Institute for Economic Studies, Keio University

Keio-IES Discussion Paper Series

Nonparametric Inference in Functional Linear Quantile Regression by RKHS Approach

Kosaku Takanashi

4 March, 2018
DP 2018-002
https://ies.keio.ac.jp/en/publications/8958/

Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office@adst.keio.ac.jp
4 March, 2018

Nonparametric Inference in Functional Linear Quantile Regression by RKHS Approach
Kosaku Takanashi
Keio-IES DP2018-002
4 March, 2018
JEL Classification: C14; C12
Keywords: Functional Linear Quantile Regression; Mosco topology; Generalized Likelihood Ratio Test; Estimation with Convex Constraint

Abstract

This paper studies an asymptotics of functional linear quantile regression in which the dependent variable is scalar while the covariate is a function. We apply a roughness regularization approach of a reproducing kernel Hilbert space framework. In the above circumstance, narrow convergence with respect to uniform convergence fails to hold, because of the strength of its topology. A new approach we propose to the lack-of-uniform-convergence is based on Mosco-convergence that is weaker topology than uniform convergence. By applying narrow convergence with respect to Mosco topology, we develop an infinite-dimensional version of the convexity argument and provide a proof of an asymptotic normality of argmin processes. Our new technique also provides the asymptotic confidence intervals and the generalized likelihood ratio hypothesis testing in fully nonparametric circumstance.

Kosaku Takanashi
Faculty of Economics, Keio University
2-15-45, Mita, Minato, Tokyo, Japan
takanasi@econ.keio.ac.jp

Nonparametric Inference in Functional Linear Quantile Regression by RKHS Approach

Kōsaku Takanashi
Faculty of Economics, Keio University

Abstract

This paper studies an asymptotics of functional linear quantile regression in which the dependent variable is scalar while the covariate is a function. We apply a roughness regularization approach of a reproducing kernel Hilbert space framework. In the above circumstance, narrow convergence with respect to uniform convergence fails to hold, because of the strength of its topology. A new approach we propose to the lack-of-uniform-convergence is based on Mosco-convergence that is weaker topology than uniform convergence. By applying narrow convergence with respect to Mosco topology, we develop an infinite-dimensional version of the convexity argument and provide a proof of an asymptotic normality of argmin processes. Our new technique also provides the asymptotic confidence intervals and the generalized likelihood ratio hypothesis testing in fully nonparametric circumstance.

1 Introduction.

Functional data have become increasingly encountered in many applications, and quantile regression, developed by Koenker and Bassett Jr (1978), offers a variety of fruitful applications for a functional data by estimating several different conditional quantiles. This paper studies an asymptotics of functional linear quantile regression in which the dependent variable is scalar while the covariate is a function. Several statistical models and methods have been developed for them: Shin and Lee (2016), Yao et al. (2017), Yuan and Cai (2010), Hall et al. (2007), Hall et al. (2006), Müller et al. (2005), Yao et al. (2005). Functional principle component analysis (FPCA) is commonly used for analyzing such models; see, Kato (2012). The success of these FPCA-based approaches, however, hinges on the availability of a good estimate of the functional principal components for the slope function; see Cai and Yuan (2012). Roughness regularization method circumvents the spacing of eigenvalues of the covariance function which is required by the FPCA method, and allow one to regularize the model complexity in a continuous manner, see Yuan and Cai (2010).

In order to construct the asymptotics of the estimator and hypothesis testing, we make the uniform convergence of the objective function to its population counterpart. In order to make the objective function satisfy the uniform convergence, we have to impose some compactness of the parameter space or entropy conditions (e.g., van der Vaart (1998)). These assumptions are rather restrictive for functional linear quantile regression models. Although since the objective function for functional linear quantile regression is convex (which is the "check function" defined in section 2), it seems that we may use the convexity lemma (e.g., Pollard (1991) and Theorem 10.8 of Rockafellar (1970)) to ensure that point-wise convergence of convex functions implies uniform convergence, however, in the infinite-dimensional case, this argument for uniform convergence may fail (see Section 3.1).

To solve the aforementioned lack-of-uniform-convergence issue, we shall propose to apply an alternative mode of convergence, *Mosco convergence*, which is weaker than uniform convergence but still strong enough to enable statistical applications. Mosco convergence of the objective function ensures the convergence of its minimizer (Attouch (1984)). We develop narrow convergence theory with respect to the Mosco metric, see also Geyer (1994), Dupacava and Wets (1988), Molchanov (2005), Knight (2003) in finite dimensional setting and Bucher et al. (2014) for epigraph convergence. There exist alternative forms of convergence that is equivalent to Mosco convergence but more easily verifiable. They include graph convergence (G-convergence) of subdifferential operators and strong convergence of resolvent. We shall explain these key concepts in Section 3. Using these equivalences, we can establish the consistency and narrow convergence of an M-estimator in an infinite-dimensional parameter space. Furthermore, Mosco convergence also ensures the invertibility of the "Hessian" operator.

The rest of this paper is organized as follows. In Section 2, we present the set-up of functional linear quantile regression model. In Section 3, we describe the Mosco convergence and introduce the narrow convergence in the Mosco topology and we derive the quadratic approximation of a convex objective function in an infinite-dimensional Hilbert space. In Section 4 we apply our techniques to functional linear quantile regression model. We also provide the asymptotic distribution of the likelihood ratio statistic. Appendixes give some technical lemmas.

Here we introduce some notations used in this paper. Let $\rightsquigarrow$ denote narrow convergence and $\xrightarrow{P}$ denote convergence in probability. We use empirical process notation: $\mathbb{G}_n\rho = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\rho(\theta, Z_i) - \mathbb{E}[\rho(\theta, Z_i)]$. We denote $\|\theta\|$ as l_2-norm or L_2-norm of an element of Hilbert space $\theta \in \mathcal{H}$. Let $\theta_n \xrightarrow{s} \theta_0$ denote convergence in strong topology, e.g., $\|\theta_n - \theta_0\| \to 0$ and $\theta_n \xrightarrow{w} \theta_0$ denote convergence in weak topology, e.g., $\langle\theta_n, \theta^*\rangle \to \langle\theta_0, \theta^*\rangle$ for all identical dual $\theta^* \in \mathcal{H}^*\ (= \mathcal{H})$. We denote the limit in weak topology as w- $\lim_{n\to\infty}\theta_n$. Let $\mathbf{1}_{(\cdot)}$ denote the indicator function.

2 The Model and Estimation Strategy

Let $Z = (Y, X)$ be a pair of a scalar response variable Y and a square integrable random function $X = \{X(t)\}_{t\in[0,1]}$ on a interval $[0,1]$. Let $Q_\tau(Y|X)$ be the τth conditional quantile function of Y given X for any $\tau \in (0,1)$ that is away from 0 and 1. The τth conditional quantile $Q_\tau(Y|X)$ can be written as a linear functional of X:

$$Q_\tau(Y|X) = \alpha_\tau + \int_0^1 X(t)\,\beta_\tau(t)\,dt, \quad \tau \in (0,1),$$

where $\bar{X}(t) = X(t) - \mathbb{E}[X(t)]$, α_τ is a scalar constant and $\beta_\tau(t)$ is a scalar function in $L^2[0,1]$. Hereafter, we consider estimating the slope function β_τ. The unknown parameter $\theta_0 = (\alpha, \beta)$ belongs to $\mathcal{H} = \mathbb{R} \times L^2[0,1]$.

Our estimation strategy is based on the method of regularization. For the detailes, see Yuan and Cai (2010), Shin and Lee (2016). We suppose $X(t)$ satisfies $\mathbb{E}\left[\int_0^1 |X(t)|^2\,dt\right] < \infty$. We take the slope function $\beta_\tau(t)$ to be an RKHS, $\mathcal{H}$, a subspace of the Hilbert space of square integralbe functions $L^2[0,1]$. We denote the inner product and the associated norm in $\mathcal{H}$ by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ respectively. Suppose we observe data $(Y_i, X_i(t))$, $1 \le i \le n$ consisting of n independent copies of $(Y, X(t))$. With them, we may estimate α_τ, β_τ via by penalization method :

$$\begin{aligned}\left(\hat{\alpha}_{\tau,n,\lambda}, \hat{\beta}_{\tau,n,\lambda}\right) &= \arg\min_{\alpha\in\mathbb{R},\beta\in\mathcal{H}} F_{\tau,n,\lambda}(\theta) \\ &\triangleq \arg\min_{\alpha\in\mathbb{R},\beta\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^n \rho_\tau\left(Y_i - \alpha - \int_0^1 X_i(t)\,\beta(t)\,dt\right) + \lambda_n J(\beta),\end{aligned} \tag{2.1}$$

where $\rho_\tau(u) = \left\{\tau - \mathbf{1}_{(u\le 0)}\right\}u$ is the check function (Koenker and Bassett Jr (1978)), λ_n is the smoothing parameter that converges to zero as $n \to \infty$ and $J(\beta)$ is a convex penalty functional on β. Obviously, the criterion function $\rho_\tau(\cdot)$ is not continuously differentiable.

