametric bootstrap	0.970	0.948	0.953	0.975	0.980

Table 3: Approximating quantile functions

	DGP1		DGP2	
	$E\|Q_{SZ} - Q_*\|$	$E\|Q_{CFG} - Q_*\|$	$E\|Q_{SZ} - Q_*\|$	$E\|Q_{CFG} - Q_*\|$
$\|\cdot\|_{L_1}$	0.341	0.334	0.194	0.197
$\|\cdot\|_{L_2}$	0.417	0.429	0.209	0.218
$\|\cdot\|_{L_\infty}$	1.039	1.667	0.363	0.567

5.2 Approximating the quantile function

We also compare our methods with the rearrangement method proposed by Chernozhukov et al. (2010).

We build the conditional quantile functions denoted by $Q_{SZ}(\cdot, x_i)$ from misspecified p quantiles using the method discussed in Section 2, where the baseline quantile function is taken to be the quantile function of $N(0,1)$. We use the quantiles $\alpha = (0.1, 0.3, 0.5, 0.7, 0.9)$, i.e. $p = 5$.

We perform linear quantile regressions for individual quantiles and construct the quantile function using the rearrangement method. The resulting quantile function is denoted by $Q_{CFG}(\cdot, x_i)$. We compare the distance between the constructed quantile functions and the true quantile function denoted by $Q_*(\cdot, x_i)$, in terms of L^1, L^2 and L^∞-norms.

We consider two DGP's for the simulated data. DGP1: $y_i = (|x_i'\beta| + 2)U_i$, where $x_i = (1, N(0,1))' \in \mathbb{R}^2$, $U_i \sim U(0,1)$ independent of x_i and $\beta = (-1, 1)'$. DGP2: $y_i = |x_i'\beta| + \varepsilon_i$, where x_i and β are as before and $\varepsilon_i \sim N(0,4)$ is independent of x_i. Following Chernozhukov et al. (2009) and Chernozhukov et al. (2010), we compute $E\|\hat{Q}(\cdot, x_i) - Q_*(\cdot, x_i)\|_{L_1}$, $E\|\hat{Q}(\cdot, x_i) - Q_*(\cdot, x_i)\|_{L_2}$ and $E\|\hat{Q}(\cdot, x_i) - Q_*(\cdot, x_i)\|_{L_\infty}$, where $\hat{Q}(\cdot, x_i)$ is either $Q_{SZ}(\cdot, x_i)$ or $Q_{CFG}(\cdot, x_i)$. The sample size is 500 and expectations of the norms are computed using 2000 random samples.

As can be seen from Table 3, our method outperforms the rearrangement in terms of L^∞-norm unde both DGP's. This is mainly due to the fact that our method provides a better approximation for the quantile function on the tails. Intuitively, for the tails, the rearrangement method builds the quantile function based on linear quantile regressions with very small or very large quantile. Such quantiles are typically estimated based on only a few observations of data and hence might be very inaccurate.

6 Empirical applications

6.1 Forecasting the distribution of stock returns

In this application, the goal is to forecast p conditional quantiles of y_t, the log excess daily stock return. Let these quantiles be $\{\alpha_j\}_{j=1}^p$. We investigate the performance of two competing specifications: (1) our model in (2) with linear specification of f_j's and (2) p linear quantile models. Let $\hat{q}_{j,t,m}^{(1)}$ and $\hat{q}_{j,t,m}^{(2)}$ denote these models' forecasts the conditional α_j-quantile of y_t with parameters estimated using the full sample. Since the objective function in quantile regressions provide a natural loss function, we conduct a quasi-likelihood ratio test by comparing the loss functions evaluated at the conditional quantile forecasts. For quantile α_j, we consider $QLR_{T,j} = \sum_{t=1}^T [\rho_{\alpha_j}(y_t - \hat{q}_{j,t,m}^{(1)}) - \rho_{\alpha_j}(y_t - \hat{q}_{j,t,m}^{(2)})]$; for the overall specification, we consider $QLR_T = \sum_{j=1}^p QLR_{T,j}$. The critical value is computed using a bootstrap procedure; see Appendix ?? for details. In this exercise, we set $\alpha_1 = 0.01$, $\alpha_2 = 0.10$, $\alpha_3 = 0.25$, $\alpha_4 = 0.5$, $\alpha_5 = 0.75$, $\alpha_6 = 0.9$ and $\alpha_7 = 0.99$ with $j_* = 4$. The explanatory variables include T-bill rate, term spread (10yr yield minus 3month yield), corporate spread (BAA yield minus AAA yield), and log of VXO. The studied is implemented using daily data from January 2, 1986 to December 30, 2010, which corresponds to 6300 time periods.

We report the results in Table 4. As we can see, our model outperforms the linear quantile specification. The difference is statistically significant at 5% level for quantiles $\alpha_1, \alpha_5, \alpha_6, \alpha_7$ as well as for the overall goodness of fit. We note that on the extreme right tail, our specification has a gain in pseudo-R^2 over 5%. The results in Table 4 also highlight the difference in predictability of different parts of the conditional distribution of the stock returns. In terms of the pseudo-R^2, the right tail appears easier to predict than the left tail and the median is almost unpredictable.

We also compare the out-of-sample forecasting performance. We repeat the same exercise, except that parameters are now estimated using a rolling window of m observations. Inference on the forecasting performance is based on we follow the methodology in Diebold and Mariano (1995) and Giacomini and White (2006). We report the results in Table 5. Our method outperforms the competing method in most of the quantiles, especially on the left tail of the distribution. This is confirmed by the pseudo-R^2. This superior forecasting performance can translate into better estimate for VaR (value-at-risk). Interestingly, our result is not due to the crossing forecasts generated by the competing model since the crossing occurs in less than 2% of the sample for the competing model. Hence, in addition to enforcing the monotonicity requirement, our model potentially better captures the nonlinear structure in the data.

Table 4: Goodness of fit for conditional quantile specification: predicting stock returns

Quantile	Pseudo-R^2 of $\hat{q}^{(1)}_{j,t}$	Pseudo-R^2 of $\hat{q}^{(2)}_{j,t}$	Difference	P-values
$\alpha_1 = 0.01$	0.264	0.242	0.022	0.042
$\alpha_2 = 0.10$	0.110	0.102	0.008	0.068
$\alpha_3 = 0.25$	0.031	0.029	0.002	0.202
$\alpha_4 = 0.50$	0.000	0.000	0.000	NA
$\alpha_5 = 0.75$	0.043	0.035	0.008	0.020
$\alpha_6 = 0.90$	0.136	0.115	0.021	0.004
$\alpha_7 = 0.99$	0.412	0.350	0.063	0.008
Overall	0.069	0.060	0.009	0.002

We estimate the parameters needed in the two models using the full sample and produce the difference in loss functions: $\rho_{\alpha_j}(y_t - \hat{q}^{(1)}_{j,t,m}) - \rho_{\alpha_j}(y_t - \hat{q}^{(2)}_{j,t,m})$ for $j \in \{1, \cdots, 7\}$. Since two methods have identical forecasts for α_4, we only compare the forecasts for the other quantiles. In the first two columns, the pseudo-R^2 for quantile α_j is computed as $1 - \left(\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \hat{q}^{(l)}_{j,t})\right) / \left(\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \bar{y}_{\alpha_j})\right)$ and the pseudo-R^2 for the overall performance is computed as $1 - \left(\sum_{j=1}^{7}\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \hat{q}^{(l)}_{j,t})\right) / \left(\sum_{j=1}^{7}\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \bar{y}_{\alpha_j})\right)$ for $l \in \{1, 2\}$, where $\bar{y}_{\alpha_j}$ is the sample α_j-quantile of y_t and $\hat{q}^{(1)}_{j,t}$ and $\hat{q}^{(2)}_{j,t}$ are computed using parameters estimated based on the full sample. We also test the hypothesis that, when evaluated at the population parameter value, the pseudo-R^2 of our method is at least as high as that of the linear model. This hypothesis is tested for each quantile and the overall performance. The p-values for the tests are reported in the last column.

We also report the point estimates together with the t-statistics in Table 6. Notice that the median behaves drastically different from other quantiles. In Table 6, none of the coefficients is statistically significant for the conditional median while the extreme quantiles are affected by variables such as log(VXO).

Lastly, we plot the estimated time series of the conditional quantiles in Figure 3. We observe that the tails of the conditional distribution vary wildly but the median of the distribution is stable. This highlights the advantage of studying the entire distribution rather than only the mean or median. For example, if the goal is to investigate whether certain information affects the distribution of asset returns, then focusing only on the median or the mean might be misleading.

6.2 Regression Discontinuity Example: Effects of Unemployment Insurance Benefit Extensions

To illustrate a potential application of our technique, we reanalyze the effect of a dramatic change in Austrian unemployment benefit rules that was initially considered by Lalive (2008). During June 1988, the maximum duration of unemployment benefits was extended from 30

Table 5: Comparison of conditional quantile forecasts

$j \setminus m$	500	1000	2000	Pseudo-R^2 of $\hat{q}^{(1)}_{j,t}$	Pseudo-R^2 of $\hat{q}^{(2)}_{j,t}$
1	-2.60	-3.31	-4.00	0.273	0.191
2	-3.54	-1.67	-3.34	0.106	0.098
3	-2.57	-2.43	-3.41	0.032	0.029
4	NA	NA	NA	0.000	0.000
5	1.77	0.72	-0.79	0.039	0.034
6	-0.78	-1.71	-1.34	0.127	0.107
7	-2.75	-1.56	-2.04	0.383	0.312

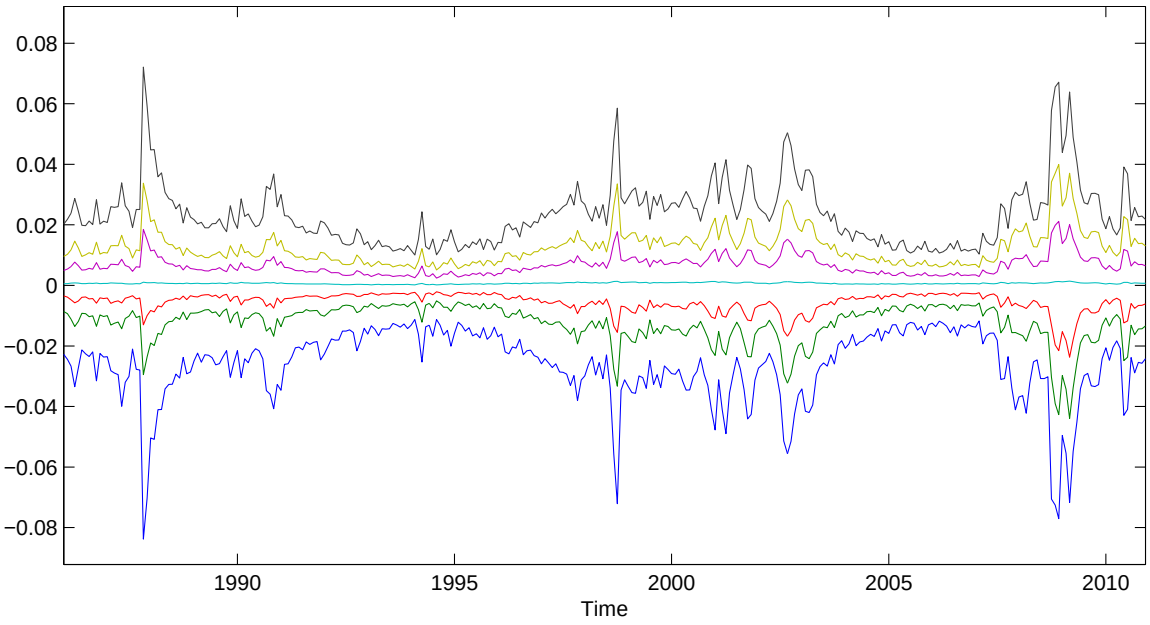
Using a rolling window of m observations, we compute the parameters needed in the two models and produce the difference in loss functions: $\rho_{\alpha_j}(y_t - \hat{q}^{(1)}_{j,t,m}) - \rho_{\alpha_j}(y_t - \hat{q}^{(2)}_{j,t,m})$ for $j \in \{1, \cdots, 7\}$. Since two methods have identical forecasts for α_4, we only compare the forecasts for the other quantiles. We compare the out-of-sample forecasting performance by testing $H_0 : E\left(\rho_{\alpha_j}(y_t - \hat{q}^{(1)}_{j,t,m}) - \rho_{\alpha_j}(y_t - \hat{q}^{(2)}_{j,t,m})\right) = 0$. In the first three columns of the table, the t-statistics computed as in Diebold and Mariano (1995) are reported. In the last two columns, we report the in-sample pseudo-R^2's, which are computed as $1 - \left(\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \hat{q}^{(1)}_{j,t})\right) / \left(\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \bar{y}_{\alpha_j})\right)$ and $1 - \left(\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \hat{q}^{(2)}_{j,t})\right) / \left(\sum_{t=1}^{n} \rho_{\alpha_j}(y_t - \bar{y}_{\alpha_j})\right)$ respectively for our method and the competing method, where $\bar{y}_{\alpha_j}$ is the sample α_j-quantile of y_t and $\hat{q}^{(1)}_{j,t}$ and $\hat{q}^{(2)}_{j,t}$ are computed using parameters estimated based on the full sample.

Table 6: Estimate of θ

	α_1	α_2	α_3	α_4	α_5	α_6	α_7
constant	**-7.47**	**-7.85**	**-8.69**	**0.00**	**-8.62**	**-7.79**	**-8.33**
	-12.63	-26.40	-33.41	-0.81	-34.06	-22.30	-18.52
T-bill	**0.01**	**-0.05**	**-0.04**	**0.00**	**-0.03**	**-0.06**	**0.05**
	0.21	-2.19	-2.18	-0.04	-1.54	-2.26	1.23
Term spread	**0.02**	**-0.06**	**-0.09**	**0.00**	**-0.03**	**-0.05**	**0.05**
	0.34	-1.95	-3.75	-1.08	-1.20	-1.18	0.89
Default spread	**0.48**	**0.07**	**-0.25**	**0.00**	**0.10**	**0.20**	**0.03**
	2.58	0.56	-2.42	-0.30	0.83	1.26	0.17
log(VXO)	**0.95**	**1.03**	**1.32**	**0.00**	**1.18**	**0.96**	**1.14**
	4.96	11.63	16.96	1.39	15.52	8.87	9.40

We use the iterative method and bootstrap procedure discussed in Section 3. For each explanatory variable, the first line corresponds to the estimates (in black bold fonts) and the second line reports the t-statistics (in blue) using 2000 bootstrap samples.

Figure 3: Estimated conditional quantiles $\alpha = \{0.01, 0.1, 0.25, 0.5, 0.75, 0.90, 0.99\}$.



