

Chapter 2

Marginal Quantiles: Asymptotics for Functions of Order Statistics

G. Jogesh Babu

*Department of Statistics,
326 Joab L. Thomas Building,
The Pennsylvania State University,
University Park, PA 16802-2111, USA**

Methods for quantile estimation based on massive streaming data are reviewed. Marginal quantiles help in the exploration of massive multivariate data. Asymptotic properties of the joint distribution of marginal sample quantiles of multivariate data are also reviewed. The results include weak convergence to Gaussian random elements. Asymptotics for the mean of functions of order statistics are also presented. Application of the latter result to regression analysis under partial or complete loss of association among the multivariate data is described.

2.1. Introduction

Data depth provides an ordering of all points from the center outward. Contours of depth are often used to reveal the shape and structure of multivariate data set. The depth of a point x in a one-dimensional data set $\{x_1, x_2, \dots, x_n\}$ can be defined as the minimum of the number of data points on one side of x (cf. [10]).

Several multidimensional depth measures $D_n(x; x_1, \dots, x_n)$ for $x \in R^k$ were considered by many that satisfy certain mathematical conditions. If the data is from a spherical or elliptic distribution, the depth contours are generally required to converge to spherical or elliptic shapes. In this paper we concentrate on marginal quantiles. They help in describing percentile contours, which lead to a description of the densities and the multivariate distributions.

This approach is useful in quickly exploring massive datasets that are

*Research supported in part by NSF grant AST-0707833.

becoming more and more common in diverse fields such as Internet traffic, large sky surveys etc. For example, several ongoing sky surveys such as the Two Micron All Sky Survey and the Sloan Digital Sky Survey are providing maps of the sky at infrared and optical wavelengths, respectively generating data sets measured in the tens of Terabytes. These surveys are creating catalogs of objects (stars, galaxies, quasars, etc.) numbering in billions, with up to a hundred measured numbers for each object. Yet, this is just a fore-taste of the much larger datasets to come from surveys such as Large Synoptic Survey Telescope. This great opportunity comes with a commensurate technological challenge: how to optimally store, manage, combine, analyze and explore these vast amounts of complex information, and to do it quickly and efficiently? It is difficult even to compute a median of massive one dimensional data. As the multidimensional case is much more complex, marginal quantiles can be used to study the structure.

In this review article we start with description of estimation methods for quantiles and density for massive streaming data. Then describe asymptotic properties of joint distribution of marginal sample quantiles of multivariate data. We conclude with recent work on asymptotics for the mean of functions of order statistics and their applications to regression analysis under partial or complete loss of association among the multivariate data.

2.1.1. Streaming Data

As described above, massive streaming datasets are becoming more and more common. The data is in the form of a continuous stream with no fixed size. Finding trends in these massive size data is very important. One cannot wait till all the data is in and stored for retrieval for statistical analysis. Even to compute median from a stored billion data points is not feasible. In this case one can think of the data as a streaming data and use low storage methods to continually update the estimate of median and other quantiles ([2] and [7]). Simultaneous estimation of multiple quantiles would aid in density estimation.

Consider the problem of estimation of p -th quantile based on a very large dataset with n points of which a fixed number, say m , points can be placed into memory for sorting and ranking. Initially, each of these points is given a weight and a score based on p . Now a new point from the dataset is put in the array and all the points in the existing array above it will have their ranks increased by 1. The weights and scores are updated for these $m + 1$ points. The point with the largest score will then be dropped from

the array, and the process is repeated. Once all the data points are run through the procedure, the data point with rank closest to np will be taken as an estimate of the p -th quantile. See [7] for the details.

Methods for estimation of several quantiles simultaneously are needed for the density estimation when the data is streaming. The method developed by [8] uses estimated ranks, assigned weights, and a scoring function that determines the most attractive candidate data points for estimates of the quantiles. The method uses a small fixed storage and its computation time is $O(n)$. Simulation studies show that the estimates are as accurate as the sample quantiles.