Similarly to Yuan and Cai (2010), we assume the penalty functional J is a squared semi-norm on $\mathscr{H}$. Let $\mathcal{H}_0$ be a finite dimensional subspace of $\mathcal{H}$ such that

$$\mathcal{H}_0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}$$

with orthonormal basis $\{\nu_1, \cdots, \nu_N\}$ and $\dim(\mathcal{H}_0) = N$. Let $\mathcal{H}_1$ be the orthogonal complement of $\mathcal{H}_0$ in $\mathcal{H}$ and $\mathcal{H}$ has an orthogonal decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. In this paper, we suppose that $J(\beta) = \|\pi_{\mathcal{H}_1}\beta\|^2_{\mathcal{H}}$, where $\pi_{\mathcal{H}_1}$ is the orthogonal projection of $\beta \in \mathcal{H}$ onto a subspace $\mathcal{H}_1$. The canonical example of penalized functional is $J(\beta) = \int_0^1 |\beta(t)|^2\,dt$ (see, for example Koenker et al. (1994), Portnoy (1997)).

Let $K(\cdot,\cdot)$ be the reproducing kernel of $\mathcal{H}_1$ such that $J(f_1) = \|f_1\|^2_K = \|f_1\|^2_{\mathcal{H}}$ for all $f_1 \in \mathcal{H}_1$. We assume that $K(\cdot,\cdot)$ is continuous and square integrable. By reproducing property, we have

$\beta(\tau) = \langle \beta(\cdot), K(u,\cdot)\rangle_{\mathcal{H}}$. The objective function (2.1) is rewritten as

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \alpha - \langle \xi_i, \beta\rangle_{\mathcal{H}}\right) + \lambda_n J(\beta)$$

where $\xi_i(t) = \int_0^1 x_i(t) K(u,t)\, du$. Then, by the representer theorem, the minimizer over β in (2.1) can be written as

$$\hat{\beta}_{\tau,n,\lambda} = \sum_{i=1}^{n} c_i \xi_i(t) + \sum_{k=1}^{N} d_k \nu_k(t)$$

where $d = (d_1, \cdots, d_N) \in \mathbb{R}^N$ and $c = (c_1, \cdots, c_n) \in \mathbb{R}^n$.

Obtaining the estimator of $\hat{\theta}_{\tau,n,\lambda} = \left(\hat{\alpha}_{\tau,n,\lambda}, \hat{\beta}_{\tau,n,\lambda}\right)$, we put an estimator of the conditional τth quantile of Y given $X = x(t)$ as

$$\hat{Q}_\tau(Y|X) = \hat{\alpha}_{\tau,n,\lambda} + \int_0^1 X(t)\, \hat{\beta}_{\tau,n\lambda}(t)\, dt$$

by plug-in method. The purpose of this paper is to derive the asymptotic statistical inference of $\hat{Q}_\tau(Y|x_0)$ for any nonrandom $x_0 \in L^2[0,1]$ and the asymptotic distribution of the penalized likelihood ratio test statistic $F_{\tau,n,\lambda}(\theta_{\tau,0,\lambda}) - F_{\tau,n,\lambda}\left(\hat{\theta}_{\tau,n,\lambda}\right)$ where $\theta_{\tau,0,\lambda}$ is the minimizer of

$$F_{\tau,0,\lambda}(\theta) \triangleq \mathbb{E}\left[\rho_\tau\left(Y - \alpha - \langle \xi, \beta\rangle_{\mathcal{H}}\right) + \lambda_n J(\beta)\right] \tag{2.2}$$

which is the population counterpart of $F_{\tau,n,\lambda}$. These results are established in a fully infinite dimensional setting.

3 Mosco Convergence and Quadratic Approximation

3.1 Lack of Uniform Convergence

Before describing our proposed techniques, let us explain a lack of uniform convergence issue for an infinite dimensional circumstance briefly. Recall that uniform convergence of the objective function to its population counterpart is follows:

$$\sup_{\theta \in \mathbb{R}\times\mathcal{H}} \left|F_{\tau,n,\lambda}(\theta) - \mathbb{E}\left[F_{\tau,n,\lambda}(\theta)\right]\right| \xrightarrow{p} 0.$$

In order to make the objective function satisfy the uniform convergence, we have to impose some compactness or entropy conditions on the parameter space $\mathbb{R}\times\mathcal{H}$ (e.g., van der Vaart (1998)). These assumptions are rather restrictive for non-differentiable convex objective function settings. It is because of the theorem by Bakhvalov (Theorem 12.1.1. of Dudley (1999)) . When $\rho = |\cdot|$ and θ is in an

infinite-dimensional space, we have

$$\sup_{\theta\in\mathbb{R}\times\mathcal{H}} \left|F_{\tau,n,\lambda}\left(\theta\right) - \mathbb{E}\left[F_{\tau,n,\lambda}\left(\theta\right)\right]\right| \geq \gamma$$

for some constant γ. The left-hand side of the inequality does not converge uniformly. The convexity lemma argument to ensure that point-wise convergence of convex functions implies uniform convergence may fail. Let $\pi_n,\ n = 1, 2, \cdots$ be the sequence of projection operators on $\mathscr{H}$ onto $E_n \subset \mathscr{H}$ where $E_n \subsetneqq E_{m>n}$. Consider a quadratic form $\langle \pi_n\theta, \theta\rangle$ for $\forall\theta \in \mathscr{H}$ that is considered as a convex function of θ. Then, as $n \to \infty$, $\langle \pi_n\theta, \theta\rangle$ converges point-wise to $\langle \theta, \theta\rangle$ but not uniformly.

To solve the aforementioned lack-of-uniform-convergence issue, we apply the *Mosco convergence*, which is weaker than uniform convergence but still strong enough to enable statistical applications. Mosco convergence of the objective function ensures the convergence of its minimizer (Attouch (1984)). If the parameter space is weakly compact, Mosco convergence of the convex objective function ensures that both empirical minimizer and empirical optimal value function will converges to the true minimizer and the true optimal value function respectively. This property makes it possible to derive the asymptotic distribution of the penalized likelihood ratio test statistic.

First, we introduce a mode of convergence, *Mosco convergence*, for proper lower semi-continuous (l.s.c.) convex functions on a real separable Hilbert space. For l.s.c. convex functions on a finite dimensional Euclidean space, point-wise convergence is equivalent to locally uniform convergence. For functions defined on an infinite-dimensional space, however, this is not the case. Mosco convergence, on the other hand, still ensures arg min convergence of l.s.c. convex functions on an infinite-dimensional space, though it is weaker than locally uniform convergence. In this section, we also provide preliminary results related to Mosco convergence for later use.

3.2 Mosco Convergence

Mosco convergence and similar concepts in a non-stochastic environment are considered in Mosco (1969), Attouch (1984) and Beer (1993). Mosco convergence is particularly useful in the context of functional optimization, making it well suited to stochastic optimization.

Definition 1. [Mosco Convergence]
Let $f_n : \mathcal{H} \to (-\infty, \infty],\ n = 1, 2, \ldots$ be a sequence of proper lower semi-continuous(l.s.c.) convex functions. f_n is said to be Mosco-convergent to the l.s.c. convex function $f : \mathcal{H} \to (-\infty, \infty]$ if and only if the following two conditions hold.
(M1) For each $\theta \in \mathcal{H}$, there exist a convergent sequence $\theta_n \xrightarrow{s} \theta$ such that $\limsup_n f_n\left(\theta_n\right) \leq f\left(\theta\right)$.
(M2) $\liminf_n f_n\left(\theta_n\right) \geq f\left(\theta\right)$ whenever $\theta_n \xrightarrow{w} \theta$.
In this paper, we let "$f_n \xrightarrow{M} f$" denote "f_n Mosco-converges to f."

