Supplementary Material for "Generative Quantile Regression with Variability Penalty"

Shijie Wang*

Department of Statistics, University of South Carolina, Columbia, SC 29208
and
Minsuk Shin
Gauss Labs, Palo Alto, CA 94301
and
Ray Bai
Department of Statistics, University of South Carolina, Columbia, SC 29208

February 28, 2024

Abstract

Section A provides additional illustrations and data applications of the proposed Penalized Generative Quantile Regression (PGQR) method. Section B presents the results from additional simulation studies and additional figures. Section C gives the proofs of all the propositions from the main manuscript. Section D conducts additional analyses of model complexity and sensitivity to the choice of hyperparameter α in the PGQR variability penalty function.

A Additional Illustrations and Real Data Analyses

A.1 Illustration: Takeuchi's Example

A popular illustrative example in the literature for nonparametric quantile estimation was given by Takeuchi et al. (2006) (henceforth known as Takeuchi's example), where

$$Y_i = \sin(\pi X_i)/(\pi X_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

*The authors gratefully acknowledge financial support from the National Science Foundation (Grant no. NSF DMS-2015528). We would like to thank Dr. Jun Liu from Harvard University for helpful comments.

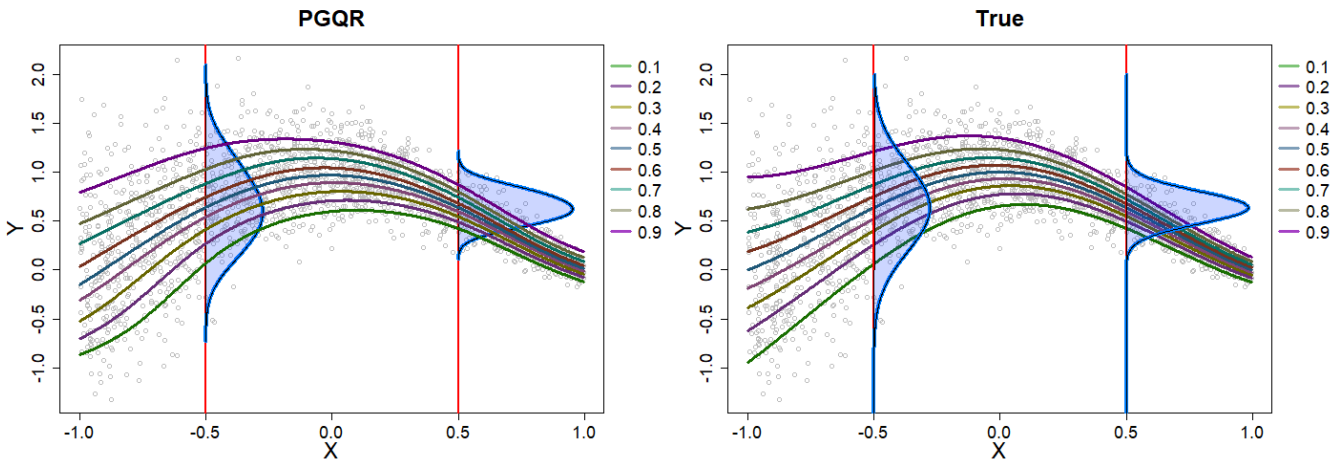

Figure 1: Using PGQR to model the conditional densities $p(Y \mid X = 0.5)$ and $p(Y \mid X = -0.5)$. The conditional quantile functions at levels $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ are also displayed.

with $x_i \sim \text{Uniform}(-1, 1)$ and $\epsilon_i \sim \mathcal{N}(0, 0.1 \exp(1 - X_i))$. By Takeuchi et al. (2006), the true τth conditional function, $\tau \in (0, 1)$, is $f_\tau(X) = \sin(\pi X)/(\pi X) + 0.1 \exp(1 - X)\Phi^{-1}(\tau)$, where $\Phi^{-1}(\cdot)$ denotes the inverse cumulative distribution function (cdf) of a standard Gaussian distribution. Moreover, the conditional density of Y given X is $Y \mid X \sim \mathcal{N}(\sin(\pi X)/(\pi X), 0.1 \exp(1 - X))$.

We made an artificial dataset of size $n = 2000$ for Takeuchi's example and applied our proposed PGQR method to it. Specifically, we aimed to estimate the conditional densities $p(Y \mid X = -0.5)$ and $p(Y \mid X = 0.5)$. The left panel of Figure 1 depicts the conditional density estimates for PGQR, which are almost identical to the true conditional densities (right panel of Figure 1). In addition, we estimated the conditional quantile function $f_\tau(x)$ at quantile levels $\tau = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. From Figure 1, we observe that the mid-range quantile functions estimated by PGQR for $\tau \in \{0.2, 0.3, \ldots, 0.8\}$ are very close to the true quantile functions. For the low and high quantile levels at $\{0.1, 0.9\}$, the PGQR estimates show a similar pattern to the true quantile functions, although there is some slight departure near $X = -1$ where there is less data.

