Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **00**, 1–30
https://doi.org/10.1093/jrsssb/qkaf080

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Original Article

Bayesian quantile estimation and regression with martingale posteriors

Edwin Fong[1] **and Andrew Yiu**[2]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
[2]School of Mathematical Sciences, University of Southampton, Southampton, UK

Address for correspondence: Edwin Fong, Department of Statistics and Actuarial Science, The University of Hong Kong, Run Run Shaw Building, Pokfulam Road, Hong Kong. Email: chefong@hku.hk

Abstract

Quantile estimation and regression within the Bayesian framework is challenging as the choice of likelihood and prior is not obvious. In this paper, we introduce a novel Bayesian nonparametric method for quantile estimation and regression based on the recently introduced martingale posterior (MP) framework. The core idea of the MP is that posterior sampling is equivalent to predictive imputation, which allows us to break free of the stringent likelihood-prior specification. We demonstrate that a recursive estimate of a smooth quantile function, subject to a martingale condition, is entirely sufficient for full nonparametric Bayesian inference. We term the resulting posterior distribution as the quantile martingale posterior (QMP), which arises from an implicit generative predictive distribution. Associated with the QMP is an expedient, MCMC-free and parallelizable posterior computation scheme, which can be further accelerated with an asymptotic approximation based on a Gaussian process. Furthermore, the well-known issue of monotonicity in quantile estimation is naturally alleviated through increasing rearrangement due to the connections to the Bayesian bootstrap. Finally, the QMP has a particularly tractable form that allows for comprehensive theoretical study, which forms a main focus of the work. We demonstrate the ease of posterior computation in simulations and real data experiments.

Keywords: Bayesian inference, martingale, quantile estimation, quantile regression

1 Introduction

Quantile estimation and regression has wide applications in fields such as econometrics and biostatistics (Koenker & Bassett Jr, 1978). The Bayesian approach has garnered attention due to the ability to fully quantify uncertainty through the posterior distribution. However, a Bayesian equivalent is not immediately obvious as the need to specify a likelihood is challenging. Yu and Moyeed (2001) utilize a 'working likelihood' based on the asymmetric Laplace distribution, where the quantile parameterizes a potentially misspecified likelihood. Yang et al. (2016) then alleviate the impact of this misspecification through an adjustment of the posterior covariance to attain asymptotic frequentist validity of the posterior credible intervals. Within the Bayesian nonparametric literature, the challenge lies in eliciting a valid nonparametric prior. Hjort and Walker (2009) introduced the quantile pyramid, which is a nonparametric prior with support on piecewise linear quantile functions. Rodrigues et al. (2019) and An and MacEachern (2024) extend the quantile pyramid to allow for the introduction of covariate dependence. Tokdar and Kadane (2012) introduce a semiparametric prior for linear quantile regression which has support on monotone curves; Yang and Tokdar (2017) and X. Chen and Tokdar (2021) then extend this to more complex covariate spaces. In general, constructing prior distributions for quantile functions is nontrivial, and posterior inference in all cases require the use of Markov chain Monte Carlo (MCMC) techniques, which can often be computationally demanding.

Received: June 4, 2024. **Revised:** November 17, 2025. **Accepted:** November 26, 2025

A recent promising class of approaches that avoids the need to work directly with a likelihood are methods which *generalize* Bayesian inference. One direction is the generalized Bayesian update of Bissiri et al. (2016), which relies on a loss function instead of a likelihood, and motivates updating through coherence. The asymmetric Laplace likelihood can be motivated in this fashion, as the likelihood is indeed proportional to the exponentiated check loss function. Another approach is to view Bayesian inference as a *predictive* task by taking advantage of connections between posterior and predictive inference, which has been explored in Berti et al. (2020); Fong et al. (2023); Fortini and Petrone (2020, 2023) and others. Particularly close to our work is the *martingale posterior* (MP) of Fong et al. (2023), where the traditional likelihood-prior construct of Bayesian inference is replaced with the elicitation of a sequence of predictive densities, which shares the motivation of the prequential approach of Dawid (1984). Given observations $Y_{1:n}$, the sequence of predictives is utilized to impute the remainder of the population, $Y_{n+1:\infty}$, from which an estimand can be computed and is then distributed according to the MP.

1.1 Our contribution

In this work, we introduce a Bayesian nonparametric method for quantile estimation and regression, motivated from the purely predictive framework of the MP (Fong et al., 2023). The core idea is to utilize a recursive estimate of the quantile function as a *generative predictive*, which is then sampled from and updated to impute $Y_{n+1:\infty}$. We will differentiate between quantile functions and quantile function *estimates*, where the first is monotonically increasing but the latter may not be. The distribution of the resulting random quantile function of $Y_{n+1:\infty}$ is then termed the *quantile martingale posterior* (QMP). The generative predictive is essentially a stochastic approximation of the quantile function with an additional coherence condition.

The QMP inherits many advantages of the MP framework. First, exact posterior computation is simple and expedient, as MCMC is not required at all. We will see later that a highly accurate approximate posterior sampling reduces computation time even further, making the imputation step negligible in time. Second, in many situations, we may not have strong prior information despite wanting to quantify posterior uncertainty. The prior distribution can thus be a nuisance to specify, motivating noninformative priors (Berger et al., 2009). This is particularly true in Bayesian nonparametrics, where the specification of the prior is both technically demanding and challenging to interpret. In contrast to traditional Bayes, the QMP is entirely data-driven and prior-free, and the model is simple to interpret due to connections to stochastic approximation.

The QMP also has unique advantages within the context of quantile estimation. The issue of monotonicity or quantile crossing is handled automatically by the imputation step in the QMP, and we rely heavily on the useful theory of increasing rearrangements. This is another benefit of working with the predictive framework and specifically with a generative predictive as in our case. Extensions to incorporate covariate dependence, e.g. for linear quantile regression, is then straightforward again due to connections to stochastic approximation, especially when compared to traditional Bayesian nonparametric priors. Finally, we will be extending beyond the conditionally identically distributed (c.i.d.) condition (Berti et al., 2004) required for the original MP, which greatly expands the possible set of models for Bayesian nonparametric inference.

In exchange for these benefits, we will immediately be faced with theoretical challenges, for which solutions form the bulk of this work. In general, theoretical study of the QMP is challenging due to the inability to rely on standard tools for Bayesian asymptotics, and we now cannot even rely on results from the c.i.d. literature. To study the existence and support of the QMP, we will leverage new tools from the Banach space valued martingale literature, which will aid us greatly. In addition, we will be able to study the weak convergence of the QMP, as well as posterior consistency and contraction in the frequentist sense, which is novel for MPs. The theoretical results have strong practical implications as they guide model elicitation, hyperparameter setting and approximate sampling. We hope these methods and tools used are also of independent interest and will be useful for future research in MPs and Bayesian inference in general. We speculate that the aforementioned theory may also be adapted to the Bayesian estimation of more general monotone functions (e.g. Chakraborty & Ghosal, 2021).

We now provide an outline the paper. In Section 2, we will review the role of increasing rearrangement in quantile estimation and the MP framework from Fong et al. (2023). We then

introduce the QMP in the unconditional setting, and provide intuition as to the various model components and sampling algorithm. Section 4 will then cover the bulk of the theory, with most derivations postponed for the Supplementary Material (SM). Section 5 will discuss the practical implications of the theory, with a focus on the setting of a few key hyperparameters and an expedient approximate posterior sampling scheme. Section 6 then extends the QMP for quantile regression, covering similar theory and practical discussions. Section 7 demonstrates the QMP in a simulation and real data example, and Section 8 concludes with future directions.

2 Quantile martingale posteriors

For ease of exposition, we first introduce the quantile martingale posterior without covariate dependence, and extend it to the quantile regression case in Section 6. For the remainder of this section, let $Y_{1:n}$ be n i.i.d. copies of the r.v. Y from an unknown sampling distribution P^* with cumulative distribution function (CDF) $P^*(y)$. In this work, we restrict ourselves to the setting where $Y \in \mathbb{R}$ is univariate; potential extensions of the QMP to the multivariate case are outlined in Section 8.3. To aid with the technical results in Section 4, we will also assume throughout that P^* has compact support, which is relatively standard in the quantile estimation literature, e.g. Chernozhukov et al. (2010), Van der Vaart (2000, Lemma 21.4(ii)).

2.1 Quantile functions and increasing rearrangement

To begin, we outline some prerequisites on the quantile function and its estimators, with a particular focus on *increasing rearrangement* (Chernozhukov et al., 2010). Readers familiar with quantile estimation may skip ahead to Section 2.2. The quantile function $Q^* : (0, 1) \to \mathbb{R}$ is the left-continuous, monotonically increasing function defined as

$$Q^*(u) = \inf\{y \in \mathbb{R} : u \leq P^*(y)\}.$$

The quantile function is particularly useful for inverse-transform sampling from P^*, which we strongly leverage in our work. In particular, given a uniform r.v. $V \sim \mathcal{U}(0, 1)$, we have that $Q^*(V) \sim P^*$. This is due to the key property that $Q^*(u) \leq y$ if and only if $u \leq P^*(y)$ for all $u \in (0, 1)$. A detailed summary of properties of quantile functions can be found in Embrechts and Hofert (2013). For the remainder of the paper, we will assume that that both $P^*(y)$ and $Q^*(u)$ are continuous.

Let Q_n be an *estimate* of the quantile function Q^* from $Y_{1:n}$. A well-known problem in quantile estimation is that $Q_n(u)$ may not be monotonically increasing on $u \in (0, 1)$, so it is not a valid quantile function. In the case of quantile regression, this is known as the *quantile crossing problem* (Bassett Jr & Koenker, 1982; Chernozhukov et al., 2010; He, 1997), where the lack of monotonicity causes quantile curves as functions of the covariates to cross one another for different values of u. Many solutions to this problem have been proposed, but we will focus particularly on increasing rearrangements, as this occurs naturally under the MP framework.

For the remainder of the paper, we will denote a potentially non-monotone quantile function estimate as Q_n. Let $Q_n^\dagger$ denote the increasing rearrangement of Q_n, which is defined as follows:

$$P_n(y) = \int_0^1 1(Q_n(u) \leq y)du, \quad Q_n^\dagger(u) = \inf\{y \in \mathbb{R} : u \leq P_n(y)\}. \tag{1}$$

$Q_n^\dagger$ is then a proper quantile function, where one can see the monotonicity as follows. For $V \sim \mathcal{U}(0, 1)$, the function P_n is the CDF of $Q_n(V)$, so $Q_n^\dagger$ is a valid quantile function and must be monotonically increasing. The connection to the bootstrap is hence obvious and of key importance—the quantile estimate Q_n gives us a means to simulate from P_n (or equivalently $Q_n^\dagger$) through the inverse transform, which forms the basis of our work. In Figure 1 (left), we show an example of rearranging a non-monotone Q_n into a monotonically increasing $Q_n^\dagger$, with corresponding P_n in Figure 1 (right). We can see that $Q_n^\dagger$ agrees with Q_n in some regions, and preserves continuity. A detailed discussion on properties of rearrangement for quantile estimation can be found in Chernozhukov et al. (2010).

There is also a close connection to rearrangement inequalities (Hardy et al., 1952), which have previously been leveraged in estimation by Chernozhukov et al. (2009) and specifically in quantile

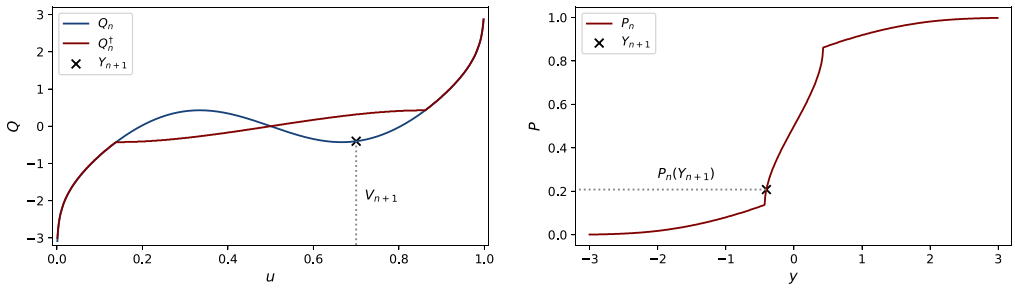

Figure 1. Plot of (Left) Q_n and rearranged $Q_n^\dagger$ and (Right) implicit P_n; for $V_{n+1} = 0.7$ which gives $Y_{n+1} \approx -0.4$. Note that $P_n(Y_{n+1}) \neq V_{n+1}$, although $P_n(Y_{n+1}) \overset{d}{=} V_{n+1}$ as both are distributed according to $\mathcal{U}(0, 1)$.

estimation/regression by Chernozhukov et al. (2010). Many useful properties of $Q_n^\dagger$ have also been shown in Chernozhukov et al. (2010), and we will outline and utilize this theory in Section 4. In particular, one can show that $Q_n^\dagger$ is always a better estimate of Q^* in terms of L^p distance as a result. Increasing rearrangement also preserves continuity properties, which will be useful for us.

2.2 Martingale posterior distributions

The MP is a generalization of the Bayesian framework introduced by Fong et al. (2023). The key notion is that Bayesian uncertainty on a parameter of interest θ arises from the unknown remainder of the population $Y_{n+1:\infty}$ that has yet to be observed. Fong et al. (2023) show that posterior sampling is equivalent to the predictive imputation of $Y_{n+1:\infty}$ given $Y_{1:n}$, followed by the computation of θ as an estimand from $Y_{1:\infty}$. This procedure is termed as *predictive resampling*, where the sequence of predictive distributions, $P_n(y) = P(Y_{n+1} \leq y \mid Y_{1:n})$, is used to sequentially impute $Y_{n+1:\infty}$, which is outlined in Algorithm 1.

Armed with this interpretation of Bayesian inference, the MP then generalizes Bayes by eliciting a general sequence of predictive distributions $\{P_n, P_{n+1}, \ldots\}$ directly as the statistical model, removing the need for a likelihood and prior, and instead relying on predictive resampling to obtain a posterior distribution on a parameter of interest. In order for the MP on θ to exist, we require the sequence P_N to converge almost surely to a random probability measure P_∞ when predictive resampling, which is ensured through a martingale condition. In particular, Fong et al. (2023) requires the following predictive coherence condition, $\mathbb{E}[P_{N+1}(y) \mid Y_{1:N}] = P_N(y)$ for each $y \in \mathbb{R}$ and all $N \geq n$. This then implies that the sequence of imputed observations $Y_{n+1:\infty}$ is c.i.d., and Berti et al. (2004) show that the c.i.d. condition is sufficient for the existence of a P_∞ which $P_N \to P_\infty$ weakly almost surely. This c.i.d. condition unfortunately greatly constrains the class of predictive distributions one can use for the MP. The MP also has close connections to the Bayesian bootstrap of Rubin (1981), which has recently had a resurgence in popularity, e.g. Fong et al. (2019); Nie and Rockova (2023). Other nonparametric MPs have been suggested in Cui and Walker (2024a), Cui and Walker (2024b) and Walker (2024). Parametric versions of the MP have also been introduced in Walker (2022) and Holmes and Walker (2023), where a parametric predictive distribution is utilized for predictive resampling. The martingale is now directly the parameter of interest θ, ensuring convergence of an estimator $\theta_N \to \theta_\infty$ instead of $P_N \to P_\infty$, which relaxes the c.i.d. condition.