The variational properties of Mosco convergence are given by the following theorem (Theorem

1.10 in Attouch (1984)), which ensures the convergence of both empirical minimizer and empirical minimum value of the objective function to the true ones. Suppose $\arg\min f_n \neq \emptyset$, and existence of $\arg\min f_n$ and $\inf f_n$ are proved in Appendix **??**.

Theorem 2. *We assume the same definitions for* $f_1, f_2, \cdots$ *and* f. *If* $f_n \overset{M}{\to} f$, *then*

$$\langle \arg\min f_n, h \rangle \to \langle \arg\min f, h \rangle \quad (\forall h \in \mathcal{H}^*),$$

in the weak topology. If there is a weakly compact set $K \subset \mathscr{H}$ *such that* $\arg\min f_n \subset K$ *for all* n, *then* $\lim_{n\to\infty} (\inf f_n) = \inf f$.

It is difficult to prove Mosco convergence directly in general settings. Fortunately, several equivalence conditions for Mosco convergence are known in the literature. One of the most convenient conditions for Mosco convergence is point-wise convergence of subdifferentials of functions.

To deal with this mode of convergence, we introduce several basic tools in convex analysis: subdifferential and resolvent. For more details and proofs on these subjects, see Aubin and Frankowska (1990). For fixed $Z \in E$ where E is an arbitrary topological space, we can define a set-valued mapping $\partial f(\theta, Z) : \Theta \times E \to \mathscr{H}$ by

$$\partial f(\theta, Z) = \left\{ \theta \in \mathscr{H} : {}^{\forall}\zeta \in \mathscr{H},\ f(\zeta, Z) \geq f(\theta, Z) + \langle \zeta - \theta,\ \theta \rangle \right\}.$$

Such $\partial f(\theta, \cdot)$ is said to be the *subdifferential* of f at θ. For each fixed θ, $\partial f(\theta, Z)$ is considered as a possibly set-valued function of Z. We may regard $\partial f(\theta, Z)$ as a generalized derivative of f at θ, for each fixed Z. If f is Gâteaux differentiable at θ and has a continuous Gâteaux derivative $\nabla f(\theta)$, then $\partial f(\theta, Z) = \nabla f(\theta, Z)$. Subdifferential operator for proper l.s.c. convex functions hold distributive law:

$$\partial (f_1 + f_2) = \partial f_1 + \partial f_2$$

where f_1 and f_2 are proper l.s.c. convex functions on $\mathscr{H}$ (see Theorem 3.16. in Phelps (1992)). When $\mathscr{H}$ is real separable, subdifferential operator is exchangeable with respect to integral (Clarke (1983) page 76.):

$$\partial f(\theta) = \partial \int_E f(\theta, Z)\, \mathbb{P}_Z(dZ) = \int_E \partial f(\theta, Z)\, \mathbb{P}_Z(dZ).$$

Subdifferetial calculus for a linear quantile are given by the following lemma:

Lemma 3. $f(\theta) = \|\theta\|_{\mathcal{H}}^2$

$$\partial f(\theta) = 2\theta \ \ (\theta \in \mathcal{H})$$

Lemma 4. *The criterion function $\rho_\tau\left(y_i - \langle x_i, \theta\rangle\right)$ is a proper l.s.c. convex function with respect to θ and has the subdifferential such that*

$$\partial\rho_\tau\left(y_i - \langle x_i, \theta\rangle\right) = \begin{cases} \left\{\tau - \mathbf{1}_{(y-\langle x,\theta\rangle\leq 0)}\right\} x, & \text{if } y - \langle x, \theta\rangle \neq 0; \\ [-1, 1]\, x, & \text{if } y - \langle x, \theta\rangle = 0. \end{cases}$$

Proof. Proof is given in Appendix A.1. □

Lemma 5. *The limit criterion $\mathbb{E}\left[\rho_\tau\left(y - \langle x, \theta\rangle\right)\right]$ is convex function and has the subdifferential*

$$\partial\mathbb{E}\left[\rho_\tau\left(y - \langle x, \theta\rangle\right)\right] = \mathbb{E}\left[\partial\rho_\tau\left(y - \langle x, \theta\rangle\right)\right],$$

and

$$\begin{aligned} \mathbb{E}\left[\partial\rho_\tau\left(y - \langle x, \theta\rangle\right)\right] &= \mathbb{E}\left[\left\{\tau - \mathbf{1}_{(y-\langle x,\theta\rangle\leq 0)}\right\} x\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\{\tau - \mathbf{1}_{(y-\langle x,\theta\rangle\leq 0)}\right\} x \middle| x\right]\right] \\ &= \mathbb{E}\left[x\left\{\tau - f_{Y|X}\left(\langle x, \theta\rangle \middle| x\right)\right\}\right]. \end{aligned} \tag{3.1}$$

In this paper, we assume that the subdifferential ∂f is selected and measurable in Z. In general, because ∂f is a set-valued mapping, the selection is not unique. Nonetheless, we can show that not only such measurable selections exist but also the set of all measurable selector $S_{\partial f}$ is identical to ∂f.

Proposition 6. *There exists a measurable selector of the subdifferential ∂f, i.e., $S_{\partial f} \neq \emptyset$. Moreover, $S_{\partial f} = \partial f$.*

Proof. Proof is given in Appendix A.2 in **?** □

Consider a map

$$R_\lambda^{\partial f}\zeta = \left\{z \in \mathscr{H} \ : \ z + \lambda\partial f\left(z\right) \ni \zeta\right\}.$$

Such a map should be single-valued (on Proposition 3.5.3 in Aubin and Frankowska (1990)). Such $R_\lambda^{\partial f}, \lambda > 0$ are called *resolvents* of ∂f and denoted by

$$^\forall\lambda > 0, \quad R_\lambda^{\partial f} = \left(I + \frac{1}{\lambda}\partial f\right)^{-1}.$$

Such map is single-valued and equal to the solution of penalized convex optimization problem:

$$R_\lambda^{\partial f}\zeta = \arg\min_z \left\{f\left(z\right) + \lambda\left\|z - \zeta\right\|_{\mathcal{H}}^2\right\}.$$

Therefore, considering the convergence of this resolvent will leads to the convergence of an estimation

problem (2.1). We write $R_\lambda^{\partial F_0}0$ as

$$R_\lambda^{\partial F_0}0 = \theta_{\tau,0,\lambda} = \arg\min_{\alpha,\beta} \mathbb{E}\left[\rho_\tau\left(Y - \alpha - \langle \xi, \beta\rangle_{\mathcal{H}}\right) + \lambda J\left(\beta\right)\right]$$

for each τ and $R_\lambda^{\partial F_n}0$ as

$$R_\lambda^{\partial F_n}0 = \theta_{\tau,n,\lambda} = \arg\min_{\alpha,\beta} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - \alpha - \langle \xi_i, \beta\rangle_{\mathcal{H}}\right) + \lambda J\left(\beta\right)$$

for each τ.

The following theorem states the equivalence between Mosco convergence and strong convergence of resolvents and G-convergence of subdifferential operators. The proofs are given in Theorem 3.26. and Theorem 3.66. of Attouch (1984).

Theorem 7. *Let $\mathscr{H}$ be a real separable Hilbert space. Let $(f_n)_{n\in\mathbb{N}}$, $f_n : \mathscr{H} \to (-\infty, \infty]$, $^\forall n \in \mathbb{N}$ be a proper l.s.c. convex function. The following statements are equivalent.*
(1) $f_n \xrightarrow{M} f_0$.
(2) $^\forall\lambda > 0$, $^\forall\theta \in \mathscr{H}$, $R_\lambda^{\partial f_n}\theta \to R_\lambda^{\partial f}\theta$ *strongly in* $\mathscr{H}$ *as* n *goes to* ∞.
(3)
$$\begin{cases} \partial f_n \xrightarrow{G} \partial f_0, \\ {}^\exists\left(\theta_0, \eta_0\right) \in \partial f_0\ {}^\exists\left(\theta_n, \eta_n\right) \in \partial f_n \text{ such that } \theta_n \xrightarrow{s} \theta_0,\ \eta_n \xrightarrow{s} \eta_0,\ f_n\left(\theta_n\right) \to f_0\left(\theta_0\right), \end{cases}$$
where $\partial f_n \xrightarrow{G} \partial f_0$ *means that, for every* $\left(\theta_0, \eta_0\right) \in \partial f_0$, *there exists a sequence* $\left(\theta_n, \eta_n\right) \in \partial f_n$ *such that* $\theta_n \to \theta_0$ *strongly in* $\mathscr{H}$, $\eta_n \to \eta_0$ *strongly in* $\mathscr{H}^*$ $(= \mathscr{H})$.