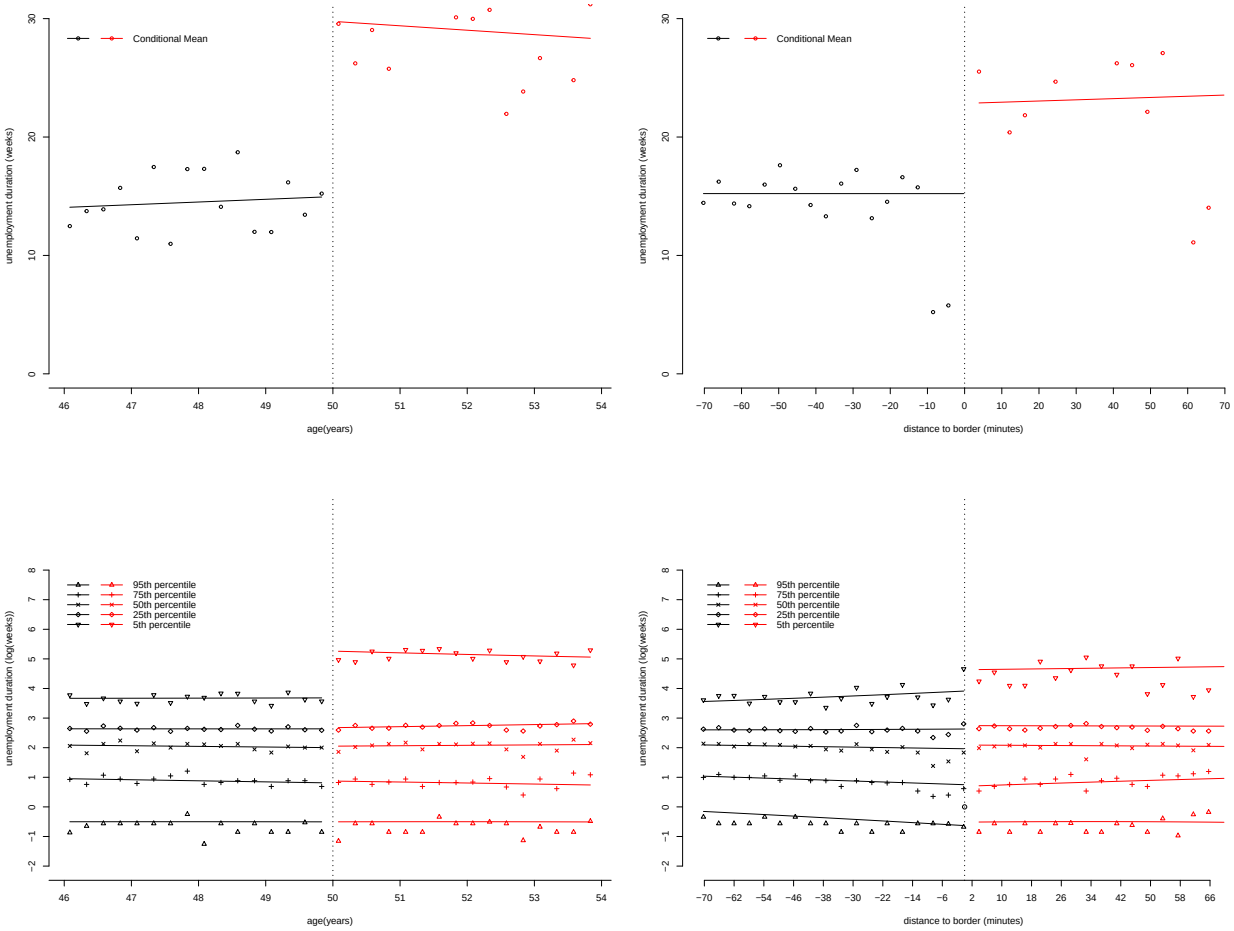
We plot the time series of $q_j\left(x_t;\hat{\theta}\right)$ for each α_j, where $\hat{\theta}$ is estimated using the full sample. In order to avoid a crowded graph, we only display the monthly series by choosing the first day of each month.

weeks to 209 weeks for individuals that were over 50 and lived in certain regions of Austria.[15] Because of the program's design, a worker laid off just after turning 50 was allowed to collect unemployment benefits for 3.5 years longer than an identical worker who was laid off a month earlier. Likewise, two identical workers on different sides of a geographical border could be eligible for different benefits even if the distance between their two residences was very small. Using administrative records from the Austrian social security database and unemployment registrar, Lalive (2008) employed a sharp regression-discontinuity design (RDD) to estimate the effect of the program on the average duration of unemployment spells.

Figure 4(a) reproduces a figure from Lalive (2008) that illustrates a striking discontinuity in the average duration of unemployment that occurs at the eligibility cutoff. The points in the figure correspond to the average duration of unemployment for men grouped by quarter of age, while the solid line corresponds to the fit of a cubic polynomial regression. Lalive (2008) estimates uses the magnitude of the discontinuity to estimate that the program increased the

[15] Individuals also needed to have worked at least 15 of the past 25 years to be eligible. See Lalive (2008) for further discussion of the program.

Figure 4: Effects of 1989 Austrian benefit extensions on unemployment durations



average duration of job search by approximately 14.8 weeks. Figure 4(b) displays conditional quantiles of the duration distribution from the same data used to construct Figure 4(a). The cubic conditional quantile curves were fit using the method outlined in Section 4.2. There is virtually no discernible difference between the 5th, 25th, 50th, and 75th quantiles at the age cutoff. Figure 2 plots the cdf and pdf conditional on ages just below and just above the age cutoff. These figures were constructed using the interpolation procedure outlined in Section 2. The distribution of unemployment duration for those eligible for the increased benefits has a thicker right tail than the distribution for those who were ineligible.

All these observations indicate that the shift in the mean in Figure 4(a) can be attributed to an increase in the right tail of the distribution rather than a uniform increase across the distribution. The average treatment effect could therefore mask significant heterogeneity in the responses to the program. For example, even though the average duration of unemployment increased by approximately 15 weeks, it is possible that the policy only affected the behavior of 5 percent of unemployed workers. Theories built to explain the relationship between unemployment benefits and unemployment duration must be able to account for the substantial heterogeneity in responses that we observe. The heterogeneity in responses also has important implications for the distributional impact of the policy. Without studying the conditional quantiles, it would have been easy to overlook these significant insights.

We also revisit some of the empirical studies in Lalive (2008) using the quantile models proposed in this paper. To be specific, we implement the following exercises. We start by estimating the proposed model with $\alpha_1 = 0.05$, $\alpha_2 = 0.25$, $\alpha_3 = 0.5$, $\alpha_4 = 0.75$ and $\alpha_5 = 0.95$ with $j_* = 3$. Following Lalive (2008), we set the covariates to be $x_i = (1, D_i, Age_i - 50, D_i(Age_i - 50))'$, where D_i is an indicator function of whether or not the individual is treated. We are interested in how the treatment changes the *distribution* of the unemployment duration. This will be referred to as the treatment effect. To study the effect of treatment at different quantiles, we consider a representative individual at 50 years old and define $x_{tr} = (1, 1, 0, 0)'$ and $x_{un} = (1, 0, 0, 0)'$, which denote the covariates for a treated and untreated representative individual, respectively. We compute the quantiles $q_j(x_{tr})$ and $q_j(x_{un})$ as well as their difference. In addition to studying the quantiles, we also estimate the distribution of the unemployment duration conditional on x_{tr} and x_{un} by interpolating the five quantiles using the first method introduced in Section 2. Since this interpolation method allows us to easy compute the conditional mean (using (7) and (8)), we can compare our results with those in Lalive (2008) who study the treatment effect on the conditional mean. In Table 7, we report the results for men under the two identification strategies in Lalive (2008): age threshold and border threshold.

Table 7: Treatment effect of extended benefit duration on unemployment duration

Panel A: age threshold

	α_1	α_2	α_3	α_4	α_5	$E(y_i \mid x_{tr})$	$E(y_i \mid x_{un})$	$E(y_i \mid x_{tr}) - E(y_i \mid x_{un})$
Coeff for D_i	0.09	0.04	0.38	0.03	1.93	31.53	11.09	20.44
	(0.12)	(0.09)	(0.54)	(0.09)	(0.15)	(2.70)	(0.49)	(2.75)
$q_j(x_{tr})$	0.43	2.38	7.78	14.52	192.37			
	(0.07)	(0.14)	(0.39)	(0.59)	(19.63)			
$q_j(x_{un})$	0.42	2.21	7.39	13.96	39.80			
	(0.07)	(0.16)	(0.36)	(0.46)	(2.94)			
$q_j(x_{tr})$-$q_j(x_{un})$	0.00	0.17	0.38	0.56	152.57			
	(0.10)	(0.22)	(0.54)	(0.78)	(19.91)			

Panel B: border threshold

	α_1	α_2	α_3	α_4	α_5	$E(y_i \mid x_{tr})$	$E(y_i \mid x_{un})$	$E(y_i \mid x_{tr}) - E(y_i \mid x_{un})$
Coeff for D_i	-0.04	0.19	0.96	0.11	1.26	28.83	13.35	15.48
	(0.12)	(0.09)	(0.59)	(0.10)	(0.19)	(3.12)	(0.76)	(3.17)
$q_j(x_{tr})$	0.48	1.99	8.08	15.58	170.12			
	(0.08)	(0.17)	(0.46)	(0.78)	(22.64)			
$q_j(x_{un})$	0.52	2.10	7.12	13.85	57.79			
	(0.06)	(0.15)	(0.38)	(0.43)	(5.20)			
$q_j(x_{tr})$-$q_j(x_{un})$	-0.04	-0.10	0.96	1.73	112.32			
	(0.10)	(0.23)	(0.59)	(0.87)	(23.07)			

The dependent variable y_i is the unemployment duration for individual i measured in weeks. We report the point estimates as well as the standard error in brackets. The standard errors are computed using the weighted bootstrap procedure discussed in Section 3. Notice that $q_j(\cdot)$ is the conditional α_j-quantile for unemployment duration.

In Table 7, we can see the heterogeneous effect of the treatment across different parts of the distribution of unemployment duration. With the age threshold as the identification, the treatment only affects the 95-percentile by more than 157 weeks. However, since only 5% of the men respond to the treatment, the overall treatment effect on the conditional mean is merely about 20 weeks, which is similar to 14.8 in column (2) of Table 2 in Lalive (2008). Using the border threshold to obtain identification, we have similar findings and the treatment effect on the conditional mean is 15.48 weeks, close to 13.62 in column (2) of Table 2 in Lalive (2008). Since the unemployment duration is a nonnegative variable, we also model its log values. We conduct the same exercise with log unemployment duration and report the results in Table 8. The results are qualitatively similar to those reported in 7.

In this empirical example, our model reveals information on the entire conditional distribution of unemployment duration and is consistent with previous studies that only consider the conditional mean. As shown in Tables 7 and 8, heterogeneity of the treatment effect on different parts of the distribution of unemployment duration make it hard to understand the whole picture based only on information on the mean.

Table 8: Treatment effect of extended benefit duration on log unemployment duration

Panel A: age threshold

	α_1	α_2	α_3	α_4	α_5	$E(e^{y_i} \mid x_{tr})$	$E(e^{y_i} \mid x_{un})$	$E(e^{y_i} \mid x_{tr}) - E(e^{y_i} \mid x_{un})$
Coeff for D_i	-0.13	-0.02	0.05	-0.02	0.90	42.92	12.62	30.29
	(0.14)	(0.07)	(0.07)	(0.09)	(0.08)	(4.82)	(0.71)	(4.90)
$e^{q_j(x_{tr})}$	0.56	2.39	7.78	14.52	193.00			
	(0.07)	(0.15)	(0.36)	(0.53)	(20.62)			
$e^{q_j(x_{un})}$	0.42	2.22	7.39	13.97	39.85			
	(0.07)	(0.17)	(0.35)	(0.49)	(3.35)			
$e^{q_j(x_{tr})} - e^{q_j(x_{un})}$	0.13	0.17	0.39	0.55	153.15			
	(0.11)	(0.23)	(0.50)	(0.73)	(20.99)			

Panel B: border threshold

	α_1	α_2	α_3	α_4	α_5	$E(e^{y_i} \mid x_{tr})$	$E(e^{y_i} \mid x_{un})$	$E(e^{y_i} \mid x_{tr}) - E(e^{y_i} \mid x_{un})$
Coeff for D_i	-0.03	0.14	0.12	-0.01	0.51	38.33	15.56	22.77
	(0.13)	(0.08)	(0.08)	(0.10)	(0.08)	(5.02)	(1.10)	(5.15)
$e^{q_j(x_{tr})}$	0.55	2.01	8.08	15.58	169.61			
	(0.07)	(0.17)	(0.46)	(0.81)	(22.10)			
$e^{q_j(x_{un})}$	0.56	2.12	7.13	13.88	58.20			
	(0.05)	(0.14)	(0.35)	(0.45)	(5.66)			
$e^{q_j(x_{tr})} - e^{q_j(x_{un})}$	-0.01	-0.11	0.95	1.70	111.41			
	(0.09)	(0.23)	(0.58)	(0.91)	(22.94)			

The dependent variable y_i is the unemployment duration for individual i measured in weeks. We report the point estimates as well as the standard error in brackets. The standard errors are computed using the weighted bootstrap procedure discussed in Section 3. Notice that $q_j(\cdot)$ is the conditional α_j-quantile for the *logarithm* of unemployment duration.

6.3 Panel Application: STW study of runs on money market mutual funds

Another setting in which the study of distributions is potentially of interest is in considering the determinants of changes in bank deposits. Lots of factors combine to generate these flows. Day-to-day transactions will cause investors to deposit and withdraw funds, leading to random, idiosyncratic variation in flows. Broad movements of investors in and out of stock/bond investments and into cash could lead to changes in deposits which are common across banks but relatively small in magnitude. In addition to these routine sources of variation, banks may occasionally be subject to runs, in which a large proportion of investors suddenly seek the return of their deposits as quickly as possible.

Imagine that we observe flows for a panel of 10 banks with identical observable characteristics. During the sample period, a solvency crisis develops, and we learn that aggregate deposits drop by 5%. Given this fact, a scenario where all banks see 5% withdrawals is quite different from a scenario where 9 banks have no change in deposits and a single bank sees 50% deposits withdrawn within a single day. One theory might predict the former outcome to be most likely, while a different theory would suggest the latter. Since the conditional expectation of the flow

distribution remains the same in both cases, an OLS regression would be unable to distinguish between the two. In this case, we gain useful insights from knowing how the cross-sectional distribution changes in response to the aggregate shock.

STW study the determinants of flows to and from money market mutual funds (MMMFs) after the failure of Lehman Brothers, a period of unprecedented stress during the financial crisis.[16] Their primary interest is in using the data in order to test the predictions of models featuring strategic complementarities: models in which an individual agent's expected payoff is positively correlated with the average actions of other agents. Many bank run models (Diamond and Dybvig, 1983) feature complementarities, since my expected payoff from running on a bank increases when I know that other investors will run as well.

STW study the behavior of investors in prime MMMFs–funds that invest primarily in short-term, corporate debt securities. Despite the fact that many of these funds hold relatively similar portfolios, they often market to very different types of investors. Some cater almost exclusively to large institutional investors, whereas other tend to market more heavily to smaller retail investors. STW argue that complementarities are likely to be stronger in funds that have a high concentration of large, well-informed investors, and perform a number of tests to exploit predetermined variation in this fraction of investors in order to test several cross sectional predictions of models with complementarities. They find strong evidence consistent with complementarities acting as a potential amplification mechanism during the crisis.

As part of their analysis, STW use our quantile regression method to characterize the evolution of the cross-sectional distribution of investor redemptions (flows to/from different funds) as the crisis unfolded. Consistent with the simple example above, their estimate a dynamic model of the conditional distribution of flows given measures of investor characteristics, measures of portfolio risk, and prior investors' redemptions. These time-varying fund characteristics are also allowed to interact with calendar time variables, which allows the distribution of flows to shift and scale over time as the crisis progressed.

Their dependent variable, Y_{it}, is the logarithm of the proportional change in daily assets under management. To do this, they use the recursive method described in section 4.1 to estimate the following model specification for the 50^{th}, 10^{th}, and 90^{th} quantiles, respectively:

$$\begin{aligned}
Y_{i,t} &= f_0(X_{i,t}, \alpha_t, \beta) = \alpha_{0,t} + X'_{i,t}\beta_0 + \epsilon^0_{i,t} && P[\epsilon^0_{i,t} < 0 | X_{i,t}] = 0.5 \\
Y_{i,t} &= f_1(X_{i,t}, \alpha_t, \beta) = \alpha_{0,t} + X'_{i,t}\beta_0 - \exp[\alpha_{1,t} + X'_{i,t}\beta_1] + \epsilon^1_{i,t} && P[\epsilon^1_{i,t} < 0 | X_{i,t}] = 0.1 \\
Y_{i,t} &= f_2(X_{i,t}, \alpha_t, \beta) = \alpha_{0,t} + X'_{i,t}\beta_0 + \exp[\alpha_{2,t} + X'_{i,t}\beta_2] + \epsilon^2_{i,t} && P[\epsilon^2_{i,t} < 0 | X_{i,t}] = 0.9.
\end{aligned} \tag{19}$$

[16] Money market funds are mutual funds that are have many bank-like features. See STW for a description of the MMMF industry, and the similarities and differences between MMMF shares and bank accounts.