Fong et al. (2023) enforce the c.i.d. condition using a nonparametric recursive update for P_N based on the bivariate copula as introduced in Hahn et al. (2018). This recursive update is inspired by the Dirichlet process mixture model, and takes the form

$$P_{N+1}^{\mathrm{MP}}(y) = (1 - \alpha_{N+1}) P_N^{\mathrm{MP}}(y) + \alpha_{N+1} H_\rho\big(P_N^{\mathrm{MP}}(y), P_N^{\mathrm{MP}}(Y_{N+1})\big), \tag{2}$$

where $H_\rho(u, v)$ is the conditional distribution of the bivariate Gaussian copula of the form

$$H_\rho(u, v) = \Phi\left\{\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1 - \rho^2}}\right\}, \tag{3}$$

Algorithm 1 Predictive resampling

Compute P_n from the observed data $Y_{1:n}$
for $b \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** N **do**
    Sample $Y_i \sim P_{i-1}$
    Update $P_i\{P_{i-1}, Y_i\}$
  end
  Evaluate $\theta_N^{(b)} = \theta(Y_{1:N})$ or $\theta(P_N)$
end
Return $\{\theta_N^{(1)}, \ldots, \theta_N^{(B)}\}$

and $\rho \in (0, 1)$ is the correlation term and Φ and Φ^{-1} are the standard normal CDF and its inverse respectively. The weights are usually chosen $\alpha_N = O(N^{-1})$ in order for the update to approach the independence copula as $N \to \infty$. Intuitively, the second term in the sum is akin to a kernel centred at Y_{N+1} as in the traditional kernel density estimate, but the main difference is that the kernel is adaptive as it depends on P_N^{MP}.

The nonparametric MP based on (2) faces a few challenges. First, estimating a probability density constrains the update due to the need to integrate to 1, or equivalently that P_N^{MP} is a valid differentiable CDF. Second, although extensions to conditional density estimation are provided in Fong et al. (2023), it is challenging to incorporate structure in the regression setting (e.g. linearity), due to the stringent c.i.d. condition. Finally, studying the asymptotic properties of the nonparametric MP based on the copula is challenging, due to working in the space of probability measures (Berti et al., 2004). We will see that the QMP alleviates these challenges faced by the nonparametric MP outlined in the previous section as the space of quantile function estimates is much easier to handle.

2.3 Quantile predictive resampling

In this section, we introduce the *quantile martingale posterior* framework, which builds on the ideas of Fong et al. (2023) to address quantile estimation. The core idea is to utilize a recursive update for an estimate of the quantile function, which serves as our predictive imputation machine. For now, assume that we have an estimate of the quantile function, $Q_n : (0, 1) \to \mathbb{R}$, computed from the i.i.d. observations $Y_{1:n}$. We will address how to obtain Q_n later, and will assume that Q_n is continuous and bounded, but not necessarily monotonic. Given Q_n, consider the following sampling scheme:

1. Simulate $V_{n+1} \sim \mathcal{U}(0, 1)$
2. Compute $Y_{n+1} = Q_n(V_{n+1})$.

Viewed in this manner, Q_n is simply a tool for simulating Y_{n+1}, and can thus be viewed as a *generative predictive sampler*. This is analogous to the approach of the generative adversarial network (Goodfellow et al., 2020), where accurate samples are generated by passing noise through a neural network instead of estimating the density. It is also not challenging to see that Y_{n+1} is in fact distributed according to P_n with the corresponding rearranged quantile function $Q_n^\dagger$, which is indeed monotonic. The quantile function estimate Q_n thus provides us a means to simulate from the rearranged predictive distribution directly, without the need to actually compute the rearrangement operator (1). This procedure is illustrated in Figure 1 (left), where we draw $V_{n+1} \sim \mathcal{U}(0, 1)$ and read off the corresponding value $Q_n(V_{n+1})$ to get a sample. The quantile function and CDF of Y_{n+1} is then $Q_n^\dagger$ and P_n, as shown in red in Figures 1 (left) and 1 (right) respectively. We will also refer to $Q_n^\dagger$ and P_n as the *implicit* quantile function and CDF respectively. We provide more intuition as to what rearrangement implies for the resulting QMP in Section 4.

Given the further specification of a recursive update $(Q_n, Y_{n+1}) \to Q_{n+1}$, and assuming appropriate conditions on the update, we will then have all the ingredients needed to sample from the QMP, which is outlined in Algorithm 2. The main difference to the original MP is that we keep track of a quantile function estimate, which can be interpreted as a generative predictive sampler,

Algorithm 2 Quantile predictive resampling

Compute Q_n from the observed data $Y_{1:n}$
for $b \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** N **do**
    Sample $V_i \sim \mathcal{U}(0, 1)$; compute $Y_i = Q_{i-1}(V_i)$
    Update $Q_i\{Q_{i-1}, Y_i\}$
  end
  Evaluate $\theta_N^{(b)} = \theta(Y_{1:N})$ or $\theta(Q_N^\dagger)$
end
Return $\{\theta_N^{(1)}, \ldots, \theta_N^{(B)}\}$

and it does not need to satisfy the monotonicity property. For now, we leave the update unspecified, but we will investigate the appropriate elicitation of the update function in detail starting in Section 3. Compared to the original MP, the class of possible predictives for the QMP is much broader, as we only require Q_n to be bounded and continuous, whereas the original MP requires estimating a probability density function. We will see in Section 4 that this relaxation allows for comprehensive theoretical study of the QMP, and Section 6 will illustrate the simplicity of incorporating covariates for conditional quantile estimation.

2.4 Martingale condition and coherence

In order for the QMP to be well-specified under the scheme of Algorithm 2, we will require an analogous condition to the c.i.d. property for the original nonparametric MP. Unsurprisingly, we find that a martingale condition is once again sufficient for existence of the MP, which corresponds to an interesting coherence property on the generative predictive.

While we will leave the technical details for Section 4, we briefly outline the martingale condition here. In particular, we require a similar condition on the estimate of the quantile:

$$\mathbb{E}\big[Q_{N+1}(u) \mid Y_{1:N}\big] = Q_N(u) \tag{4}$$

for each $u \in (0, 1)$ for all $N \geq n$. Here, the conditional expectation is over $Y_{N+1} = Q_N(V_{N+1})$, so we are averaging over the r.v. $V_{N+1} \sim \mathcal{U}(0, 1)$. Under assumptions on the recursive update, we show in Section 4 that the limiting empirical distribution of $Y_{n+1:\infty}$ converges to some P_∞ weakly almost surely, which has a corresponding random quantile function $Q_\infty^\dagger$. This kind of convergence also has close connections to exchangeability. The QMP is then the distribution of $Q_\infty^\dagger$ or P_∞ (or appropriate functionals thereof). The theory requires technical tools from the function-valued martingales and rearrangement operator literature, but intuitively, the above weak convergence implies that the QMP over the unknown quantile function exists. Furthermore, we will see that the additional flexibility gained in working with quantile functions instead of CDFs will allow us to quantify the convergence of $Q_N^\dagger$ to $Q_\infty^\dagger$ more precisely.

Previously, Fong et al. (2023) highlighted that the c.i.d. condition was equivalent to predictive coherence, as the posterior mean of the predictive CDF $P_\infty(y)$ is equal to the initial estimate $P_n(y)$. In the QMP case, we will instead have a kind of *generative* coherence. To interpret this, suppose we would like to draw a sample $\widetilde{Y} \sim P_\infty$, which requires drawing $V \sim \mathcal{U}(0, 1)$ and plugging it into the limiting generative sampler $\widetilde{Y} = Q_\infty(V)$. From (4), we have that $\mathbb{E}[Q_\infty(v) \mid Y_{1:n}, v] = Q_n(v)$ for each $v \in (0, 1)$, where the conditional expectation is over the imputed $Y_{n+1:\infty}$. This suggests that the posterior distribution of the whole generative sampler $Q_\infty(\cdot)$ is unbiased, and we term this property as generative coherence.

3 Recursive quantile estimator

3.1 Stochastic approximation

We now introduce a novel recursive update to estimate continuous quantile functions. Recursive updates are particularly well-suited for the QMP, as it gives us both a means for predictive resampling and for ensuring the necessary martingale condition, which will we discuss in depth shortly.

The motivation is based on the connection between recursive methods and *stochastic approximation* (Lai, 2003). Hahn et al. (2018) and Fong et al. (2023) highlight the interpretation of (2) as a stochastic approximation of the CDF/density, and Walker (2022), Holmes and Walker (2023) and Fortini and Petrone (2025) rely on a stochastic gradient descent approach to update the parameter θ_N in the parametric predictive Bayesian context.

We take a similar approach here, leveraging a stochastic approximation estimate of the quantile function, which has also been investigated in works such as Aboubacar and Thiam (2014), Kohler et al. (2014) and L. Chen et al. (2023) in the non-Bayesian setting. One can define the quantile at $u \in (0, 1)$ as $Q^*(u) = \arg\min_q \int \rho_u(y-q)dP^*(y)$ where $\rho_u(z) = z(u - 1(z \leq 0))$ is the familiar check loss. Although the check loss is not differentiable at $z = 0$, one can still utilize the sub-gradient, and define the recursive update

$$Q_{n+1}(u) = Q_n(u) + \alpha_{n+1}\big[u - 1\big(Y_{n+1} \leq Q_n(u)\big)\big] \tag{5}$$

which is studied in L. Chen et al. (2023). In the above, α_n is a sequence of decreasing weights chosen so that

$$\sum_{i=1}^{\infty} \alpha_i = \infty, \quad \sum_{i=1}^{\infty} \alpha_i^2 < \infty \tag{6}$$

as is standard in stochastic approximation, where the first condition ensures initial conditions are forgotten and the second condition on α_i ensures the algorithm converges. One can show that stochastic approximation yields a consistent estimator for the minimizer under additional assumptions such as convexity of the objective and boundedness of the gradients; further details on these conditions can be found in Lai (2003). For the QMP, we will consider consistency in full detail in Section 4.

There are however two main issues with (5) that cause it to be unsuitable for the QMP, which we address now. First, we are interested in the case where Q^* is continuous, whilst (5) will recover a discontinuous estimate of the quantile. Second, for the purposes of the QMP, there is the subtle but important point that (5) does not imply a martingale for $Q_n(u)$ under the quantile predictive resampling, which will be important for showing the existence of the QMP.

3.2 Recursive copula update

We now describe a recursive estimate of the quantile function which returns both continuous curves and satisfies the required martingale condition. To begin, we highlight the connection between the recursive update of the predictive CDF based on the bivariate Gaussian copula as shown in (2) and the empirical distribution and Bayesian bootstrap. The empirical distribution can be written recursively as

$$P_{N+1}^{\mathrm{MP}}(y) = (1 - \alpha_{N+1})P_N^{\mathrm{MP}}(y) + \alpha_{N+1}1\big(Y_{N+1} \leq y\big)$$

where $\alpha_N = N^{-1}$. By comparing the above update to (2), we see that the indicator term $1(Y_{N+1} \leq y)$ corresponds to the term $H_\rho(P_N^{\mathrm{MP}}(y), P_N^{\mathrm{MP}}(Y_{N+1}))$. In fact, we have that $\lim_{\rho \to 1} H_\rho(u, v) = 1(v \leq u)$. As a result, (2) can be viewed as a smoothed version of the empirical distribution update.

Inspired by this connection, we apply the same intuition to extend (5) into a form that is suitable for the QMP. Our suggested recursive update of the quantile function estimate is then

$$Q_{N+1}(u) = Q_N(u) + \alpha_{N+1}\big[u - H_{\rho_{N+1}}\big(u, P_N(Y_{N+1})\big)\big], \tag{7}$$

where P_N is the rearranged CDF function of Q_N, and α_N satisfies (6). We postpone discussion on the sequence $\rho_N \in (0, 1)$ except for requiring that $\rho_N \to 1$ as N increases, which is a key difference between the QMP and the regular MP, as the bandwidth ρ is kept fixed in the latter. Intuitively, the update (7) is akin to a Bayesian analogue of a recursive kernel-smoothed quantile estimator (e.g. Aboubacar & Thiam, 2014) which arises naturally from a stochastic optimization viewpoint.

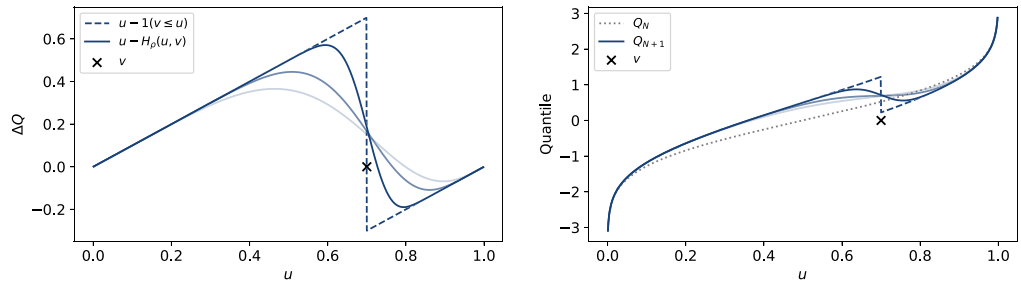

Figure 2. Plot of (Left) $[u - H_\rho(u, v)]$ and (Right) updated quantile estimate $Q_{N+1}(u)$ and old $Q_N(u)$ (fx2); for $v = 0.7$ with $\rho = (0.9, 0.95, 0.99)$ (fx5,fx4,fx1) and $[u - 1(v \le u)]$ (fx3).

Figure 2 (left) illustrates the form of $[u - H_\rho(u, v)]$ for increasing values of ρ, which we see approaches the limiting case $[u - 1(v \le u)]$. We can thus directly view $[u - H_\rho(u, v)]$ as a continuous relaxation of $[u - 1(v \le u)]$. Figure 2 (right) then illustrates the effect of updating with an observation with $P_N(Y_{N+1}) = v$.

Unlike the non-Bayesian case, much care is needed to ensure the coherence condition discussed in Section 2.4 is satisfied. To this end, the rearrangement step is crucial for obtaining the martingale under predictive resampling, as Q_N may not be monotonic. We highlight the key property that both $[u - H_\rho(u, v)]$ and $[u - 1(v \le u)]$ are not monotonic, so it is possible for Q_{N+1} to not be monotonic even if Q_N is. This is illustrated in Figure 2 (right), where for ρ close to 1, we have non-monotonicity of the updated Q_{N+1}.

To understand the importance of rearrangement for the martingale condition, we focus on the step function case, and contrast between $[u - 1(Y_{N+1} \le Q_N(u))]$ versus $[u - 1(P_N(Y_{N+1}) \le u)]$, where the first case is from (5) and the latter is from (7) with $\rho \to 1$. If Q_N is a proper quantile function, i.e. it is motonically increasing and left-continuous, then we have $Y_{N+1} \le Q_N(u) \Leftrightarrow P_N(Y_{N+1}) \le u$. In this case, it is thus clear that the two updates are equivalent. However, when Q_N is not monotonic, the two updates will differ. To see why the latter update is more suitable, consider $Y_{N+1} \sim P_N$ where P_N is continuous. Under predictive resampling, we have $P_N(Y_{N+1}) \sim \mathcal{U}(0, 1)$, so in the latter case we have

$$\mathbb{E}[u - 1(P_N(Y_{N+1}) \le u) \mid Y_{1:N}] = u - \int_0^1 1(v \le u)\, dv = 0.$$

In the first case however, we have $\mathbb{E}[u - 1(Y_{N+1} \le Q_N(u)) \mid Y_{1:N}] = u - P_N(Q_N(u)) \ne 0$. The issue arises as $P_N(Q_N(u)) \ne u$ when Q_N is not monotonic, and the size of the deviation is related to how non-monotonic Q_N is. Finally, the above logic extends to the smooth case, where one can show that

$$\mathbb{E}\left[u - H_{\rho_{N+1}}(u, P_N(Y_{N+1})) \mid Y_{1:N}\right] = u - \int_0^1 H_{\rho_{N+1}}(u, v)dv = 0.$$

This follows as $\int_0^{v'} H_\rho(u, v)dv = C_\rho(u, v')$ where C_ρ is the bivariate Gaussian copula, and taking $v' \to 1$ returns $C_\rho(u, 1) = u$. As a result, the recursive update (7) satisfies the required martingale condition from Section 2.4 when P_N is continuous. This once again highlights the bivariate copula as a versatile building block for Bayesian nonparametrics, especially for smooth functions.