Remark 8. Statement (3) in Theorem 7 is called G-convergence of monotone operators. This states that point-wise convergence of all measurable selectors of subdifferential operators is equivalent to Mosco convergence of functionals. When the subdifferential is calculable, point-wise convergence of measurable selectors are easy to verify.

Remark 9. From the foregoing theorems: theorem 2, proposition 6 and theorem 7, it will be seen that the law of large numbers(LLN) of subdifferential $\partial\rho\left(\theta\right)$ implies the Mosco convergence. From lemma 13 and the LLN in Banach spaces for each sequence of measurable selectors of $\partial\rho\left(\theta\right)$, we have the LLN of subdiffential $\partial\rho\left(\theta\right)$:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \partial\rho\left(\theta, Z_i\right) &\xrightarrow{P} \mathbb{E}\left[\partial\rho\left(\theta, Z\right)\right] \\ &= \partial\mathbb{E}\left[\rho\left(\theta, Z\right)\right]. \end{aligned}$$

Thus this fact establish the consistency of local functional estimation.

(2) in the above theorem 7 give a metric that induces the Mosco convergence. Based on resolvet, Attouch (1984) (p. 365) gives a metric that induces graph convergence on the space of subdifferential

operators:

$$d_G\left(\partial f, \partial g\right) \triangleq \sum_{k\in\mathbb{N}} \frac{1}{2^k} \inf\left\{1, \left\| R_{\lambda_0}^{\partial f}\theta_k - R_{\lambda_0}^{\partial g}\theta_k \right\|\right\},$$

for any subdifferential operators ∂f and ∂g where λ_0 is taken strictly positive and $\{\theta_k; k \in \mathbb{N}\}$ is a dense subset of $\mathscr{H}$. This metric d_G induces the Mosco convergence topology and is complete. Convergence in d_G are equivalent to the convergence results in (1)$\sim$(3) in Theorem 7.

Hoffman-Jørgensen weak convergence theory performs in a metric space. Generally, epi-convergence does not usually work with a metric but a semi-metric. Even if functions f, g are different each other, it is possible f epi-converges to g (see, Section 3 in Bucher et al. (2014)). Fortunately, in the case where the functional space is constituted by convex functions, we can obtain a metric space as described above. We shall define a weak convergence in the following way.

Definition 10. [Mosco Convergence in Distribution]
A sequence of random elements f_n in the space of proper l.s.c. convex functions $\mathscr{H} \to (-\infty, \infty]$ is said to be Mosco converges in distribution to the random element f_0 in the space of proper l.s.c. convex functions if $f_n \rightsquigarrow f_0$ with metric d_G. We use the notation $f_n \overset{M}{\rightsquigarrow} f_0$.

3.3 Second Order Differentiability

In typical situations, we assume that the function F_0 has a quadratic expansion at θ_0 and their Hessian is often supposed to be continuously invertible (Theorem 3.3.1. of van der Vaart and Wellner (1996)). In an infinite-dimensional case, the assumption that the Hessian operator is continuously invertible is harder to ascertain. However, if the convex function F_0 has a *generalized second order differentiability* (defined later), its "generalized Hessian" is continuously invertible.

Define the Young-Fenchel conjugate f^* of convex function f as

$$f^*\left(\eta\right) \triangleq \sup_{\theta}\left(\langle\eta, \theta\rangle - f\left(\theta\right)\right).$$

The conjugate f^* has a strong link between a convex function f in the second order differentiability. Recall the case of a convex function defined on finite dimensional parameters. A convex function f defined on the Euclid space $\mathbb{R}^d$ is second order differentiable and the Hessian $\nabla^2 f\left(\theta\right)$ of f at θ is non-degenerate. Then the conjugate function f^* is second order differentiable at $y = \nabla f\left(\theta\right)$, and its Hessian $\nabla^2 f^*\left(\eta\right)$ at y is the inverse of $\nabla^2 f\left(\theta\right)$, i.e.,

$$\nabla^2 f\left(\theta\right) = \left(\nabla^2 f^*\left(\eta\right)\right)^{-1}.$$

In order to maintain a duality-type of this relation in an infinite-dimensional space, we shall define the second order differential concepts based on Mosco convergence. Mosco convergence ensures the

continuity of this type of conjugation (Kato (1989) and Borwein and Noll (1994)).

Define *second difference quotient* of f at $\theta \in \mathscr{H}$ relative to $\eta^* \in \partial f(\theta)$ as

$$\Delta_{f,\theta,\eta,t}(h) \triangleq \frac{f(\theta+th) - f(\theta) - t\langle \eta^\star, h\rangle}{t^2}$$

and define a *purely quadratic* continuous convex function as

$$q(h) \triangleq \frac{1}{2}\langle Vh, h\rangle,$$

where V is a closed symmetric positive linear operator. f is said to have *generalized second order differentiability* at θ relative to $\eta^\star \in \partial f(\theta)$ if there exists a purely quadratic function q such that the second order difference quotient $\Delta_{f,\theta,\eta,t}(\cdot)$ converges to $q(\cdot)$ in the Mosco sense, i.e.,

$$\Delta_{f,\theta,\eta,t}(h) \xrightarrow[t\downarrow 0]{M} q(h).$$

The closed symmetric positive linear operator V is called the *generalized Hessian* of f at θ relative to $\eta \in \partial f(\theta)$.

Mosco convergence is invariant under Young-Fenchel conjugation, so that Mosco convergence of $\Delta_{f,\theta,\eta,t}(h)$ is equivalent to Mosco convergence of $(\Delta_{f,\theta,\eta,t}(h))^* = \Delta_{f*,\eta,\theta,t}(h)$. And generalized Hessian of f^* at η relative to $\theta \in \partial f^*(\eta)$ is V^{-1}.

Next, we derive sufficient conditions under which the objective function of stochastic optimization has generalized second order differentiability. ∂f is called *weak* Gâteaux differentiable* at θ if there exists a bounded linear operator $T : \mathscr{H} \to \mathscr{H}^*$ such that

$$\lim_{t\to 0}\frac{1}{t}(\eta_t^* - \eta^*) = Vh,$$

in the weak* sense for any fixed $h \in \mathscr{H}$ and all $\eta_t^* \in \partial f(\theta+th)$, $\eta^* \in \partial f(\theta)$ where $\partial f(\theta)$ must consist of a single element η^*. We use the notation $T = \nabla\partial f(\theta)$ for the operator T. For the generalized differentiability, we quote the following result of Borwein and Noll (1994).

Theorem 11. *(a variant of Proportion 6.4. of Borwein and Noll (1994))*
Let $(Z, \mathcal{Z}, \mathbb{P}_Z)$ be a probability space and $\Theta \subseteq \mathscr{H}$ be a separable Hilbert space. Suppose $\rho : \Theta \times Z \to (-\infty, \infty]$ is measurable on $(Z, \mathcal{Z}, \mathbb{P}_Z)$ and convex at any $\theta \in \Theta$ and define a closed convex integral functional f on $\Theta \subset \mathscr{H}$ as

$$f(\theta) = \int_Z \rho(\theta, z)\, d\mathbb{P}_Z(z).$$

Then f is generalized second order differentiable at θ if and only if $\partial\rho$ is weak Gâteaux differentiable*

and

$$\operatorname*{ess\,sup}_{z\in Z}\left|\nabla\partial\rho\left(\theta,z\right)\right|<\infty.$$

Therefore, in order to obtain invertibility of "generalized Hessian", we impose the following assumption on ρ:

Assumption. *A*
$\partial\rho\left(\cdot\right)$ is weak Gâteaux differentiable at θ_0 and*

$$\operatorname*{ess\,sup}_{(y.x)}\left|\lim_{t\to0}\frac{\partial\rho_\tau\left(y_i-\left\langle x_i,\theta_0+th\right\rangle\right)-\partial\rho_\tau\left(y_i-\left\langle x_i,\theta_0\right\rangle\right)}{t}\right|<\infty.$$

This assumption is a "low-level" condition which are sufficient for locally asymptotically quadratic at θ_0 than that of Geyer (1994). Of course, this result is attributed to the convexity of the objective function.