This specification allows for set of three (potentially unobservable) aggregate shocks, α_t, that allow the overall shape of the cross-section to change over time. Looking at the first row of (19), the model for the median is a linear panel median regression with time fixed-effects. This is STW's model for "common shocks" that hit all funds. $\alpha_{0,t}$ is a common aggregate shock, which shifts the median flow for all funds.

Now let's turn to the tails. An equivalent way to write (19) is

$$Y_{i,t} = X'_{i,t}\beta_0 + \alpha_{0,t} - D_{i,t}\exp[X'_{i,t}\beta_1 + \alpha_{1,t}]\eta_{i,t} + (1 - D_{i,t})\exp[X'_{i,t}\beta_2 + \alpha_{2,t}]\eta_{i,t}, \tag{20}$$

where $\eta_{i,t}$ is a nonnegative random variable with $P[\eta_{i,t} < 1|X_{i,t}] = 0.8$ and $D_{i,t}$ is a Bernoulli random variable which equals 1 with probability 0.5. Very little needs to be assumed about the idiosyncratic shock, $\eta_{i,t}$, except that it satisfies the above conditional quantile restriction, though STW argue that the data suggest that $\eta_{i,t}$ is well-approximated by a scaled exponential random variable. One way to think about this setup is that on each date we flip a coin to determine whether a fund gets hit with a "good shock" ($D_{it} = 0$) or a "bad shock" ($D_{it} = 1$). The standard deviation of the good shock is proportional to $\exp[\alpha_{1,t}+X'_{i,t}\beta_1]$ and the standard deviation of the bad shock is proportional to $\exp[\alpha_{2,t}+X'_{i,t}\beta_2]$. $\exp[\alpha_{1,t}]$ is an aggregate shock which scales up the standard deviation of the bad shock for all funds, while $\exp[\alpha_{2,t}]$ scales up the standard deviation for the good shock. As such, aggregate factors can independently affect the shape of the distribution for all funds, "lucky" funds, and "unlucky" funds, adding quite a bit of flexibility to the model.

Table 9 presents the estimated coefficients from this analysis. STW partition the sample into two subperiods and allow the coefficients on fund characteristics to change over each. The first is the "early crisis" period, 9/10-9/16, during which above average outflows began but prior to the announcement that the Reserve Primary Fund had "broken the buck".[17] While further discussion of these coefficient estimates may be found in the main text and online appendix of STW, we wish to highlight several insights which emerge from the estimation, many of which feature less prominently when one restricts analysis only to estimates of conditional means.

STW simulate from this dynamic model in order to characterize the effects of changing various fund characteristics on the distribution of cumulative outflows during the course of the week following the failure of Lehman Brothers. Due to the non-crossing property of the conditional

[17] In exchange for many regulatory requirements, MMMFs are allowed to round their share prices to the nearest half-cent. In practice, this means that the share prices of essentially all MMMFs were held fixed at $1.00 essentially always prior to September 17, 2008. At the end of the trading day on September 16th, the Reserve Primary Fund, which had owned Lehman Brothers commercial paper and suffered heavy outflows, suspended redemptions and notified investors that the share prices would be marked downwards. This procedure is referred to in the MMMF industry as "breaking the buck." New regulations, which will take effect at the end of 2016, are eliminating this structure.

Table 9: STW Panel Quantile Regression Coefficients

	Panel A Common (Median) Exposure		Panel B Left Tail Exposure		Panel C Right Tail Exposure	
Variable	Early Crisis	Peak Crisis	Early Crisis	Peak Crisis	Early Crisis	Peak Crisis
% Sophisticated$_{i,t-1}$	-0.0015 **	-0.0041 **	0.3004 ***	0.3364 **	0.1207	0.0322
	[0.024]	[0.028]	[0.008]	[0.033]	[0.326]	[0.320]
Average gross yield$_{i,t-1}$	-0.0007	-0.0055 ***	0.0820 *	0.0801	0.0144	0.1333
	[0.180]	[0.000]	[0.081]	[0.174]	[0.433]	[0.158]
Log flow std. dev.$_{i,t-1}$	-0.0007	-0.0047 **	0.5830 ***	0.4671 ***	0.4845 ***	0.4789 ***
	[0.199]	[0.030]	[0.000]	[0.004]	[0.000]	[0.002]
Log total fund assets$_{i,t-1}$	-0.0024 ***	-0.0095 ***	0.0180	0.1584	0.0070	0.2488 **
	[0.000]	[0.000]	[0.216]	[0.188]	[0.335]	[0.037]
$y_{i,t-1} - \bar{y}_{t-1} > 0$	0.1109	0.3274 **				
	[0.107]	[0.016]				
$y_{i,t-1} - \bar{y}_{t-1} < 0$	0.2627 ***	0.4527 ***				
	[0.004]	[0.002]				
$\lvert y_{i,t-1} - \bar{y}_{t-1} \rvert$			0.0556 *	-0.0139	0.0949 **	0.0763 *
			[0.089]	[0.493]	[0.043]	[0.063]
N	615	367	615	367	615	367
Pseudo-R^2 (50,10,90)	0.053	0.186	0.284	0.326	0.155	0.052

This table, which reproduces table C1 in STW, presents the coefficients from estimating equation (19) via quantile regression using the recursive method. The dependent variable ($y_{i,t}$) is the daily log difference in fund-level assets under management for prime institutional funds, in percentage points (i.e., $\times$ 100). Panel A, on the left, reports β_0, which controls the conditional median and shifts all quantiles symmetrically. Panel B, in the middle, reports β_1, which governs the width of the left tail (the distance between the median and the 10th percentile). Panel C, on the right, reports β_2, which controls the width of the right tail (the distance between the 90th percentile and the median). All three sets of coefficients are allowed to vary over two different periods in 2008: 9/10-9/16 Early Crisis and 9/17-9/19 Peak Crisis, respectively. More detailed variable descriptions may be found in Table A1 of STW. In addition to the coefficients in the table, models include time dummies to capture the common shocks, $\alpha_{0,t}$, $\alpha_{1,t}$, and $\alpha_{2,t}$. Numbers in brackets are one-sided bootstrapped p-values clustered at the fund level. With the exception of lagged flows, all variables are divided by their (cross-sectional) standard deviations.

quantile estimates, these simulations are always well-defined. Results are summarized in Table 6 below.

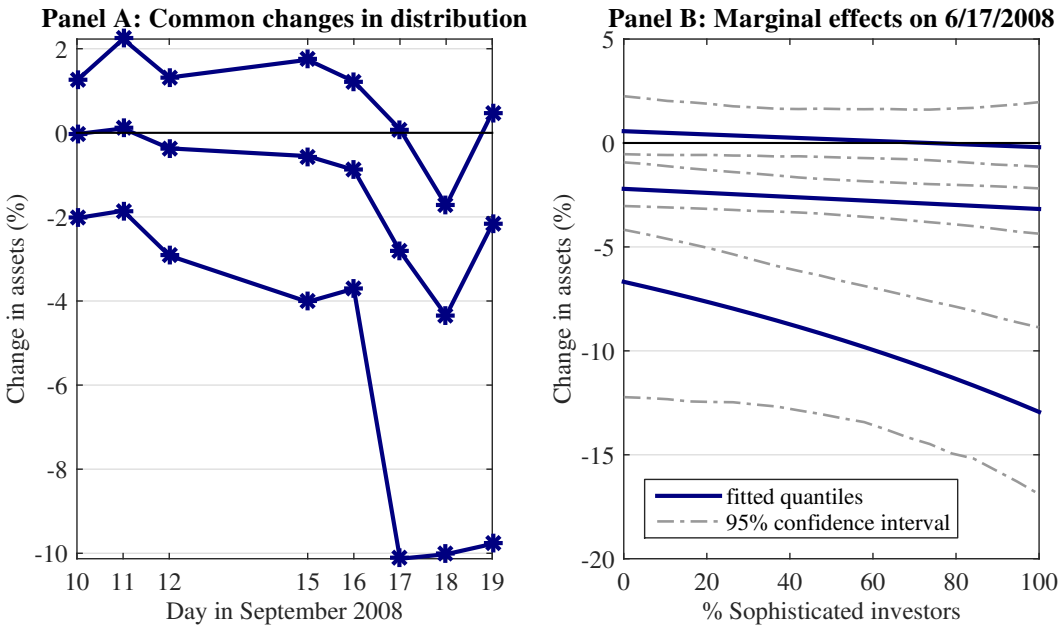


Figure 5: Level curves of conditional quantiles from dynamic model

This table, which is derived from the STW estimates presented in Table 9, plots the level curves of the fitted 10th, 50th, and 90th quantiles of daily percentage changes in assets under management on 9/17/2008–the day with the peak outflows during week following the Lehman failure–as a function of the fraction of sophisticated investors. All other variables are held fixed at their sample means, including lagged flows. See STW for full descriptions of each of the variables. Bootstrapped 95% confidence intervals are given by dashed lines.

Table 10: Marginal effects of fund characteristics on cumulative flow quantiles

		Cumulative Flow Quantile				
Variable	Value	1%	5%	10%	50%	90%
	$f(\bar{x})$	-52.02	-41.30	-35.81	-17.24	0.62
% Sophisticated	$f(\bar{x}+\sigma_x)$	-62.30	-50.29	-44.18	-21.96	-0.47
	$f(\bar{x}-\sigma_x)$	-42.22	-32.87	-28.19	-12.93	2.59
	Difference	-20.08 ***	-17.42 ***	-15.99 ***	-9.03 ***	-3.06
	p-value	[0.003]	[0.002]	[0.001]	[0.001]	[0.170]
	p-value vs. median	[0.020]	[0.019]	[0.018]	-	[0.046]
Average gross yield	$f(\bar{x}+\sigma_x)$	-55.04	-44.07	-38.51	-19.18	0.03
	$f(\bar{x}-\sigma_x)$	-48.92	-38.42	-33.04	-15.17	1.54
	Difference	-6.12 *	-5.65 **	-5.47 **	-4.02 **	-1.50
	p-value	[0.052]	[0.038]	[0.030]	[0.012]	[0.221]
	p-value vs. median	[0.181]	[0.173]	[0.167]	-	[0.144]
Log flow std. dev.	$f(\bar{x}+\sigma_x)$	-63.18	-50.45	-43.72	-19.36	10.86
	$f(\bar{x}-\sigma_x)$	-42.28	-33.14	-28.64	-14.42	-2.18
	Difference	-20.90 ***	-17.31 ***	-15.08 ***	-4.94 **	13.04 **
	p-value	[0.000]	[0.000]	[0.000]	[0.013]	[0.014]
	p-value vs. median	[0.000]	[0.000]	[0.000]	-	[0.000]
Log fund total assets	$f(\bar{x}+\sigma_x)$	-56.93	-45.97	-40.31	-20.57	-0.32
	$f(\bar{x}-\sigma_x)$	-46.90	-36.32	-31.08	-13.49	2.74
	Difference	-10.02 **	-9.64 **	-9.23 **	-7.08 ***	-3.06
	p-value	[0.042]	[0.019]	[0.011]	[0.001]	[0.163]
	p-value vs. median	[0.257]	[0.235]	[0.221]	-	[0.090]

This table, which reproduces Table 7 from STW, shows the impact of explanatory variables on cumulative flow distributions (as a percentage of initial assets) for prime institutional share classes (aggregated to the fund level) for the September 15-19 period. These estimates are obtained by simulating from an estimated dynamic quantile panel regression model for daily flows that is further described in an appendix. Columns report the 1st, 5th, 10th, 50th, and 90th quantiles of the cumulative flow distributions, respectively. STW begin by fixing each of the explanatory variables at its average, assuming that the initial value of lagged flows equals the prime institutional category average. Then, STW report the impact on the simulated flow distribution of adding and subtracting one standard deviation to each explanatory variable, as well as p-values for a test of whether the difference in the simulated quantiles is statistically significant, obtained by using the bootstrapped distribution of parameter estimates from our model, as well as the p-value of whether the marginal effect is significantly different at a given quantile, relative to the marginal effect at the median (using the bootstrapped distribution).

7 Conclusion

This paper proposes a simple but flexible parametric method for estimating multiple conditional quantiles. By construction, the estimated quantiles will satisfy the monotonicity requirement which must hold for any distribution by construction, so, in contrast to many benchmark methods, they are not susceptible to the well-known quantile crossing problem. Rather than directly modeling the level of each individual quantile, we begin with a single quantile (usually the median), and then add or subtract sums of nonnegative functions (quantile spacings) to obtain the other quantiles. Our approach is thus a natural extension of the location-scale paradigm that permits higher order moments (e.g., skewness and kurtosis) to vary. Two estimation methods are discussed in detail, and we characterize the limiting behavior of each, establishing consistency, asymptotic normality, and the validity of bootstrap inference. The latter method, under an additional "linear index" assumption, respects monotonicity but preserves the computational tractability of standard linear quantile regression. We propose a simple interpolation method which generates a mapping from a finite number of quantiles to a probability density function. Simulation exercises demonstrate that the estimators perform well in finite samples. Finally, three applications demonstrate the utility of the method in time-series (forecasting), cross-sectional, and panel settings.

A Proofs

Let $P_n = n^{-1}\sum_{i=1}^n \delta_{Z_i}$ be the empirical measure and $P_n^* = n^{-1}\sum_{i=1}^n \xi_i\delta_{Z_i}$ the weighted bootstrap probability measure, where δ_{Z_i} is the Dirac measure of Z_i and Z_i denotes the ith observation in the sample. We also define the empirical processes $\mathbb{G}_n = n^{1/2}(P_n - P)$ and $\mathbb{G}_n^* = n^{1/2}(P_n^* - P)$. Notice that for any function f, $\mathbb{G}_n^* f = \mathbb{G}_n \xi f$. We follow van der Vaart and Wellner (1996) and Kosorok (2007) and adopt the notations for empirical processes. For a class of functions $\mathcal{F}$, $\|\mathbb{G}_n\|_\mathcal{F} = \sup_{f\in\mathcal{F}}|\mathbb{G}_n f|$ and similarly $\|\mathbb{G}_n^*\|_\mathcal{F} = \sup_{f\in\mathcal{F}}|\mathbb{G}_n^* f| = \sup_{f\in\mathcal{F}}|\mathbb{G}_n \xi f|$. For a random variable X, $\|X\|_{P,r}$ denotes $(E_P|X|^r)^{1/r}$ with $r > 0$.

A.1 Proof of Theorem 1

We start by verifying the following property of the check function.

Lemma A.1. *Let $\tau \in (0,1)$, $a, b \in \mathbb{R}$ and $\psi_\tau(a) = \tau - \mathbf{1}_{\{a<0\}}$. Then $|\rho_\tau(a+b) - \rho_\tau(a) - b\psi_\tau(a)| \leq 2|b|\mathbf{1}_{\{|a|\leq|b|\}}$ and $|\rho_\tau(a+b) - \rho_\tau(a)| \leq 3|b|$.*

Proof. By the definition of $\rho_\tau(\cdot)$, we have

$$\rho_\tau(a+b) - \rho_\tau(a) - b\psi_\tau(a) = (a+b)\left(\mathbf{1}_{\{a<0\}} - \mathbf{1}_{\{a+b<0\}}\right).$$

When $\mathbf{1}_{\{a+b<0\}} \neq \mathbf{1}_{\{a<0\}}$, we have two possibilities: (1) $a+b<0$ and $a \geq 0$; (2) $a+b \geq 0$ and $a<0$. In case (1), $0 \leq a < -b$; in case (2), $-b \leq a < 0$. Thus, in both cases, $|a| \leq |b|$. It means that $|\mathbf{1}_{\{a+b<0\}} - \mathbf{1}_{\{a<0\}}| \leq \mathbf{1}_{\{|a|\leq|b|\}}$ and thus,

$$\begin{aligned} |\rho_\tau(a+b) - \rho_\tau(a) - b\psi_\tau(a)| &\leq |a+b|\left|\mathbf{1}_{\{a+b<0\}} - \mathbf{1}_{\{a<0\}}\right| \\ &\leq (|a|+|b|)\,\mathbf{1}_{\{|a|\leq|b|\}} \\ &\leq 2\,|b|\,\mathbf{1}_{\{|a|\leq|b|\}}. \end{aligned}$$

The first part follows. The second part holds by $|2\mathbf{1}_{\{|a|\leq|b|\}}| + |\psi_\tau(a)| \leq 3$. □

Proof of Theorem 1. By Example 3.2.22 of van der Vaart and Wellner (1996) or Theorem 2.13 of Kosorok (2007), it suffices to check the following conditions.