3.3 Posterior sampling from the QMP

A nice property of the QMP is that rearrangement is automatically handled during predictive resampling. To see this, we revisit the quantile predictive resampling scheme, where $Y_{N+1} = Q_N(V_{N+1})$ for $V_{N+1} \sim \mathcal{U}(0, 1)$, resulting in $Y_{N+1} \sim P_N$. The recursive quantile update only relies on Y_{N+1} through $P_N(Y_{N+1})$, and again we have $P_N(Y_{N+1}) \sim \mathcal{U}(0, 1)$ if P_N is continuous. To carry out one step of predictive resampling, it is then simply a matter of simulating $V_{N+1} \sim \mathcal{U}(0, 1)$ and computing

$$Q_{N+1}(u) = Q_N(u) + \alpha_{N+1}\left[u - H_{\rho_{N+1}}\left(u, V_{N+1}\right)\right]. \tag{8}$$

Once again, posterior sampling only depends on the simulation of uniform r.v.s, which is extremely cheap, and does not require complex MCMC schemes. In practice, the update (8) truncated at some reasonably large $N \gg n$ is sufficient for convergence to Q_∞. Similar to the examples in Fong et al. (2023), one can assess convergence of Q_N by monitoring the stability of the L^2 distance between Q_N and the initial Q_n; more details can be found in Section D.3 of the online supplementary material. Looking ahead, we will shortly see in Section 5.3 that a truncation may not even be necessary as we can identify the limiting law of $Q_\infty - Q_N$.

The advantages of predictive resampling over traditional Bayes is clear and outlined in Fong et al. (2023), which we now recap. First, we can completely avoid issues of mixing and serial computation that faces MCMC, relying only on uniform r.v.s and simple computations to provide i.i.d. posterior samples. The update (8) is also particularly easy to parallelize, which we discuss further in Section 3.5. Finally, the only source of approximation comes from a relatively harmless truncation step. Interestingly, sampling from the QMP has additional advantages over the regular MP due to working in the space of quantile function estimates. The first is that samples from P_∞ can be obtained directly when quantile predictive resampling, as we are working with a generative predictive. Specifically, if we replace u in (8) with $U \sim \mathcal{U}(0, 1)$, then computing the recursive update will transform $Y_N = Q_N(U) \sim P_N$ into $Y_{N+1} = Q_{N+1}(U) \sim P_{N+1}$. In practice, we can thus initialize a vector $U_{1:T} \overset{\text{iid}}{\sim} \mathcal{U}(0, 1)$ and pass it through the update (8) for $N \geq n + 1$. A second benefit is that theoretical study will be more comprehensive compared to the regular MP, as the QMP only requires u and V_{N+1} as inputs for the update. In particular, we will be able to show a weak convergence result which allows even faster approximate sampling from the QMP based on a Gaussian process (GP).

3.4 Coherency vs. consistency of the initial estimate Q_n

We have thus far motivated the update rule (7) for Q_N precisely due to both its connection to quantile estimation and the satisfaction of the key martingale condition. The latter martingale condition ensures the generative predictive Q_∞ is not biased under predictive resampling. This plays a similar role to the c.i.d. condition in Fong et al. (2023), and we provide a comparison to other martingale conditions found in the MP literature in Section E.1 of the online supplementary material. We also highlight here again that Q_N need not be monotonic even if it is derived from a quantile estimate; the non-Bayesian recursive update (5) similarly produces non-monotonic quantile estimates. This is not an issue for predictive resampling as Q_N is only a vehicle for drawing predictive samples.

As (7) is our learning rule, it is natural and coherent to utilize the same update rule to compute the initial estimate Q_n from the i.i.d. observations $Y_{1:n} \overset{\text{iid}}{\sim} P^*$, starting from some pre-specified Q_0. The initial estimate Q_n is of great importance as it where we begin predictive resampling, and governs the central tendency of the QMP. This tether between estimation and predictive resampling is also discussed in Fong et al. (2023). One can then interpret Bayesian inference as applying the update (7) up until the final 'real' observation Y_n, then imputing $Y_{n+1:\infty}$ from Q_n once we 'run out' of observed data points.

A technical caveat with the above ideal case is that it is challenging to prove frequentist consistency of Q_n at Q^* when applying the update rule (7) to i.i.d. samples (though we suspect it holds). This property is important for the frequentist properties of the QMP, which we will study in Section 4.2.1. Fortunately, it is much easier to establish frequentist consistency if we apply a small tweak to (7), namely an additional *rearrangement* step:

$$Q_{i+1}(u) = Q_i^\dagger(u) + \alpha_{i+1}\left[u - H_{\rho_{i+1}}\left(u, P_i(Y_{i+1})\right)\right], \tag{9}$$

for $i = 1, \ldots, n-1$. The only change is the additional rearrangement of Q_i to $Q_i^\dagger$ before adding the next update term. This change then allows us to establish that $Q_n^\dagger$ is consistent at Q^*, although it disrupts the martingale condition satisfied by (7). The discrepancy between (7) and (9) is however only of technical importance, and we find it to be of little practical significance as (7) and (9)

tend to produce very similar estimates of Q_n as long as ρ_i is chosen to not approach 1 too quickly. The intuition for this practical similarity can be seen in Figure 2 (right), where a sufficiently smooth update (smaller ρ) prevents Q_{N+1} from becoming non-monotonic. In Section E.2 of the online supplementary material, we demonstrate the practical equivalence of (7) and (9) in a simulation example, and provide additional technical intuition for the proof of consistency. As such, we prefer to interpret (7) as the principled coherent updating rule, and view (9) as a technical necessity.

In summary, we recommend the following compromise: utilize the update (9) for $Y_{1:n} \overset{\text{iid}}{\sim} P^*$ to obtain the initial $Q_n^\dagger$, then carry out predictive resampling with (7) for imputing $Y_{n+1:\infty}$ starting at $Q_n^\dagger$. Under this scheme, (9) will guarantee frequentist consistency while (7) will ensure the martingale condition of the QMP. This slight mismatch appears to be the small technical price to pay for working with flexible quantile function estimates, which are known to have issues related to monotonicity of estimates as earlier discussed.

3.5 Algorithm summary

We now summarize the QMP method, and postpone the setting of ρ_i and α_i and approximate sampling to Section 5. Algorithms 3 and 4 below illustrate the full process of obtaining the QMP. Like with the regular MP, there is a distinct separation of estimation and obtaining uncertainty, which is more akin to frequentist methods. In practice, it may be desirable to average the output of Algorithm 3 over multiple permutations of the data (e.g. 10) if it is desirable for the initial estimate $Q_n^\dagger$ to be permutation invariant. Due to the expediency of the update, this is not too restrictive computationally, and no permutation-averaging is required for predictive resampling due to asymptotic exchangeability (discussed in Section 4.1.2). We will require a grid of u-values on which we compute the quantile estimates, and this also governs the 'resolution' of our samples. We find that a grid of 200 evenly spaced points from [0, 1] works well in practice. The number of future samples N can be set by monitoring the convergence of Q_N, and we see that $N \approx n + 5{,}000$ is sufficient in practice. Algorithm 4 can be easily executed in an embarrassingly parallel manner, where each sequence $V_{n+1:N}^{(j)}$ can be sampled beforehand and passed to each worker. In addition to the embarrassingly parallel nature of the QMP, the computation of (8) across the grid of u values and over B independent samples involve only simple computations without complex control flow (unlike MCMC), so leveraging GPU compute can further reduce computation time. However, we will see in Section 5.3 than Algorithm 4 can be approximated even more quickly using a GP.

4 Theory

For the original MP, asymptotic theory was challenging due to the complex dependence in the update. Interestingly, the lack of dependence on the predictive of the first input into $H_\rho(u, v)$ helps to simplify the theory. We distinguish between two asymptotic regimes under the MP framework. The first is the convergence of $P_N \to P_\infty$ from predictive resampling, starting at $N = n + 1$, which we term *predictive* asymptotics. This is closely connected to Doob's consistency theorem (Doob, 1949), and is discussed nicely in Fortini and Petrone (2025). The second is the classical *frequentist* asymptotics, where we study the convergence of the MP or relevant estimates (such as P_n) as $n \to \infty$, where n is the number of i.i.d. observations $Y_{1:n} \overset{\text{iid}}{\sim} P^*$. We will now investigate both for the QMP. Full derivations are postponed to the online supplementary material, although we provide proof outlines when they are particularly insightful.

4.1 Predictive asymptotics

To study the predictive asymptotics of the QMP, we will rely on the theory of *function-valued* martingales (Pisier, 2016). Although the theory is technical, the results and conditions are insightful and simple to interpret. We begin this subsection with prerequisite theory from functional analysis, with details deferred to Section A in the online supplementary material. As the space of possible of quantile function estimates Q_N is quite large due to not requiring monotonicity, we will have sufficient structure to borrow powerful results from functional analysis. Let B be a Banach space of real-valued functions $f : (0, 1) \to \mathbb{R}$ with norm $\|\cdot\|_B$, which Q_N will belong to. In particular, we will work with two very useful spaces that lend themselves to easy study of recursive updates for Q_N. The first is the $L^2((0, 1))$ space, which consists of square-integrable

Algorithm 3 Estimation of quantile function

Initialize Q_0
Data is $Y_1, \ldots, Y_n$
for $i \leftarrow 1$ **to** n **do**
  Compute $V_i = P_{i-1}(Y_i)$
  $Q_i(u) = Q_{i-1}^{\dagger}(u) + \alpha_i[u - H_{\rho_i}(u, V_i)]$
end
Return $Q_n^{\dagger}$

Algorithm 4 QMP sampling

Initialize $Q_n^{\dagger}$ from Algorithm 3
for $b \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** N **do**
    Draw $V_i^{(b)} \sim \mathcal{U}(0, 1)$
    $Q_i^{(b)}(u) = Q_{i-1}^{(b)}(u) + \alpha_i[u - H_{\rho_i}(u, V_i^{(b)})]$
  end
end
Return $\left\{Q_N^{\dagger\,(1)}, \ldots, Q_N^{\dagger\,(B)}\right\}$

functions with norm $\|f\|_2 = \sqrt{\int f(u)^2 du}$. We write the L^2 distance between two elements $f, g \in L^2((0, 1))$ as $d_2(f, g) = \|f - g\|_2$. The second is the Sobolev space $H^1((0, 1))$ consisting of functions $f \in L^2((0, 1))$ which are weakly differentiable with weak derivative $f' \in L^2((0, 1))$, which shares properties with the regular derivative. A very useful property in the 1-dimensional case is that if $f \in H^1((0, 1))$, then f is equal almost everywhere to an absolutely continuous function. The norm in the Sobolev space $H^1((0, 1))$ is then $\|f\|_{1,2} = \sqrt{\|f\|_2^2 + \|f'\|_2^2}$, with corresponding distance $d_{1,2}(f, g) = \|f - g\|_{1,2}$. Both L^2 and H^1 are Hilbert spaces, which will allow us to apply function-valued martingale convergence theorems easily.

Through Algorithm 4, Q_N will evolve randomly, so we require a probability space on B-valued objects. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space. A r.v. in this case is a function $f : \Omega \to B$ which is Bochner measurable and takes values in B, so realizations of the r.v. are functions, that is $f(\omega) \in B$ for $\omega \in \Omega$. We write $L^p(\Omega, \mathcal{F}, \mathbb{P}; B)$ or $L^p(B)$ as the space of Bochner measurable functions with $\mathbb{E}[\|f\|_B^p] = \int \|f\|_B^p d\mathbb{P} < \infty$ for some $1 \leq p < \infty$, where we will mostly be using $p = 2$. The norm of this space is defined as $\|f\|_{L^p(B)} = (\mathbb{E}[\|f\|_B^p])^{1/p}$, and functions that are equal a.e. are identified. In our use cases, expectations within this space can be evaluated pointwise on the function, so the condition (4) is enough to ensure Q_N is a function-valued martingale. Details regarding (conditional) expectations are in Section A.1 of the online supplementary material.

4.1.1 Existence and support of the QMP

We now study the convergence of the sequence $Q_{n+1}, Q_{n+2}, \ldots$ under quantile predictive resampling with Algorithm 4, which will inform us on properties of the QMP. The main theorem we will use is the convergence theorem for Banach space valued martingales, which we cover in detail in Section A.1 of the online supplementary material.

We will need the following assumptions on the initial quantile function $Q_n^{\dagger} : (0, 1) \to \mathbb{R}$ from which we start predictive resampling, as well as an assumption on the copula update.

Assumption 1 ((Bounded in L^2)). $Q_n^{\dagger}$ satisfies $\|Q_n^{\dagger}\|_2 < \infty$.

Assumption 2 ((Weak derivatives bounded in L^2)). $Q_n^\dagger$ is weakly differentiable with weak derivative $q_n^\dagger$ which satisfies $\|q_n\|_2 < \infty$, so $\|Q_n^\dagger\|_{1,2} < \infty$.

Assumption 3 ((Learning rate)). The learning rate sequence takes the form $\alpha_i = a(i+1)^{-1}$ for some $a \in (0, \infty)$ for $i \geq 1$.

Assumption 4 ((Bandwidth)). The bandwidth sequence takes the form $\rho_i = \sqrt{1 - ci^{-k}}$ where $0 < k < 1$ and $0 < c < 1$ for $i \geq 1$.

Intuitively, Assumptions 1 and 2 ensure that the initial $Q_n^\dagger$ is sufficiently well-behaved. Assumption 3 satisfies (6) which is standard for stochastic approximation. Assumption 4 ensures that ρ_N does not approach 1 too quickly, i.e. the smoothness of the update function does not decrease too quickly. All the above assumptions are verifiable in practice, as they only pertain to the initial estimate or the learning rule.

Proposition 1 Under Assumptions 1 and 3, there exists a random function Q_∞ with realizations in $L^2((0, 1))$ such that $d_2(Q_N, Q_\infty) \to 0$ a.s.

Proof Outline. We rely on the martingale convergence theorem for Banach spaces as given in Theorem A1 in the online supplementary material. By construction, we have (4) so Q_N is a martingale. The main condition to check is that $\sup_{N \geq n} \mathbb{E}[\|Q_N\|_2^2] < \infty$, which is detailed in the online supplementary material. □

The above proposition thus guarantees the existence of the QMP, which is the distribution of Q_∞. Under relatively weak constraints on the predictive update, we can say much more about the support of the QMP.

Theorem 1 Under Assumptions 1–4, there exists a random function Q_∞ with realizations in $H^1((0, 1))$ such that $d_{1,2}(Q_N, Q_\infty) \to 0$ a.s.

Proof Outline. The key here is that the Assumption 4 on the bandwidth prevents the expected Sobolev norm from diverging to infinity, i.e. $\sup_{N \geq n} \mathbb{E}[\|Q_N\|_{1,2}^2] < \infty$. This allows us to apply online supplementary material, Theorem A1 as we did in Proposition 1. □

Corollary 1 Under Assumptions 1–4, realizations of Q_∞ are absolutely continuous on $(0, 1)$ a.s., up to the equivalence class of $H^1((0, 1))$.