While the estimated quantiles are useful and informative on their own, it is often more useful to have information about the density as well. The probability density function can give a more intuitive picture of such characteristics as the skewness of the distribution or the number of modes. Any of the many standard curve fitting methods can now be employed to obtain an estimate of the cumulative distribution function.

The procedure is also useful in the approximation of the unknown underlying cumulative distribution function by fitting a cubic spline through the estimates obtained by this extension. The derivative of this spline fit provides an estimate of the probability density function.

The concept of convex hull peeling is useful in developing procedures for median in 2 or more dimensions. The convex hull of a dataset is the minimal convex set of points that contains the entire dataset. The convex hull based multivariate median is obtained by successively peeling outer layers until the dataset cannot be peeled any further. The centroid of the resulting set is taken as the multivariate median. Similarly, a multivariate interquartile range is obtained by successively peeling convex hull surfaces until approximately 50% of the data is contained within a hull. This hull is then taken as the multivariate interquartile range. See [5] for a nice overview. This procedure requires assumptions on the shape of the density. To avoid this one could use joint distribution of marginal quantiles to find the multidimensional structure.

2.2. Marginal Quantiles

Babu & Rao (cf. [3]) derived asymptotic results on marginal quantiles and quantile processes. They also developed tests of significance for population medians based on the joint distribution of marginal sample quantiles. Joint asymptotic distribution of the sample medians was developed by [9]; see

also [6], where they assume the existence of the multivariate density. On the other hand Babu & Rao work with a much weaker assumption, the existence of densities of univariate marginals alone.

2.2.1. Joint Distribution of Marginal Quantiles

Let F denote a k -dimensional distribution function and let F_j denote the j -th marginal distribution function. The quantile functions of the marginals are defined by:

$$F_j^{-1}(u) = \inf\{x : F_j(x) \geq u\}, \text{ for } 0 < u < 1.$$

Thus $F_j^{-1}(u)$ is u -th quantile of the j th marginal.

Let X_1, \dots, X_n be independent random vectors with common distribution F , where $X_i = (X_{i1}, \dots, X_{ik})$. Hence F_j is the distribution of X_{ij} . To get the joint distribution of sample quantiles, let $0 < q_1, \dots, q_k < 1$. Let δ_j denote the density of F_j at $F_j^{-1}(q_j)$ and let $\hat{\theta}_j$ denote the q_j -th sample quantile based on the j -th coordinates X_{1j}, \dots, X_{nj} of the sample. [3] obtained the following theorem.

Theorem 2.1. *Let F_j be twice continuously differentiable in a neighborhood of $F_j^{-1}(q_j)$ and $\delta_j > 0$. Then the asymptotic distribution of*

$$\sqrt{n}(\hat{\theta}_1 - F_1^{-1}(q_1), \dots, \hat{\theta}_k - F_k^{-1}(q_k))$$

is k -variate Gaussian distribution with mean vector zero and variance-covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} q_1(1-q_1)\delta_1^{-2} & \sigma_{12} & \cdots & \sigma_{1k} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \cdots & q_k(1-q_k)\delta_k^{-2} \end{pmatrix},$$

where for $i \neq j$, $\sigma_{ij} = (F_{ij}(F_i^{-1}(q_i), F_j^{-1}(q_j)) - q_i q_j) / (\delta_i \delta_j)$.

The proof uses Bahadur's representation of the sample quantiles (see [4]).

In practice σ_{ij} can be directly estimated using bootstrap method,

$$\widehat{\sigma}_{ij} = E^*(n(\theta_i^* - \hat{\theta}_i)(\theta_j^* - \hat{\theta}_j)),$$

where θ_j^* denotes the bootstrapped marginal sample quantile and E^* denotes the expectation under the bootstrap distribution function. An advantage of the bootstrap procedure is that it avoids density estimation altogether.