(a) $\|\hat{\theta}_{general} - \theta_0\| = o_P(1)$.

(b) For some function $\dot{H}(\cdot)$ with $\|\dot{H}\|_{P,2} < \infty$, we have that $|h_i(\theta_1) - h_i(\theta_2)| \leq \dot{H}(z_i)\|\theta_1 - \theta_2\|$ for $\theta_1, \theta_2 \in \Theta$.

(c) $E|(h_i(\theta) - h_i(\theta_0) - v_i(\theta_0)(\theta-\theta_0)|^2 = o(\|\theta-\theta_0\|^2)$ and $\|v_i(\theta_0)\|_{P,2} < \infty$.

Notice that claim (a) follows by Assumption 1 and Corollary 3.2.3 of van der Vaart and Wellner (1996). By Lemma A.1,

$$|h_i(\theta_1) - h_i(\theta_2)| \leq 3\sum_{j=1}^{p} |q_j(x_i;\theta_1) - q_j(x_i;\theta_2)| \leq \left(3\sum_{j=1}^{p} \sup_\theta \|\nabla_\theta q_j(x_i;\theta)\|\right)\|\theta_1 - \theta_2\|.$$

Since $E\sup_\theta \|\nabla_\theta q_j(x_i;\theta)\|^2 < \infty$ for each j, claim (b) follows. By Lemma A.1, we have

$$\begin{aligned} &\left|h_i(\theta) - h_i(\theta_0) - \sum_{j=1}^{p} \left(q_j(x_i;\theta_0) - q_j(x_i;\theta)\right)\psi_{\alpha_j}(y_i - q_j(x_i;\theta_0))\right| \\ &\qquad \leq 2\sum_{j=1}^{p} |q_j(x_i;\theta_0) - q_j(x_i;\theta)|\,\mathbf{1}\left\{|y_i - q_j(x_i;\theta_0)| \leq |q_j(x_i;\theta_0) - q_j(x_i;\theta)|\right\} \end{aligned}$$

Let $k_{i,j}(\theta) = q_j(x_i;\theta) - q_j(x_i;\theta_0)$. Since $|k_{i,j}(\theta)| \leq \sup_\theta \|\nabla_\theta q_j(x_i;\theta)\|\|\theta - \theta_0\|$, $\|\sup_\theta \|\nabla_\theta q_j(x_i;\theta)\|\|_{P,2+\delta} < \infty$ and $y_i - q_j(x_i;\theta_0)$ has bounded p.d.f around zero, Holder's

inequality implies that $E|k_{i,j}(\theta)\mathbf{1}\{|y_i - q_j(x_i;\theta_0)| \le k_{i,j}(\theta)\} = o(\|\theta-\theta_0\|^2)$, meaning that

$$E\left|h_i(\theta) - h_i(\theta_0) - \sum_{j=1}^{p}\left(q_j(x_i;\theta_0) - q_j(x_i;\theta)\right)\psi_{\alpha_j}(y_i - q_j(x_i;\theta_0))\right|^2 = o(\|\theta-\theta_0\|^2). \quad (21)$$

Also by the boundedness of $\psi_\alpha(\cdot)$, we have

$$\left\|\sum_{j=1}^{p}\left(q_j(x_i;\theta_0) - q_j(x_i;\theta)\right)\psi_{\alpha_j}(y_i - q_j(x_i;\theta_0)) - v_i(\theta_0)(\theta-\theta_0)\right\|_{P,2}$$
$$\le \sum_{j=1}^{p}\|q_j(x_i;\theta) - q_j(x_i;\theta_0) - \nabla_\theta q_j(x_i;\theta_0)(\theta-\theta_0)\|_{P,2} = o(\|\theta-\theta_0\|). \quad (22)$$

By (21) and (22), $E|(h_i(\theta) - h_i(\theta_0) - v_i(\theta_0)(\theta-\theta_0)|^2 = o(\|\theta-\theta_0\|^2)$. Notice that $\|v_i(\theta_0)\|_{P,2} \le \sum_{j=1}^{p}\|\nabla_\theta q_j(x_i;\theta_0)\|_{P,2} < \infty$. Claim (c) follows. The proof is complete. □

A.2 Proof of Theorem 2 and Corollary 1

We impose the following regularity condition for Theorem 2 and Corollary 1.

Assumption 2. *For $j \in \{j_*, \cdots, p-1\}$, define $\beta = (\theta_{j^*}, \cdots, \theta_j)$ and $\gamma = \theta_{j+1}$, along with their pseudo-true values $\beta_* = (\theta_{j^*,*}, \cdots, \theta_{j,*})$ and $\gamma_* = \theta_{j+1,*}$. Let $\Gamma = \Theta_{j+1}$ and $B = \Theta_{j^*} \times \cdots \times \Theta_j$. Define $\varepsilon_i(\beta) = y_i - q_{U,j}(x_i;\theta_{j^*}, \cdots, \theta_j)$ and $m_i(\gamma,\beta) = \rho_\tau\left(g_j^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)\right)\mathbf{1}_{\{\varepsilon_i(\beta)>0\}}$, $w_i(\gamma) = f_{j+1}(x_i;\gamma)$ and $\tau = (\alpha_{j+1}-\alpha_j)/(1-\alpha_j)$. Suppose that the following hold:*
(i) $\Gamma \times B$ is a compact set.
(ii) There exists a constant $C_1 > 0$ such that $\forall(\gamma,\beta) \in \Gamma \times B$, $\varepsilon_i(\beta)$, $\|w_i(\gamma)\|$, $\|\nabla_\gamma w_i(\gamma)\|$ and $\|\nabla_\beta \varepsilon_i(\beta)\|$ lie in $[-C_1, C_1]$ with probability one and the p.d.f's of $\varepsilon_i(\beta)$ and of $\varepsilon_i(\beta) - g(w_i(\gamma))$ are bounded by C_1.
(iii) There exists some constant $\lambda > 2$ such that $\sup_{(\gamma,\beta)\in\Gamma\times B}\|\rho_\tau(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma))\mathbf{1}\{\varepsilon_i(\beta) > 0\}\|_{P,\lambda} < \infty$.
(iv) $M(\gamma,\beta)$ is twice continuously differentiable in (γ,β) over $\Gamma \times B$ and $V_* = \nabla_{\gamma\gamma}M(\gamma_*,\beta_*)$ is nonsingular, where $M(\gamma,\beta) = Em_i(\gamma,\beta)$.
(v) $\int_{-1}^{1}|g_j^{-1}(|x|)|^q dx < \infty$ for some $q > 1$.

We impose analogous conditions for $j \in \{1, \cdots, j_* - 1\}$.

In the rest of the paper, we maintain the following notations:

$$
\begin{aligned}
&m_i(\gamma,\beta) = \rho_\tau\left(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)\right)\mathbf{1}\{\varepsilon_i(\beta) > 0\}\\
&\dot{m}_i(\gamma,\beta) = \nabla_\gamma w_i(\gamma)\psi_\tau\left(\varepsilon_i(\beta) - g(w_i(\gamma))\right)\mathbf{1}\{\varepsilon_i(\beta) > 0\}\\
&R_i(\gamma,\beta) = m_i(\gamma,\beta) - m_i(\gamma_*,\beta) - (\gamma-\gamma_*)'\dot{m}_i(\gamma_*,\beta)\\
&m_{(n),i}(\gamma,\beta) = \rho_\tau\left(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)\right)\mathbf{1}\{\varepsilon_i(\beta) > a_n\}\\
&\dot{m}_{(n),i}(\gamma,\beta) = \nabla_\gamma w_i(\gamma)\psi_\tau\left(\varepsilon_i(\beta) - g(w_i(\gamma))\right)\mathbf{1}\{\varepsilon_i(\beta) > a_n\}\\
&R_{(n),i}(\gamma,\beta) = m_{(n),i}(\gamma,\beta) - m_{(n),i}(\gamma_*,\beta) - (\gamma-\gamma_*)'\dot{m}_{(n),i}(\gamma_*,\beta)\\
&M(\gamma,\beta) = Em_i(\gamma,\beta)\\
&M_n(\gamma,\beta) = Em_{(n),i}(\gamma,\beta)\\
&\hat{M}_n(\gamma,\beta) = n^{-1}\sum_{i=1}^n m_{(n),i}(\gamma,\beta)\\
&\hat{M}^*_n(\gamma,\beta) = n^{-1}\sum_{i=1}^n \xi_i m_{(n),i}(\gamma,\beta),
\end{aligned}
\tag{23}
$$

Empirical processes of functions m_i, $\dot{m}_i$, R_i, $m_{(n),i}$, $\dot{m}_{(n),i}$ and $R_{(n),i}$ will be denoted without the subscript i. For example, $\mathbb{G}_n R(\gamma,\beta) = n^{-1/2}\sum_{i=1}^n[R_i(\gamma,\beta) - ER_i(\gamma,\beta)]$ and $\mathbb{G}_n\xi R_{(n)}(\gamma,\beta) = n^{-1/2}\sum_{i=1}^n[\xi_i R_{(n),i}(\gamma,\beta) - ER_{(n),i}(\gamma,\beta)]$, etc. We first establish auxilary results and then prove Theorem 2 and Corollary 1 at the end of this section.

Condition 1. There exists constants $A_1, A_2, A_3, A_4 \in (0,\infty)$ and $\lambda \in (1,\infty]$ such that the following hold:

(1) $\forall(\gamma,\beta) \in \Gamma\times B$, $\varepsilon_i(\beta)$, $\|\nabla_\beta\varepsilon_i(\beta)\|$, $\|w_i(\gamma)\|$ and $\|\nabla_\gamma w_i(\gamma)\|$ lie in $[-A_1, A_1]$ with probability one and the p.d.f of $\varepsilon_i(\beta)$ is bounded by A_1.

(2) $\forall(\gamma,\beta) \in \Gamma\times B$, the p.d.f of $\varepsilon_i(\beta) - g(w_i(\gamma))$ is bounded by A_2.

(3) $\forall(\gamma,\beta) \in \Gamma\times B$, $\|\gamma\| \le A_3$ and $\|\beta\| \le A_3$.

(4) $\sup_{(\gamma,\beta)\in\Gamma\times B}\|\rho_\tau(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma))\mathbf{1}\{\varepsilon_i(\beta) > 0\}\|_{P,2\lambda} \le A_4$.

Lemma A.2. *There exists $\lambda > 1$ such that $\forall c > 0$, $\int_{-c}^{c}|g^{-1}(|x|)|^\lambda dx < \infty$. If Condition 1 holds, then $\sup_{(\gamma,\beta)\in\Gamma\times B}|M_n(\gamma,\beta) - M(\gamma,\beta)| = o(1)$.*

Proof. Notice that $M_n(\gamma,\beta) - M(\gamma,\beta) = \|\delta_{n,t}(\gamma,\beta)\|_{P,1}$, where $\delta_{n,t}(\gamma,\beta) = \rho_\tau\left(g^{-1}(|\varepsilon_i(\beta)|) - w_i(\gamma)\right)\mathbf{1}\{0 < \varepsilon_i(\beta) \le a_n\}$. By Holder's inequality and the fact that

$\rho_\tau(x) \leq |x|$,

$$\begin{aligned}
\|\delta_{n,t}(\gamma,\beta)\|_{P,1} &\leq \|g^{-1}(|\varepsilon_i(\beta)|) - w_i(\gamma)\|_{P,\lambda} \|\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\}\|_{P,\pi} \\
&\leq \left(\|g^{-1}(|\varepsilon_i(\beta)|)\|_{P,\lambda} + \|w_i(\gamma)\|_{P,\lambda}\right) \|\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\}\|_{P,\pi} \\
&\leq (Ca_n)^{1/\pi} \left(\|g^{-1}(|\varepsilon_i(\beta)|)\|_{P,\lambda} + \|w_i(\gamma)\|_{P,\lambda}\right), \qquad (24)
\end{aligned}$$

where the last line follows by the bounded p.d.f of $\varepsilon_i(\beta)$ and $C > 0$ is a constant that upper bounds the p.d.f of $\varepsilon_i(\beta)$. Under Condition 1, $\|g^{-1}(|\varepsilon_i(\beta)|)\|_{P,\lambda}^{\lambda} = E|g^{-1}(|\varepsilon_i(\beta)|)|^{\lambda} \leq \int_{-A_1}^{A_1} |g^{-1}(|x|)|^{\lambda} A_1 dx < M$ for some constant M that does not depend on β. Since w_i has bounded support and Γ is compact, $\|w_i(\gamma)\|_{P,\lambda}$ is bounded by a finite constant. By (24) and $a_n \to 0$, $\sup_{\gamma,\beta} \|\delta_{n,t}(\gamma,\beta)\|_{P,1} \to 0$. The result follows. □

Lemma A.3. *Let Condition 1 hold. Suppose that $a_n = o(n^{-1/2})$, $\tilde{\beta} - \beta_* = O_P(n^{-1/2})$ and w_i has bounded support and the smallest eigenvalue of $\nabla_{\gamma\gamma} M(\gamma_*, \beta_*)$ is positive. Then, for any $d_n = o_P(1)$,*

$$n\left[M_n(\gamma_* + d_n, \tilde{\beta}) - M_n(\gamma_*, \tilde{\beta})\right] = o(n\|d_n\|^2) + o_P(n^{1/2}\|d_n\|) + nd_n' \left(D_{\gamma\beta}(\gamma_*, \beta_*)(\tilde{\beta} - \beta_*)\right) + nd_n' \Omega(\gamma_*, \beta_*) d_n/2,$$

where $\Omega(\gamma, \beta) = \nabla_{\gamma\gamma} M(\gamma, \beta)$.