In other words, the above theorem and corollary implies that samples of Q_∞ from the QMP are absolutely continuous and thus differentiable almost everywhere a.s. We have thus managed to identify the support of the QMP by leveraging the Sobolev space, which is crucial if absolute continuity of the quantile function estimate is desired. However, we have only studied the quantile estimate Q_N, which may not be monotonic. Since the actual object of interest is the implicit quantile function or CDF $Q_N^\dagger / P_N$, the question is whether we can say anything about the QMP distribution over those. Fortunately the answer is yes, due to the regularizing effect of the rearrangement operator. To first study the convergence of $Q_N^\dagger$, we will need the following well-known proposition on rearrangement:

Proposition 2 ((Chernozhukov et al., 2009; Lorentz, 1953)). Let f, g be any two functions $[0, 1] \to C$ for some bounded subset $C \subset \mathbb{R}$ with increasing rearrangements $f^\dagger, g^\dagger$ respectively. We then have $d_2(f^\dagger, g^\dagger) \leq d_2(f, g)$.

Consider the case where $g^\dagger = Q^*$ is a proper quantile function. The above proposition then states that the rearrangement of Q_N to $Q_N^\dagger$ can only improve the estimate (Chernozhukov

et al., 2009). Furthermore, the rearrangement procedure does not hurt the smoothness of the function, which implies the following result.

Proposition 3 Under Assumptions 1–4, there exists a random function $Q^\dagger_\infty$ with realizations in $H^1((0,1))$ such that $d_2(Q^\dagger_N, Q^\dagger_\infty) \to 0$ a.s, where realizations of $Q^\dagger_N$ and $Q^\dagger_\infty$ are proper monotonically increasing quantile functions.

We highlight that $Q^\dagger_\infty$ is absolutely continuous on $(0, 1)$ up to the equivalence class of $H^1((0,1))$ a.s., which follows from the well-known property that rearrangement preserves absolute continuity (Theorem A2 in the online supplementary material). We thus have that the predictive quantile function $Q^\dagger_N$ converges to an absolutely continuous random quantile function $Q^\dagger_\infty$ a.s. In other words, the rearrangement operator does not significantly affect the predictive asymptotics of the QMP, although we highlight that in general the posterior mean of the QMP is not $Q^\dagger_n$. Note that the convergence is only in the L^2 norm, as strengthening the convergence to hold in the Sobolev norm is trickier. A technical result states that a variant of the rearrangement operator is continuous in the Sobolev space $H^1((0,1))$ (Almgren Jr & Lieb, 1989; Coron, 1984), so it is likely possible for our case. However, this is stronger than what we require, so we leave this for future work.

Summary. Although the results are technical, the intuition is hopefully clear. We have utilized the function-valued martingale convergence theorem to show that the quantile estimate Q_N converges a.s. (in the norm of the respective Banach space) to a random Q_∞ under Algorithm 4. The regularizing behaviour of the rearrangement operator then assures us that the implicit proper quantile functions $Q^\dagger_N$ also converge in L^2 a.s. to a random proper quantile function $Q^\dagger_\infty$. This guarantees the existence of the QMP, which is precisely the distribution of $Q^\dagger_\infty$. Under additional smoothness assumptions on $Q^\dagger_n$ and the bandwidth sequence ρ_N, we can then leverage the Sobolev space to show that the support of the QMP is on proper quantile functions which are absolutely continuous on $(0, 1)$.

4.1.2 QMP over probability measures and asymptotic exchangeability

So far, we have only been working in the quantile space, but it is interesting to study the QMP on the more familiar space of probability measures. This will also allow us to make statements on the convergence of limiting probability distributions as studied in Berti et al. (2004) and Fong et al. (2023). Fortunately, we can leverage a simple connection between L^2 convergence of quantile functions and weak convergence.

Proposition 4 Under Assumptions 1 and 3, there exists a random probability measure P_∞ on $\mathbb{R}$ such that $P_N \to P_\infty$ in Wasserstein-2 distance a.s., which further implies $P_N \to P_\infty$ weakly a.s. Under the additional Assumptions 2 and 4, P_∞ corresponds to an absolutely continuous $Q^\dagger_\infty$.

Proof. As $d_2(Q^\dagger_N, Q^\dagger_\infty)$ is exactly the Wasserstein-2 distance between P_N and P_∞, where P_∞ is computed from Q_∞, we have that $P_N \to P_\infty$ in Wasserstein-2 distance a.s. As the Wasserstein distance metrizes weak convergence in $\mathbb{R}$ (e.g. Villani, 2009, Theorem 6.8), we have the above result. □

As we have a.s. weak convergence of P_N to a random probability measure, we can make the usual statements on limiting empirical distributions and asymptotic exchangeability.

Corollary 2 Under Assumptions 1 and 3, the sequence $(Y_{n+1}, Y_{n+2}, \ldots)$ arising from Algorithm 4 is asymptotically exchangeable. Furthermore, the empirical distribution of $(Y_{n+1}, Y_{n+2}, \ldots, Y_N)$ converges weakly to P_∞ a.s. as $N \to \infty$.

Both results in the above corollary arise due to the a.s. weak convergence of P_N to some P_∞, which implies that this convergence of the predictive distribution is sufficient for Bayesian inference (Cui & Walker, 2024a; Fong et al., 2023). The QMP distribution over any functional is then simply the push-forward of $\theta(P_\infty)$. The c.i.d. condition is a very convenient means to attain this convergence, but the above two results highlight that it is by no means necessary. Unlike in

Fong et al. (2023), which relies on the c.i.d. condition, here we instead rely on a martingale condition on the potentially non-monotonic quantile estimate. A keen reader may notice that we have not assured absolute continuity on the probability measure P_∞, which would then imply the existence of a probability density function. Unfortunately the absolute continuity and non-strict monotonicity of $Q^\dagger_\infty$ is not enough to guarantee this, as any flat regions of $Q^\dagger_\infty$ could be mapped to an atom for P_∞. However, absolute continuity of $Q^\dagger_\infty$ allows us to guarantee that P_∞ does not have any gaps in its support, and in practice we also see that P_∞ is continuous a.s.

4.1.3 Gaussian process

Having established the existence of Q_∞ which is distributed according to the QMP, a natural question is to investigate the properties of $Q_\infty - Q_N$ as we take $N \to \infty$ in Algorithm 4. This is closely related to the study carried out in Fortini and Petrone (2020), Fortini and Petrone (2023) and Fortini and Petrone (2025), but we will require some technical tools from empirical process theory as we would like to study the entire function Q_∞. One surprising consequence of the theory to come is the simplicity of the law of $Q_\infty - Q_N$, which allows us to accelerate sampling from the QMP even further. We now introduce the results before discussing their implications.

For the rest of this section, we will assume that α_N takes the form given in Assumption 3. To begin, we first discuss the object of study. We will focus on quantifying the convergence of Q_N to Q_∞, as this is much more tractable than the rearranged case. Specifically, we are interested in the law of the *random* function $S_N = Q_\infty - Q_{N-1}$ as $N \to \infty$, which we suspect to be Gaussian due to the summative form of (7). More concretely, let us define the random function

$$S_N(u) = Q_\infty(u) - Q_{N-1}(u) = \sum_{i=N}^{\infty} \alpha_i\big(u - H_{\rho_i}(u, V_i)\big) \tag{10}$$

where $V_i \overset{\text{iid}}{\sim} \mathcal{U}(0, 1)$ for all $i \geq N$. We highlight to the reader again that S_N has an additive form and in particular consists of a sum of independent terms. As an aside, one concern may be that the distribution of S_N does not depend on observed data (through $Q^\dagger_n$). However, we can quell these concerns by drawing a connection to the Bayesian bootstrap, where the random Dirichlet weights $w_{1:n}$ do not depend on the data at all, but the *location* of observations contribute to the posterior. In the QMP case, S_N plays the role of the Dirichlet weights, and the initial function $Q^\dagger_n$ plays the role of the observations' locations.

This independent form of S_N is in fact a strength of the QMP compared to the traditional MP, as it allows us to much more easily leverage central limit theorems for the sum of independent functions. Armed with this, we can study the convergence of the whole function, which depends on technical empirical process theory that we defer to Section A.3 in the online supplementary material. In particular, the independent form of $S_N(u)$ allows us to easily verify an asymptotic tightness condition and marginal convergence to a Gaussian distribution using the Lindeberg-Feller central limit theorem (CLT), which gives the following result.

Theorem 2 Under Assumptions 3 and 4, the function $\sqrt{N}S_N$ converges weakly in $\ell^\infty((0, 1))$ to $\mathbb{G}_a$, where $\mathbb{G}_a$ is a zero-mean GP with covariance function $\mathbb{E}[\mathbb{G}_a(u)\,\mathbb{G}_a(u')] = a^2(\min\{u, u'\} - uu')$.

Proof Outline. Asymptotic tightness of $\sqrt{N}S_N$ is shown in Theorem A5 in the online supplementary material. We also show in the online supplementary material that any finite collection of points of $\sqrt{N}S_N(u)$ converges to a Gaussian distribution using the Lindeberg-Feller CLT, which together with asymptotic tightness is sufficient for weak convergence to the GP. □

This covariance function is a^2 times the Brownian bridge covariance, which is unsurprising as this arises in the asymptotics for traditional quantile estimation as well. We conclude this section with a brief discussion of the implications of the above, and postpone a detailed demonstration for Section 5. Following Fortini and Petrone (2020), we note that the above gives us a measure of contraction of Q_N to Q_∞, which is quantified by $\sqrt{N}$ term pre-multiplying S_N. More interesting for us however, is the ability to approximate Algorithm 4 with the above GP, which we dedicate

Section 5.3 to. A remaining question is whether the rearranged $Q_\infty^\dagger$ satisfies a similar result. Our conjecture is that it may hold, but it is challenging to extend the proof due to an issue of the centring function. Nonetheless, as we are primarily interested in posterior sampling, we can still utilize the asymptotic normality to sample Q_∞ which then gives the implied $Q_\infty^\dagger$.

4.2 Frequentist asymptotics

We now address the frequentist properties of QMP, which requires a different set of technical tools, but relies on similar recursive arguments such as martingale theory. We will shortly see that posterior consistency and contraction rates can be shown for the QMP, where the $L^2((0, 1))$ Hilbert space and rearrangement theory aid us greatly. The proofs depend critically on the consistency and the convergence rate of the initial $Q_n^\dagger$. However, the latter properties depend on somewhat more technical tools from the stochastic approximation literature. We hope to distinguish this in the discussion below.

To begin, we introduce the setup which is slightly different to the previous subsection. Let $Y_{1:n} \overset{\text{iid}}{\sim} P^*$ where P^* has the corresponding quantile function Q^*, and we consider the case as $n \to \infty$. Following the discussion in Section 3.4, we study the frequentist properties of the QMP obtained through applying Algorithm 3 to the i.i.d. observations $Y_{1:n}$ to obtain the initial $Q_n^\dagger$, followed by predictive resampling with Algorithm 4 in order to obtain $Q_\infty^\dagger$. The QMP is then the distribution of $Q_\infty^\dagger$ conditional on $Y_{1:n}$.

4.2.1 Posterior consistency

Posterior consistency is a crucial property of a Bayesian model which in our context states that the posterior distribution concentrates on the true Q^* from which the data is i.i.d. This is much stronger than Doob's consistency theorem, which only holds a.s. with respect to the prior and is closely connected to the previously discussed predictive asymptotics. Posterior consistency usually hinges on the Kullback-Leibler (KL) property of the prior distribution (Ghosal & Van der Vaart, 2017, Chapter 6), which states that the prior allocates non-zero mass to a KL ball around the truth. Within the martingale posterior context, no such prior distribution exists, so we must develop novel tools for posterior consistency. Fong et al. (2023) showed consistency of the posterior mean of the MP, but did not make any statements on the entire posterior distribution. We will now show this for the QMP case, which requires the following conditions.

Assumption 5 ((Lipschitz quantile function)). Assume that P^* has a quantile function Q^* which is M-Lipschitz continuous on $[0, 1]$, where M is a constant. Furthermore, Q_0 is chosen to be Lipschitz continuous.

A sufficient condition for the above is that P^* has compact support, and P^* is continuously differentiable with strictly positive derivative on its support. This is relatively standard in quantile estimation, e.g. Van der Vaart (2000, Lemma 21.4). Note that Assumption 5 implies the support of P^* is bounded. While this assumption can likely be weakened, it drastically simplifies the proof of consistency, and is likely necessary for the contraction rate result to follow. We now present the consistency result for the initial estimate $Q_n^\dagger$.

Theorem 3 Under Assumptions 3, 4 and 5, we have that $d_2(Q_n^\dagger, Q^*) \to 0$ a.s.$[P^*]$ under Algorithm 3.

Proof Outline. The proof has similar components to the proofs of consistency in Hahn et al. (2018) and Fong et al. (2023), but require additional tools specialized to quantile functions and rearrangement. We show that $d_2(Q_n^\dagger, Q^*)$ is an *almost supermartingale* in the sense of Robbins and Siegmund (1971). The bandwidth condition ensures that (9) approaches a variant of the step update (5). The condition $\sum \alpha_n^2 < \infty$ prevent the errors from accumulating so $d_2(Q_n^\dagger, Q^*)$ converges a.s. The Lipschitz assumption on Q^* and $\sum \alpha_n = \infty$ guarantee that the distance converges to 0 a.s. We also highlight that the rearrangement inequality in Proposition 2 is crucial in handling the rearrangement step after updating with each data point. □

Let us now write $Q^{\dagger}_{n\infty}$ as the random function obtained from Algorithm 4 starting at $Q^{\dagger}_{n}$ for each n, where the additional index n on $Q^{\dagger}_{n\infty}$ is to indicate the dependence on the initial $Q^{\dagger}_{n}$. A novel contribution of our work is that consistency of $Q^{\dagger}_{n}$ can be used to show consistency of the entire QMP, which follows from an application of Markov's inequality and Proposition 2.

Theorem 4 Under Assumptions 3, 4 and 5, for any $\varepsilon > 0$, the QMP from Algorithms 3 and 4 satisfies

$$\Pi_n\left(Q^{\dagger}_{n\infty} : d_2\left(Q^{\dagger}_{n\infty}, Q^*\right) \geq \varepsilon \mid Y_{1:n}\right) \to 0 \quad \text{a.s.}[P^*]$$

Proof Outline. We follow a similar approach to Example 8.5 from Ghosal and Van der Vaart (2017). As $d_2^2(Q^{\dagger}_{n\infty}, Q^*) \leq d_2^2(Q_{n\infty}, Q^*)$ from Proposition 2, we have from Markov's inequality that

$$\Pi_n\left(Q^{\dagger}_{n\infty} : d_2\left(Q^{\dagger}_{n\infty}, Q^*\right) \geq \varepsilon \mid Y_{1:n}\right) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left[d_2^2\left(Q_{n\infty}, Q^*\right) \mid Y_{1:n}\right].$$

We decompose $d_2^2(Q_{n\infty}, Q^*)$ into a posterior variance component $\mathbb{E}[d_2^2(Q_{n\infty}, Q^{\dagger}_{n}) \mid Y_{1:n}]$, a point estimate component $d_2(Q^{\dagger}_{n}, Q^*)$ and a cross-term. The posterior variance is sent to 0 by the sequence α_N, so posterior consistency depends only on consistency of $Q^{\dagger}_{n}$, which is guaranteed by Theorem 3. □

Once again, the connections of the L^2 distance between quantile functions and the Wasserstein metric suggest that the QMP over P_∞ is consistent at P^* in the Wasserstein metric; posterior asymptotics in this metric space has also been studied by Chae et al. (2021).

4.2.2 *Posterior contraction rate*

A more challenging but informative result is the posterior contraction rate, which quantifies how quickly the QMP concentrates on the true Q^*. Once again, we will rely on the convergence rate of $Q^{\dagger}_{n}$ to the truth, but we have only managed to show results for quite stringent additional assumptions, given below.