A.2 Clinical Application: Discovering Hidden Subpopulations

As discussed in the main manuscript, PGQR aims to simultaneously generate samples from multiple conditional quantiles $Q_{Y|\boldsymbol{X}}(\tau)$ of $p(Y \mid \boldsymbol{X})$ at different quantile levels $\tau \in (0, 1)$. An automatic byproduct of joint *nonparametric* quantile regression (as opposed to linear quantile regression) is

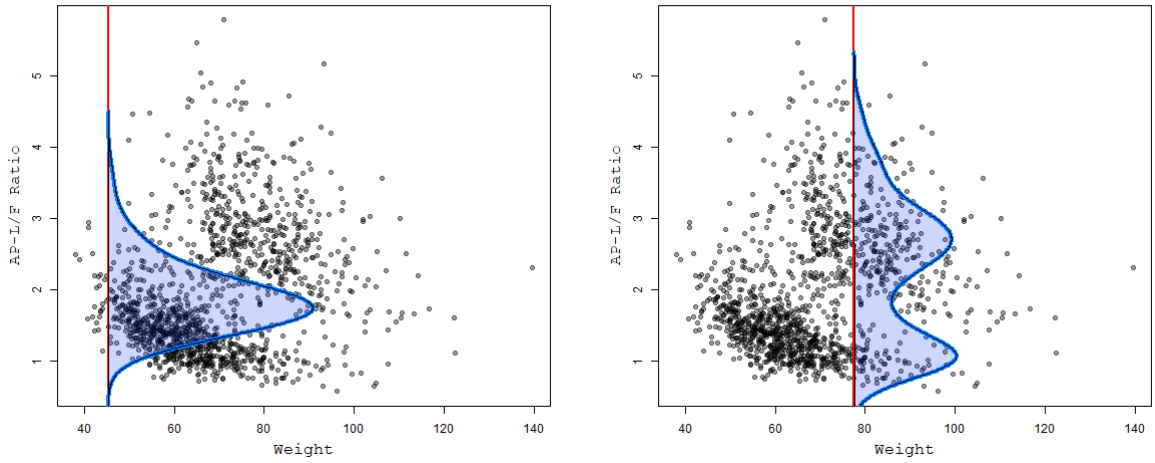

Figure 2: Using PGQR to model the conditional density of AP-L/F ratio given weight in older adults. Left panel: Weight = 45.4 kg, Right panel: Weight = 77.5 kg.

that if the conditional quantiles $Q_{Y|\boldsymbol{X}}(\tau)$ are estimated well for a large number of quantiles, then we can also infer the *entire* conditional distribution for Y given $\boldsymbol{X}$.

To demonstrate the clinical utility of our method, we apply our proposed PGQR method to a real dataset on body composition and strength in older adults (RoyChoudhury and Xu, 2020). The data was collected over a period of 12 years for 1466 subjects as part of the Rancho Bernardo Study (RBS), a longitudinal observational cohort study. We are interested in modeling the appendicular lean/fat (AP-L/F) ratio, i.e.

$$\text{AP-L/F Ratio} = \frac{\text{Weight on legs and arms}}{\text{Fat weight}},$$

as a function of weight (kg). Accurately predicting the AP-L/F ratio is of practical clinical interest, since the AP-L/F ratio provides information about limb tissue quality and is used to diagnose sarcopenia (age-related, involuntary loss of skeletal muscle mass and strength) in adults over the age of 30 (Evans, 2010; Scafoglieri et al., 2017).

Figure 2 plots the approximated conditional density of AP-L/F ratio given weight of 45.4 kg (left panel) and 77.5 kg (right panel) under the PGQR model. We see evidence of data heterogeneity (actually, depending on an unobserved factor of gender), as the estimated conditional density is unimodal when the weight of older adults is 45.4 kg but *bimodal* when the weight of older adults is 77.5 kg. In short, our method *discovers* the presence of two heterogeneous subpopulations of

	Simulation 4			Simulation 5			Simulation 6		
Method	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)
PGQR ($\alpha = 1$)	0.15	0.07	0.95 (4.33)	0.004	**0.0001**	0.93 (0.42)	7.20	57.75	0.79(17.72)
PGQR ($\alpha = 5$)	**0.14**	0.08	**0.96 (4.50)**	0.005	0.03	**0.99 (1.01)**	7.20	56.74	0.79(18.80)
GCDS	0.23	0.05	0.87 (3.39)	**0.002**	0.0011	0.73 (0.27)	7.44	64.63	**0.80**(18.41)
deep-GCDS	0.43	0.20	0.76 (2.88)	0.004	0.0022	**0.99 (0.56)**	11.9	85.60	0.61(13.46)
WGCS	0.92	0.15	0.79 (4.41)	0.640	0.2372	0.70 (1.15)	9.58	93.52	0.51(11.54)
FlexCoDE-NNR	0.23	0.003	0.92 (3.82)	0.069	0.0168	0.89 (0.27)	**1.57**	**35.46**	0.37(8.38)
FlexCoDE-SAM	0.23	**0.002**	0.93 (3.83)	0.043	0.0130	0.93 (4.04)	1.76	50.72	0.09(5.56)
FlexZBoost	0.45	0.06	0.82 (3.50)	1.244	0.2824	0.81 (4.13)	19.1	40.36	0.58(15.5)
RFCDE	0.24	0.003	0.94 (3.97)	0.067	0.0292	0.92 (3.97)	3.12	67.33	0.28(6.61)

Table 1: Table reporting the PMSE for the conditional expectation and standard deviation, as well as the coverage rate (Cov) and average width of the 95% prediction intervals, for Simulations 4 through 6. Results were averaged across 20 replicates.

adults that weigh around 78 kg. In contrast, mean regression (e.g. simple linear regression or nonparametric mean regression) of AP-L/F ratio given weight might obscure the presence of two modes and miss the fact that weight affects AP-L/F ratio differently for these two clusters of adults.

B More Simulation Results

B.1 Additional Simulation Studies

In addition to the three simulations described in Section 5.1 of the main manuscript, we also conducted simulation studies under the following scenarios:

- **Simulation 4: Nonlinear function with an interaction term and one irrelevant covariate.** $Y_i = 0.5\log(10 - X_{i1}^2) + 0.75\exp(X_{i2}X_3/5) - 0.25|X_{i4}/2| + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$. Note that there is a (nonlinear) interaction between X_2 and X_3, while X_5 is irrelevant.