Proof. Let $d_n = o_P(1)$ and $\delta_{n,t}(\gamma,\beta) = m_{(n),i}(\gamma,\beta) - m_i(\gamma,\beta)$. Notice that $\delta_{n,t}(\gamma,\beta) = \rho_\tau\left(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)\right)\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\}$ and for large n,

$$\begin{aligned}
\delta_{n,t}(\gamma_1,\beta) - \delta_{n,t}(\gamma_2,\beta) &= \left[\rho_\tau\left(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma_1)\right) - \rho_\tau\left(g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma_2)\right)\right]\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\} \\
&= \left(w_i(\gamma_2) - w_i(\gamma_1)\right)\psi_\tau\left(\varepsilon_i(\beta) - g(w_i(\gamma_2))\right)\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\} \\
&\quad + \left[\mathbf{1}\{g(w_i(\gamma_1)) \leq \varepsilon_i(\beta) \leq g(w_i(\gamma_2))\} + \mathbf{1}\{g(w_i(\gamma_2)) \leq \varepsilon_i(\beta) \leq g(w_i(\gamma_1))\}\right]\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\} \\
&= \left(w_i(\gamma_2) - w_i(\gamma_1)\right)\psi_\tau\left(\varepsilon_i(\beta) - g(w_i(\gamma_2))\right)\mathbf{1}\{0 < \varepsilon_i(\beta) \leq a_n\},
\end{aligned}$$

where the last line follows by noticing that $g(w_i(\gamma))$ is bounded below by a positive constant and $a_n = o(1)$. Since w_i and $\|\nabla_\gamma w_i(\gamma)\|$ have bounded support and $\varepsilon_i(\beta)$ has bounded p.d.f, it follows that, for large n,

$$\begin{aligned}
&\sup_{\|\gamma_1 - \gamma_2\| \leq \eta, \beta \in B} \left|n\left[M_n(\gamma_1,\beta) - M(\gamma_1,\beta)\right] - n\left[M_n(\gamma_2,\beta) - M(\gamma_2,\beta)\right]\right| \\
&\qquad \leq \sup_{\|\gamma_1 - \gamma_2\| \leq \eta, \beta \in B} E|\delta_{n,t}(\gamma_1,\beta) - \delta_{n,t}(\gamma_2,\beta)| \leq C_0 a_n \eta, \qquad (26)
\end{aligned}$$

where $C_0 > 0$ is a constant. Hence,

$$\begin{aligned} n\left[M_n(\gamma_* + d_n, \tilde{\beta}) - M_n(\gamma_*, \tilde{\beta})\right] &= n\left[M_n(\gamma_* + d_n, \tilde{\beta}) - M(\gamma_* + d_n, \tilde{\beta})\right] - n\left[M_n(\gamma_*, \tilde{\beta}) - M(\gamma_*, \tilde{\beta})\right] \\ &\quad + n\left[M(\gamma_* + d_n, \tilde{\beta}) - M(\gamma_*, \tilde{\beta})\right] \\ &= o_P(n^{1/2}\|d_n\|) + n\left[M(\gamma_* + d_n, \tilde{\beta}) - M(\gamma_*, \tilde{\beta})\right], \end{aligned} \tag{27}$$

where the last line follows by (26) and $n^{1/2}a_n = o(1)$. Let $D_\gamma(\gamma, \beta) = \nabla_\gamma M(\gamma, \beta)$ and $D_{\gamma\beta}(\gamma, \beta) = \nabla_{\gamma\beta} M(\gamma, \beta)$. By Taylor's theorem,

$$\begin{aligned} n\left[M(\gamma_* + d_n, \tilde{\beta}) - M(\gamma_*, \tilde{\beta})\right] &= nd_n' D_\gamma(\gamma_*, \tilde{\beta}) + \frac{n}{2} d_n' \Omega(\tilde{\gamma}_n, \tilde{\beta}) d_n \\ &= nd_n'\left(D_\gamma(\gamma_*, \beta_*) + D_{\gamma\beta}(\gamma_*, \dot{\beta}_n)(\tilde{\beta} - \beta_*)\right) + \frac{n}{2} d_n' \Omega(\tilde{\gamma}_n, \tilde{\beta}) d_n \quad (28) \\ &\overset{(i)}{=} o_P(n^{1/2}\|d_n\|) + nd_n'\left(D_{\gamma\beta}(\gamma_*, \beta_*)(\tilde{\beta} - \beta_*)\right) + \frac{n}{2} d_n' \Omega(\tilde{\gamma}_n, \beta_*) d_n \quad (29) \end{aligned}$$

where $\tilde{\gamma}_n = \gamma_* + \alpha_n d_n$ for some $\alpha_n \in [0, 1]$ and $\dot{\beta}_n = \beta_* + b_n(\tilde{\beta} - \beta_*)$ for some $b_n \in [0, 1]$. Here, (i) follows by $D_\gamma(\gamma_*, \beta_*) = 0$, $D_{\gamma\beta}(\gamma_*, \dot{\beta}_n) = D_{\gamma\beta}(\gamma_*, \beta_*) + o_P(1)$ and $\tilde{\beta} - \beta_* = O_P(n^{-1/2})$. Hence, the desired result follows by (27), (29) and $\Omega(\tilde{\gamma}_n, \beta_*) = \Omega(\gamma_*, \beta_*) + o_P(1)$. □

Lemma A.4. *Let $\hat{Q}_n(\theta)$ and $Q_n(\theta)$ be two stochastic processes on Θ_n. Let $\hat{\theta}$ and θ_n satisfy $\hat{Q}_n(\hat{\theta}) \le \inf_{\theta\in\Theta_n} \hat{Q}_n(\theta) - o_P(n^{-1})$ and $Q_n(\theta_n) \le \inf_{\theta\in\Theta_n} Q_n(\theta) - o_P(n^{-1})$. Suppose that there exist $Z_n = O_P(1)$ and $V_n = O_P(1)$ such that for any $d_n = o_P(1)$, $n\left[\hat{Q}_n(\theta_n + d_n) - \hat{Q}_n(\theta_n)\right] = o_P(n\|d_n\|^2) + o_P(n^{1/2}\|d_n\|) + o_P(1) + n^{1/2}d_n' Z_n + nd_n' V_n d_n/2$. Suppose that there exists a constant $c > 0$ such that $P(\lambda_{\min}(V_n) > c) \to 1$. If $\|\hat{\theta} - \theta_n\| = o_P(1)$, then $\hat{\theta} - \theta_n = -n^{-1/2}V_n^{-1}Z_n + o_P(n^{-1/2})$.*

Proof. Let $h_n = \hat{\theta} - \theta_n$. Since $P(\lambda_{\min}(V_n) > c) \to 1$, $nh_n' V_n h_n \ge (o_P(1) + c/2)\|h_n\|^2$ with probability approaching one. Hence, $n^{1/2}h_n' Z_n + nh_n' V_n h_n/2 \ge (o_P(1) + c/4)\|h_n\|^2 + n^{1/2}O_P(\|h_n\|)$ with probability approaching one. Notice that $o_P(n\|h_n\|^2) + o_P(n^{1/2}\|h_n\|) + o_P(1) + n^{1/2}h_n' Z_n + nh_n' V_n h_n/2 = n\left[\hat{Q}_n(\theta_n + h_n) - \hat{Q}_n(\theta_n)\right] \le o_P(1)$. It follows that, with probability approaching one, $o_P(1) \ge (o_P(1) + c/4)\|h_n\|^2 + n^{1/2}O_P(\|h_n\|)$. Completing the square, we have $(c/4 + o_P(1))\left(n^{1/2}\|h_n\| + O_P(1)\right)^2 \le O_P(1)$. Hence, $h_n = O_P(n^{-1/2})$.

Let $\tilde{h}_n = -n^{-1/2}V_n^{-1}Z_n$. By assumption, $\tilde{h}_n = O_P(n^{-1/2})$ and hence, $n\left[\hat{Q}_n(\theta_n + \tilde{h}_n) - \hat{Q}_n(\theta_n)\right] = -Z_n' V_n^{-1} Z_n/2 + o_P(1)$. Since $h_n = O_P(n^{-1/2})$, $n\left[\hat{Q}_n(\theta_n + h_n) - \hat{Q}_n(\theta_n)\right] = o_P(1) + n^{1/2}h_n' Z_n + nh_n' V_n h_n/2$. Notice that $n\left[\hat{Q}_n(\theta_n + \tilde{h}_n) - \hat{Q}_n(\theta_n)\right] + o_P(1) \ge n\left[\hat{Q}_n(\theta_n + h_n) - \hat{Q}_n(\theta_n)\right]$. It follows that $o_P(1) \ge 2n^{1/2}h_n' Z_n + nh_n' V_n h_n + Z_n' V_n^{-1} Z_n$. Since $2n^{1/2}h_n' Z_n + nh_n' V_n h_n + Z_n' V_n^{-1} Z_n =$

$\left(n^{1/2}V_n h_n + Z_n\right)' V_n^{-1} \left(n^{1/2}V_n h_n + Z_n\right)$, we have $n^{1/2}V_n h_n + Z_n = o_P(1)$. The desired result follows. □

Lemma A.5. *If $\mathcal{F}$ is uniformly bounded by a constant K and $\mathcal{G}$ is uniformly bounded by one, then $N_{[]}\left(2K\varepsilon, \mathcal{F}\cdot\mathcal{G}, L_r(Q)\right) \leq N_{[]}\left(K\varepsilon, \mathcal{F}, L_r(Q)\right) N_{[]}\left(\varepsilon, \mathcal{G}, L_r(Q)\right)$.*

Proof. Notice that $N_{[]}\left(2K\varepsilon, \mathcal{F}\cdot\mathcal{G}, L_r(Q)\right) = N_{[]}\left(2\varepsilon, (\mathcal{F}/K)\cdot\mathcal{G}, L_r(Q)\right)$ and $\mathcal{F}/K$ is uniformly bounded by one. The result follows by Lemma 9.25 of Kosorok (2007). □

Lemma A.6. *For $\delta_1, \delta_2 > 0$, let $\Gamma_1 = \{\gamma \mid \|\gamma - \gamma_*\| \leq \delta_1\}$, $B_2 = \{\beta \mid \|\beta - \beta_*\| \leq \delta_2\}$, $\mathcal{F}_n = \{f_{n,t}(\gamma,\beta) \mid (\gamma,\beta) \in \Gamma_1 \times B_2\}$ and $\mathcal{K}_n = \{\mathbf{1}\{h_{n,t}(\beta) > 0\} \mid \beta \in B_2\}$. Suppose that the following hold:*
(i) There exists a constant $C_1 > 0$ such that $\forall(\gamma_1,\beta_1),(\gamma_2,\beta_2) \in \Gamma_1 \times B_2$, we have $|f_{n,t}(\gamma_1,\beta_1) - f_{n,t}(\gamma_2,\beta_2)| \leq C_1\|\gamma_1 - \gamma_2\| + s_n C_1\|\beta_1 - \beta_2\|$ and $|h_{n,t}(\beta_1) - h_{n,t}(\beta_2)| \leq C_1\|\beta_1 - \beta_2\|$, where the sequence s_n is positive.
(ii) $f_{n,t}(\gamma,\beta) \geq 0$ $\forall(\gamma,\beta) \in \Gamma_1 \times B_2$.
(iii) There exist constants $C_2 > 0$ such that the p.d.f of $h_{n,t}(\beta)$ is bounded by C_2.
(iv) There exist constants $C_3 \in (0,\infty)$ and $\lambda \in (1,\infty]$ with $\sup_{(\gamma,\beta)\in\Gamma_1\times B_2} \|f_{n,t}(\gamma,\beta)\|_{P,2\lambda} \leq C_3$.

Then there exists constants $M_0, M_1, M_2, \pi > 0$ depending only on $\dim\gamma$, $\dim\beta$ and the constants above, such that $\lambda^{-1} + \pi^{-1} = 1$ and $\forall z > 0$,

$$N_{[]}\left(z, \mathcal{F}_n\cdot\mathcal{K}_n, L_2(P)\right) \leq M_0\left[\left(\delta_1 z^{-1}\right)^{M_1} \vee 1\right]\left[\left(s_n\delta_2 z^{-1}\right)^{M_2} \vee 1\right]\left[\left(\delta_2 z^{-2\pi}\right)^{M_2} \vee 1\right].$$

Proof. Fix an arbitrary $z > 0$. Define $x_1 = z/(6C_1)$, $x_2 = z/(6C_1 s_n)$ and $x_3 = z^{2\pi}(C_1C_2)^{-1}(3C_3)^{-2\pi}$. Let $\{\gamma_j \mid j = 1,\cdots,n_{\gamma,x_1}\}$, $\{\beta_j \mid j = 1,\cdots,n_{\beta,x_2}\}$ and $\{\bar\beta_j \mid j = 1,\cdots,n_{\beta,x_3}\}$ be an x_1-net in Γ_1, an x_2-net in B_2 and an x_3-net in B_2, respectively. Then the brackets $\{[f_{n,t,L,j_1,j_2} \vee 0, f_{n,t,U,j_1,j_2}] \mid j_1 = 1,\cdots,n_{\gamma,x_1}\ j_2 = 1,\cdots,n_{\beta,x_2}\}$ cover $\mathcal{F}_n$, where $f_{n,t,U,j_1,j_2} = f_{n,t}(\gamma_{j_1},\beta_{j_2}) + x_1C_1 + x_2C_1$ and $f_{n,t,L,j_1,j_2} = f_{n,t}(\gamma_{j_1},\beta_{j_2}) - x_1C_1 - x_2C_1$. Also the brackets $\{[k_{n,t,L,j}, k_{n,t,U,j}] \mid j = 1,\cdots,n_{\beta,x_3}\}$ cover $\mathcal{K}_n$, where $k_{n,t,U,j} = \mathbf{1}\{h_{n,t}(\bar\beta_j) + x_3C_1 > 0\}$ and $k_{n,t,L,j} = \mathbf{1}\{h_{n,t}(\bar\beta_j) - x_3C_1 > 0\}$. Notice that $\|k_{n,t,U,j} - k_{n,t,L,j}\|_{P,2\pi} \leq \|\mathbf{1}\{|h_t(\bar\beta_j)| \leq C_1x_3\}\|_{P,2\lambda} \leq (C_2C_1x_3)^{1/(2\pi)}$, where the last inequality follows by the bounded p.d.f of $h_{n,t}(\beta)$.

Then the brackets $\{[(f_{n,t,L,j_1,j_2} \vee 0)k_{n,t,L,j_3}, f_{n,t,U,j_1,j_2}k_{n,t,U,j_3}]\}_{j_1,j_2,j_3}$ cover $\mathcal{F}_n\cdot\mathcal{K}_n$. Moreover,

$$\begin{aligned}
&\|f_{n,t,U,j_1,j_2}k_{n,t,U,j_3} - (f_{n,t,L,j_1,j_2} \vee 0)k_{n,t,L,j_3}\|_{P,2} \\
&\leq \|f_{n,t,U,j_1,j_2} - (f_{n,t,L,j_1,j_2} \vee 0)\|_{P,2} + \|(f_{n,t,L,j_1,j_2} \vee 0)(k_{n,t,U,j_3} - k_{n,t,L,j_3})\|_{P,2}
\end{aligned}$$

$$\overset{(i)}{\leq} \|f_{t,L,j_1} - (f_{n,t,L,j_1,j_2} \vee 0)\|_{P,2} + \|f_{n,t,L,j_1,j_2} \vee 0\|_{P,2\lambda} \|k_{n,t,U,j_3} - k_{n,t,L,j_3}\|_{P,2\pi}$$
$$\overset{(ii)}{\leq} \|f_{t,L,j_1} - f_{n,t,L,j_1,j_2}\|_{P,2} + \|f_{n,t}(\gamma_{j_1}, \beta_{j_2})\|_{P,2\lambda} \|k_{n,t,U,j_3} - k_{n,t,L,j_3}\|_{P,2\pi}$$
$$\overset{(iii)}{\leq} 2x_1 C_1 + 2s_n x_2 C_1 + C_3 (C_1 C_2 x_3)^{1/(2\pi)} \overset{(iv)}{\leq} z,$$

where (i) follows by Holder's inequality and (ii) follows by $|f_{t,L,j_1} - (f_{n,t,L,j_1,j_2} \vee 0)| \leq |f_{t,L,j_1} - f_{n,t,L,j_1,j_2}|$ and $|f_{n,t,L,j_1,j_2} \vee 0| \leq |f_{n,t}(\gamma_{j_1}, \beta_{j_2})|$, (iii) follows by $\|f_{n,t}(\gamma_{j_1}, \beta_{j_2})\|_{P,2\lambda} \leq C_3$ and $\|k_{n,t,U,j} - k_{n,t,L,j}\|_{P,2\pi} \leq (C_2 C_1 x_3)^{1/(2\pi)}$ and (iv) follows by the definitions of x_1, x_2 and x_3. Hence, $N_{[]}(z, \mathcal{F}_n \cdot \mathcal{K}_n, L_2(P)) \leq n_{\gamma,x_1} n_{\beta,x_2} n_{\beta,x_3}$. Notice that $n_{\gamma,x_1} \leq K_1(\delta_1 x_1^{-1})^{K_2} \vee 1$ and $n_{\beta,x_2} \leq K_3(\delta_2 x_2^{-1})^{K_4} \vee 1$. The desired result follows by plugging in the definitions of x_1, x_2 and x_3. □

Lemma A.7. *Suppose that $\forall \theta_1, \theta_2 \in \Theta$, $|f_t(\theta_1) - f_t(\theta_2)| \leq C_1 \|\theta_1 - \theta_2\|$. Let $C_2, C_3 \in (0, \infty)$ be constants such that $\forall \theta \in \Theta$, the p.d.f of $f_t(\theta)$ is bounded by C_2 and the diameter of Θ is bounded by C_3. Then $N_{[]}(x, \mathcal{F}, L_2(P)) \leq M_1(C_2 x^{-2})^{M_2} \vee 1$, where $\mathcal{F} = \{\mathbf{1}\{f_t(\theta) > 0\} \mid \theta \in \Theta\}$ and $M_1, M_2 > 0$ are constants depending only on C_1, C_2 and $\dim \theta$. The same conclusion holds if $\mathcal{F}$ is replaced by $\mathcal{F}' = \{\mathbf{1}\{f_t(\theta) \geq 0\} \mid \theta \in \Theta\}$.*