2.2.2. Weak Convergence of Quantile Process

We now describe the weak limits of the entire marginal quantile processes. For $(q_1, \dots, q_k) \in (0, 1)^k$, define the sample quantile process,

$$Z_n(q_1, \dots, q_k) = \sqrt{n} \left(\delta_1(\hat{\theta}_1 - F_1^{-1}(q_1)), \dots, \delta_k(\hat{\theta}_k - F_k^{-1}(q_k)) \right).$$

The following theorem is from Section 4 of [3].

Theorem 2.2. *Suppose for $j = 1, \dots, k$, the marginal d.f. F_j is twice differentiable on (a_j, b_j) , where*

$$\begin{aligned} -\infty &\leq a_j = \sup\{x : F_j(x) = 0\} \\ \infty &\geq b_j = \inf\{x : F_j(x) = 1\}. \end{aligned}$$

Further suppose that the first two derivatives F_j' and F_j'' of F_j satisfy the conditions

$$F_j' \neq 0 \text{ on } (a_j, b_j),$$

$$\max_{1 \leq j \leq k} \sup_{a_j < x < b_j} F_j(x)(1 - F_j(x)) \frac{|F_j''(x)|}{(F_j'(x))^2} < \infty,$$

and F_j' is non-decreasing (non-increasing) on an interval to the right of a_j (to the left of b_j). Then $Z_n(q_1, \dots, q_k)$ converges weakly to a Gaussian random element (W_1, \dots, W_k) on $C[0, 1]^k$.

Thus, each marginal of Z_n converges weakly to a Brownian bridge. The covariance of the limiting Gaussian random element is given by

$$E(W_i(t)W_j(s)) = P(F_i(X_{1i}) \leq t, F_j(X_{1j}) \leq s) - ts.$$

2.3. Regression under Lost Association

[11] developed a method of estimation of linear regression coefficients when the association among the paired data is partially or completely lost. He considered the simple linear regression problem

$$Y_i = \alpha + \beta U_i + \epsilon_i,$$

where U_i are independent identically distributed (i.i.d.) with mean μ and standard deviation σ_U , the residual errors ϵ_i are i.i.d with mean zero and

standard deviation σ_ϵ . Further, $\{U_i\}$ and $\{\epsilon_i\}$ are assumed to be independent sequences. If Π_n denotes the set of all permutations of $\{1, \dots, n\}$, then it is natural to find estimators $\hat{\alpha}, \hat{\beta}$ of α, β that minimize

$$h(\alpha, \beta) = \min_{\pi \in \Pi_n} \sum_{i=1}^n (Y_{\pi(i)} - \alpha - \beta U_i)^2.$$

[11] has shown that the permutation that minimizes h is free from α, β .

The main difficulty is the computational complexity. As there are $n!$ permutations, conceivably it requires that many computations. [11] has shown that $\hat{\beta}$ depends only on two permutations. In particular, he has shown that

$$\frac{1}{n} \sum_{i=1}^n U_{(i)} Y_{(i)} \text{ and } \frac{1}{n} \sum_{i=1}^n U_{(i)} Y_{(n-i+1)}$$

appear in the definition of $\hat{\beta}$. Hence the results on their limits are needed to obtain the asymptotics for $\hat{\beta}$. This would aid in the estimation of the bias of $\hat{\beta}$. Further testing of hypothesis or obtaining confidence intervals for $\hat{\beta}$ require limiting distribution of

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n U_{(i)} Y_{(i)}.$$

See the Example 2.2 in the last section.

In the next section we present some work in progress on the strong law of large numbers and central limit theorems for means of general functions of order statistics. These results would aid in establishing

$$\hat{\beta} \xrightarrow{a.e.} \beta_1 = \text{sign}(\beta) \sqrt{\beta^2 + \sigma_\epsilon^2 \sigma_U^{-1}}.$$

2.4. Mean of Functions of Order Statistics

This section is based on the current research by [1]. We present some recent results on strong law of large numbers and the central limit theorem for the means of functions of order statistics. Let \mathbf{X}_i and X_{ij} be as in Section 2.2.1. Let $X_{n:i}^{(j)}$ denote the i -th order statistic of $\{X_{1j}, \dots, X_{nj}\}$. Suppose ϕ is a measurable function on \mathbb{R}^k and the function γ defined by

$$\gamma(u) = \phi(F_1^{-1}(u), \dots, F_k^{-1}(u)), \quad 0 < u < 1,$$

is integrable on $(0, 1)$.