- **Simulation 5: Very small conditional variance.** $Y_i = \beta X_i + \epsilon_i, i = 1, \ldots, n$, where $\beta = 1$ and $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 0.01)$.

- **Simulation 6: Error term dependent on norm of predictor $\boldsymbol{X}$.** $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i$, where $\boldsymbol{\beta} \in \mathbb{R}^5$ is equispaced between $[-2, 2]$, $\boldsymbol{X}_i \sim \text{Uniform}[-1, 1]^5$ and $\epsilon_i \sim \mathcal{N}(0, \exp(0.5\|X_i\|_1))$.

The results from these three simulations averaged across 20 replicates are shown in Table 1.

One may be concerned whether the regularized PGQR overestimates the conditional variance when the true conditional density has a very *small* variance. To illustrate the flexibility of PGQR,

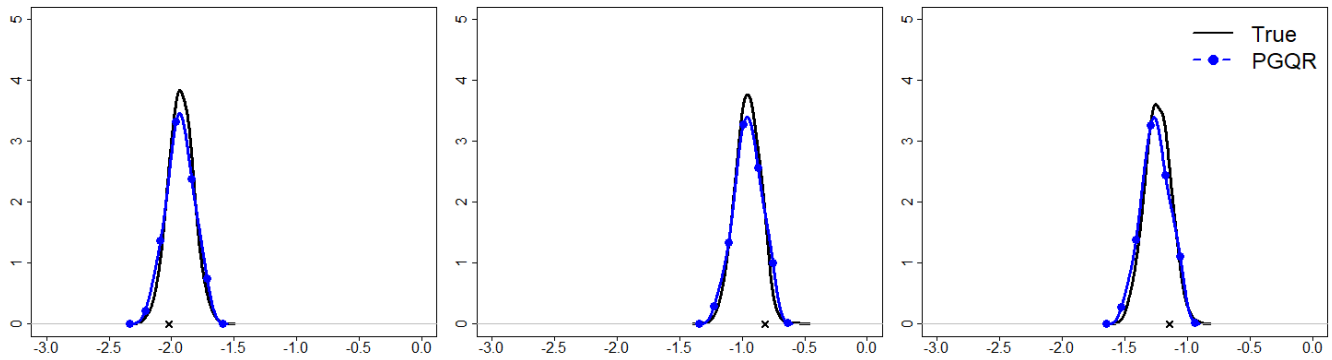

Figure 3: Plots of the estimated PGQR ($\alpha = 1$) conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for three different test observations from one replication of Simulation 5. The optimal $\lambda^\star$ is chosen by our tuning parameter selection method in Section 4.2 of the main manuscript.

Figure 3 plots the estimated conditional densities for three test observations from one replication of Simulation 5. Recall that in Simulation 5, the true conditional variance is very small ($\sigma^2 = 0.01$). With the optimal $\lambda^\star$ selected using the method introduced in Section 4.2 of the main manuscript, Figure 3 shows that the estimated PGQR conditional density *still* manages to capture the Gaussian shape while matching the true variance of 0.01. If the true conditional variance is very small (as in Simulation 5), then PGQR selects a tiny $\lambda^\star \approx 0$. In this scenario, PGQR only applies a small amount of variability penalization and thus does not overestimate the variance.

In Simulation 6, we investigated the especially challenging case when the error term ϵ is dependent on ℓ_1 norm of predictor $\boldsymbol{X}$. This simulation setting is inspired by simulation M2 in Moon et al. (2021) and is a variant of an example from Appendix A of Takeuchi et al. (2006). In Figure 4, we see that the variance of true conditional distribution varies with $\|\boldsymbol{X}\|_1$. We observe that PGQR was able to capture of true conditional distribution in some cases where $\|\boldsymbol{X}\|_1$ is not large (e.g. $\|\boldsymbol{X}\|_1 < 4$). However, PGQR ($\alpha = 1$) is incapable of estimating very large variance in the third graph (upper panel) when $\|\boldsymbol{X}\|_1 = 6.4$. In this case, the true conditional density is very flat and thus difficult for all of the deep generative methods to estimate well. It is worth noting that the other state-of-the art methods, GCDS and WGCS, struggled even more than PGQR in this heavy heteroscedasticity scenario. Figure 8 also indicates that PGQR had better average performance than NMQN and MCQRNN for multiple quantile estimation.