3.4 Quadratic Approximation

A common starting point in developing an asymptotic distribution theory for an M-estimator is to define a centered stochastic process based on the objective function. We may define such a centered stochastic process as

$$H_{\tau,n,\lambda}\left(t\right)\triangleq n\left[F_{\tau,n,\lambda}\left(\theta_0+\frac{1}{\sqrt{n}}t\right)-F_{\tau,0,\lambda}\left(\theta_0\right)\right], \tag{3.2}$$

where $F_{\tau,0,\lambda}\left(\theta\right)=\mathbb{E}\left[\rho_\tau\left(Y-Q_\tau\left(Y|X\right)\right)+\lambda J\left(\beta\right)\right]$. And

$$Q_{\tau,0,\lambda}\left(t\right)\triangleq\left\langle t,W\right\rangle+\frac{1}{2}\left\langle\left(V+\lambda I\right)t,t\right\rangle, \tag{3.3}$$

where W is an $N\left(\mathbf{0},A\right)$ random vector in a Hilbert space and V is a "Hessian" operator. Note that $t=\sqrt{n}\left(\theta_{\tau,n,\lambda}-\theta_{\tau,0,\lambda}\right)$ minimizes $H_{\tau,n,\lambda}\left(t\right)$. $H_{\tau,n,\lambda}\left(\theta,t\right)$ is interpreted as the log likelihood ratio for hypothesis testing against the local alternative, i.e., $\mathcal{H}_0:\ \theta=\theta_{\tau,0,\lambda}$; $\mathcal{H}_1:\ \theta=\theta_{\tau,0,\lambda}+\frac{1}{\sqrt{n}}t$. Also define auxiliary stochastic process as

$$\begin{aligned}
G_{\tau,n,\lambda}\left(t\right)&\triangleq n\left\langle\frac{1}{\sqrt{n}}t,\partial F_{\tau,n,\lambda}\left(\theta_{\tau,0,\lambda}\right)\right\rangle+n\left[F_{\tau,0,\lambda}\left(\theta_{\tau,0,\lambda}+\frac{1}{\sqrt{n}}t\right)-F_{\tau,0,\lambda}\left(\theta_{\tau,0,\lambda}\right)\right],\\
G'_{\tau,n,\lambda}\left(t\right)&\triangleq n\left\langle\frac{1}{\sqrt{n}}t,\partial F_{\tau,n,\lambda}\left(\theta_{\tau,0,\lambda}\right)\right\rangle+\frac{1}{2}\left\langle\left(V+\lambda I\right)t,t\right\rangle.
\end{aligned}$$

We also impose the following assumption. Considering Proposition 6 : the set of all measurable selectors of a subdifferential coincides with its own subdifferential, we denote any measurable selector

of $\partial\rho(\cdot)$ as itself.

Assumption. *B*
Every measurable selector in $\partial\rho(\theta, Z)$ has a bounded variance: $\forall\theta \in \Theta$, $\mathbb{E}\left[\|\partial\rho(\theta, Z)\|^2\right] < \infty$, and there is a sequence of measurable selectors satisfying a central limit theorem in the Hilbert space:

$$\mathbb{G}_n \partial\rho(\theta_0, Z) \rightsquigarrow N(0, A),$$

for some trace class covariance operator A.

Proposition 12.

1. $H_{\tau,n,\lambda}(t)$ Mosco-converges to $G'_{\tau,n,\lambda}(t)$ in probability.
2. $G'_{\tau,n,\lambda}(t)$ converges in law to $Q_{\tau,0,\lambda}(t)$. Then, $H_{\tau,n,\lambda}(t)$ Mosco-converge in law to $Q_{\tau,0,\lambda}(t)$.

Proof. See Appendix section A.2. □

Next, we will also show convergence of the minimizer of $H_{\tau,n,\lambda}$ to that of $Q_{\tau,0,\lambda}$, provided that the minimizer is almost surely unique. This follows from the following lemma.

Lemma 13. *The minimizer of the function $Q_{\tau,0,\lambda}(t) = \langle t, W\rangle + \frac{1}{2}\langle (V+\lambda I)t, t\rangle$ is single valued.*

Proof. Let $t_0 = \arg\min_t Q_{\tau,0,\lambda}(t)$. Suppose there exists $t_1 (\neq t_0)$ such that

$$\langle t_1, W\rangle + \frac{1}{2}\langle (V+\lambda I)t_1, t_1\rangle = \langle t_0, W\rangle + \frac{1}{2}\langle (V+\lambda I)t_0, t_0\rangle = \alpha.$$

Then,

$$\begin{aligned}
&\left\langle \frac{t_1+t_0}{2}, W\right\rangle + \frac{1}{2}\left\langle (V+\lambda I)\frac{t_1+t_0}{2}, \frac{t_1+t_0}{2}\right\rangle \\
<&\frac{1}{2}\langle t_1, W\rangle + \frac{1}{2}\langle t_0, W\rangle + \frac{1}{2}\left(\frac{1}{2}\langle (V+\lambda I)t_1, t_1\rangle + \frac{1}{2}\langle (V+\lambda I)t_0, t_0\rangle\right) \\
=&\frac{1}{2}\alpha + \frac{1}{2}\alpha = \alpha.
\end{aligned}$$

This means $Q_{\tau,0,\lambda}\left(\frac{t_1+t_0}{2}\right) < \alpha$, which is contradiction. □

4 Main Results

4.1 Asymptotic Normaltiy

Next, we show that the reparametrized objective function admits a certain quadratic expansion. Note that objective function of quantile regression is

$$\begin{aligned} F(\theta) &= \mathbb{E}\left[\rho_\tau\left(Y - \langle x, \theta\rangle\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\rho_\tau\left(Y - \langle x, \theta\rangle\right)|x\right]\right]. \end{aligned}$$

Then, quantile regression objective function $F(\theta)$ is generalized second order differentiable at θ if and only if $\partial\mathbb{E}\left[\rho_\tau\left(Y - \langle x, \theta\rangle\right)|x\right]$ is weak* Gâteaux differentiable and

$$\operatorname*{ess\,sup}_{x\in X} \left|\nabla\partial\mathbb{E}\left[\rho_\tau\left(Y - \langle x, \theta\rangle\right)|x\right]\right| < \infty.$$

From (3.1), weak* Gâteaux differentiability of $\partial\mathbb{E}\left[\rho_\tau\left(Y - \langle x, \theta\rangle\right)|x\right]$ at θ is equivalent to the Gâteaux differentiability of the distribution function $F_e\left(q - \langle x, \theta\rangle\,|x\right)$ at θ. If the distribution function $F_e\left(q - \langle x, \theta\rangle\,|x\right)$ is Gâteaux differentiable at θ, essential boundedness of

$$\operatorname*{ess\,sup}_{x\in X} \left|\nabla\partial\mathbb{E}\left[\left|Y - \langle x, \theta\rangle\right|\,|X\right]\right| < \infty$$

will be automatically satisfied.

We apply the previous results to consider the asymptotic distribution of $\sqrt{n}\left\langle\hat{\theta}_{\tau,n,\lambda} - \theta_{\tau,0,\lambda}, \theta^*\right\rangle$ in the weak topology.

Proposition 14. *Asymptotic Normality*
Let W be an $N(0, A)$ distribution. Under Assumption A and B, we obtain the asymptotic distribution of $\sqrt{n}\left\langle\hat{\theta}_{\tau,n,\lambda} - \theta_{\tau,0,\lambda}, \theta^*\right\rangle$ as following;

$$\sqrt{n}\left\langle\hat{\theta}_{\tau,n,\lambda} - \theta_{\tau,0,\lambda}, \theta^*\right\rangle \rightsquigarrow \left\langle V^{-1}W, \theta^*\right\rangle \qquad \forall\theta^* \in \Theta$$

where V^{-1} is generalized Hessian of Young-Fenchel conjugate of $F_{\tau,0,\lambda}(\theta)$.

Proof. From Proposition 12, $H_{\tau,n,\lambda}\left(\theta_0, \hat{t}\right)$ converges weakly to $Q_{\tau,0,\lambda}(t)$ in Mosco topology. Applying a.s. representation theorem(Theorem1.10.4 in van der Vaart and Wellner (1996)) we get an almost sure representation $H_{\tau,n,\lambda} \xrightarrow{M} Q_{\tau,0,\lambda}\ a.s.$. By Theorem 7 we have

$$\lim_{n\to\infty}\left(\arg\min H_{\tau,n,\lambda}\right) \to \arg\min Q_N\ a.s.$$

in the weak topology. This provide

$$\sqrt{n}\left\langle \hat{\theta}_{\tau,n,\lambda}-\theta_{\tau,0,\lambda},\theta^{*}\right\rangle \rightsquigarrow\left\langle V^{-1}W,\theta^{*}\right\rangle \qquad \forall\theta^{*}\in\Theta.$$

□

For the implement, we need a consistent estimator of the generalized Hessian. From the fact of the properties of the generalized differential, the natural candidates are

$$\lim_{h_n\to 0}\frac{1}{k_n}\left(\hat{\eta}_{k_n}^{*}-\hat{\eta}^{*}\right)$$

in the weak* sense for any fixed $h\in\mathscr{H}$ and all $\hat{\eta}_{k_n}^{*}\in\partial f\left(\hat{\theta}+k_nh\right),\hat{\eta}^{*}\in\partial f\left(\hat{\theta}\right)$.