Theorem 2.3. Suppose F_j are continuous, $\phi(F_1^{-1}(u_1), \dots, F_k^{-1}(u_k))$ is continuous in the neighborhood of the diagonal $u_1 = u, \dots, u_k = u, 0 < u < 1$, and for some A and $0 < c_0 < 1/2$,

$$|\phi(F_1^{-1}(u_1), \dots, F_k^{-1}(u_k))| \leq A \left(1 + \sum_{j=1}^k |\gamma(u_j)| \right),$$

whenever $(u_1, \dots, u_k) \in (0, c_0)^k \cup (1 - c_0, 1)^k$. Then

$$\frac{1}{n} \sum_{i=1}^n \phi(X_{n:i}^{(1)}, \dots, X_{n:i}^{(k)}) \xrightarrow{a.e.} \int_0^1 \gamma(y) dy.$$

For example, in the two dimensional case,

$$\phi(F_1^{-1}(u), F_2^{-1}(v)) = \min(u, v)^{-\alpha} (1 - \max(u, v))^{-\alpha}$$

with $0 < \alpha < \frac{1}{2}$ satisfies the conditions of Theorem 2.3.

To establish asymptotic normality, we require

$$\lim_{u \rightarrow 0^+} \sqrt{u} (|\gamma(u)| + |\gamma(1-u)|) = 0,$$

square integrability of partial derivatives ψ_j ,

$$\psi_j(u) = \frac{\partial \phi(F_1^{-1}(u_1), \dots, F_k^{-1}(u_k))}{\partial u_j} \Big|_{(u, \dots, u)},$$

and some smoothness conditions on ψ_j and $\phi(F_1^{-1}(u_1), \dots, F_k^{-1}(u_k))$, in addition to the conditions of Theorem 2.3.

Theorem 2.4. Assume for any pair $(1 \leq j \neq r \leq k)$, the joint distribution $F_{j,r}$ of (X_{ij}, X_{ir}) is continuous. Under regularity assumptions that include the conditions mentioned above, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\phi(X_{n:i}^{(1)}, \dots, X_{n:i}^{(k)}) - \int_0^1 \gamma(y) dy) \xrightarrow{dist} N(0, \sigma^2),$$

where

$$\begin{aligned} \sigma^2 = & 2 \sum_{1 \leq j \neq r \leq k} \int_0^1 \int_0^1 [F_{j,r}(F_j^{-1}(x), F_r^{-1}(y)) - xy] \psi_j(x) \psi_r(y) dx dy \\ & + 2 \sum_{j=1}^k \int_0^1 \int_0^y x(1-y) \psi_j(x) \psi_j(y) dx dy. \end{aligned}$$

Details are in [1].

2.5. Examples

The above results are illustrated with two of examples.

Example 2.1. Let X_i be as in Section 2.2.1. Let the marginals X_{ij} be uniformly distributed and let $\phi(u_1, \dots, u_k) = u_1^{a_1} \dots u_k^{a_k}$, for some $a_j \geq 1$. Then

$$\frac{1}{n} \sum_{i=1}^n (X_{n:i}^{(1)})^{a_1} \dots (X_{n:i}^{(k)})^{a_k} \xrightarrow{a.e.} \frac{1}{a_1 + \dots + a_k + 1},$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left[(X_{n:i}^{(1)})^{a_1} \dots (X_{n:i}^{(k)})^{a_k} - \frac{1}{a_1 + \dots + a_k + 1} \right] \xrightarrow{dist.} N(0, \sigma^2),$$

where

$$\sigma^2 = 2 \sum_{1 \leq j < r \leq k} a_j a_r E(X_{1j} X_{1r})^M + \frac{(2M-3)(M^2-2)}{2M+1} \sum_{j=1}^k a_j^2,$$

and $M = a_1 + \dots + a_k$.