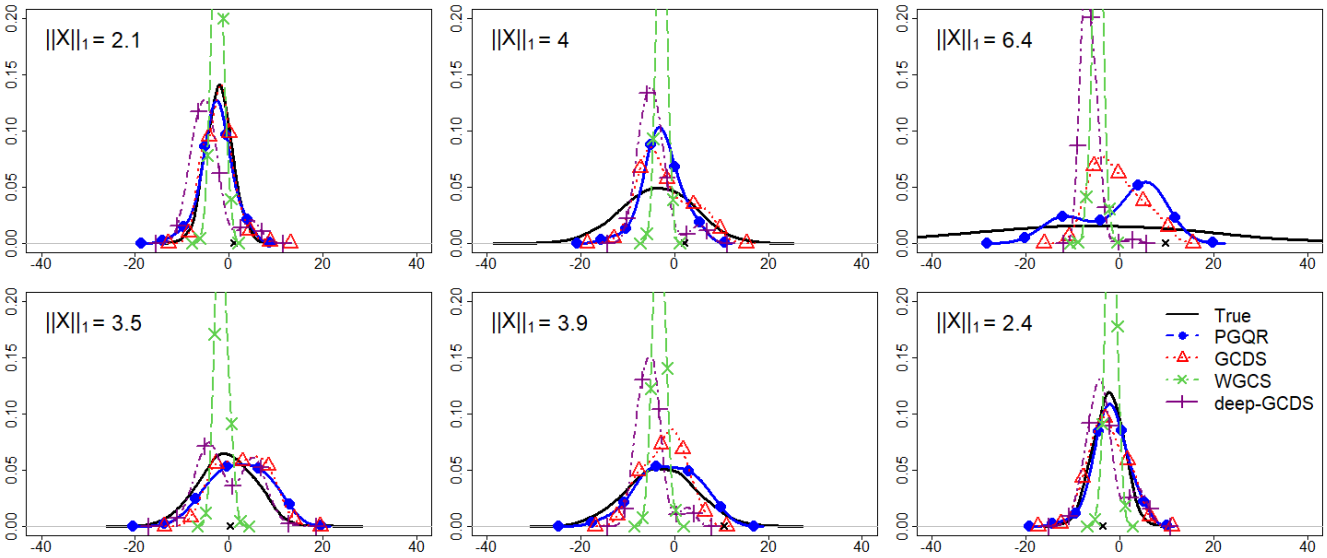

Figure 4: Plots of the estimated PGQR ($\alpha = 1$) conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for six different test observations from one replication of Simulation 6.

B.2 Additional Figures

Here, we provide additional figures from one replication each of Simulations 1, 2, and 4 (with $\alpha = 1$ in PGQR). Figures 5-7 illustrate that PGQR (blue solid line with filled circles) is better able to estimate the true conditional densities (solid black line) than GCDS, WGCS, and deep-GCDS (dashed lines). In particular, PGQR does a better job of capturing critical aspects of the true conditional distributions such as multimodality, heteroscedasticity, and skewness.

- **Simulation 1: Multimodal and heteroscedastic.**

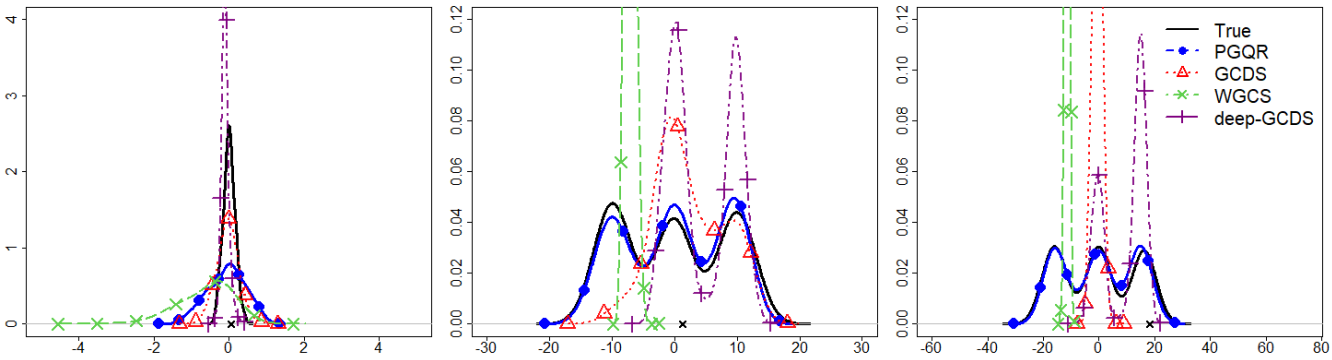

Figure 5: Plots of the estimated conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for three different test observations from one replication of Simulation 1.

- **Simulation 2: Mixture of left-skewed and right-skewed.**

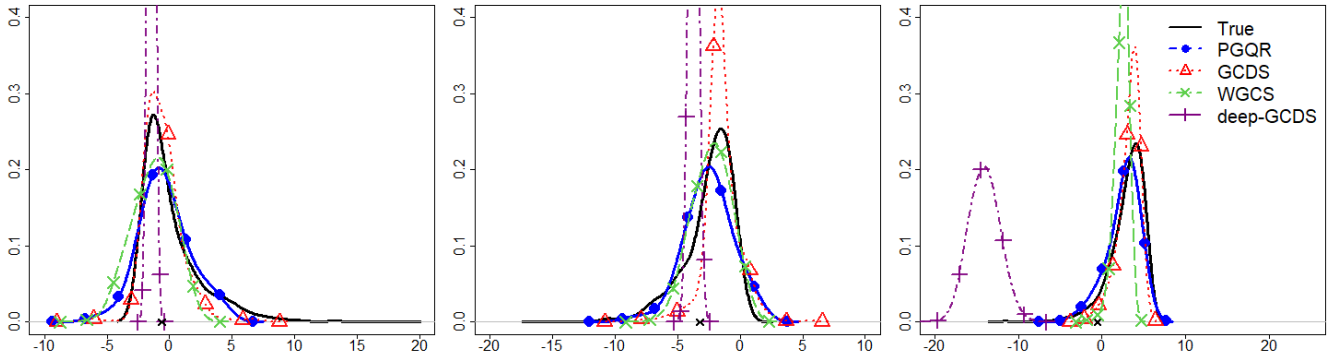

Figure 6: Plots of the estimated conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for three different test observations from one replication of Simulation 2.