Proof. Let $\{\theta_j\}_{j=1}^{n_z}$ be a z-net in Θ. Then the brackets $\{[\mathbf{1}\{f_{t,L,j} > 0\}, \mathbf{1}\{f_{t,U,j} > 0\}]\}_{j_1,j_2}$ cover $\mathcal{F}$, where $f_{t,L,j} = f_t(\theta_j) - C_1 z$ and $f_{t,U,j} = f_t(\theta_j) + C_1 z$. Notice that $\|\mathbf{1}\{f_{t,L,j} > 0\} - \mathbf{1}\{f_{t,U,j} > 0\}\|_{P,2} = \|\mathbf{1}\{|f_t(\theta_j)| \leq C_1 z\}\|_{P,2} \leq \sqrt{C_1 C_2 z}$. Hence, $N_{[]}(\sqrt{C_1 C_2 z}, \mathcal{F}, L_2(P)) \leq n_z \leq K_1(C_3 z^{-1})^{K_2} \vee 1$ for some constants $K_1, K_2 > 0$ depending only on $\dim \theta$. The result follows by a change of variable. The same argument applies to $\mathcal{F}'$. □

Lemma A.8. *Let Condition 1 hold. Let $\mathcal{F}_n(\delta) = \{R_{(n),i}(\gamma_* + h_1, \beta_* + h_2) \mid \|h_1 + h_2\| \leq \delta\}$. Then there exist constants $M_1, M_2, M_3, M_4, M_5, M_6 > 0$ depending only on $\dim \beta$, $\dim \gamma$ and the constants in Condition 1 such that $\forall z > 0$,*

$$N_{[]}(z, \mathcal{F}_n(\delta), L_2(P)) \leq \left[M_1 \left(\delta z^{-1}\right)^{M_2} \vee 1\right] \left[M_3 \left(\delta z^{-2}\right)^{M_4} \vee 1\right] \left[M_5 \left(\delta z^{-2}\right)^{M_6} \vee 1\right].$$

Proof. Simple computations yield $R_{(n),i}(\gamma, \beta) = R_i(\gamma, \beta)\mathbf{1}\{\varepsilon_i(\beta) > a_n\}$ and

$$R_i(\gamma, \beta) = \left|g^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)\right| \left[\mathbf{1}\{g(w_i(\gamma)) \leq \varepsilon_i(\beta) < g(w_i(\gamma))\} + \mathbf{1}\{g(w_i(\gamma)) \leq \varepsilon_i(\beta) < g(w_i(\gamma))\}\right]. \tag{30}$$

Notice that for large n, $R_{(n),i}(\cdot, \cdot) = R_i(\cdot, \cdot)$. Hence, in the rest of the proof, we prove the result for $R_i(\gamma, \beta)$ and $\mathcal{F}(\delta) = \{R_i(\gamma_* + h_1, \beta_* + h_2) \mid \|h_1 + h_2\| \leq \delta\}$.

Notice that in the expression for R_i in (30), we can replace $g^{-1}(\cdot)$ with its truncated version,

$g_c^{-1}(\cdot)$, where $g_c^{-1}(x) = g^{-1}((x \vee C_1) \wedge C_2))$ and $[C_1, C_2]$ is a finite length interval that contains the support of $w_i(\gamma)$ for each $\gamma \in \Gamma$. Observe that there exists $C_0 > 0$ depending only on C_1 and C_2 such that $\forall x_1, x_2 \in \mathbb{R}$, $|g_c^{-1}(x_1) - g_c^{-1}(x_2)| \leq C_0|x_1 - x_2|$.

Let $\mathcal{F}_1(\delta) = \{f_t(\gamma_* + h_1, \beta_* + h_2) \mid \|h_1 + h_2\| \leq \delta\}$, $\mathcal{F}_2(\delta) = \{\mathbf{1}\{f_t(\gamma_* + h_1, \beta_* + h_2) \leq 0\} \mid \|h_1 + h_2\| \leq \delta\}$ and $\mathcal{F}_3(\delta) = \{\mathbf{1}\{f_t(\gamma_*, \beta_* + h_2) > 0\} \mid \|h_1 + h_2\| \leq \delta\}$, where $f_t(\gamma, \beta) = g_c^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)$. Notice that $|f_t(\gamma_* + h_{1,1}, \beta_* + h_{2,1}) - f_t(\gamma_* + h_{1,2}, \beta_* + h_{2,2})| \leq C_3(\|h_{1,1} - h_{1,2}\| + \|h_{2,1} - h_{2,2}\|)$, where C_3 is a constant depending only on C_0 and the constants in Condition 1. By Theorem 2.7.11 of van der Vaart and Wellner (1996), $N_{[]}(2C_3 z, \mathcal{F}_1(\delta), L_2(P)) \leq K_1(\delta z^{-1})^{K_2} \vee 1$, where $K_1, K_2 > 0$ are constants depending only on $\dim \gamma$ and $\dim \beta$. By Lemma A.7, $N_{[]}(z, \mathcal{F}_2(\delta), L_2(P)) \leq K_3(\delta z^{-2})^{K_4} \vee 1$ and $N_{[]}(z, \mathcal{F}_3(\delta), L_2(P)) \leq K_5(\delta z^{-2})^{K_6} \vee 1$ for constants K_3, K_4, K_5, K_6 depending only on $\dim \beta$, $\dim \gamma$ and the constants in Condition 1. Since $\mathcal{F}_1$ is uniformly bounded and $\mathcal{F} \subset \mathcal{F}_1 \cdot \mathcal{F}_2 + \mathcal{F}_1 \cdot \mathcal{F}_3$, the result follows by Lemma A.5, Lemma 9.25 in Kosorok (2007) and a change of variables. □

Lemma A.9. *Let Condition 1 hold. Suppose that* $\sup_{x \geq a_n} |dg^{-1}(x)/dx| = O(n^a)$ *for some* $a < 1$. *Then* $\forall K \in (0, \infty)$, $\sup_{(\gamma,\beta) \in \Gamma \times B_n} \left|n^{-1/2}\mathbb{G}_n m_{(n)}(\gamma, \beta))\right| = o_P(1)$, *where* $B_n = \{\beta \mid \|\beta - \beta_*\| \leq n^{-1/2}K\}$.

Proof. Notice that in the definition of $m_{(n),i}$, $g^{-1}(\cdot)$ can be equivalently replaced by $g_n^{-1}(\cdot)$ with $g_n^{-1}(x) = g^{-1}(x \vee a_n)$. One can easily verify that for $(\gamma_1, \beta_1), (\gamma_2, \beta_2) \in \Gamma \times B_n$, we have

$$\left|\rho_\tau(g_n^{-1}(\varepsilon_i(\beta_1)) - w_i(\gamma_1)) - \rho_\tau(g_n^{-1}(\varepsilon_i(\beta_2)) - w_i(\gamma_2))\right| \leq C\|\gamma_1 - \gamma_2\| + s_n C\|\beta_1 - \beta_2\|,$$

where $s_n = \sup_{x \geq a_n} |dg^{-1}(x)/dx|$ and $C > 0$ is a constant such that $\|\nabla_\gamma w_i(\gamma)\| \leq C$, $\|w_i\| \leq C$ and $\|\nabla_\beta \varepsilon_i(\beta)\| \leq C$ a.s. We can apply Lemma A.6 with $\mathcal{F}_n = \{\rho_\tau(g_n^{-1}(\varepsilon_i(\beta)) - w_i(\gamma)) \mid (\gamma, \beta) \in \Gamma_1 \times B_n\}$ and $\mathcal{K}_n = \{\mathbf{1}\{\varepsilon_i(\beta) - a_n > 0\} \mid \beta \in B_n\}$, where $\Gamma_1 = \{\gamma \mid \|\gamma - \gamma_*\| \leq \delta_1\}$ and δ_1 is equal to the diameter of Γ. We can choose $F_{n,t} = m_{(n),i}(\gamma_{n,*}, \beta_{n,*}) + \delta_1 C_1 + n^{-1/2}s_n KC$ to be an envelope function of $\mathcal{D}_n = \{m_{(n),i}(\gamma, \beta) \mid (\gamma, \beta) \in \Gamma_1 \times B_n\}$. By Theorem 2.14.2 of van der Vaart and Wellner (1996),

$$\begin{aligned}
n^{-1/2}E\|\mathbb{G}_n\|_{\mathcal{D}_n} &\lesssim n^{-1/2}\|F_{n,t}\|_{P,2} \int_0^1 \sqrt{1 + \log N_{[]}\left(z\|F_{n,t}\|_{P,2}, \mathcal{D}_n, L_2(P)\right)} dz \\
&\overset{(i)}{\leq} n^{-1/2}\|F_{n,t}\|_{P,2} \int_0^1 \sqrt{1 + \log\left\{M_0\left[(\delta_1 z^{-1})^{M_1} \vee 1\right]\left[\left(s_n n^{-1/2} K z^{-1}\right)^{M_2} \vee 1\right]\left[\left(n^{-1/2}Kz^{-2\pi}\right)^{M_2} \vee 1\right]\right\}} dx \\
&\overset{(ii)}{\leq} \underbrace{n^{-1/2}\|F_{n,t}\|_{P,2} \int_0^1 \sqrt{1 + \log\left\{M_0\left[(\delta_1 z^{-1})^{M_1} \vee 1\right]\right\}} dz}_{T_1}
\end{aligned}$$

$$+\underbrace{n^{-1/2}\|F_{n,t}\|_{P,2}\int_0^1\sqrt{\log\left[\left(s_n n^{-1/2}Kz^{-1}\right)^{M_2}\vee 1\right]}dz}_{T_2}$$
$$+\underbrace{n^{-1/2}\|F_{n,t}\|_{P,2}\int_0^1\sqrt{\log\left[\left(n^{-1/2}Kz^{-2\pi}\right)^{M_2}\vee 1\right]}dz}_{T_3}, \tag{31}$$

where (i) follows by Lemma A.6 and (ii) follows by the elementary inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b\ge 0$ (used twice).

Notice that $n^{-1/2}\|F_{n,t}\|_{P,2} = O(n^{-1/2}) + O(n^{-1}s_n) = O(n^{-\min\{1-a,1/2\}})$. Since $\int_0^1\sqrt{1+\log\left\{M_0\left[(\delta_1 z^{-1})^{M_1}\vee 1\right]\right\}}dz<\infty$, $T_1=o(1)$. Similarly, $T_3=o(1)$. Notice that

$$\begin{aligned}
T_2 &= n^{-1/2}\|F_{n,t}\|_{P,2}\int_0^1\sqrt{0\vee\left[M_2\log\left(n^{-1/2}s_nK\right)-M_2\log z\right]}dz\\
&\overset{(i)}{\le} n^{-1/2}\|F_{n,t}\|_{P,2}\sqrt{0\vee\left[M_2\log\left(n^{-1/2}s_nK\right)\right]}+n^{-1/2}\|F_{n,t}\|_{P,2}\int_0^1\sqrt{-M_2\log z}dz\\
&\overset{(ii)}{=} O(n^{-\min\{1-a,1/2\}})\sqrt{0\vee\left[M_2\log\left(O(n^{a-1/2})\right)\right]}+O(n^{-\min\{1-a,1/2\}})=o(1),
\end{aligned}$$

where (i) follows by the elementary inequalities $0\vee(a+b)\le(0\vee a)+b$ for $b\ge 0$ and $a\in\mathbb{R}$ and $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b\ge 0$, and (ii) follows by $\int_0^1\sqrt{-M_2\log z}dz<\infty$ and $s_n=O(n^a)$. Hence, all the above along with (31) imply that $n^{-1/2}E\|\mathbb{G}_n\|_{\mathcal{D}_n}=o(1)$. The desired result follows by Markov's inequality. □

Lemma A.10. *Under the conditions and notations of Lemma A.9, $\forall K\in(0,\infty)$, $\sup_{(\gamma,\beta)\in\Gamma\times B_n}\left|n^{-1/2}\mathbb{G}_n^* m_{(n)}(\gamma,\beta))\right|=o_P(1)$.*

Proof. The proof is almost the same as that of Lemma A.9 with $\mathbb{G}_n$ replaced by $\mathbb{G}_n^*$. The only difference in the proof is that in deriving (31) in the proof of Lemma A.9, we now use Lemma 21.9 of Kosorok (2007), instead of Theorem 2.14.2 of van der Vaart and Wellner (1996). □

Lemma A.11. *Let Condition 1 hold. Then $E\sup_{(h_1,h_2)\in B(\delta)}\left|\mathbb{G}_n R_{(n)}(\gamma_*+h_1,\beta_*+h_2)\right|\le M\delta^{6/5}$ for small enough $\delta>0$, where $M>0$ is a constant depending only on the constants in Condition 1.*

Proof. By simple computations, it can be easily show that an envelope function for $\mathcal{F}_n(\delta)=\{R_{(n),i}(\gamma_*+h_1,\beta_*+h_2)\mid\|h_1\|+\|h_2\|\le\delta\}$ is $F_t = C_0\delta\mathbf{1}\{|\varepsilon_i(\beta_*)-g(w_i(\gamma_*))|\le C_1\delta\}$ for some constants $C_0,C_1>0$ depend only on the constants in Condition 1. Notice that

$\|F_t\|_{P,2} \leq C_2\delta^{3/2}$ with $C_2 = C_0\sqrt{A_1C_1}$. By Theorem 2.14.2 of van der Vaart and Wellner (1996), we have

$$
\begin{aligned}
E\left(\sup_{(h_1,h_2)\in B(\delta)}\left|\mathbb{G}_n R_{(n)}(\gamma_* + h_1, \beta_* + h_2)\right|\right) &\lesssim \|F_t\|_{P,2}\int_0^1 \sqrt{1 + \log N_{[]}\left(z\|F_t\|_{P,2}, \mathcal{F}_n(\delta), L_2(P)\right)}dz \\
&\overset{(i)}{\leq} \underbrace{\|F_t\|_{P,2}\int_0^1 \sqrt{1 + \log\left[M_1\left(\delta\|F_t\|_{P,2}^{-1}z^{-1}\right)^{M_2} \vee 1\right]}dz}_{T_1} \\
&+ \underbrace{\|F_t\|_{P,2}\int_0^1 \sqrt{\log\left[M_3\left(\delta\|F_t\|_{P,2}^{-2}z^{-2}\right)^{M_4} \vee 1\right]}dz}_{T_2} \\
&+ \underbrace{\|F_t\|_{P,2}\int_0^1 \sqrt{\log\left[M_5\left(\delta\|F_t\|_{P,2}^{-2}z^{-2}\right)^{M_6} \vee 1\right]}dz}_{T_3},
\end{aligned} \tag{32}
$$

where $M_1, M_2, M_3, M_4, M_5, M_6 > 0$ depending only on $\dim\beta$, $\dim\gamma$ and the constants in Condition 1. Here, (i) follows by Lemma A.8 and the elementary inequalities $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for $a, b \geq 0$ (applied twice).