Note that the limit in this example does not depend on the joint distribution of X_{1j} . In particular, if $a_1 = a_2 = 1$, we obtain that both $\frac{1}{n} \sum_{i=1}^n X_{n:i}^{(1)} X_{n:i}^{(2)}$ and $\frac{1}{n} \sum_{i=1}^n (X_{n:i}^{(1)})^2$ converge to the same limit $E(X_{11}^2) = \frac{1}{3}$ a.e.

Example 2.2. (*Regression with lost associations.*) Let $\{(X_i, Y_i), 1 \leq i \leq n\}$ be i.i.d. bivariate normal random vectors with correlation ρ , means μ_1, μ_2 , and standard deviations σ_1, σ_2 . Let the marginal distributions of X_1 and Y_1 be denoted by F and G . Clearly,

$$G^{-1}(F(x)) = \mu_2 + \frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Then

$$\frac{1}{n} \sum_{i=1}^n X_{n:i} Y_{n:i} \xrightarrow{a.e.} \int_0^1 F^{-1}(u) G^{-1}(u) du = E(X_1 G^{-1}(F(X_1)))$$

$$= \mu_1 \mu_2 + \sigma_1 \sigma_2,$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_{n:i} Y_{n:i} - \mu_1 \mu_2 - \sigma_1 \sigma_2) \xrightarrow{dist.} N(0, \sigma^2),$$

where

$$\sigma^2 = \mu_1^2 \sigma_2^2 + \mu_2^2 \sigma_1^2 + (1 + \rho^2) \sigma_1^2 \sigma_2^2 + 2\rho \mu_1 \mu_2 \sigma_1 \sigma_2.$$

Regression under broken samples are considered in Section 2.3, where it is indicated that the regression coefficients depend only on

$$\frac{1}{n} \sum_{i=1}^n X_{n:i} Y_{n:i} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n X_{n:i} Y_{n:(n-i+1)}.$$

Acknowledgment

I would like to thank an anonymous referee for the comments which helped in improving the presentation.

References

- [1] Babu, G. J., Bai, Z. D., Choi, K.-P. and Mangalam, V. (2008). Limit theorems for functions of marginal quantiles. Submitted.
- [2] Babu, G. J. and McDermott, J. P. (2002). Statistical methodology for massive datasets and model selection. In *Astronomical Data Analysis II* (Jean-Luc Starck and Fionn D. Murtagh, Eds.), Proceedings of SPIE. 4847 228–237.
- [3] Babu, G. J. and Rao, C. R. (1988). Joint asymptotic distribution of marginal quantiles and quantile functions in samples from a multivariate population. *J. Multivariate Anal.* 27 15–23.
- [4] Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* 37 577–580.
- [5] Barnett, V. (1981). *Interpreting Multivariate Data*. Wiley, New York.
- [6] Kuan, K. S. and Ali, M. M. (1980). Asymptotic distribution of quantiles from a multivariate distribution. *Multivariate Statistical Analysis* (Proceedings of Conference at Dalhousie University), Halifax, Nova Scotia, 1979, 109–120. North-Holland, Amsterdam.
- [7] Liechty, J. C., Lin, D. K. J. and McDermott, J. P. (2003). Single-pass low-storage arbitrary quantile estimation for massive datasets. *Statistics and Computing.* 13 91–100.
- [8] McDermott, J. P., Babu, G. J., Liechty, J. C. and Lin, D. K. J. (2007). Data skeletons: simultaneous estimation of multiple quantiles for massive streaming datasets with applications to density estimation. *Statistics and Computing.* 17 311–321.
- [9] Mood, A. M. (1941). On the joint distribution of the medians in samples from a multivariate population. *Ann. Math. Statist.* 12 268–278.
- [10] Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, Canadian Mathematical Congress, Montreal, 2 523–531.
- [11] Mangalam, V. (2006). The regression under lost association. Private Communication.