- **Simulation 4: Nonlinear function with an interaction term and one irrelevant covariate.**

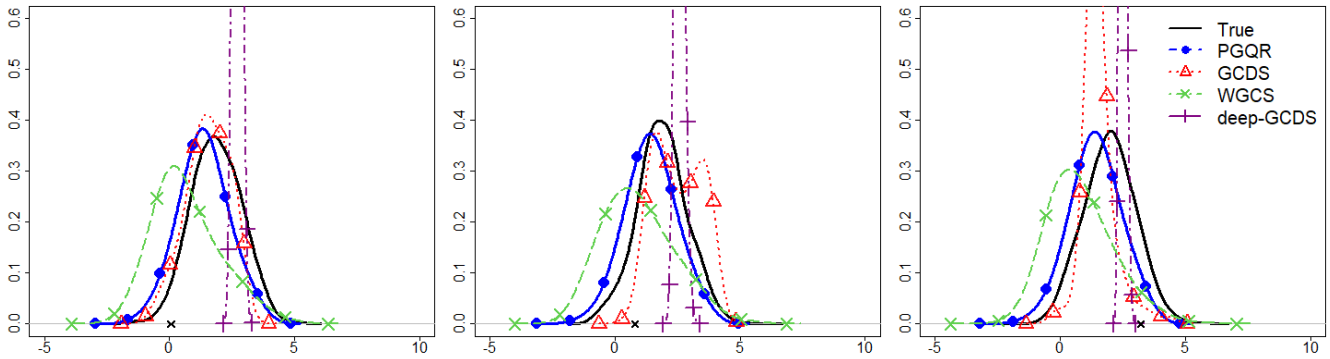

Figure 7: Plots of the estimated conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for three different test observations from one replication of Simulation 4.

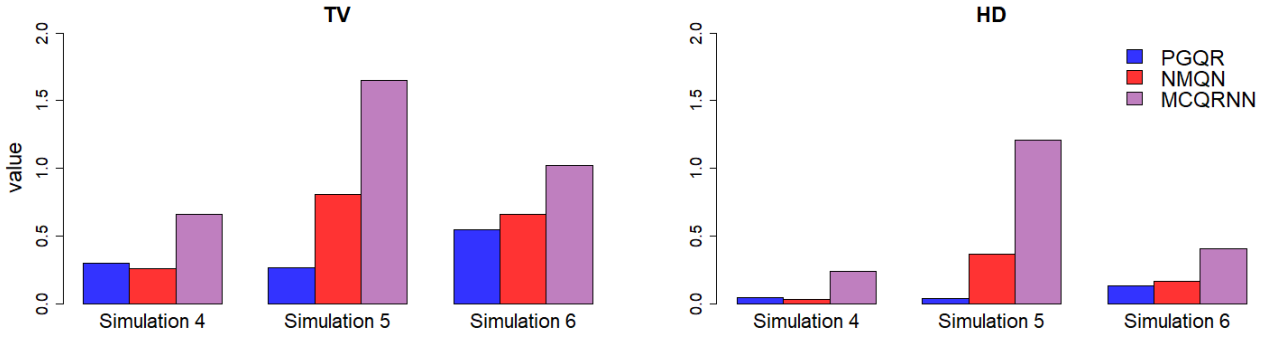

Figure 8: Barplots of the average total variation distance (TV) and Hellinger distance (HD) across 20 replicates evaluated at $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ for Simulations 4 through 6.

Figure 8 depicts the quantile estimation performance of PGQR compared to NMQN and MCQRNN in Simulations 4 through 6. In Simulation 4, NMQN performed slightly better than PGQR, but the performance of the two methods was very comparable. However, in Simulations 5 and 6, PGQR outperformed NMQN, with lower average total variation distance (TV) and Hellinger distance (HD). Both PGQR and NMQN outperformed MCQRNN in Simulations 4 through 6.

C Proofs of Propositions

Proof of Proposition 2.1. Suppose that for some $\epsilon > 0$, there exists a set $\mathcal{C} \subset (0,1)$ with $P_\tau(\mathcal{C}) > \epsilon$ such that for some $i^* \in \{1,\ldots,n\}$, $\widehat{g}_\tau(\boldsymbol{X}_{i^*}) \neq \widehat{G}(\boldsymbol{X}_{i^*},\tau)$ for all $\tau \in \mathcal{C}$. Then we can construct another optimal generator $\widetilde{G}$ such that

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\tau\left[\rho_\tau\big(y_i - \widehat{G}(\boldsymbol{X}_i,\tau)\big)\right] + \mathbb{E}_{\widetilde{\tau},\widetilde{\tau}'}\left\{\mathrm{pen}_{\lambda,\alpha}\left(\widehat{G}(\boldsymbol{X}_i,\widetilde{\tau}),\widehat{G}(\boldsymbol{X}_i,\widetilde{\tau}')\right)\right\} \\ \geq\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_\tau\left[\rho_\tau\big(y_i - \widetilde{G}(\boldsymbol{X}_i,\tau)\big)\right] + \mathbb{E}_{\widetilde{\tau},\widetilde{\tau}'}\left\{\mathrm{pen}_{\lambda,\alpha}\left(\widetilde{G}(\boldsymbol{X}_i,\widetilde{\tau}),\widetilde{G}(\boldsymbol{X}_i,\widetilde{\tau}')\right)\right\}, \end{aligned}$$

where

$$\widetilde{G}(\boldsymbol{X}_{i^*},\tau) = \begin{cases} \widehat{G}(\boldsymbol{X}_{i^*},\tau) \text{ for } \tau \notin \mathcal{C}, \\ \widehat{g}_\tau(\boldsymbol{X}_{i^*}) \text{ for } \tau \in \mathcal{C}. \end{cases}$$

This is a contradiction due to the fact that $\widehat{g}_\tau$ is the minimizer of $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \widehat{G}(\boldsymbol{X}_i,\tau)) + \mathbb{E}_{\widetilde{\tau},\widetilde{\tau}'}\{\mathrm{pen}_{\lambda,\alpha}(\widehat{G}(\boldsymbol{X}_i,\widetilde{\tau}),\widehat{G}(\boldsymbol{X}_i,\widetilde{\tau}'))\}$. □

Proof of Proposition 2.2. Suppose that for all $\lambda \geq 0$, all $\tau \in (0,1)$, and all $i \in \{1,\ldots,n\}$,

$$\widehat{G}(\boldsymbol{X}_i,\tau) = Y_i.$$

Then the penalty part in the PGQR loss (6) attains $\mathbb{E}_{\tau,\tau'}\{\mathrm{pen}_{\lambda,\alpha}(\widehat{G}(\boldsymbol{X}_i,\tau),\widehat{G}(\boldsymbol{X}_i,\tau'))\} = \lambda\log(\alpha)$, while the first term in (6) satisfies $\mathbb{E}_\tau[\rho_\tau\big(y_i - \widehat{G}(\boldsymbol{X}_i,\tau)\big)] = 0$. As a result, the total loss is $\lambda\log(\alpha)$.