Assumption 6 ((Lipschitz quantile functions, learning rate and bandwidth)). Suppose Assumption 5 holds, and additionally that $\alpha_i = a(i+1)^{-1}$ for $a > M/2$ and the bandwidth satisfies $\rho_i = \sqrt{1 - ci^{-k}}$ for $k > 4$ and $c \in (0, 1)$.

Theorem 5 Under Assumption 6, we have that for any $0 < \delta < 1$, Algorithm 3 satisfies

$$n^{\delta} d_2^2\left(Q^{\dagger}_{n}, Q^*\right) \to 0 \quad \text{a.s.}[P^*]$$

Proof Outline. The proof follows a similar argument to Aboubacar and Thiam (2014), where we extend the consistency proof to show that $n^{\delta} d_2^2(Q^{\dagger}_{n}, Q^*)$ is an almost supermartingale.

Theorem 6 Under Assumption 6, the sequence $\varepsilon_n = n^{-\delta/2}$ for any $0 < \delta < 1$ is a valid posterior contraction rate for the QMP from Algorithms 3 and 4, that is for any finite $K > 0$, we have

$$\Pi_n\left(Q^{\dagger}_{n\infty} : d_2\left(Q^{\dagger}_{n\infty}, Q^*\right) \geq K\varepsilon_n \mid Y_{1:n}\right) \to 0 \quad \text{a.s.}[P^*]$$

Proof Outline. The proof continues from that of Theorem 4. The posterior variance is $O(n^{-1})$ due to the sequence α_n, so we just require the convergence rate of $Q^{\dagger}_{n}$ as provided by Theorem 5. □

Although it is encouraging that obtaining a posterior contraction rate is possible for the QMP, the assumptions on a and k are not conducive for good performance in practice, as we will see in Section 5. In particular, we require $k < 1$ for smoothness, and M can be very large if P^* has light tails, greatly inflating posterior variance. As a result, we do not suggest the usage of the above Theorem 6 for hyperparameter setting. We believe it likely that the condition on k is an artefact of the proof, and suspect it may be relaxed. However, it is possible that the assumption on $a > M/2$ is necessary. One potential solution is to only consider the posterior contraction rate of the quantile function on a subset of $(0, 1)$, which can decrease the required lower bound on a. Another potential remedy for this impractical setting of a is to use a functional learning rate $a(u)$ based on a density estimate, which we discuss in Section 8.1.

5 Hyperparameters and approximate posterior sampling

Although the theory just introduced is technical, we now shed light on the practical utility of the above theory and its extensions for practical selection of hyperparameters and approximate posterior sampling.

5.1 Learning rate

The sequence α_i is extremely important for both consistency of $Q_n^\dagger$ and the amount of uncertainty obtained when predictive resampling. As a reminder, we will let $\alpha_i = a(i+1)^{-1}$, as in Assumption 3, where $a \in (0, \infty)$ which we will refer to as the learning rate. The above sequence clearly satisfies (6) due the rate of $\alpha_i \to 0$. The choice of the learning rate a however requires much care, as it directly controls the magnitude of the posterior uncertainty. Perhaps surprisingly, a default choice for a can be justified by considering the asymptotic posterior variance of a low-dimensional functional of the QMP, which we now discuss. This works well in practice across a general range of settings.

Consider the mean of P_∞, which can be written as $\mu_\infty = \int_0^1 Q_\infty^\dagger(u)du = \int_0^1 Q_\infty(u)du$, where the last equality can be seen from the integral preserving property of increasing rearrangement (Lemma A1 in the online supplementary material). This allows us to work with Q_∞ directly instead of $Q_\infty^\dagger$, where the latter is much more challenging due to its non-linearity. The following proposition quantifies the posterior mean and asymptotic variance of μ_∞.

Proposition 5 Let $\mu_n = \int Q_n^\dagger(u)du$ for $\{Q_n^\dagger\}_{n\geq 1}$ from Algorithm 3, and $\mu_{n\infty} = \int_0^1 Q_{n\infty}^\dagger(u)du$ where $Q_{n\infty}^\dagger$ arises from Algorithm 4 starting from $Q_n^\dagger$. Under Assumptions 3 and 4, we have $E[\mu_{n\infty} \mid Y_{1:n}] = \mu_n$ for each $n \geq 1$, and $n\,\mathbb{E}[(\mu_{n\infty} - \mu_n)^2 \mid Y_{1:n}] \to a^2/12$ a.s.$[P^*]$.

If μ_n is the sample mean for $Y_{1:n} \overset{\text{iid}}{\sim} P^*$ with mean μ^*, then it has the asymptotic variance σ^2/n, where σ^2 is the variance of $Y \sim P^*$. A natural matching of the asymptotic variance of the QMP to the sample mean then involves setting $a = \sqrt{12}\sigma$. Although we cannot guarantee that $Q_n^\dagger$ gives an efficient estimate of μ_n, this serves as a simple default choice which works well in practice, and one can also regard our suggestion as a lower bound on a. In this case, we can actually show asymptotic normality by leveraging Theorem 2, which we defer to Section C.1 of the online supplementary material.

One potential weakness of the QMP is that a is only a scalar, so we have to choose a single low-dimensional functional for which we want to match asymptotic variances. Nonetheless, our above suggestion based on the mean functional works well in practice. As discussed after Theorem 6, the above issue can also be potentially alleviated with a functional learning rate $a(u)$ which we discuss in Section 8.1, but requires a separate density estimate. More generally, the learning rate for martingale posteriors remains an important open topic of research.

5.2 Bandwidth sequence

The bandwidth sequence ρ_i governs the smoothness of the update, and there are two competing effects. First, we would like $\rho_i \to 1$ so $H_{\rho_i}(u, v)$ approaches an indicator function, which is required for frequentist consistency in the L^2 norm (Theorem 4). This is akin to the condition required for kernel density estimation, and we see in practice that having $\rho_i \to 1$ also improves the initial $Q_n^\dagger$. On the other hand, Theorem 1 assures us that posterior samples of the quantile function from the QMP are only absolutely continuous if ρ_i does not approach 1 too quickly.

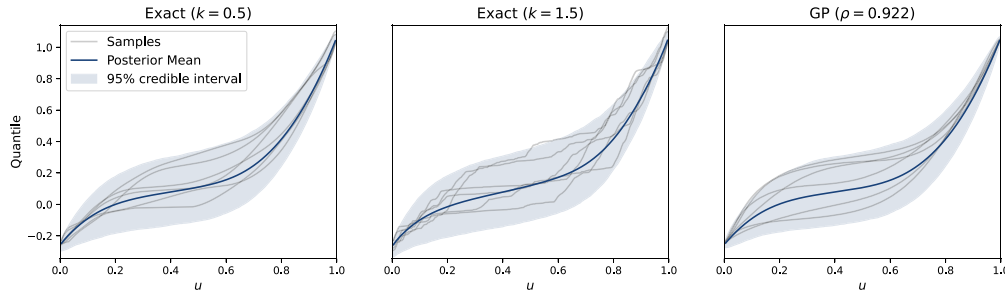

Figure 3. Posterior samples, mean and 95% credible intervals of $Q^\dagger$ for (Left) $k = 0.5$; (Middle) $k = 1.5$; (Right) GP approximation; all plots are with initial $Q_n(u) = 4(u - 0.4)^3 + 0.2u$, $n = 10$, $N = n + 5{,}000$, $a \approx 0.95$ and $c = 0.5$; generating $B = 5{,}000$ exact and approximate samples required 15 s and 0.2 s respectively.

Furthermore, under the assumption of $Q^*(u)$ being differentiable, if ρ_i approaches 1 too quickly, then the weak derivatives $q_n^\dagger$ do not approximate the derivative of q^* well. A slower convergence of $\rho_i \to 1$ also results in fewer violations of monotonicity when applying Algorithm 3. The importance of Assumption 4 for Corollary 1 is illustrated in Figure 3 (left, middle), where we see that QMP samples of $Q_N^\dagger$ are smooth for $k = 0.5$, but non-smooth for $k = 1.5$.

Our suggestion is thus to set the bandwidth sequence as $\rho_i = \sqrt{1 - ci^{-k}}$ as in Assumption 4, where $c \in (0, 1)$ and $k \in (0, 1)$ are two hyperparameters. This form arises naturally from the proofs of Theorems 1 and 4. Although both theorems are satisfied for any $k \in (0, 1)$, we find the choice of $k = 0.5$ to work well in practice which balances between smoothness of the QMP and attaining L^2 consistency. We then suggest setting the constant $c \in (0, 1)$ in a data-adaptive manner, which allows fine-tuning of the smoothness of the initial $Q_n^\dagger$ to the specific dataset. As we have $\rho_1 = \sqrt{1 - c}$ and $\rho_n = \sqrt{1 - cn^{-0.5}}$, the constant c controls the initial value ρ_1 which increases monotonically to $\sqrt{1 - cn^{-0.5}}$ as $i \to n$.

To choose c, we suggest maximizing the prequential log score due to its connections to the marginal likelihood (Dawid, 1984; Fong & Holmes, 2020; Gneiting & Raftery, 2007). In particular, the prequential log score is easy to compute in our setting, as we have $\sum_{i=1}^n \log [p_{i-1}(Y_i)] = -\sum_{i=1}^n \log [q_{i-1}^\dagger(P_{i-1}(Y_i))]$ where $q_i^\dagger$ is the weak derivative of $Q_i^\dagger$. The existence of $q_i^\dagger$ is guaranteed by the absolute continuity of $Q_i^\dagger$ and Theorem A2 in the online supplementary material. The choice of the above is justified as we should rely on $q_i^\dagger$ in some way to set c, as relying on $Q_i^\dagger$ alone (e.g. with the L^2 norm) will not guarantee smooth estimates. We can compute $q_i^\dagger$ easily with finite differences, and $P_{i-1}(Y_i)$ is already computed for our update.

5.3 Approximate posterior sampling

This subsection is dedicated to utilizing Theorem 2 in order to drastically accelerate quantile predictive resampling. For $S_n(u) = Q_\infty(u) - Q_n(u)$, we essentially have that $a^{-1}\sqrt{n+1}S_n \overset{d}{\approx} \mathbb{G}$ for sufficiently large n, where $\mathbb{G} \sim \mathcal{GP}(0, (\min\{u, u'\} - uu'))$ is the Brownian bridge. Unlike in the case of Fortini and Petrone (2020) and Fortini and Petrone (2023) and the regular MP, the distribution of $\mathbb{G}$ does not depend on any random quantities, which arises from working with the quantile instead of the distribution, and allows easier sampling. Furthermore, we only require realizations of Q_∞ to lie in $L^2((0, 1))$ or $H^1((0, 1))$, which is much simpler than needing realizations to be valid probability measures as in the regular MP case. It thus seems reasonable to approximate sampling Q_∞ with $\widetilde{Q}_\infty = Q_n^\dagger + a\,\mathbb{G}/\sqrt{n+1}$. Algorithmically, this involves drawing a sample from a Brownian bridge, then scaling it by $a/\sqrt{n}$ and adding it to the initial $Q_n^\dagger$. The immediate downside to this approach is that samples of $\widetilde{Q}_\infty$ will not be smooth (i.e. in $H^1((0, 1))$) even if Q_∞ is from Theorem 1, due to the a.s. nowhere differentiability of paths from a Brownian bridge.

To remedy this, we propose the following alternative approximation:

$$\widetilde{Q}_\infty = Q_n^\dagger + a\mathbb{G}_{\rho_{n+1}}/\sqrt{n+1},$$

where $\mathbb{G}_\rho$ is a zero-mean GP with covariance function $k_\rho(u, u') = C_{\rho^2}(u, u') - uu'$ and $C_\rho(u, u')$ is the bivariate normal copula. To justify this above choice, we have the following theorem.

Theorem 7 Let $S_n = Q_\infty - Q_n$ and let $\widetilde{S}_n = a\,\mathbb{G}_{\rho_{n+1}}/\sqrt{n+1}$ be the approximation as defined above, and suppose Assumptions 3 and 4 hold true. The covariance function of S_n, which we write as $k_n(u, u') := \mathbb{E}[S_n(u)\,S_n(u')]$, satisfies the following for all $u, u' \in (0, 1)$ and $n \geq 1$:

$$k_{\rho_{n+1}}(u, u') \leq r_n^{-1} k_n(u, u') \leq \min\{u, u'\} - uu',$$

where $r_n = \sum_{i=n+1}^{\infty} \alpha_i^2 \approx a^2(n+1)^{-1}$, and both $k_{\rho_{n+1}}(u, u')$ and $r_n^{-1}k_n(u, u')$ converge to $\min\{u, u'\} - uu'$ as $n \to \infty$. Furthermore, realizations of $\widetilde{S}_n$ lie in $H^1((0, 1))$ a.s., and $a^{-1}\sqrt{n}\,\widetilde{S}_n$ converges weakly in $\ell^\infty((0, 1))$ to the Brownian bridge $\mathbb{G}$.

From the above, we have that $\widetilde{Q}_\infty$ and Q_∞ have the same distribution asymptotically when suitably normalized, which happens as $\rho_n \to 1$. Furthermore, realizations of $\widetilde{Q}_\infty$ (and thus $\widetilde{Q}^\dagger_\infty$) lie in the same Sobolev space a.s. This occurs as the true covariance function k_n lies in between $k_{\rho_{n+1}}$ and that of the Brownian motion in terms of smoothness, where we prefer $k_{\rho_{n+1}}$ to k_n as the former is much cheaper to compute. The above theorem thus justifies the choice of $\widetilde{Q}_\infty$ as a suitable approximation to Q_∞. The above inequality actually suggests that sample paths of $\widetilde{Q}_\infty$ may be slightly smoother than that of Q_∞. In practice, this effect disappears quickly with increasing n as the inequality is very tight even for moderate n.

This approximate sampling scheme is given in Algorithm 5, where drawing from the GP is very cheap and detailed in Section D.2 of the online supplementary material. In practice, this approximation works extremely well, as we illustrate in Figure 3 (right). Both samples and credible intervals of $\widetilde{Q}^\dagger_\infty$ are visually very similar to $Q^\dagger_N$ even for $n = 10$. Furthermore, generating $B = 5{,}000$ posterior samples required 15s and 0.2s for the exact and approximate case respectively, which indicates a substantial speedup. We will see further demonstration of the computational gains and similar results in later in the illustrations.

6 Quantile regression

Having established the framework and theory for the QMP, we now introduce the QMP in the quantile regression setting, which is a natural extension. This is in contrast to the usual intricacies involved in specifying nonparametric prior distributions with covariate dependence. We will focus on the linear case, and leave discussion of potential directions for the non-linear case to Section 8.2.

To begin, we assume that $\{Y_i, X_i\}_{i=1,\ldots,n} \overset{\text{iid}}{\sim} P^*(y, x)$, where $Y \in \mathbb{R}$ and $X \in \mathcal{X} \subset \mathbb{R}^p$. The conditional distribution $P^*(y \mid x)$ is assumed to have a quantile function which varies linearly, that is $Q^*(u \mid x) = \beta^*(u)^T x$, where $\beta^*(u) : (0, 1) \to \mathbb{R}^p$ is the true unknown coefficients. As we can write $\beta^*(u) = \arg\min_\beta \int \rho_u(y - \beta^T x) dP^*(y, x)$, this immediately suggests a quantile regression version of (7):

$$\beta_{n+1}(u) = \beta_n(u) + \alpha_{n+1}\big[u - H_{\rho_{n+1}}\big(u, P_n(Y_{n+1} \mid X_{n+1})\big)\big]X_{n+1}, \tag{11}$$

where $P_n(y \mid x) = \int_0^1 1(Q_n(u \mid x) \leq y) du$. We now utilize the above for the QMP for quantile regression.