Notice that $\log\left[M_1\left(\delta\|F_t\|_{P,2}^{-1}z^{-1}\right)^{M_2} \vee 1\right] \leq \log M_1 + M_2(0 \vee \log\delta) - M_2\log z - M_2\log\|F_t\|_{P,2}$. Since for small $\delta > 0$ $\int_0^1 \sqrt{1 + \log M_1 + M_2(0 \vee \log\delta) - M_2\log z}dz$ is bounded by a constant that does not depend on δ, it follows that $T_1 \leq M_7\|F_t\|_{P,2} + \|F_t\|_{P,2}\sqrt{-M_2\log\|F_t\|_{P,2}}$ for some $M_7 > 0$ depending only on M_1 and M_2. Notice that $\|F_t\|_{P,2}\sqrt{-M_2\log\|F_t\|_{P,2}} \leq \|F_t\|_{P,2}^{4/5}\sqrt{-M_2\|F_t\|_{P,2}^{2/5}\log\|F_t\|_{P,2}}$. Since $x \mapsto x^{2/5}\log x$ is bounded on the interval $(0, c)$ $\forall c > 0$, there exists a constant $M_8 > 0$ depending only on M_2 and M_7 such that $T_1 \leq M_8\|F_t\|_{P,2}^{4/5}$. Since $\|F_t\|_{P,2} \leq C_2\delta^{3/2}$, $T_1 \leq M_8C_2^{4/5}\delta^{6/5}$. Similar arguments yield $T_2 \leq M_9\delta^{6/5}$ and $T_3 \leq M_{10}\delta^{6/5}$ for constants $M_9, M_{10} > 0$ depending only on M_3, M_4, M_5, M_6. The result follows. □

Lemma A.12. *Suppose that there exists a constant $K > 0$ such that $\forall\delta > 0$ small enough, $E\sup_{(h_1,h_2)\in B(\delta)}\left|\mathbb{G}_n R_{(n)}(\gamma_* + h_1, \beta_* + h_2)\right| \leq K\delta^{1+a}$, where $B(\delta) = \{(h_1, h_2) \mid \|h_1\| + \|h_2\| \leq \delta\}$. Then for any $h_{n,1} = o_P(1)$ and $h_{n,2} = o_P(1)$, we have*

$$\mathbb{G}_n R_{(n)}(\gamma_* + h_{n,1}, \beta_* + h_{n,2}) = o_P(\|h_{n,1}\|) + o_P(\|h_{n,2}\|).$$

Proof. We use the so-called "peeling device" discussed in Section 5.3 of van de Geer (2000).

Fix an arbitrary small $\varepsilon > 0$. Let $\delta > 0$ be a small enough number to be chosen later. Define $B_j(\delta) = \{(h_1, h_2) \mid 2^{-j}\delta \leq \|h_1\| + \|h_2\| < 2^{1-j}\delta\}$. Notice that $B(\delta) = \bigcup_{j=1}^{\infty} B_j(\delta)$. Then

$$
\begin{aligned}
P\left(\sup_{(h_1,h_2)\in B(\delta)} \left|\frac{\mathbb{G}_n R_{(n)}(\gamma_* + h_1, \beta_* + h_2)}{\|h_1\| + \|h_2\|}\right| > \varepsilon\right) &\overset{(i)}{\leq} \sum_{j=1}^{\infty} P\left(\sup_{(h_1,h_2)\in B_j(\delta)} \left|\frac{\mathbb{G}_n R_{(n)}(\gamma_n + h_1, \beta_* + h_2)}{\|h_1\| + \|h_2\|}\right| > \varepsilon\right) \\
&\overset{(ii)}{\leq} \sum_{j=1}^{\infty} P\left(\sup_{(h_1,h_2)\in B_j(\delta)} \left|\mathbb{G}_n R_{(n)}(\gamma_* + h_1, \beta_* + h_2)\right| > 2^{-j}\delta\varepsilon\right) \\
&\overset{(iii)}{\leq} \sum_{j=1}^{\infty} \frac{E \sup_{(h_1,h_2)\in B_j(\delta)} \left|\mathbb{G}_n R_{(n)}(\gamma_* + h_1, \beta_* + h_2)\right|}{2^{-j}\delta\varepsilon} \\
&\leq \sum_{j=1}^{\infty} \frac{K(2^{1-j}\delta)^{1+a}}{\varepsilon 2^{-j}\delta} = \left(\varepsilon^{-1} K 2^{1+a} \sum_{j=1}^{\infty} 2^{-ja}\right) \delta^a,
\end{aligned}
$$

where (i) follows by the sub-additivity of probability measures, (ii) follows by the definition of $H_j(\delta)$ and (iii) holds by Markov's inequality.

Since $\varepsilon^{-1} K 2^{1+a} \sum_{j=1}^{\infty} 2^{-ja} < \infty$, we can choose small enough δ such that

$$
\limsup_{n\to\infty} P\left(\sup_{(h_1,h_2)\in B(\delta)} \left|\frac{\mathbb{G}_n R_{(n)}(\gamma_* + h_1, \beta_* + h_2)}{\|h_1\| + \|h_2\|}\right| > \varepsilon\right) \leq \varepsilon.
$$

Since $h_{n,1} = o_P(1)$ and $h_{n,2} = o_P(1)$, $P((h_{n,1}, h_{n,2}) \in H(\delta)) \to 1$ and hence $\limsup P\left(|\mathbb{G}_n R_{(n)}(\gamma_* + h_{n,1}, \beta_* + h_{n,2})| > (\|h_{n,1}\| + \|h_{n,2}\|)\varepsilon\right) \leq \varepsilon$. Since $\varepsilon > 0$ is arbitrary, the desired result follows. □

Lemma A.13. *Let Condition 1 hold. Then for any $h_{n,1} = o_P(1)$ and $h_{n,2} = o_P(1)$, we have $\mathbb{G}_n R_{(n)}(\gamma_* + h_{n,1}, \beta_* + h_{n,2}) = o_P(\|h_{n,1}\|) + o_P(\|h_{n,2}\|)$ and $\mathbb{G}_n^* R_{(n)}(\gamma_* + h_{n,1}, \beta_* + h_{n,2}) = o_P(\|h_{n,1}\|) + o_P(\|h_{n,2}\|)$.*

Proof. The first claim follows by Lemmas A.11 and A.12. For the second claim, it suffices to notice that the conclusions in Lemmas A.11 and A.12 hold with $\mathbb{G}_n$ replaced by $\mathbb{G}_n^*$. To see this, notice that all the arguments in the proof of these two lemmas still hold, except that in deriving (32) in the proof of Lemma A.11, we now use Lemma 21.9 of Kosorok (2007), instead of Theorem 2.14.2 of van der Vaart and Wellner (1996). □

Lemma A.14. *Let Condition 1 hold. If $a_n = O(n^{-c})$ for $c \in (0, \infty)$, then $\forall K_0 \in (0, \infty)$,*

$$
\sup_{\|h\|\leq K_0} \left\|\mathbb{G}_n\left(\dot{m}_{(n)}(\gamma_*, \beta_* + n^{-1/2}h) - \dot{m}(\gamma_*, \beta_*)\right)\right\| = o_P(1).
$$

Proof. Let $B_n = \{\beta \mid \|\beta - \beta_*\| \leq n^{-1/2}K_0\}$ and $\mathcal{F}_n = \{\dot{m}_{(n),i}(\gamma_*, \beta) - \dot{m}_{(n),i}(\gamma_*, \beta_*) \mid \beta \in B_n\}$. Let $K > 0$ be large enough such that $\sup_{\beta \in B_n} \|\beta\| \leq K$. Notice that . Let $\mathcal{F}_{n,1} = \{\mathbf{1}\{\varepsilon_i(\beta) > a_n\} \mid \beta \in B_n\}$ and $\mathcal{F}_{n,2} = \{\mathbf{1}\{\varepsilon_i(\beta) < g(w_i(\gamma))\} \mid \beta \in B_n\}$. Fix $x > 0$.

Notice that $\dot{m}_{(n),i}(\gamma, \beta) = \nabla_\gamma w_i(\gamma)\mathbf{1}\{\varepsilon_i(\beta) > a_n\}\tau - \nabla_\gamma w_i(\gamma)\mathbf{1}\{\varepsilon_i(\beta) > a_n\}\mathbf{1}\{\varepsilon_i(\beta) < g(w_i(\gamma))\}$. Let $k_{n,t,1}(\beta) = \varepsilon_i(\beta) - a_n$ and $\mathcal{K}_{n,1} = \{\mathbf{1}\{k_{n,t,1}(\beta) > 0\} \mid \beta \in B_n\}$. By Lemma A.7, $N_{[]}(x, \mathcal{F}_{n,1}, L_2(P)) \leq N_{[]}(x, \mathcal{K}_{n,1}, L_2(P)) \leq M_1(n^{-1/2}K_0x^{-1})^{M_2} \vee 1$, where $M_1, M_2 > 0$ are constants depending only on A_1 and $\dim \beta$. Let $k_{t,2}(\beta) = g(w_i(\gamma_*)) - \varepsilon_i(\beta)$ and $\mathcal{K}_{n,2} = \{\mathbf{1}\{k_{t,2}(\beta) > 0\} \mid \beta \in B_n\}$. The same reasoning yields $N_{[]}(x, \mathcal{F}_{n,2}, L_2(P)) \leq N_{[]}(x, \mathcal{K}_{n,2}, L_2(P)) \leq M_1(n^{-1/2}K_0x^{-1})^{M_2} \vee 1$. By the bounded support of $\|w_i\|$ and Lemma A.5, $N_{[]}(4\tau A_1 x, (w_i \cdot \tau) \cdot \mathcal{F}_{n,1}, L_2(P)) \leq N_{[]}(2x, \mathcal{F}_{n,1}, L_2(P)) \leq N_{[]}(x, \mathcal{F}_{n,1}, L_2(P))$ and

$$N_{[]}(4A_1x, \nabla_\gamma w_i(\gamma_*)\cdot\mathcal{F}_{n,1}\cdot\mathcal{F}_{n,2}, L_2(P)) \leq N_{[]}(2x, \mathcal{F}_{n,1}\cdot\mathcal{F}_{n,2}, L_2(P)) \leq N_{[]}(x, \mathcal{F}_{n,1}, L_2(P))N_{[]}(x, \mathcal{F}_{n,2}, L_2(P)).$$

Hence,

$$\begin{aligned} N_{[]}(8\tau A_1 x, \mathcal{F}_n, L_2(P)) &= N_{[]}(8\tau A_1 x, \mathcal{F}_n + \dot{m}_{(n),i}(\gamma_*, \beta_*), L_2(P)) \\ &\overset{(i)}{\leq} N_{[]}(4\tau A_1 x, (\nabla_\gamma w_i(\gamma_*) \cdot \tau) \cdot \mathcal{F}_{n,1}, L_2(P))N_{[]}(4A_1 x, \nabla_\gamma w_i(\gamma_*) \cdot \mathcal{F}_{n,1} \cdot \mathcal{F}_{n,2}, L_2(P)) \\ &\leq M_1^3 \left(n^{-1/2}K_0x^{-1}\right)^{3M_2} \vee 1, \end{aligned}$$

where (i) follows by Lemma 9.25 of Kosorok (2007) and the observation that $\mathcal{F}_n + \dot{m}_{(n),i}(\gamma_*, \beta_*) \subset \nabla_\gamma w_i(\gamma_*) \cdot \tau \cdot \mathcal{F}_{n,1} - \nabla_\gamma w_i(\gamma_*) \cdot \mathcal{F}_{n,1} \cdot \mathcal{F}_{n,2}$. A change of variable yields that $\forall x > 0$, $N_{[]}(x, \mathcal{F}_n, L_2(P)) \leq M_3(n^{-1/2}K_0x^{-1})^{M_4} \vee 1$, where $M_3 = (8\tau A_1)^{3M_2}M_1^3$ and $M_4 = 3M_2$.

By simple computations, one can verify that $F_{n,t} = A_1\mathbf{1}\{|\varepsilon_i(\beta_*) - g(w_i(\gamma_*))| \leq n^{-1/2}K_0A_1\} + A_1\mathbf{1}\{|\varepsilon_i(\beta_*) - a_n| \leq n^{-1/2}K_0A_1\}$ is an envelope function for $\mathcal{F}_n$. Notice that $\|F_{n,t}\|_{P,2} \leq A_1\sqrt{A_2A_1K_0n^{-1/2}} + A_1\sqrt{A_1(n^{-1/2}K_0A_1 + a_n)} = o(1)$. By Theorem 2.14.2 of van der Vaart and Wellner (1996),

$$\begin{aligned} E\|\mathbb{G}_n\|_{\mathcal{F}_n} &\lesssim \|F_{n,t}\|_{P,2}\int_0^1 \sqrt{1 + \log N_{[]}(x\|F_{n,t}\|_{P,2}, \mathcal{F}_n, L_2(P))}dx \\ &\overset{(i)}{\leq} \|F_{n,t}\|_{P,2}\int_0^1 \sqrt{1 + 0 \vee \log\left[M_3\left(n^{-1/2}K_0x^{-1}\|F_{n,t}\|_{P,2}^{-1}\right)^{M_4}\right]}dx \\ &\leq \|F_{n,t}\|_{P,2}\int_0^1 \sqrt{(D - M_4\log x) + 0 \vee \left(-M_4\log n^{1/2}\|F_{n,t}\|_{P,2}\right)}dx \end{aligned}$$

$$\overset{(ii)}{\leq} \underbrace{\|F_{n,t}\|_{P,2}\int_0^1 \sqrt{(D - M_4 \log x)}dx}_{T_1} + \underbrace{\|F_{n,t}\|_{P,2}\sqrt{0 \vee \left(-M_4 \log n^{1/2}\|F_{n,t}\|_{P,2}\right)}}_{T_2} \quad (33)$$

where $D = 1 + \log M_3 - 0 \vee (M_4 \log K_0)$. Here, (i) follows by $N_{[]}(x, \mathcal{F}_n, L_2(P)) \leq M_3(n^{-1/2}K_0x^{-1})^{M_4} \vee 1$ and (ii) follows by the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$.

Since $\int_0^1 \sqrt{(D - M_4 \log x)}dx < \infty$, $T_1 = O(\|F_{n,t}\|_{P,2}) = o(1)$. Since $a_n = O(n^{-c})$, we have $\|F_{n,t}\|_{P,2} = O(n^{-\min\{1/4, c/2\}})$ and $|\log n^{1/2}\|F_{n,t}\|_{P,2}| = |1/2 - \min\{1/4, c/2\}| \log n$, implying $T_2 = o(1)$. Hence, $E\|\mathbb{G}_n\|_{\mathcal{F}_n} = o(1)$ and, by Markov's inequality,

$$\sup_{\|h\| \leq K_0} \left\| \mathbb{G}_n \left(\dot{m}_{(n)}(\gamma_*, \beta_* + n^{-1/2}h) - \dot{m}_{(n)}(\gamma_*, \beta_*) \right) \right\| = \|\mathbb{G}_n\|_{\mathcal{F}_n} = o_P(1). \quad (34)$$

Let $X_{n,t} := \dot{m}_{(n),i}(\gamma_*, \beta_*) - \dot{m}_i(\gamma_*, \beta_*)$. Notice that $X_{n,t} = -\nabla_\gamma w_i(\gamma_*)\psi_\tau(\varepsilon_i(\beta_*) - w_i(\gamma_*))\mathbf{1}\{0 < \varepsilon_i(\beta_*) \leq a_n\}$ and $E\|X_{n,t}\|^3 = o(1)$. By Lyapunov's CLT, $\mathbb{G}_n X_{n,t} = o_P(1)$. This, combined with (34), implies the desired result. □

Lemma A.15. *Under the conditions of Lemma A.14, $\forall K_0 \in (0, \infty)$, $\sup_{\|h\| \leq K_0} \left\| \mathbb{G}_n^* \left(\dot{m}_{(n)}(\gamma_*, \beta_* + n^{-1/2}h) - \dot{m}(\gamma_*, \beta_*) \right) \right\| = o_P(1)$.*

Proof. The proof is almost the same as in Lemma A.14 with $\mathbb{G}_n$ replaced by $\mathbb{G}_n^*$. The only difference in proof is that in deriving (33) in the proof of Lemma A.14, we now use Lemma 21.9 of Kosorok (2007), instead of Theorem 2.14.2 of van der Vaart and Wellner (1996). □

Theorem 3. *Let Condition 1 hold. Suppose that the following also hold:*
(i) $\hat{M}_n(\hat{\gamma}, \hat{\beta}) \leq \inf_{\gamma \in \Gamma} \hat{M}_n(\gamma, \hat{\beta}) + o_P(n^{-1})$ and $\hat{\beta} = \beta_* + O_P(n^{-1/2})$.
(ii) $M(\gamma, \beta)$ is twice continuously differentiable in (γ, β) over $\Gamma \times B$.
(iii) $V_* = \nabla_{\gamma\gamma} M(\gamma_*, \beta_*)$ is nonsingular.
(iv) $a_n = O(n^{-c})$ for some $c \in (0, \infty)$, $\sup_{x \geq a_n} |dg^{-1}(x)/dx| = O(n^a)$ for some $a < 1$ and $\int_{-1}^1 |g^{-1}(|x|)|^p dx < \infty$ for some $p > 1$.