Since the case of $\widehat{G}(\boldsymbol{X}_i,\tau) = Y_i$ is included in cases of $\mathrm{Var}_\tau\{\widehat{G}(\boldsymbol{X}_i,\tau)\} = 0$, we focus on the

variance. When there exists some $i \in \{1, \ldots, n\}$ and some $\tau \in (0, 1)$ such that

$$\mathrm{Var}_\tau \left\{ \widehat{G}(\boldsymbol{X}_i, \tau) \right\} > 0,$$

the resulting total loss can be made less than $\lambda \log(\alpha)$ by choosing an appropriate $\lambda > 0$. This contradicts the fact that $\widehat{G}$ is the minimizer as in (6). □

Proof of Proposition 4.2. It is trivial that $F_Q(W) := P(Q \leq W \mid W) \sim \mathrm{Uniform}(0, 1)$ when $Q \overset{d}{=} W$. Without loss of generality, we assume that F_Q is invertible. If F_Q is not invertible, then we replace $F_Q^{-1}(W)$ below with $F_Q^{-}(W) := \inf\{x : F_Q(x) \geq W\}$, and the result still holds. We shall show that $F_Q(W) \sim \mathrm{Uniform}(0, 1)$ implies that the two distributions for Q and W are identical. Suppose that $F_Q(W)$ follows a standard uniform distribution. Then,

$$x = P(F_Q(W) < x) = P(W < F_Q^{-1}(x)) = F_Q(F_Q^{-1}(x)),$$

which implies that $F_W(F_W^{-1}(x)) = x$. Thus, it follows that the distributions of Q and W are identical. □

D Analyses of Model Complexity and Choice of α

D.1 Model Complexity Analysis

The estimated conditional quantile function $\widehat{G}(\boldsymbol{X}, \tau)$ depends on estimation of two sub-networks $g_c(\tau)$ and $g_{uc}(\boldsymbol{X})$. To avoid vanishing variability (4), we proposed a novel variability penalty (5) which essentially controls $\|\partial G(\boldsymbol{X}, \tau)/\partial \tau\|$. However, the estimated $\widehat{G}(\boldsymbol{X}, \tau)$ also depends on model complexity $\|\partial G(\boldsymbol{X}, \tau)/\partial \boldsymbol{X}\|$. This indicates that tuning the model complexity of $g_{uc}(\boldsymbol{X})$ might be another potential way to solve the vanishing variability problem.

In Section 2.2, we conducted a simple simulation study under the model, $Y_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i, i = 1, \ldots, n$, where $\boldsymbol{X}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{20})$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, the coefficients in $\boldsymbol{\beta}$ are equispaced over $[-2, 2]$, and the sample size is $n = 2000$. Figure 2 of the main manuscript illustrated that GQR (without any variability penalty) can encounter severe vanishing variability – essentially collapsing to a

point mass – even when dealing with a simple linear model. We also showed in Section 2.2 that PGQR (i.e. GQR with a variability penalty function) avoids the vanishing variability by controlling $\|\partial G(\boldsymbol{X}, \tau)/\partial \tau\|$. In this example, both GQR and PGQR were constructed by a feedforward neural network (FNN) with three hidden layers and 1000 neurons per hidden layer. Here, GQR is heavily overparameterized, i.e. the number of learnable parameters in the FNN is much larger than the number of training samples. As discussed in Section 2.2, the main motivation for overparameterization is that it improves the generalization and robustness of the model (Allen-Zhu et al., 2019; Zhang et al., 2021; Soltanolkotabi et al., 2019; Montanari and Zhong, 2022). But in the case of nonparameteric quantile estimation, overparameterization can also lead to more severe vanishing variability. This is because the GQR loss (3) actually achieves the minimum value of zero when $\widehat{G}(\boldsymbol{X}_i, \tau) = Y_i, i = 1, \ldots, n$, and a very complex model is likely to perfectly interpolate the observed responses.

Therefore, instead of using a variability penalty on an overparameterized model, it is also very natural to consider controlling the model complexity $\|\partial G(\boldsymbol{X}, \tau)/\partial \boldsymbol{X}\|$ by simply choosing a simpler FNN structure $g_{uc}(\boldsymbol{X})$. A simpler model would not be overparameterized, and therefore, it is a promising alternative way to avoid vanishing variability. To investigate whether tuning model complexity indeed helps to avoid this phenomenon, we considered two different (simpler) FNN settings for the simple linear example that we presented in Section 2.2. We applied these simple FNN structures to (non-penalized) GQR so we could see the impact of $\|\partial G(\boldsymbol{X}, \tau)/\partial \boldsymbol{X}\|$ on vanishing variability.