6.1 Quantile predictive resampling

The predictive resampling scheme for the quantile regression setting is a straightforward extension of Section 2.3. The key extra ingredient is that we will use the empirical distribution for predictive resampling $X_{n+1:\infty}$, which is equivalent to the Bayesian bootstrap as suggested in Fong et al. (2023). This is particularly natural in our setting, where we are mainly interested in $P^*(y \mid x)$ or $Q^*(u \mid x)$. Quantile predictive resampling then consists of first drawing $X_{N+1} \sim \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$, then simulating $V_{N+1} \sim \mathcal{U}(0, 1)$ and computing

Algorithm 5 Approximate QMP sampling with GPs

Initialize $Q_n^\dagger$ from Algorithm 3; set $\rho_{n+1} = \sqrt{1 - c(n+1)^{-k}}$
for $b \leftarrow 1$ **to** B **do**
  Draw $S^{(b)} \sim \mathcal{GP}(0, C_{\rho_{n+1}^2}(u, u') - uu')$
  Compute $\widetilde{Q}_\infty^{(b)} = Q_n^\dagger + a\, S^{(b)}/\sqrt{n+1}$
end
Return $\{\widetilde{Q}_\infty^{\dagger\,(1)}, \ldots, \widetilde{Q}_\infty^{\dagger\,(B)}\}$

$$\beta_{N+1}(u) = \beta_N(u) + \alpha_{N+1}\big[u - H_{\rho_{N+1}}(u, V_{N+1})\big]X_{N+1}, \tag{12}$$

which again is a martingale. The simple uniform r.v. again arises as $P_N(Y_{N+1} \mid X_{N+1}) \sim \mathcal{U}(0, 1)$ if $Y_{N+1} \sim P_N(\cdot \mid X_{N+1})$. For the covariates, it will be simpler computationally to draw $w_{1:n} \sim \mathrm{Dir}(1, \ldots, 1)$, followed by $X_{n+1:N} \overset{\text{iid}}{\sim} \sum_{i=1}^n w_i \delta_{X_i}$. To draw actual samples of the observations Y_{N+1} given X_{N+1}, we can analogously compute $Y_{N+1} = \beta_{N+1}(V_{N+1})^T X_{N+1}$ which is straightforward. For the initial estimate, we can once again just apply the update (11) on the i.i.d. observables $Y_{1:n}$. Our implicit quantile function $Q_i^\dagger(u \mid x)$ is then the increasing rearrangement of $\beta_i(u)^T x$ at each value of x.

6.2 Predictive asymptotics

6.2.1 Martingale

It is not too hard to verify that we once again have a pointwise martingale condition, that is

$$\mathbb{E}\big[\beta_{N+1}(u) \mid Y_{1:N}, X_{1:N}\big] = \beta_N(u),$$

which we can see by first computing the conditional expectation of Y_{N+1} given X_{N+1}, which returns $\beta_N(u)$, so the additional expectation over X_{N+1} does not affect this. Looking at (12), we see that each component $\beta_{Nj}(u)$ for $j \in \{1, \ldots, p\}$ is a function-valued martingale as before, with the additional term due to the covariates. It is thus not too difficult to show the following.

Theorem 8 Under Assumptions 3, 4, online supplementary material, A1 and A2, there exists a random vector function $\beta_\infty(u)$ with realizations in $H^1((0, 1))^p$ such that $\beta_N(u)$ satisfies $d_{1,2}(\beta_{Nj}, \beta_{\infty j}) \to 0$ for each component $j \in \{1, \ldots, p\}$ a.s. Furthermore, each component of the realizations of $\beta_\infty(u)$ is absolutely continuous on $(0, 1)$ a.s.

Proof Outline. For each dimension $j \in \{1, \ldots, p\}$, we can apply the same derivation as in Theorem 1, with the key difference that the update term for $\beta_{Nj}(u)$ is scaled by X_{Nj}. □

In order to study the result of the rearrangement process, we now study the conditional quantile function estimate directly, which satisfies $Q_N(u \mid x) = \beta_N(u)^T x$ for each $x \in \mathcal{X}$. The implicit conditional quantile function is then the increasing rearrangement of $Q_N(u \mid x)$ for each x, which we write as $Q_N^\dagger(u \mid x)$.

Proposition 6 Under Assumptions 3, 4, online supplementary material, A1 and A2, for each $x \in \mathcal{X}$, there exists a random function $Q_\infty^\dagger(u \mid x)$ with realizations in $H^1((0, 1))$ such that $d_2(Q_N^\dagger(\cdot \mid x), Q_\infty^\dagger(\cdot \mid x)) \to 0$ a.s., where $Q_N^\dagger(u \mid x)$ and $Q_\infty^\dagger(u \mid x)$ are proper monotonically increasing quantile functions. Furthermore, $Q_\infty^\dagger(u \mid x)$ is the increasing rearrangement of $Q_\infty(u \mid x) = \beta_\infty(u)^T x$ a.s.

Proof Outline. Since $Q_N(u \mid x)$ is just a weighted sum of $\beta_{Nj}(u)$, which are elements in a Banach space, the continuous mapping theorem can be used to show $Q_N(u \mid x) \to \beta_\infty(u)^T x$ a.s. The rearrangement step is then analogous to Theorem 1. □

We remark that once again, since $Q^\dagger_\infty(u \mid x)$ is in $H^1((0, 1))$, it can be identified almost everywhere with an absolutely continuous conditional quantile function. One could also make similar statements on the weak convergence of the conditional distributions.

An interesting phenomenon due to the nonlinearity of rearrangement is that even if $Q_N(u \mid x)$ is linear in x, the rearranged $Q^\dagger_N(u \mid x)$ may no longer be so. We do not view this as problematic, as $\beta_N(u)$ is simply a vehicle for drawing conditional samples of $Y \mid x$, and is not a parameter in the traditional sense. This is analogous to $Q_N(u)$ being a vehicle for drawing samples of Y, where $Q_N(u)$ need not be monotonic. If one were interested in the posterior of the parameters of linear quantile regression, one could compute these as functionals of the imputed $\{Y_{n+1:\infty}, X_{n+1:\infty}\}$; see Buja et al. (2019) for a similiar discussion in linear regression. Furthermore, Proposition 2 guarantees us that $Q^\dagger_N(u \mid x)$ will always be closer to $Q^*(u \mid x)$ in L^2 compared to $Q_N(u \mid x)$, so the induced nonlinearity cannot impede the quality of prediction. Interestingly, we can still say something about the QMP over the regression function $\mathbb{E}[Y \mid X]$. Let us define $\mathbb{E}_\infty[Y \mid x] := \int_0^1 Q^\dagger_\infty(u \mid x)du$, so realizations of $\mathbb{E}_\infty[Y \mid x]$ are samples of the regression function from the QMP. We then have the below, which follows from the equimeasurable property of rearrangement.

Proposition 7 Under Assumptions 3, 4, online supplementary material, A1 and A2, the QMP has support over linear regression functions, that is realizations of $\mathbb{E}_\infty[Y \mid x]$ are linear functions of x a.s.

6.2.2 Gaussian process

We can again study the asymptotic normality, this time focusing on the vector $\beta_n(u)$. Consider the difference

$$S_N(u, j) = \sum_{i=N}^{\infty} \alpha_i \big[u - H_{\rho_i}(u, V_i)\big] X_{ij}$$

for $j \in \{1, \ldots, p\}$ and $u \in (0, 1)$, where X_{ij} is the j-th entry of X_i. All of the results in this subsection will be conditional on the Bayesian bootstrap weights $w_{1:n}$ and $X_{1:n}$. Similar to the non-regression case, we can use the Cramér-Wold device to help us study the joint convergence of S_N for an arbitrary finite collection of points. Combining the above with asymptotic tightness, we can again extend the finite-dimensional joint convergence to uniform convergence with respect to $\mathcal{F} = (0, 1) \times \{1, \ldots, p\}$.

Theorem 9 Under Assumptions 3, 4, online supplementary material, A1 and A2, conditional on $w_{1:n}$, $\sqrt{N}S_N$ converges weakly in $\ell^\infty(\mathcal{F})$ to $\mathbb{G}_a$ almost surely, where $\mathbb{G}_a$ is a zero-mean GP with covariance function $\mathbb{E}[\mathbb{G}_a(u, j), \mathbb{G}_a(u', j')] = a^2[\sum_{k=1}^n w_k X_{kj} X_{kj'}](\min\{u, u'\} - uu')$.

We are then free to replace the covariance function of the limiting GP with $C_{\rho^2_{n+1}}(u, u') - uu'$ for approximate sampling as before, giving the covariance function

$$k_{\rho_{n+1}}(\{u, j\}, \{u', j'\}; w_{1:n}) = \left[\sum_{k=1}^n w_k X_{kj} X_{kj'}\right]\left(C_{\rho^2_{n+1}}(u, u') - uu'\right). \tag{13}$$

6.3 Frequentist asymptotics

In the quantile regression setting, the frequentist asymptotics of the QMP is unfortunately more challenging. The main challenge is that the rearrangement $Q^\dagger_n(u \mid x)$ does not preserve linearity of the rearranged conditional quantile, so we do not necessarily have a corresponding vector

$\beta_n^\dagger(u)$. As a result, we cannot use an analogous rearranged update like in Section 3.4. We are however able to show an analogous posterior consistency result in the case where $\rho = 1$, which we detail in Section C.2 of the online supplementary material, as this special case lends itself more easily to a consistent estimate. However, this does not extend easily to the $\rho \neq 1$ case. Nonetheless, (12) works well in practice, and for sufficiently slow rate of $\rho_i \to 1$, we find that $Q_n^\dagger(u \mid x) = Q_n(u \mid x)$ anyways. We thus conjecture that it will also satisfy posterior consistency, and we leave this for future work.

6.4 Practical considerations

In the quantile regression case, the same considerations as Section 5 can be made, where the added complications are that we also need to handle the random covariates.

6.4.1 Approximate posterior sampling

As outlined in Section 6.3, a rearranged version of the update is not obvious, so we opt for Algorithm 6 to estimate the initial β_n. In the interest of space, we jump straight to the approximate sampling procedure in Algorithm 7, with the exact case in Algorithm A1 of the online supplementary material. Once again, the GP approximation is extremely expedient, and drawing from a GP with kernel (13) is covered in Section D.2 of the online supplementary material.

6.4.2 Hyperparameters

The quantile regression case has the same hyperparameters, i.e. the learning rate a and the bandwidth sequence ρ_i. Fortunately, the bandwidth sequence works exactly as before, where we set the value of c according to $\sum_{i=1}^n p_{i-1}(Y_i \mid X_i)$ which can be computed analogously. We thus turn our focus on the learning rate. Once again, we can consider the asymptotic posterior variance of a low-dimensional functional. In this case, we can look at the marginal posterior mean and asymptotic covariance matrix on the linear regression coefficients, $\bar{\beta}_\infty = \int \beta_\infty(u) du$.

Algorithm 6 Estimation of quantile regression coefficients

Initialize β_0
Data is $(Y_1, X_1), \ldots, (Y_n, X_n)$
for $i \leftarrow 1$ **to** n **do**
  Compute $V_i = P_{i-1}(Y_i \mid X_i)$
  $\beta_i(u) = \beta_{i-1}(u) + \alpha_i[u - H_{\rho_i}(u, V_i)]X_i$
end
Return β_n

Algorithm 7 Approximate QMP Sampling for Quantile Regression with GPs

Initialize β_n from Algorithm 6
Compute $\rho_{n+1} = \sqrt{1 - c(n+1)^{-k}}$
for $b \leftarrow 1$ **to** B **do**
  Draw $w_{1:n}^{(b)} \sim \text{Dirichlet}(1, \ldots, 1)$
  Draw $S_{1:p}^{(b)} \sim \mathcal{GP}(0, k_{\rho_{n+1}}(\{u, j\}, \{u', j'\}; w_{1:n}^{(b)}))$
  Compute $\widetilde{\beta}_\infty^{(b)} = \beta_n + a\, S_{1:p}^{(b)} / \sqrt{n}$
end
Return $\{\widetilde{\beta}_\infty^{(1)}, \ldots, \widetilde{\beta}_\infty^{(B)}\}$

Proposition 8 For $n \geq 1$, let $\bar{\beta}_n := \int \beta_n(u)du$ for $\{\beta_n\}_{n\geq 1}$ arising from online supplementary material, Algorithm A1, and suppose that $X_{1:n} \overset{\text{iid}}{\sim} P^*(x)$ with $\Sigma_x = \mathbb{E}[X_i X_i^T]$. Let $\bar{\beta}_{n\infty} = \int_0^1 \beta_{n\infty}(u)du$ where $\beta_{n\infty}$ arises from online supplementary material, Algorithm A1 starting from β_n. Under Assumptions 3, 4 and online supplementary material, A3, we then have $\mathbb{E}[\bar{\beta}_{n\infty} \mid Y_{1:n}] = \bar{\beta}_n$ for each $n \geq 1$, and

$$n\mathbb{E}\left[(\bar{\beta}_{n\infty} - \bar{\beta}_n)(\bar{\beta}_{n\infty} - \bar{\beta}_n)^T \mid Y_{1:n}, X_{1:n}\right] \to (a^2/12)\Sigma_x \quad \text{a.s.}[P^*].$$

We assume the covariates and response are standardized, so the intercept is 0 for simplicity, and online supplementary material, Assumption A3 ensures Σ_x is non-singular. The asymptotic covariance matrix of the least squares estimate of $\hat{\beta}_n$ in linear regression is $\sigma^2\Sigma_x^{-1}/n$, where σ^2 is the variance of the residuals from the linear model. We can once again attempt a matching of asymptotic covariances, but matching the entire covariance matrix is not possible with a scalar a. Instead, we can match the determinants of the covariance matrices, which can be interpreted as matching the generalized variance (Wilks, 1932). This then gives the setting $a = \sqrt{12}\sigma(\det \Sigma_x)^{-1/p}$, where we can estimate σ and Σ_x from the data. This default choice appropriately inflates the posterior variance in the presence of highly correlated covariates and as the dimension of x increases, and works well in practice. Another potential approach to setting a, which we do not investigate here, is to match the traces of the covariance matrices. Analogous to the unconditional case, we can also adopt a u-specific and dimension-specific learning rate, $a_j(u) = a_j a(u)$, at the cost of having to depend on a separate density estimate of the residuals. We provide a brief discussion in Section E.5 of the online supplementary material, but leave a detailed investigation for future work.

7 Illustrations

We now illustrate the QMP for quantile estimation and regression on a simulation and real dataset respectively. All methods are implemented in JAX (Bradbury et al., 2018) in Python, and executed on an Apple M2 Pro CPU. Due to the parallel nature of the QMP, significant acceleration is possible on a GPU, but we use a CPU to illustrate the speed-up attained by the GP approximation. For all examples, we truncate exact predictive resampling at $N = n + 5{,}000$, when n is the size of the observed dataset; convergence diagnostics for justifying this choice can be found in Section D.3 of the online supplementary material.

7.1 Simulations

In this section, we demonstrate the method and practical performance for unconditional quantile estimation under different sample sizes, as well as comparing the computation time of exact and approximate sampling schemes. Let $Y_{1:n} \overset{\text{iid}}{\sim} P^*$, where P^* has the associated quantile function $Q^*(u) = 4(u - 0.4)^3 + 0.2u$. We consider two sample sizes, $n = 50$ and $n = 500$, and compare the QMP distributions. For estimation, we initialize with $Q_0(u) = y_{\min} + (y_{\max} - y_{\min})u$, which implies a uniform distribution over the range of the observations, and is appropriate here as we know the range of y is bounded. We average over 10 permutations of the data to compute $Q_n^\dagger$. We follow the guidance of Sections 5.1 and 5.2, and set c by maximizing the prequential log score (also averaged over 10 permutations) on a grid of $c \in (0, 1)$ values of size 20. For both exact and approximate predictive resampling, we draw $B = 5000$ independent posterior samples. Convergence diagnostics for exact predictive resampling in this example can be found in Section E.3 of the online supplementary material. For all examples, we compute the quantile function estimates on a uniform grid on [0, 1] of size 200.