Then

$$\hat{\gamma} - \gamma_* = -n^{-1/2}V_*^{-1}\left(\mathbb{G}_n \dot{m}(\gamma_*, \beta_*) + n^{1/2} D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta} - \beta_*) \right) + o_P(n^{-1/2}),$$

where $D_{\gamma\beta}(\gamma, \beta) = \nabla_{\gamma\beta} M(\gamma, \beta)$.

Proof. Let $D_{n,\gamma\gamma}(\gamma, \beta) = \nabla_{\gamma\gamma} M_n(\gamma, \beta)$, $D_{n,\gamma\beta}(\gamma, \beta) = \nabla_\beta D_{n,\gamma}(\gamma, \beta)$, $D_{n,\gamma}(\gamma, \beta) = \nabla_\gamma M_n(\gamma, \beta)$ and $D_{\gamma\beta}(\gamma, \beta) = \nabla_{\gamma\beta} M(\gamma, \beta)$. Define $Z_n := \mathbb{G}_n \dot{m}_{(n)}(\gamma_{n,*}, \hat{\beta}) + n^{1/2} D_{n,\gamma}(\gamma_{n,*}, \hat{\beta})$.

Notice that for any random sequence $h_n = o_P(1)$, we have

$$\begin{aligned}
& n\left[\hat{M}_n(\gamma_* + h_n, \hat{\beta}) - \hat{M}_n(\gamma_*, \hat{\beta})\right] \\
= \; & n^{1/2}\mathbb{G}_n\left(m_{(n)}(\gamma_* + h_n, \hat{\beta}) - m_{(n)}(\gamma_*, \hat{\beta})\right) + n\left[M_n(\gamma_* + h_n, \hat{\beta}) - M_n(\gamma_*, \hat{\beta})\right] \\
= \; & n^{1/2}\mathbb{G}_n\left(m_{(n)}(\gamma_* + h_n, \hat{\beta}) - m_{(n)}(\gamma_*, \hat{\beta}) - h_n'\dot{m}_{(n)}(\gamma_*, \hat{\beta})\right) + n^{1/2}h_n'\mathbb{G}_n\dot{m}_{(n)}(\gamma_*, \hat{\beta}) \\
& + n\left[M_n(\gamma_* + h_n, \hat{\beta}) - M_n(\gamma_*, \hat{\beta})\right] \\
\overset{(i)}{=} \; & o_P(n^{1/2}\|h_n\|) + o_P(1) + n^{1/2}h_n'\mathbb{G}_n\dot{m}_{(n)}(\gamma_*, \hat{\beta}) + n\left[M_n(\gamma_* + h_n, \hat{\beta}) - M_n(\gamma_*, \hat{\beta})\right] \\
\overset{(ii)}{=} \; & o_P(n^{1/2}\|h_n\|) + o_P(1) + n^{1/2}h_n'\mathbb{G}_n\dot{m}_i(\gamma_*, \beta_*) + n\left[M_n(\gamma_* + h_n, \hat{\beta}) - M_n(\gamma_*, \hat{\beta})\right] \\
\overset{(iii)}{=} \; & o_P(n^{1/2}\|h_n\|) + o_P(1) + n^{1/2}h_n'\mathbb{G}_n\dot{m}_i(\gamma_*, \beta_*) + o(n\|h_n\|^2) + nh_n'\left(D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta} - \beta_*)\right) + \frac{n}{2}h_n'V_*h_n \\
= \; & o_P(n^{1/2}\|h_n\|) + o_P(1) + o(n\|h_n\|^2) + n^{1/2}h_n'Z_n + \frac{n}{2}h_n'V_*h_n, \qquad (35)
\end{aligned}$$

where $Z_n = \mathbb{G}_n\dot{m}(\gamma_*, \beta_*) + n^{1/2}D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta} - \beta_*)$. Here, (i) follows by Lemma A.13, which implies $\mathbb{G}_n R_{(n)}(\gamma_{n,*} + h_n, \hat{\beta}) = o_P(\|h_n\|) + o_P(\|\hat{\beta} - \beta_{n,*}\|)$. (ii) follows by Lemma A.14 and (iii) follows by Lemma A.3.

Let $S_n(\gamma) = \hat{M}_n(\gamma, \hat{\beta})$ and $Q(\gamma) = M(\gamma, \beta_*)$. By the triangular inequality, $\sup_{\gamma\in\Gamma}|S_n(\gamma) - Q_n(\gamma)| \le \sup_{\gamma\in\Gamma}|\hat{M}_n(\gamma, \hat{\beta}) - M_n(\gamma, \hat{\beta})| + \sup_{\gamma\in\Gamma}|M_n(\gamma, \hat{\beta}) - M(\gamma, \hat{\beta})| + \sup_{\gamma\in\Gamma}|M(\gamma, \hat{\beta}) - M(\gamma, \beta_*)|$. By Lemma A.9, $\sup_{\gamma\in\Gamma}|\hat{M}_n(\gamma, \hat{\beta}) - M_n(\gamma, \hat{\beta})| = o_P(1)$. By Lemma A.2, $\sup_{\gamma\in\Gamma}|M_n(\gamma, \hat{\beta}) - M(\gamma, \hat{\beta})| = o_P(1)$. By the uniform continuity (implied by the differentiability) of $M(\cdot, \cdot)$ and $\hat{\beta} - \beta = O_P(n^{-1/2})$, we have $\sup_{\gamma\in\Gamma}|M(\gamma, \hat{\beta}) - M(\gamma, \beta_*)| = o_P(1)$. Hence, $\sup_{\gamma\in\Gamma}|S_n(\gamma) - Q_n(\gamma)| = o_P(1)$ and, by the nonsingularity of V_*, $\forall\eta > 0$, $M(\gamma_*, \beta_*) < \inf_{\|\gamma-\gamma_*\|>\varepsilon} M(\gamma, \beta_*)$. It follows, by Corollary 3.2.3 of van der Vaart and Wellner (1996), that $\hat{\gamma} = \gamma_* + o_P(1)$. Then the result follows by (35) and Lemma A.4. □

Theorem 4. *Let the conditions of Theorem 3 hold. If* $\inf_{\gamma\in\Gamma}\hat{M}_n^*(\gamma, \hat{\beta}^*) \ge \hat{M}_n^*(\hat{\gamma}^*, \hat{\beta}^*) + o_P(n^{-1})$ *and* $\hat{\beta}^* = \beta_* + O_{P^*}(n^{-1/2})$, *then*

$$\hat{\gamma}^* - \gamma_* = -n^{-1/2}V_*^{-1}\left(\mathbb{G}_n^*\dot{m}(\gamma_*, \beta_*) + n^{1/2}D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta}^* - \beta_*)\right) + o_P(n^{-1/2}).$$

Proof. The proof is essentially the same as that of Theorem 3 with $\mathbb{G}_n$ replaced by $\mathbb{G}_n^*$. For any random sequence $h_n = o_P(1)$, we can show

$$n\left[\hat{M}_n^*(\gamma_* + h_n, \hat{\beta}^*) - \hat{M}_n^*(\gamma_*, \hat{\beta}^*)\right] = o_P(n^{1/2}\|h_n\|) + o_P(1) + o(n\|h_n\|^2) + n^{1/2}h_n'Z_n^* + \frac{n}{2}h_n'V_*h_n, \qquad (36)$$

$Z_n^* = \mathbb{G}_n^*\dot{m}(\gamma_*, \beta_*) + n^{1/2}D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta}^* - \beta_*)$. The arguments for the above display are the

same as those for (35) in the proof of Theorem 3, except that $\mathbb{G}_n$ is replaced by $\mathbb{G}_n^*$ and that we invoke Lemma A.15, instead of Lemma A.14.

We show that $\hat{\gamma}^* = \gamma_* + o_P(1)$ using the same argument as in Theorem 3, except that $\hat{M}_n(\gamma, \hat{\beta})$ is replaced by $\hat{M}_n^*(\gamma, \hat{\beta}^*)$ and that we invoke Lemma A.10, instead of Lemma A.9. Then the desired result follows by (36) and Lemma A.4. □

Proof of Theorem 2. The proof proceeds by induction. We start from $j = j^*$ and then move to the tails. Under the notations of Assumption 2, we apply Theorems 3 and 4. It follows that

$$\hat{\gamma} - \gamma_* = -n^{-1/2} V_*^{-1} \left(\mathbb{G}_n \dot{m}(\gamma_*, \beta_*) + n^{1/2} D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta} - \beta_*) \right) + o_P(n^{-1/2}) \tag{37}$$

and

$$\hat{\gamma}^* - \gamma_* = -n^{-1/2} V_*^{-1} \left(\mathbb{G}_n^* \dot{m}(\gamma_*, \beta_*) + n^{1/2} D_{\gamma\beta}(\gamma_*, \beta_*)(\hat{\beta}^* - \beta_*) \right) + o_P(n^{-1/2}). \tag{38}$$

The asymptotic normality follows by (37). The bootstrap validity follows by Theorem 21.7 of Kosorok (2007), together with (37) and (38). □

Proof of Corollary 1. The result follows by Theorem 2, together with Theorems 3.9.4 and 3.9.11 of van der Vaart and Wellner (1996). □

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens, 2002, Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings, *Econometrica* 70, 91–117.

Adrian, Tobias, and Markus K Brunnermeier, 2011, Covar, Technical report, National Bureau of Economic Research.

Angrist, Joshua, Victor Chernozhukov, and Iván Fernández-Val, 2006, Quantile regression under misspecification, with an application to the us wage structure, *Econometrica* 74, 539–563.

Athey, Susan, and Guido W Imbens, 2006, Identification and inference in nonlinear difference-indifferences models, *Econometrica* 431?497.

Barbe, Philippe, and Patrice Bertail, 2012, *The weighted bootstrap*, volume 98 (Springer Science & Business Media).

Belloni, Alexandre, Victor Chernozhukov, et al., 2011a, L1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics* 39, 82–130.

Belloni, Alexandre, Victor Chernozhukov, and Iván Fernández-Val, 2011b, Conditional quantile processes based on series or many regressors .

Bloom, Nicholas, 2009, The impact of uncertainty shocks, *Econometrica* 77, 623–685.

Bondell, Howard D, Brian J Reich, and Huixia Wang, 2010, Noncrossing quantile regression curve estimation, *Biometrika* 97, 825–838.

Buchinsky, Moshe, 1994, Changes in the us wage structure 1963-1987: Application of quantile regression, *Econometrica: Journal of the Econometric Society* 405–458.

Cenesizoglu, Tolga, and Allan G Timmermann, 2008, Is the distribution of stock returns predictable?, *Available at SSRN 1107185* .

Chernozhukov, Victor, Ivan Fernandez-Val, and Alfred Galichon, 2009, Improving point and interval estimators of monotone functions by rearrangement, *Biometrika* asp030.

Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon, 2010, Quantile and probability curves without crossing, *Econometrica* 78, 1093–1125.

Chernozhukov, Victor, and Christian Hansen, 2005, An iv model of quantile treatment effects, *Econometrica* 73, 245–261.

Chernozhukov, Victor, and Han Hong, 2003, An mcmc approach to classical estimation, *Journal of Econometrics* 115, 293–346.

Covas, Francisco B, Ben Rump, and Egon Zakrajšek, 2014, Stress-testing us bank holding companies: A dynamic panel quantile regression approach, *International Journal of Forecasting* 30, 691–713.

Decker, Ryan A, John Haltiwanger, Ron S Jarmin, and Javier Miranda, 2016, Where has all the skewness gone? the decline in high-growth (young) firms in the us, *European Economic Review* 86, 4–23.

Dette, Holger, and Stanislav Volgushev, 2008, Non-crossing non-parametric estimates of quantile curves, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 609–627.

Diebold, Francis X, and Roberto S Mariano, 1995, Comparing predictive accuracy, *Journal of Business & Economic Statistics* 253–263.

Engle, Robert F, and Simone Manganelli, 2004, Caviar: Conditional autoregressive value at risk by regression quantiles, *Journal of Business & Economic Statistics* 22, 367–381.

Ghysels, Eric, Alberto Plazzi, and Rossen Valkanov, 2016a, Why invest in emerging markets? the role of conditional return asymmetry, *The Journal of Finance* .

Ghysels, Eric, Alberto Plazzi, and Rossen I. Valkanov, 2016b, Why invest in emerging markets? the role of conditional return asymmetry, *Social Science Research Network Working Paper Series* .

Giacomini, Raffaella, and Halbert White, 2006, Tests of conditional predictive ability, *Econometrica* 74, 1545–1578.

Gosling, Amanda, Stephen Machin, and Costas Meghir, 2000, The changing distribution of male wages in the uk, *The Review of Economic Studies* 67, 635–666.

Gouriéroux, Christian, and Joann Jasiak, 2008, Dynamic quantile models, *Journal of econometrics* 147, 198–205.

Granger, Clive WJ, 2010, Some thoughts on the development of cointegration, *Journal of Econometrics* 158, 3–6.

Guvenen, Fatih, Serdar Ozkan, and Jae Song, 2014, The nature of countercyclical income risk, *Journal of Political Economy* 122, 621–660.

He, Xuming, 1997, Quantile curves without crossing, *The American Statistician* 51, 186–192.

Hinkley, David V, 1975, On power transformations to symmetry, *Biometrika* 62, 101–111.

Kehrig, Matthias, 2015, The cyclical nature of the productivity distribution, *Earlier version: US Census Bureau Center for Economic Studies Paper No. CES-WP-11-15* .

Kelly, Bryan, and Hao Jiang, 2014, Tail risk and asset prices, *Review of Financial Studies* 27, 2841–2871.

Kim, Tae-Hwan, and Halbert White, 2003, Estimation, inference, and specification testing for possibly misspecified quantile regression, *Advances in Econometrics* 17, 107–132.

Koenker, Roger, 2005, *Quantile regression*, number 38 (Cambridge university press).

Koenker, Roger, and Gilbert Bassett, 1978, Regression quantiles, *Econometrica* 46, 33–50.

Kosorok, Michael R, 2007, *Introduction to empirical processes and semiparametric inference* (Springer Science & Business Media).

Lalive, Rafael, 2008, How do extended benefits affect unemployment duration? a regression discontinuity approach, *Journal of Econometrics* 142, 785–806.

Ma, Shuangge, and Michael R Kosorok, 2005, Robust semiparametric m-estimation and the weighted bootstrap, *Journal of Multivariate Analysis* 96, 190–217.

Machado, José AF, and José Mata, 2005, Counterfactual decomposition of changes in wage distributions using quantile regression, *Journal of applied Econometrics* 20, 445–465.

Mammen, Enno, 1991, Nonparametric regression under qualitative smoothness assumptions, *The Annals of Statistics* 741–759.

Qu, Zhongjun, and Jungmo Yoon, 2015, Nonparametric estimation and inference on conditional quantile processes, *Journal of Econometrics* 185, 1–19.

Schmidt, Lawrence DW, Allan G Timmermann, and Russ Wermers, 2016, Runs on money market mutual funds, *American Economic Review (forthcoming)* .

van de Geer, Sara A., 2000, *Empirical Processes in M-Estimation (Cambridge Series in Statistical and Probabilistic Mathematics)* (Cambridge University Press).

van der Vaart, AW, and Jon Wellner, 1996, *Weak Convergence and Empirical Processes: With Applications to Statistics* (Springer Science & Business Media).

White, Halbert, Tae-Hwan Kim, and Simone Manganelli, 2015, {VAR} for var: Measuring tail dependence using multivariate regression quantiles, *Journal of Econometrics* 187, 169 – 188.

White Jr, Halbert L, Tae-Hwan Kim, and Simone Manganelli, 2008, Modeling autoregressive conditional skewness and kurtosis with multi-quantile caviar .