To be more specific, we denote the model GQR_1 as (non-penalized) GQR fit with a simple FNN with two hidden layers, each having 50 hidden neurons. This setting is similar to the network structure in Zhou et al. (2023). The model GQR_2 is (non-penalized) GQR fit with an even simpler FNN architecture with only one hidden layer and five hidden neurons. This setting is similar to that considered by Cannon (2018) and Moon et al. (2021). We compared GQR_1 and GQR_2 to PGQR on the same out-of-sample test data, where the optimal penalty parameter $\lambda^\star$ in PGQR was chosen according to Algorithm 1 in the main manuscript. These results are displayed in Figure 9 (GQR_1 vs. PGQR) and Figure 10 (GQR_2 vs. PGQR).

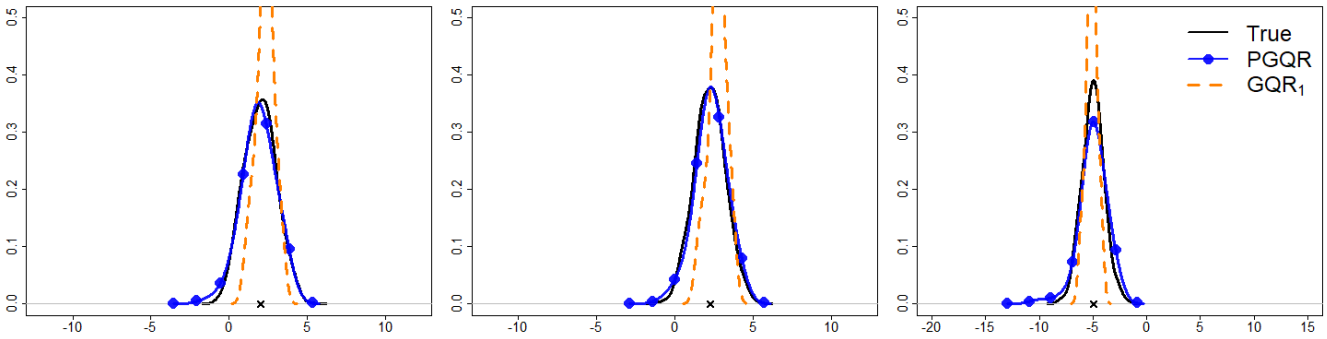

Figure 9: Plots of the estimated conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for three different test observations under GQR_1 and PGQR.

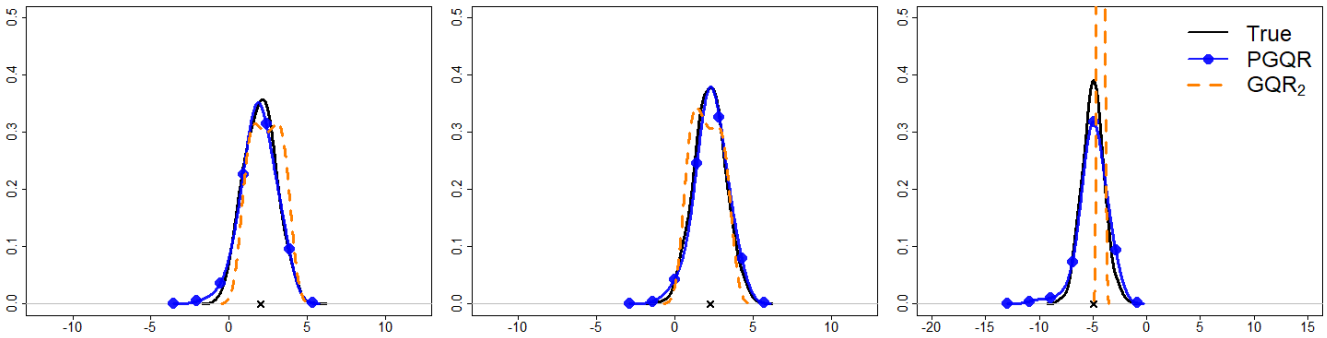

Figure 10: Plots of the estimated conditional densities $p(Y \mid \boldsymbol{X}_{\text{test}})$ for three different test observations under GQR_2 and PGQR.

It is obvious from Figure 9 that the simpler model GQR_1 (non-penalized GQR with two hidden layers and 50 nodes per hidden layer) mitigates vanishing variability a little bit, but it still suffers from this problem by underestimating the true variance. Looking at the results for GQR_2 (non-penalized GQR with one hidden layer and 5 hidden nodes) in Figure 10, the first two plots indicate that an even simpler FNN structure helps to avoid vanishing variability phenomenon for these particular test points. However, the third plot in Figure 10 shows that GQR_2 can *still* encounter the vanishing variability problem for other test observations, with significant variance underestimation.

From this simple example, we can see that even though we applied an extremely shallow FNN, vanishing variability can still occur. However, reducing the FNN complexity does appear to make vanishing variability less pronounced. Although a very simple model might not result in *exact* interpolation of the training labels, it nevertheless does *not* entirely fix variance underestimation. In short, controlling $\|\partial G(\boldsymbol{X}, \tau)/\partial \boldsymbol{X}\|$ by tuning the FNN complexity for $g_{uc}(\boldsymbol{X})$ *fails* to completely

avoid vanishing variability. This simple example demonstrates the distinct need to use a variability penalty to control $\|\partial G(\boldsymbol{X}, \tau)/\partial \tau\|$ (rather than just $\|\partial G(\boldsymbol{X}, \tau)/\partial \boldsymbol{X}\|$).

Figures 9 and 10 also show that the overparameterized PGQR model does a better job recovering the true conditional density $p(Y \mid \boldsymbol{X})$ – and therefore also estimates the true conditional quantile functions better – than the (non-penalized) GQR models with simpler FNN structures $g_{uc}(\boldsymbol{X})$. This may be because a simple neural network inevitably has less expressive power than a more complex one. By using (overparameterized) PGQR with a variability penalty, we are not only free of vanishing variability, but we *also* fully realize the well-known benefits of overparameterization (Allen-Zhu et al., 2019; Zhang et al., 2021; Soltanolkotabi et al., 2019; Montanari and Zhong, 2022).