The selection of c and estimation of $Q_n^\dagger$ for $n = 50$ and $n = 500$ required 0.7 s and 1.4 s respectively, where c is chosen to be 0.6 and 0.75 respectively. We highlight that tuning c can be easily parallelized if desired. In both sample sizes, exact predictive resampling required 15 seconds, whereas approximate predictive resampling with the GP only required 0.15 s, which is a very significant speed-up. In Figure 4, we plot the QMP mean and 95% credible intervals for $Q_\infty^\dagger(u)$ and $\theta = \mathbb{E}[Y]$ for the two simulated sample sizes, with the empirical quantile estimate and true Q^* for reference. As the exact and approximate QMP are visually indistinguishable, we only plot the

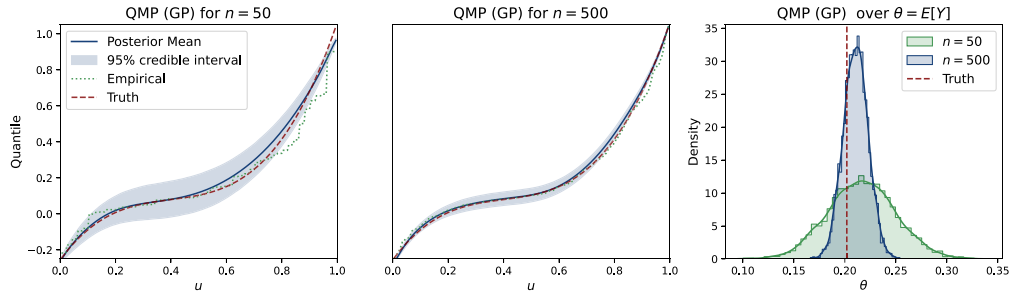

Figure 4. QMP over $Q_\infty^\dagger$ for (Left) $n = 50$; (Middle) $n = 500$; (Right) QMP over $\theta = \mathbb{E}[Y]$; we only show the GP approximation as it is visually indistinguishable from exact sampling.

latter in the interest of space in the main paper, with the exact QMP in Section E.3 of the online supplementary material. We can see that the posterior mean is monotonic and smooth, and is regularized towards the initial linear Q_0 compared to the empirical quantile estimate. As n increases, the posterior mean approaches the truth, and the credible intervals shrink and capture the truth for central values of u but seem to be anticonservative for values of u close to 0 or 1. As addressed by Proposition 5, the learning rate a is chosen based on the asymptotic variance for the mean functional, which manifests as conservative and anticonservative credible intervals for the central and tail quantiles respectively. This is an inherent limitation of the scalar learning rate, and we discuss a potential extension on the QMP to address this in Section 8.1. We see in the Figure 4 (right) that the posterior distribution for θ concentrates as n increases. We highlight that the posterior on θ appears to be biased upwards, which occurs due to Q_0 regularizing the QMP in a manner similar to a prior. As Q_0 implies a uniform distribution on the support, the initial estimate of $\theta = \mathbb{E}[Y]$ before seeing any data is approximately 0.66, which is the centre of the support. Finally, we provide a further simulation example in Section E.3.2 of the online supplementary material where P^* no longer has compact support. In this case, the results are similar but the tail effect is more noticeable and the choice of Q_0 is more important.

7.2 Cyclone dataset

Following Tokdar and Kadane (2012) and An and MacEachern (2024), we now demonstrate the QMP for quantile regression in a real dataset based on a tropical cyclone intensity dataset from Elsner et al. (2008). The dataset[1] consists of $n = 2097$ tropical cyclones and their respective lifetime maximum wind speeds from the years 1981–2006. Covariates include the year, basin, latitude, and age of the cyclone; see the Supplementary Information of Elsner et al. (2008) for more details. Both Tokdar and Kadane (2012) and An and MacEachern (2024) studied a subset of tropical cyclones in the North Atlantic (NA) basin ($n = 291$), with the year as the single covariate, and identified an increasing trend.

For the QMP, we initialize Q_0 by setting $\beta_{0j}(u) = 0$ for $j \in \{1, \ldots, p\}$ and only set the intercept term $\beta_{00}(u)$ to be non-zero, which corresponds to initializing $Q_0(u \mid x) = Q_0(u)$. We set $\beta_{00}(u)$ to be the line interpolating the lower and upper quartile of y, which will reduce the impact of outliers on Q_0 compared to using the whole range of y. For both data sizes, we average over 10 permutations, but this could be reduced for large n as there is less sensitivity to data ordering. Once again, we choose $c \in (0, 1)$ by maximizing the prequential log score on a grid of size 20, and estimate $\beta(u)$ on a grid on $[0, 1]$ of size 200. We standardize all covariates and the response, and rescale after estimation. For the results, we again only present the GP approximation, as the posterior samples are visually indistinguishable from the exact sampler; this comparison is provided in Section E.4 of the online supplementary material.

As a main benchmark, we compare to the dependent quantile pyramids (DQP) method of An and MacEachern (2024), and utilize the author's MCMC implementation in `C++`. We choose the DQP as our main comparator as it is closest to the QMP in terms of flexibility and generality

[1] https://myweb.fsu.edu/jelsner/temp/Data.html

compared to other Bayesian methods, producing a full joint distribution over $\{Q(u \mid x), u \in (0, 1), x \in \mathcal{X}\}$ and allowing for complex dependence structure. An and MacEachern (2024) also demonstrate that the DQP performed the most consistently across a wide range of empirical settings compared to Bayesian competitors. We also compare to non-Bayesian linear quantile regression with the `quantreg` package (Koenker et al., 2018) for each u independently followed by increasing rearrangement, as the the QMP can be regarded as a smooth recursive version of this estimator. For reference, we further provide nonlinear quantile regression fits for the full dataset with $n = 2{,}097$ using random forests (QRF) (Meinshausen & Ridgeway, 2006) and neural networks (QRNN) (Taylor, 2000), using the `quantile-forest` (Johnson, 2024) and `qrnn` (Cannon, 2018) packages respectively.

We first analyze the subset of tropical cyclones within the NA basin ($n = 291$) with a single covariate and the lifetime maximum wind speed as the response. MCMC for the DQP required 26 min to generate $B = 10{,}000$ posterior samples after a burn-in period of 10,000 samples. This long burn-in is necessary due to slow mixing, and note that the posterior samples may still be highly correlated without thinning. The QMP required 1s for tuning the hyperparameter ($c = 0.95$) and estimating $Q_n^\dagger$. A further 33 s for exact predictive resampling or 0.4 s for approximate predictive resampling was needed to generate $B = 10{,}000$ independent QMP samples. Not only is this orders of magnitude faster than MCMC, both exact and approximate predictive resampling are inherently parallelizable, and can be efficiently accelerated using GPUs if desired (Fong et al., 2023). Furthermore, the samples produced are independent and convergence concerns are minor, unlike MCMC where mixing is always a concern. The effect of truncating at a final N is relatively harmless for exact predictive resampling, and the approximate sampler with the GP does not even require truncation. This highlights the scalability of the QMP due to not relying on MCMC.

Figure 5 (left) illustrates the posterior means and 95% credible intervals of $\beta_{\infty,1}(u)$ corresponding to the year, in comparison to that from the DQP. In general, we caution against interpreting $\beta_{1\infty}(u)$ directly, as we generally do not have $Q_\infty^\dagger(u \mid x) = \beta_\infty(u)^T x$. However, in this specific case, the smoothness of the update resulted in no rearrangement being required for both $Q_n(u \mid x)$ and $Q_\infty(u \mid x)$. We see here that the exact and approximate QMP are numerically indistinguishable, so again it seems that n is already sufficiently large for the GP approximation to hold. In comparison to the DQP, we see that the QMP has wider credible intervals within the centre but narrower in the tails, and the QMP posterior mean is also more regularized towards the initial $\beta_{01}(u) = 0$ than the DQP. Figure 5 (right) illustrates the exact and approximate QMP over the linear regression coefficient $\bar{\beta}_\infty$, where again the exact and approximate QMP are very similar.

Figure 6 (left) illustrates the posterior mean and 95% credible intervals for the conditional quantile functions for the earliest and latest year, i.e. $Q_\infty^\dagger(u \mid x = 1981)$ and $Q_\infty^\dagger(u \mid x = 2016)$. As mentioned earlier, in this specific case, no rearrangement was necessary as the updates are sufficiently smooth. The conditional quantiles are smooth and monotonic, and again deviate from the quantile regression estimate for values of u near 1, as it is regularized more towards the linear $Q_0(u)$. We see that there is an increasing trend in maximum speed with year, with a more significant difference for small and large values of u. Figure 6 (right) illustrates the posterior means and 95% credible intervals of quantile regression curves $Q_\infty^\dagger(u = u^* \mid x)$ at $u^* \in \{0.25, 0.50, 0.75\}$, which in this case are linear and non-crossing. Again, we see that there is an increasing linear trend which is larger for values of u near 1.

We now study the full data set ($n = 2{,}097,\ p = 3$) with the year, latitude and cyclone age as covariates, where we exclude the basin indicator due to strong collinearity with latitude. We do not compute the DQP posterior due to the computational expense. The QMP required 5.2 s for tuning the hyperparameter ($c = 0.95$) and estimating $Q_n^\dagger$, which can be accelerated if fewer data permutations are used. Exact and approximate predictive resampling then required a further 42 s and 0.7 s respectively, where again we only display results for the approximate QMP as they are visually indistinguishable from the exact sampler. In this case, the cyclone age is the most significant predictor of maximum speed. Figure 7 (left) shows the posterior mean and 95% credible intervals for $Q(u = u^* \mid x)$ for $u^* \in \{0.25, 0.50, 0.75\}$, where we fix the year and latitude at the respective sample means and only vary age. We see that the credible intervals are tighter as n is much larger, and again the QMP agrees with the pointwise QR.

Figure 7 (right) illustrates the equivalent fits using the QRF and QRNN. The QRNN appears to produce relatively linear quantile regression curves, as does the QRF for Age < 50. However, the QRF plateaus for Age ≥ 50, which likely occurs as only 4% of the cyclones have age larger than

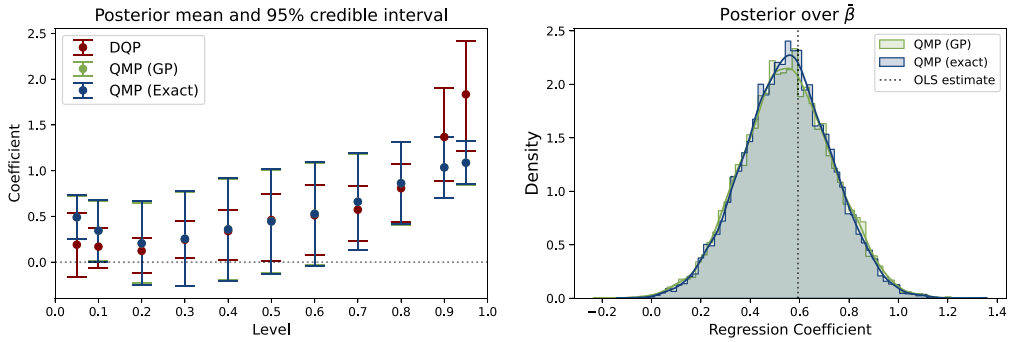

Figure 5. Tropical cyclone maximum speeds in the NA basin ($n = 291$): (Left) Posterior mean and 95% credible intervals for $\beta_{1\infty}(u)$ from the exact and approximate QMP and DQP; (Right) Posterior distribution of $\bar{\beta}_\infty$ for the exact and approximate QMP.

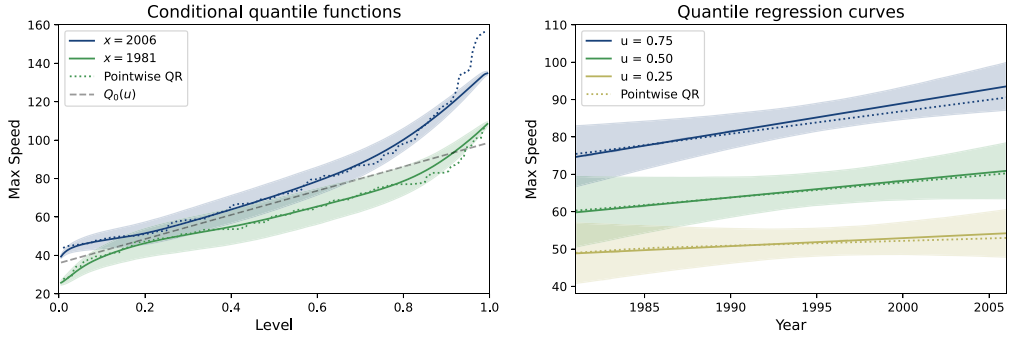

Figure 6. Tropical cyclone maximum speeds in the NA basin ($n = 291$): (Left) Posterior mean and 95% credible intervals for $Q(u \mid x = 1981)$ and $Q(u \mid x = 2006)$ from the approximate QMP; (Right) Posterior mean and 95% credible intervals for $Q^{\dagger}_{\infty}(u = u^* \mid x)$ for $u^* \in \{0.25, 0.50, 0.75\}$ from the approximate QMP.

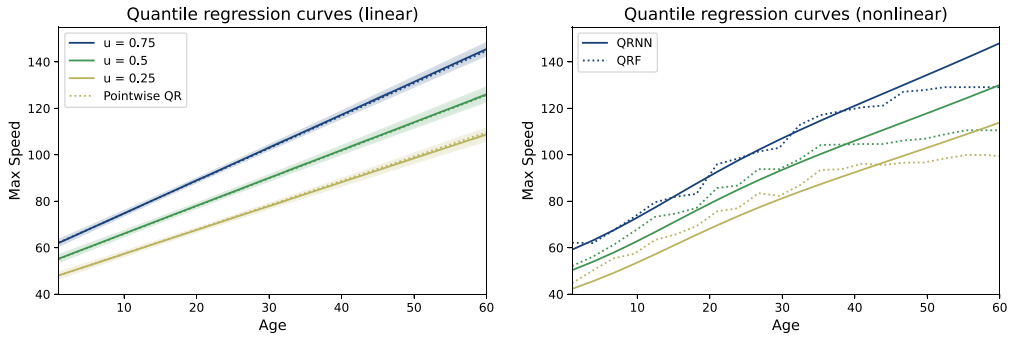

Figure 7. Tropical cyclone maximum speeds globally ($n = 2097$): (Left) Posterior mean and 95% credible intervals for $Q^{\dagger}_{\infty}(u = u^* \mid x)$ for $u^* \in \{0.25, 0.50, 0.75\}$ from the QMP and pointwise QR; (Right) Estimates of $Q^{\dagger}_{\infty}(u = u^* \mid x)$ for $u^* \in \{0.25, 0.50, 0.75\}$ from the QRNN and QRF.

50, so the random forest may be extrapolating poorly. We demonstrate in Section E.4 of the online supplementary material however that the nonlinear methods tend to slightly outperform the linear methods in terms of average quantile loss on cross-validated held-out data, indicating the presence of slight nonlinearity. In addition to increased computation time, a key disadvantage of the nonlinear methods is the difficulty in obtaining uncertainty estimates; we consider the extension of the QMP to utilize nonlinear methods in Section 8.2. Finally, in Section E.4 of the online supplementary material, we include additional results on the smaller dataset $n = 291$, where we find that the linear methods tend to outperform the nonlinear methods in this smaller sample.