4.2 Confidence Interval.

This subsection consider a confidence interval for the conditional quantile. We consider the plug-in estimate $\hat{Y}=\hat{\alpha}_{\tau}+\int_0^1 x_0\left(t\right)\hat{\beta}_{\tau}\left(t\right)dt$. By proposition 14 with the Delta method, we obtain the proposition below on the pointwise confidence interval where the asymptotic estimation bias is assumed to be removed by undersmoothing.

Corollary 15. *Suppose Assumptions A1, A2 and A3 are satisfied. Then*

$$\frac{\sqrt{n}}{\left\langle x_0V^{-1}A,x_0\right\rangle}\left(\hat{Y}_0-Y_0\right)\rightsquigarrow N\left(0,1\right)\qquad {}^{\forall}x_0\in L^2\left[0,1\right].$$

Hence, the $(1-\alpha)$ *confidence inter val for* Y_0 *is*

$$\left[\hat{Y}_0\pm\frac{1}{\sqrt{n}}W_{\alpha/2}\left\langle x_0V^{-1}A,x_0\right\rangle Y_0\right],$$

where $W_{\alpha/2}$ *is the* $\left(1-\frac{1}{2}\alpha\right)$*-quantile of standard normal distribution.*

4.3 Estimation with Convex Constraints

In this subsection we consider a stochastic optimization of a parameter constrained to some convex set in $\mathscr{H}$. To avoid subtlety of the asymptotics of constrained estimator, we concentrate our attention on the convex constrained. Define an objective function with convex constraint $G\left(\theta\right)$ from $\mathscr{H}$ to $(-\infty,\infty]$ by

$$G_n\left(\theta\right)=F_n\left(\theta\right)+\Psi_A\left(\theta\right)$$

where Ψ_A is defined by

$$\Psi_A(\theta) = \begin{cases} 0 & (\theta \in A) \\ \infty & (\theta \notin A) \end{cases}$$

and A is convex subset of parameter space Θ in a Hilbert space $\mathscr{H}$.

Lemma 16 ("Optimization Theory" Indicator Function)**.** *The indicator function Ψ_A is defined by*

$$\Psi_A(\theta) = \begin{cases} 0 & (\theta \in A) \\ \infty & (\theta \notin A) \end{cases}$$

where the set A is a convex subset of Θ. The normal cone $N_A(a)$ is defined by

$$N_A(a) = \left\{\theta^* \in \mathscr{H} : \langle \theta - a, \theta^* \rangle \leqq 0, \ {}^{\forall}\theta \in A\right\}.$$

Then, $N_A(a) = \partial\Psi_A(a)$, where $N_A(a)$ is such that $0 \in N_A(a)$.

Proof.

$$\begin{aligned} \theta^\star \in \partial\Psi_A(a) &\Leftrightarrow \Psi_A(a) + \langle \theta - a, \theta^\star \rangle \leqq \Psi_A(\theta) \ (\forall\theta \in A) \\ &\Leftrightarrow \langle \theta - a, \theta^\star \rangle \leqq \Psi_A(\theta) \ (\forall\theta \in A) \\ &\Leftrightarrow \langle \theta - a, \theta^\star \rangle \leqq 0 \ (\forall\theta \in A) \\ &\Leftrightarrow \theta^\star \in N_A(a) \end{aligned}$$

Then, $N_A(a) = \partial\Psi_A(a)$. □

Because F_n and Ψ_A are convex function, $G_n(\theta)$ are also convex function with respect to θ for all n. Let $A_n = \sqrt{n}(A - \theta_0)$. Redefine (3.2), (3.3) as

$$\begin{aligned} H_n^{A_n}(\theta, t) &\triangleq n\left[F_n\left(\theta + \frac{1}{\sqrt{n}}t\right) - F_n(\theta)\right] + \Psi_{A_n}(t) \\ Q_0^A(t) &\triangleq \langle t, Z \rangle + \frac{1}{2}\langle Vt, t \rangle + \Psi_{T_A(\theta_0)}(t) \end{aligned}$$

where $T_A(\theta_0)$ is tangent cone:

$$T_A(\theta_0) = \limsup_{\tau \downarrow 0} \frac{A - \theta_0}{\tau}.$$

The following corollary follows from proposition 12 and proposition **??**.

Corollary 17. *Suppose Ψ_{A_n} Mosco-converges to $\Psi_{T_{A(\theta_0)}}$. Then, $H_n^{A_n}(t)$ Mosco-converges in law to $Q_0^A(t)$.*

4.4 Asymptotics of Likelihood Ratio Test

Using the previous proposition 12, we derive the asymptotic distribution of the likelihood ratio statistics. Let $A_n = \sqrt{n}(\Theta - \theta_0)$ and $A_{n,0} = \sqrt{n}(\Theta_0 - \theta_0)$. The likelihood ratio statistic is written by the form

$$\Lambda_n = \inf_{t \in A_n} H_{\tau,n,\lambda}(\theta_0, t) - \inf_{t \in A_{n,0}} H_{\tau,n,\lambda}(\theta_0, t).$$

By the previous proposition 12, for large n, the likelihood ratio process is similar to the same as in the normal experiment. By the Mosco convergence argument in theorem 7, if the parameter space is weakly compact, the empirical optimal value of convex function achieves the true optimal.

Assumption. *C*
The parameter set Θ is weakly compact. In a Hilbert space setting $\Theta \subset \mathscr{H}$, weakly compactness is equal to boundedness: for all $\theta \in \Theta$,there exists constant C such that $\|\theta\| \leq C$.

Lemma 18. *Let W be an $N(0, A)$ distribution and repeat (3.2);*

$$H_{\tau,n,\lambda}(\theta, t) = n\left[F_{\tau,n,\lambda}\left(\theta + \frac{1}{\sqrt{n}}t\right) - F_{\tau,0,\lambda}(\theta)\right].$$

Let $\hat{t} = \sqrt{n}\left(\hat{\theta}_n - \theta_0\right)$ denote this minimizer. Under Assumption A-C, the asymptotic distribution of the optimal value function

$$H_{\tau,n,\lambda}\left(\theta_0, \hat{t}\right) = n\left[F_{\tau,n,\lambda}\left(\hat{\theta}_n\right) - F_{\tau,n,\lambda}(\theta_0)\right]$$

is the distribution of $Q_{\tau,0,\lambda}\left(\hat{t}\right)$.

Proof. From Proposition 12, $H_{\tau,n,\lambda}\left(\theta_0, \hat{t}\right)$ converges weakly to $Q_{\tau,0,\lambda}(t)$ in Mosco topology. Applying a.s. representation theorem (theorem 1.10.4 in van der Vaart and Wellner (1996)) we get an almost sure representation $H_{\tau,n,\lambda} \overset{Mosco}{\longrightarrow} Q_{\tau,0,\lambda}$ a.s.. By Theorem 2 and Assumption C, we have

$$\lim_{n\to\infty}(\inf H_{\tau,n,\lambda}) = \inf Q_{\tau,0,\lambda}.$$

This provide the optimal value of function $H_{\tau,n,\lambda}$ converges weakly to $Q_{\tau,0,\lambda}$. □

From the result of lemma 16 and lemma 18, we obtain the asymptotic distribution of the optimal

value function

$$H^{A}_{\tau,n,\lambda}\left(\theta_0,\hat{t}\right) \rightsquigarrow Q_N\left(\hat{t}\right).$$

The above result yields the asymptotic distribution of the likelihood ratio statistics Λ_n. The proof strategy is based on van der Vaart (1998), Chapter 16, Theorem 16.7.

Proposition 19. *Assume the parameter spaces Θ and Θ_0 is convex. And assume Assumption A-C. If the sets A_n and $A_{n,0}$ converge to sets A and A_0, then the sequence of likelihood ratio statistics Λ_n converges under $\theta_0 + \frac{t}{\sqrt{n}}$ in distribution to*

$$\left\|V^{-\frac{1}{2}}W + V^{\frac{1}{2}}t\left(\in A_{n,0}\right)\right\|^2 - \left\|V^{-\frac{1}{2}}W + V^{\frac{1}{2}}t\left(\in A_{n}\right)\right\|^2$$

where W is an $N\left(\mathbf{0},A\right)$ random vector.