D.2 Sensitivity Analysis of PGQR to the Choice of α

As mentioned in Section 2.2 of the main manuscript, we choose to fix $\alpha > 0$ in the variability penalty (5). The main purpose of α is to ensure that the logarithmic term in the penalty is always well-defined. In this section, we conduct a sensitivity analysis to the choice of α. To do this, we generated data using the same settings from Simulations 1 through 6. We then fit PGQR with eight different choices for $\alpha \in \{0.5, 1, 5, 10, 20, 30, 40, 50\}$ and evaluated the performance of these eight PGQR models on out-of-sample test data. Table 2 shows the results from our sensitivity analysis averaged across 20 replicates.

We found that in Simulations 1 through 4 and Simulation 6, PGQR was not particularly sensitive to the choice of α. PGQR was somewhat more sensitive to the choice of α in Simulation 5 (i.e. when the true conditional variance is very small), with larger values of α leading to higher PMSE for the conditional expectation and conditional standard deviation. In practice, we recommend fixing $\alpha = 1$ as the default α for PGQR to perform well. This choice of $\alpha = 1$ leads to good empirical performance across many different scenarios.

	Simulation 1			Simulation 2			Simulation 3		
α	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)
0.5	0.39	0.41	0.95 (23.60)	0.43	0.18	0.91 (7.69)	0.36	0.09	0.89 (5.93)
1	0.42	0.34	0.95 (23.49)	0.38	0.11	0.93 (8.14)	0.30	0.09	0.92 (6.61)
5	0.36	0.31	0.95 (23.41)	0.31	0.07	0.94 (8.83)	0.25	0.06	0.96 (6.60)
10	0.32	0.30	0.95 (23.40)	0.29	0.06	0.95 (9.02)	0.25	0.07	0.95 (6.55)
20	0.32	0.27	0.95 (23.74)	0.28	0.07	0.94 (9.11)	0.22	0.06	0.95 (6.52)
30	0.32	0.35	0.95 (24.20)	0.29	0.07	0.94 (9.09)	0.21	0.07	0.95 (6.45)
40	0.36	0.49	0.96 (24.42)	0.28	0.07	0.94 (8.91)	0.21	0.07	0.94 (6.45)
50	0.33	0.34	0.97 (24.15)	0.29	0.07	0.95 (9.15)	0.25	0.06	0.95 (6.63)

	Simulation 4			Simulation 5			Simulation 6		
α	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)	$\mathbb{E}(Y \mid \boldsymbol{X})$	$\text{sd}(Y \mid \boldsymbol{X})$	Cov (Width)
0.5	0.14	0.01	0.96 (3.95)	0.004	0.0001	0.92 (0.38)	6.88	59.65	0.78 (17.08)
1	0.15	0.07	0.95 (4.33)	0.004	0.0001	0.93 (0.42)	7.20	57.75	0.79 (17.72)
5	0.14	0.08	0.96 (4.50)	0.005	0.03	0.99 (1.01)	7.20	56.74	0.79 (18.80)
10	0.13	0.08	0.95 (4.39)	0.007	0.06	0.99 (1.33)	6.63	58.71	0.78 (18.14)
20	0.13	0.09	0.95 (4.49)	0.007	0.08	0.99 (1.64)	6.33	59.26	0.78 (17.97)
30	0.14	0.12	0.95 (4.39)	0.008	0.08	0.99 (1.69)	7.07	57.88	0.79 (18.72)
40	0.13	0.07	0.95 (4.29)	0.007	0.08	0.99 (1.69)	6.86	57.77	0.79 (18.70)
50	0.15	0.14	0.94 (4.42)	0.009	0.10	0.99 (1.81)	6.69	58.25	0.78 (18.55)

Table 2: PGQR results with different choices of $\alpha \in \{0.5, 1.0, 5.0, 10.0, 20.0, 30.0, 40.0, 50.0\}$ in the variability penalty for Simulations 1 through 6. This table reports the average PMSE for the conditional expectation and standard deviation, as well as the coverage rate (Cov) and the average width of the 95% prediction intervals. Results were averaged across 20 replicates.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 6158–6169. Curran Associates, Inc.

Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, 32(11):3207–3225.

Evans, W. J. (2010). Skeletal muscle loss: cachexia, sarcopenia, and inactivity. *The American Journal of Clinical Nutrition*, 91(4):1123S–1127S.

Montanari, A. and Zhong, Y. (2022). The interpolation phase transition in neural networks:

Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816 – 2847.

Moon, S. J., Jeon, J.-J., Lee, J. S. H., and Kim, Y. (2021). Learning multiple quantiles with neural networks. *Journal of Computational and Graphical Statistics*, 30(4):1238–1248.

RoyChoudhury, A. and Xu, C. (2020). A dataset on body composition, strength and performance in older adults. *Data in Brief*, 29:105103.

Scafoglieri, A., Clarys, J. P., Bauer, J. M., Verlaan, S., Van Malderen, L., Vantieghem, S., Cederholm, T., Sieber, C. C., Mets, T., and Bautmans, I. (2017). Predicting appendicular lean and fat mass with bioelectrical impedance analysis in older adults with physical function decline – the provide study. *Clinical Nutrition*, 36(3):869–875.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. (2019). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769.

Takeuchi, I., Le, Q., Sears, T., Smola, A., et al. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(45):1231–1264.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zhou, X., Jiao, Y., Liu, J., and Huang, J. (2023). A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, 118(543):1837–1848.