8 Discussion and extensions

In this paper, we introduce the quantile martingale posterior (QMP), which is a method for non-parametric Bayesian quantile estimation/regression based on a solely predictive framework, where we focus on the smooth case. Model specification only requires an estimate of the (conditional) quantile function, which does not need to be monotonic, as we rely on increasing rearrangement which naturally arises from predictive resampling. One main advantage of the QMP compared to the traditional Bayesian approach is that we no longer need to specify a likelihood or a prior distribution, which is complex in the quantile estimation/regression case. Another key advantage is computational cost—we can carry out exact posterior sampling without MCMC, where we are orders of magnitude faster and free of convergence challenges. By relying on an asymptotic Gaussian process approximation of the QMP, we can accelerate posterior sampling even further. However, this gain in flexibility of model specification and computational speed comes at a cost of being less 'automatic' than traditional Bayesian inference. Significant effort is needed to show the existence, support and consistency/contraction rate of the QMP, and there are still some gaps in the theory for the regression case. Furthermore, careful specification of the learning rate and bandwidth sequence are needed to achieve good results, which is a limitation of the recursive approach.

The QMP also has some unique strengths compared to the original MP, as the space of quantile function estimates is easier to work with than the space of probability measures. This nicer function space allows us to more straightforwardly quantify posterior support, consistency and contraction. We are also able to easily incorporate covariate dependence in a structured manner, e.g. using linear models, which is challenging for the original MP. However, a main limitation of the QMP relative to the original MP is the restriction to univariate data on a compact support, which is not needed for the original MP.

We now discuss some potential future directions to alleviate some of the limitations of the QMP.

8.1 Functional learning rates

Throughout the paper, we hinted at the inherent limitation of a scalar learning rate a, resulting in sub-optimal estimation of the quantile function near $u = 0$ and $u = 1$, as well as the need to inflate posterior uncertainty for central values of u to compensate for anticonservative uncertainty in the tails. A potential extension of the QMP to tackle this limitation is to introduce a functional learning rate $a(u)$ which depends on u, allowing for a slower and faster learning rate in the centre and tails respectively. In Section E.5 of the online supplementary material, we show that under some assumptions on $a(u)$, this does not affect posterior consistency. We also conjecture that attaining a posterior contraction rate of n^{-1} can be attained under more reasonable hyperparameter settings, but leave this for future work. To guide the setting of $a(u)$, we note that the asymptotic variance of the empirical quantile estimate is equal to $u(1-u)q^*(u)^2$ (Van der Vaart, 2000), where $q^*(u) = 1/p^*(Q^*(u))$ is the quantile density function. This hints at an appropriate choice of $a(u) = q^*(u)$, which is also suggested in Aboubacar and Thiam (2014). One downside of this approach is the need to separately estimate a density function, which is somewhat unsatisfying from a coherence point of view. Furthermore, the posterior uncertainty of the QMP will be very sensitive to the tails of the estimated density, as posterior variance will be proportional to the reciprocal of the density, and the tails are difficult to estimate. Finally, while beyond the scope of this work, we highlight that the setting of the functional learning rate $a(u)$ is closely connected to asymptotic frequentist validity of the resulting posterior credible intervals. Similar to Yang et al. (2016), one can interpret the learning rate $a(u)$ as a quantile-specific correction term in order to match the pointwise asymptotic posterior variance to that of the frequentist estimator. In the online supplementary material, we also explore an example where we estimate p^* using a kernel density estimate, but leave a proper investigation for future work.

8.2 Other generative predictives

While we motivate the choice of the generative predictives Q_N using quantile estimation, it would be interesting to investigate whether other recursive updates for the generative sampler that preserve the martingale exist. One future direction is to utilize score functions corresponding to other losses to update Q_N; score functions for predictive Bayes are considered in Cui and Walker (2025) and Fortini and Petrone (2025). Another direction is whether one can leverage the dual

formulation of quantile estimation/regression (Gibbs et al., 2025; Gutenbrunner & Jurecková, 1992) to design new recursive updates for Q_N.

In the quantile regression setting, a natural extension to the QMP is to leverage nonlinear estimators from machine learning, such as random forests or neural networks as outlined in Section 7.2. This would involve replacing X_{n+1} in (11) with the gradient of a nonlinear function estimator, and would allow us to quantify uncertainty which is often a challenge with nonlinear models.

8.3 Multivariate data

In this paper, we focused on the case where y is univariate and the conditional quantiles are linear in x. However, the predictive asymptotics naturally extend to the multivariate case, as demonstrated in Theorems 8 and 9. As a result, we expect a multivariate extension of the QMP to retain the same theoretical advantages compared to the MP, as we can work in a larger space of (vector-valued) functions instead of being constrained to probability density functions.

In the multivariate setting, the main challenge is then to design a meaningful recursive update, as multivariate generalizations of the quantile function are non-trivial. However, utilizing flexible deep generative models for the QMP could offer an exciting avenue of future work. For example, generative adversarial networks (Goodfellow et al., 2020) draw high-dimensional samples from complex distributions by passing simple random variables through complex nonlinear functions, in a fashion similar to our quantile predictive resampling scheme. Extensions to increasing rearrangement within the multivariate case may also be of interest, e.g. as studied in Carlier et al. (2016) and Rosenberg et al. (2022).

Acknowledgments

We thank Hyoin An for providing the code for the DQP method which we used for our experiments. We also thank the Joint Editor, the Associate Editor and two anonymous reviewers for their feedback, which has helped substantially improve the paper.

Conflicts of interest: None declared.

Funding

AY received funding from Novo Nordisk during part of this work. **EF** receives funding from the Research Grants Council of Hong Kong through the the Early Career Scheme (Grant No. 27304424) and the General Research Fund (Grant No. 17306925).

Data availability

All datasets used in this paper are publicly available. Code for reproducing the simulation results in the paper can be found at https://github.com/edfong/qmp.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Aboubacar A., & Thiam B. (2014). A smoothing stochastic algorithm for quantile estimation. *Statistics & Probability Letters*, *93*(4), 116–125. https://doi.org/10.1016/j.spl.2014.06.016

Almgren Jr F. J., & Lieb E. H. (1989). Symmetric decreasing rearrangement is sometimes continuous. *Journal of the American Mathematical Society*, *2*(4), 683–773. https://doi.org/10.1090/jams/1989-02-04

An H., & MacEachern S. N. (2024). A process of dependent quantile pyramids. *Journal of Nonparametric Statistics*, 1–25. https://doi.org/10.1080/10485252.2024.2305811

Bassett Jr G., & Koenker R. (1982). An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, *77*, 407–415. https://doi.org/10.1080/01621459.1982.10477826

Berger J. O., Bernardo J. M., & Dongchu S. (2009). The formal definition of reference priors. *Annals of Statistics*, *37*(2), 905–938. https://doi.org/10.1214/07-AOS587

Berti P., Pratelli L., & Rigo P. (2004). Limit theorems for a class of identically distributed random variables. *Annals of Probability*, *32*(3A), 2029–2052. https://doi.org/10.1214/009117904000000676

Berti P., Dreassi E., Pratelli L., & Rigo P. (2020). A class of models for Bayesian predictive inference. *Bernoulli*, *27*(1), 702–726. https://doi.org/10.3150/20-BEJ1255

Bissiri P. G., Holmes C. C., & Walker S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B*, *78*(5), 1103–1130. https://doi.org/10.1111/rssb.12158

Bradbury J., Frostig R., Hawkins P., Johnson M. J., Leary C., Maclaurin D., & Wanderman-Milne S. (2018). JAX: Composable transformations of Python+NumPy programs. http://github.com/google/jax.

Buja A., Brown L., Berk R., George E., Pitkin E., Traskin M., Zhang K., & Zhao L. (2019). Models as approximations I. *Statistical Science*, *34*(4), 523–544. https://doi.org/10.1214/18-STS693

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. https://doi.org/10.1007/s00477-018-1573-6

Carlier G., Chernozhukov V., & Galichon A. (2016). Vector quantile regression: An optimal transport approach. *Annals of Statistics*, *44*(3), 1165–1192. https://doi.org/10.1214/15-AOS1401

Chae M., De Blasi P., & Walker S. G. (2021). Posterior asymptotics in Wasserstein metrics on the real line. *Electronic Journal of Statistics*, *15*(2), 3635–3677. https://doi.org/10.1214/21-EJS1869

Chakraborty M., & Ghosal S. (2021). Coverage of credible intervals in nonparametric monotone regression. *Annals of Statistics*, *49*(2), 1011–1028. https://doi.org/10.1214/20-AOS1989

Chen L., Keilbar G., & Wu W. B. (2023). Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, *24*(91), 1–25.

Chen X., & Tokdar S. T. (2021). Joint quantile regression for spatial data. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *83*(4), 826–852. https://doi.org/10.1111/rssb.12467

Chernozhukov V., Fernandez-Val I., & Galichon A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, *96*(3), 559–575. https://doi.org/10.1093/biomet/asp030

Chernozhukov V., Fernández-Val I., & Galichon A. (2010). Quantile and probability curves without crossing. *Econometrica: Journal of the Econometric Society*, *78*(3), 1093–1125. https://doi.org/10.3982/ECTA7880

Coron J. (1984). The continuity of the rearrangement in $W^{1,p}(\mathbb{R})$. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, *11*(1), 57–85.

Cui F., & Walker S. G. (2024a). A Bayesian bootstrap for mixture models. *Bayesian Analysis*, *1*(1), 1–28. https://doi.org/10.1214/24-BA1498

Cui F., & Walker S. G. (2024b). 'Martingale posterior distributions for log-concave density functions', arXiv, arXiv:2401.14515, preprint: not peer reviewed.

Cui F., & Walker S. G. (2025). 'Martingale posteriors from score functions', arXiv, arXiv:2501.01890, preprint: not peer reviewed.

Dawid A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, *147*(2), 278–290. https://doi.org/10.2307/2981683

Doob J. L. (1949). Application of the theory of martingales. In *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin–3 Juillet 1948)* (pp. 23–27). Paris CNRS.

Elsner J. B., Kossin J. P., & Jagger T. H. (2008). The increasing intensity of the strongest tropical cyclones. *Nature*, *455*(7209), 92–95. https://doi.org/10.1038/nature07234

Embrechts P., & Hofert M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, *77*(3), 423–432. https://doi.org/10.1007/s00186-013-0436-7

Fong E., & Holmes C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489–496. https://doi.org/10.1093/biomet/asz077

Fong E., Lyddon S., & Holmes C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 1952–1962). PMLR. http://proceedings.mlr.press/v97/fong19a.html.

Fong E., Holmes C., & Walker S. G. (2023). Martingale posterior distributions. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *85*(5), 1357–1391. https://doi.org/10.1093/jrsssb/qkad005

Fortini S., & Petrone S. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. *Journal of the Royal Statistical Society: Series B*, *82*(4), 1087–1114. https://doi.org/10.1111/rssb.12385

Fortini S., & Petrone S. (2023). Prediction-based uncertainty quantification for exchangeable sequences. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220142. https://doi.org/10.1098/rsta.2022.0142

Fortini S., & Petrone S. (2025). Exchangeability, prediction and predictive modeling in Bayesian statistics. *Statistical Science*, *40*(1), 40–67. https://doi.org/10.1214/24-STS965

Ghosal S., & Van der Vaart A. (2017). *Fundamentals of nonparametric Bayesian inference* (Vol. *44*). Cambridge University Press.

Gibbs I., Cherian J. J., & Candès E. J. (2025). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *87*(4), qkaf008. https://doi.org/10.1093/jrsssb/qkaf008

Gneiting T., & Raftery A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437
Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., & Bengio Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622
Gutenbrunner C., & Jurecková J. (1992). Regression rank scores and regression quantiles. *Annals of Statistics*, *20*(1), 305–330. https://doi.org/10.1214/aos/1176348524
Hahn P. R., Martin R., & Walker S. G. (2018). On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, *113*(523), 1085–1093. https://doi.org/10.1080/01621459.2017.1304219
Hardy G. H., Littlewood J. E., & Pólya G. (1952). *Inequalities*. Cambridge University Press.
He X. (1997). Quantile curves without crossing. *American Statistician*, *51*(2), 186–192. https://doi.org/10.1080/00031305.1997.10473959
Hjort N. L., & Walker S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *Annals of Statistics*, *37*, 105–131. https://doi.org/10.1214/07-AOS553
Holmes C. C., & Walker S. G. (2023). Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220143. https://doi.org/10.1098/rsta.2022.0143
Johnson R. A. (2024). Quantile-forest: A python package for quantile regression forests. *Journal of Open Source Software*, *9*(93), 5976. https://doi.org/10.21105/joss
Koenker R., & Bassett Jr G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, *46*(1), 33–50. https://doi.org/10.2307/1913643
Koenker R., Portnoy S., Ng P. T., Zeileis A., Grosjean P., & Ripley B. D. (2018). Package 'quantreg'. *Reference manual available at R-CRAN:* https://cran.rproject.org/web/packages/quantreg/quantreg.pdf.
Kohler M., Krzyżak A., & Walk H. (2014). Nonparametric recursive quantile estimation. *Statistics & Probability Letters*, *93*(4), 102–107. https://doi.org/10.1016/j.spl.2014.06.007
Lai T. L. (2003). Stochastic approximation. *Annals of Statistics*, *31*(2), 391–406. https://doi.org/10.1214/aos/1051027873
Lorentz G. G. (1953). An inequality for rearrangements. *The American Mathematical Monthly*, *60*(3), 176–179. https://doi.org/10.2307/2307574
Meinshausen N., & Ridgeway G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*(6), 983–999.
Nie L., & Rockova V. (2023). Bayesian bootstrap spike-and-slab lasso. *Journal of the American Statistical Association*, *118*(543), 2013–2028. https://doi.org/10.1080/01621459.2022.2025815
Pisier G. (2016). *Martingales in Banach spaces* (Vol. *155*). Cambridge University Press.
Robbins H., & Siegmund D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics* (pp. 233–257). Elsevier.
Rodrigues T., Dortet-Bernadet J.-L., & Fan Y. (2019). Pyramid quantile regression. *Journal of Computational and Graphical Statistics*, *28*(3), 732–746. https://doi.org/10.1080/10618600.2019.1575225
Rosenberg A. A., Vedula S., Romano Y., & Bronstein A. 2022. Fast nonlinear vector quantile regression. In *The Eleventh International Conference on Learning Representations*.
Rubin D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, *9*(1), 130–134. https://doi.org/10.1214/aos/1176345338
Taylor J. W. (2000). A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *Journal of Forecasting*, *19*(4), 299–311. https://doi.org/10.1002/(ISSN)1099-131X
Tokdar S. T., & Kadane J. B. (2012). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis*, *7*(1), 51–72. https://doi.org/10.1214/12-BA702
Van der Vaart A. W. (2000). *Asymptotic statistics* (Vol. *3*). Cambridge University Press.
Villani C. (2009). *Optimal transport: Old and new* (Vol. *338*). Springer.
Walker S. G. (2022). A new look at Bayesian uncertainty. In *Handbook of statistics* (Vol. *47*, pp. 83–101). Elsevier.
Walker S. G. (2024). Martingale posterior distributions for cumulative hazard functions. *Scandinavian Journal of Statistics*, *51*(3), 936–955. https://doi.org/10.1111/sjos.v51.3
Wilks S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*(3-4), 471–494. https://doi.org/10.1093/biomet/24.3-4.471
Yang Y., & Tokdar S. T. (2017). Joint estimation of quantile planes over arbitrary predictor spaces. *Journal of the American Statistical Association*, *112*(519), 1107–1120. https://doi.org/10.1080/01621459.2016.1192545
Yang Y., Wang H. J., & He X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, *84*(3), 327–344. https://doi.org/10.1111/insr.v84.3
Yu K., & Moyeed R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, *54*(4), 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9