Proof. By Lemma 18 and simple algebra

$$\begin{aligned}
\Lambda_n =& \inf_{t\in A_n} H_{\tau,n,\lambda}\left(\theta_0,t\right) - \inf_{t\in A_{n,0}} H_{\tau,n,\lambda}\left(\theta_0,t\right) \\
=& 2\inf_{t\in A_n}\left(n\left\langle\frac{1}{\sqrt{n}}t,\partial F_n\left(\theta_0\right)\right\rangle + \frac{1}{2}\left\langle Vt,t\right\rangle\right) \\
& - 2\inf_{t\in A_{n,0}}\left(n\left\langle\frac{1}{\sqrt{n}}t,\partial F_n\left(\theta_0\right)\right\rangle + \frac{1}{2}\left\langle Vt,t\right\rangle\right) + o_P\left(1\right) \\
=& \left\|V^{-\frac{1}{2}}\mathbb{G}_n\partial\rho\left(\theta_0\right) + V^{\frac{1}{2}}\hat{t}\left(\in A_{n,0}\right)\right\|^2 - \left\|V^{-\frac{1}{2}}\mathbb{G}_n\partial\rho\left(\theta_0\right) + V^{\frac{1}{2}}\hat{t}\left(\in A_{n}\right)\right\|^2 + o_P\left(1\right)
\end{aligned}$$

the proposition follows by the continuous mapping theorem. □

Consider a likelihood ratio statistics for testing the value of $\langle\theta_0,x_0\rangle$ at any $x_0 \in E$. For some prespecified point (x_0,c), we consider the following hypothesis:

$$H_0 : \langle\theta_0,x_0\rangle \leq 0 \quad \text{vs.} \quad H_1 : \langle\theta_0,x_0\rangle > 0.$$

The objective function under the null constrained is defined as

$$F_{\tau,n,\lambda}\left(\theta^{H_0}\right) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \left\langle x_i,\theta^{H_0}\right\rangle\right) + \frac{\lambda}{2}\left\|\theta^{H_0}\right\|$$

where $\theta^{H_0} \in H_0 = \{\theta\in\Theta : \langle\theta_0,x_0\rangle \leq 0\}$. Note that the set H_0 is convex. We define the generalized likelihood ratio test statistic as

$$\Lambda_n = F_{\tau,n,\lambda}\left(\hat{\theta}^{H_0}\right) - F_{\tau,n,\lambda}\left(\hat{\theta}_n\right),$$

where $\hat{\theta}^{H_0}$ is the M-estimator under convex constraint:

$$\hat{\theta}^{H_0} = \arg \min_{\theta^{H_0} \in H_0} F_{\tau,n,\lambda}\left(\theta^{H_0}\right).$$

If the null the interior of the hypothesis H_0 contains the true parameter θ_0, the sequence of Λ_n converges to zero in distribution. This means that an error of the first kind converges to zero under that the null hypothesis is true. If the true parameter θ_0 belongs to the boundary: $\langle \theta_0, x_0 \rangle = 0$, the sets $\sqrt{n}\left(\Theta_0 - \theta_0\right)$ converge to the $H_0 = \{\theta : \langle \theta, x_0 \rangle \leq 0\}$. The sequence of Λ_n converges in distribution to the distribution of the square distance of a standard normal vector to the half-space $V^{\frac{1}{2}} H_0 = \left\{\theta : \left\langle \theta, V^{-\frac{1}{2}} x_0 \right\rangle \leq 0\right\}$, that is the distribution of $(W \vee 0)^2$.

A Appendix

A.1 Proof of Subdifferential Calculus of $\rho = |y - \langle x, \theta \rangle|$

Here we show the subdifferential calculus of $\rho = |y - \langle x, \theta \rangle|$. We use the following lemma.

Lemma 20. *The subdifferential of* $\|\theta\| = \langle \theta, \theta \rangle$ *is* $\partial \|\theta\| = \{\theta\}$, $\theta \in \mathscr{H}$.

Proof. For $\theta \in \mathscr{H}$,

$$\langle \eta, \theta \rangle - \langle \theta, \theta \rangle = \langle \eta - \theta, \theta \rangle, \quad \eta \in \mathscr{H},$$

then $\partial \|\theta\| = \{\theta\}$. □

Proposition (Subdifferential Calculus of $\rho = |y - \langle x, \theta \rangle|$)**.** *The criterion function* $\rho(\theta, Z) = |y - \langle x, \theta \rangle|$ *is a proper l.s.c. convex function and has the subdifferential such that*

$$\partial \rho(\theta, Z) = \begin{cases} sgn\left(y - \langle x, \theta \rangle\right) x, & if\ y - \langle x, \theta \rangle \neq 0; \\ [-1, 1]\, x, & if\ y - \langle x, \theta \rangle = 0. \end{cases}$$

Proof. Let $t \in [-1, 1]$, $\theta = tx$. For all $\zeta \in \mathscr{H}$,

$$\langle tx, \zeta - \theta \rangle = t \langle x, \zeta \rangle - ty \leq t\left|\langle x, \zeta \rangle - y\right| \leq |t|\left|\langle x, \zeta \rangle - y\right| \leq \left|\langle x, \zeta \rangle - y\right|.$$

Then, $\theta = tx \in \partial \rho\left(y - \langle x, \theta \rangle = 0\right)$ and $[-1, 1]\, x \subset \partial \rho\left(y - \langle x, \theta \rangle = 0\right)$.

Next, we shall show the inverse inclusion: $\partial \rho\left(y - \langle x, \theta \rangle = 0\right) \subset [-1, 1]\, x$. Let $\theta \in \partial \rho\left(y - \langle x, \theta \rangle = 0\right)$ and assume $\theta \neq x$. From $\theta \in \partial \rho\left(y - \langle x, \theta \rangle = 0\right)$, we have

$$|y - \langle x, \zeta \rangle| \geq \langle \zeta - \theta,\ \theta \rangle, \qquad {}^{\forall}\zeta \in \mathscr{H}. \tag{A.1}$$

From now on, set $H = \{\eta \in \mathscr{H} : \langle x, \eta\rangle = y\}$ and $G = \{\eta \in \mathscr{H} : \langle \eta, \theta\rangle = \langle \theta, \theta\rangle\}$, we shall show that $H = G$. When $\dim(\mathscr{X}) = 1$, $H = G = \left\{\frac{y}{x^*}\right\}$. Assume $\dim\{\mathscr{H}\} > 2$. First $\eta \in H \Rightarrow \eta \in G$, pick $\eta \in H$: $\langle x, \eta\rangle = y$ we have $\eta = \theta$, so $\langle \eta, \theta\rangle = \langle \theta, \theta\rangle$. Then, $H \subset G$. We shall show the inverse inclusion $G \subset H$. Assume $\eta \in G$ and $\eta \notin H$. Because $\theta \neq x$, there exists $u \in \mathscr{H}$ such that $\langle \theta, u\rangle \neq y$. Put $p = \langle x, \eta\rangle u - \langle x, u\rangle \eta + \theta$, because u and η are linear independent, $p \neq \theta$. On the other hand

$$\begin{aligned}\langle x, p\rangle &= \langle x, \langle x, \eta\rangle u - \langle x, u\rangle \eta + \theta\rangle \\ &= \langle x, \eta\rangle \langle x, u\rangle - \langle x, u\rangle \langle x, \eta\rangle + y \\ &= y.\end{aligned}$$

This is contradiction, therefore $G \subset H$. Finally, we have $G = H$.

Now, set

$$x' \triangleq \zeta - \frac{y - \langle x, \zeta\rangle}{y - \langle x, v\rangle}(v - \theta), \qquad {}^{\forall}\zeta \in \mathscr{H},$$

Then, we have

$$\begin{aligned}\langle x, x'\rangle &= \langle x, \zeta\rangle - \frac{y - \langle x, \zeta\rangle}{y - \langle x, v\rangle}\langle x, v - \theta\rangle \\ &= \langle x, \zeta\rangle - \frac{y - \langle x, \zeta\rangle}{y - \langle x, v\rangle}(\langle x, v\rangle - y) \\ &= \langle x, \zeta\rangle + y - \langle x, \zeta\rangle \\ &= y.\end{aligned}$$

Furthermore $x' \in H \Rightarrow x' \in G$. Therefore,

$$\begin{aligned}\langle \theta, \theta\rangle &= \langle \theta, x'\rangle \\ &= \langle \theta, \zeta\rangle - \frac{y - \langle x, \zeta\rangle}{y - \langle x, v\rangle}\langle \theta,\ v - \theta\rangle \\ &= \langle \theta, \zeta\rangle - \frac{\langle \theta,\ v - \theta\rangle}{y - \langle x, v\rangle}(y - \langle x, \zeta\rangle) \\ &= \langle \theta, \zeta - \theta\rangle - \frac{\langle \theta,\ v - \theta\rangle}{y - \langle x, v\rangle}(y - \langle x, \zeta\rangle) \\ &= \langle \theta, \zeta - \theta\rangle - \frac{\langle \theta,\ v - \theta\rangle}{y - \langle x, v\rangle}(\langle x, \theta\rangle - \langle x, \zeta\rangle),\end{aligned}$$

and we get $\langle \theta, \zeta - \theta\rangle = t\langle x, \zeta - \theta\rangle$ where $t = \frac{\langle \theta,\ v - \theta\rangle}{y - \langle x, v\rangle} \neq 0$. Because of (A.1), $\langle \theta,\ v - \theta\rangle \leq$

$|y - \langle x, v\rangle|$ and

$$\begin{aligned} -\langle \theta,\ v-\theta\rangle = \langle \theta,\ \theta - v\rangle &\leq |-\langle x, \theta - v\rangle| \\ &= |\langle x, v\rangle - y| \\ &= |y - \langle x, v\rangle|, \end{aligned}$$

Since $\langle \theta, \zeta - \theta\rangle \neq 0$,$\langle x, \zeta - \theta\rangle \neq 0$. We have $|\langle \theta,\ v-\theta\rangle| \leq |y - \langle x, v\rangle|$, $|t| \leq 1$. Therefor, $\partial\rho\left(y - \langle x, \theta\rangle = 0\right) \subset [-1, 1]\, x$. □

A.2 Proof of Proposition 12

Proof. The proof is the same as in **?**. A minor difference is a subdifferential culculus. So, we shall prove the first statement only. In order that $H_{\tau,n,\lambda}(t)$ converges in Mosco to $G_{\tau,n,\lambda}(t)$, we will apply Theorem 7 to $H_{\tau,n,\lambda}(t)$ and $G_{\tau,n,\lambda}(t)$. All we have to do is to show the graph convergence of the subdifferential $\partial H_{\tau,n,\lambda}(t)$ to $\partial G_{\tau,n,\lambda}(t)$ in probability. Considering proposition 6, we denote any measurable selector of $\partial\rho(\cdot)$ as itself in the following proof below. Calculate subdifferential of $H_{\tau,n,\lambda}, G_{\tau,n,\lambda}$ with respect to t, we obtain

$$\begin{aligned} \partial H_{\tau,n,\lambda}(t) &= \sqrt{n}\partial F_{\tau,n,\lambda}\left(\theta_0 + \frac{1}{\sqrt{n}}t\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t\right) + \frac{\lambda_n}{\sqrt{n}}\left(\theta_0 + \frac{1}{\sqrt{n}}t\right), \\ \partial G_{\tau,n,\lambda}(t) &= \sqrt{n}\partial F_n(\theta_0) + \sqrt{n}\partial F_0\left(\theta_0 + \frac{1}{\sqrt{n}}t\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial\rho(\theta_0, Z_i) + \frac{\lambda_n}{\sqrt{n}}(\theta_0) + \sqrt{n}\mathbb{E}\left[\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right]. \end{aligned}$$

Recall $\partial f_n \overset{G}{\to} \partial f_0$ means that for every $(\theta_0, \eta_0) \in \partial f_0$, there exists a sequence $(\theta_n, \eta_n) \in \partial f_n$ such that $\theta_n \to \theta_0$ strongly in $\mathscr{H}$, $\eta_n \to \eta_0$ strongly in $\mathscr{H}^*$ $(= \mathscr{H})$. $\partial H_n \overset{G}{\to} \partial G_n$ means that there exists a sequence of measurable selectors of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right)$ such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z_i\right) \to \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\partial\rho(\theta_0, Z_i) + \sqrt{n}\mathbb{E}\left[\partial\rho\left(\theta_0 + \frac{1}{\sqrt{n}}t, Z\right)\right],$$

strongly in $\mathscr{H}$. Later is the same as in **?** □

Acknowledgements

This research is supposed by grant-in-aid for JSPS Fellows (DC1, 20137989).

References

Attouch, H. (1984): *Variational Convergence for Functions and Operators*, Pitman Publishing.

Aubin, J. and H. Frankowska (1990): *Set-Valued Analysis*, Birkhauser.

Beer, G. (1993): *Topologies on Closed and Closed Convex Sets*, Kluwer Academic Publishing.

Borwein, J. and D. Noll (1994): "Second Order Differentiability of Convex Functions in Banach Spaces," *Trans. Amer. Math. Soc.*, 342, 43–81.

Bucher, A., J. Segers, and S. Volgushev (2014): "When Uniform Weak Convergence Fails: Empirical Processes for Dependence Functions and Residuals via Epi- and Hypographs," *Ann. Statist.*, 42, 1598–1634.

Cai, T. T. and M. Yuan (2012): "Minimax and adaptive prediction for functional linear regression," *Journal of the American Statistical Association*, 107, 1201–1216.

Clarke, F. (1983): *Optimization and Nonsmooth Analysis*, New York: Wiley.

Dudley, R. (1999): *Uniform Central Limit Theorems*, Cambridge University Press.

Dupacava, J. and R. Wets (1988): "Asymptotic Behavior of Statistical Estimators and of Optimal Solution of Stochastic Optimization Problems," *Ann. Statist.*, 16, 1517–1549.

Geyer, C. (1994): "On the Asymptotics of Constrained M-Estimation," *Ann. Statist.*, 22, 1993–2010.

Hall, P., J. L. Horowitz, et al. (2007): "Methodology and convergence rates for functional linear regression," *The Annals of Statistics*, 35, 70–91.

Hall, P., H.-G. Müller, J.-L. Wang, et al. (2006): "Properties of principal component methods for functional and longitudinal data analysis," *The annals of statistics*, 34, 1493–1517.

Kato, K. (2012): "Estimation in Functional Linear Quantile Regression," *Ann. Statist*, 40, 3108–3136.

Kato, N. (1989): "On the Second Derivatives of Convex Functions on Hilbert Spaces," *Proc. Amer. Math. Soc.*, 106, 697–705.

Knight, K. (2003): "Epi-convergence in distribution and stochastic equi-semicontinuity," *Unpublished Manuscript*.

Koenker, R. and G. Bassett Jr (1978): "Regression quantiles," *Econometrica: journal of the Econometric Society*, 33–50.

Koenker, R., P. Ng, and S. Portnoy (1994): "Quantile smoothing spline," *Biometrika*, 81, 673–680.

Molchanov, I. (2005): *Theory of Random Sets*, Springer.

Mosco, U. (1969): "Convergence of Convex Sets and of Solutions of Variational Inequalities," *Adv. Math.*, 3, 510–585.

Müller, H.-G., U. Stadtmüller, et al. (2005): "Generalized functional linear models," *the Annals of Statistics*, 33, 774–805.

Phelps, R. (1992): *Convex Functions, Monotone Operators and Differentiability*, Springer, 2nd ed.

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

Portnoy, S. (1997): "Local asymptotics for quantile smoothing splines," *Annals of statistics*, 25, 414–434.

Rockafellar, R. T. (1970): *Convex analysis*, Princeton university press.

Shin, H. and S. Lee (2016): "An RKHS approach to robust functional linear regression," *Statistica Sinica*, 255–272.

van der Vaart, A. (1998): *Asymptotic statistics*, Cambridge University Press.

van der Vaart, A. and J. Wellner (1996): *Weak convergence and empirical processes*, Springer.

Yao, F., H.-G. Müller, J.-L. Wang, et al. (2005): "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, 33, 2873–2903.

Yao, F., S. Sue-Chee, and F. Wang (2017): "Regularized partially functional quantile regression," *Journal of Multivariate Analysis*, 156, 39–56.

Yuan, M. and T. T. Cai (2010): "A reproducing kernel Hilbert space approach to functional linear regression," *The Annals of Statistics*, 38, 3412–3444.