Simultaneous quantiles of several variables (and their role in missing data imputation)

Snigdhansu Chatterjee

School of Statistics, University of Minnesota

Partially joint work with:
N. Mukhopadhyay, Virginia Commonwealth U.

UMN-CS Workshop, August 16, 2011
Suppose $X \in \mathbb{R}$ is a random variable.

For any $\alpha \in (0, 1)$, the $\alpha^{th}$ quantile $Q_{\alpha}$ is the number below which $X$ is observed with probability $\alpha$.

A bit more precisely: $Q_{\alpha} = \inf\{q : P[X \leq q] \geq \alpha\}$.

If $X$ is continuous, there is a one-to-one relationship between $\alpha$ and $Q_{\alpha}$.

We should not use this co-ordinate-wise for multivariate data, since this ignores all dependency patterns and is statistically inferior.
(Univariate quantiles: an alternative view) Recall that the \textit{median} is the unique minimizer of $\mathbb{E}|X - q|$.

(An extension) The $\alpha^{th}$ quantile $Q_\alpha$ is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}$.

(Alternative notation) The $\beta^{th}$ quantile $Q_\beta$ is the unique minimizer of $\mathbb{E}\{|X - q| + \beta(X - q)\}$, for every $\beta \in (-1, 1)$. Identify $\beta = 2\alpha - 1$.

(Chaudhuri’s geometric quantiles) For random vector $X \in \mathbb{R}^p$, for every $u \in B_p = \{x : ||x|| < 1\}$, the $u^{th}$ quantile $Q(u)$ is defined as the minimizer of

$$
\psi_u(q) = \mathbb{E} [||X - q|| + < u, X - q >].
$$
Univariate to multivariate quantiles

- **(Univariate quantiles: an alternative view)** Recall that the median is the unique minimizer of $E|X - q|$.

- **(An extension)** The $\alpha^{th}$ quantile $Q_\alpha$ is the unique minimizer of $E\{|X - q| + (2\alpha - 1)(X - q)\}$.

- **(Alternative notation)** The $\beta^{th}$ quantile $Q_\beta$ is the unique minimizer of $E\{|X - q| + \beta(X - q)\}$, for every $\beta \in (-1, 1)$. Identify $\beta = 2\alpha - 1$.

- **(Chaudhuri’s geometric quantiles)** For random vector $X \in \mathbb{R}^p$, for every $u \in B_p = \{x : ||x|| < 1\}$, the $u^{th}$ quantile $Q(u)$ is defined as the minimizer of

$$\psi_u(q) = E[||X - q|| + < u, X - q >] .$$
(Univariate quantiles: an alternative view) Recall that the median is the unique minimizer of $\mathbb{E}|X - q|$. 

(An extension) The $\alpha^{th}$ quantile $Q_\alpha$ is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)|\}$. 

(Alternative notation) The $\beta^{th}$ quantile $Q_\beta$ is the unique minimizer of $\mathbb{E}\{|X - q| + \beta(X - q)|\}$, for every $\beta \in (-1, 1)$. Identify $\beta = 2\alpha - 1$.

(Chaudhuri's geometric quantiles) For random vector $X \in \mathbb{R}^p$, for every $u \in \mathcal{B}_p = \{x : \|x\| < 1\}$, the $u^{th}$ quantile $Q(u)$ is defined as the minimizer of

$$\psi_u(q) = \mathbb{E}[\|X - q\| + < u, X - q >].$$
Univariate to multivariate quantiles

- **(Univariate quantiles: an alternative view)** Recall that the median is the unique minimizer of $\mathbb{E}|X - q|$.

- **(An extension)** The $\alpha^{th}$ quantile $Q_\alpha$ is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}$.

- **(Alternative notation)** The $\beta^{th}$ quantile $Q_\beta$ is the unique minimizer of $\mathbb{E}\{|X - q| + \beta(X - q)\}$, for every $\beta \in (-1, 1)$. Identify $\beta = 2\alpha - 1$.

- **(Chaudhuri’s geometric quantiles)** For random vector $X \in \mathbb{R}^p$, for every $u \in B_p = \{x : \|x\| < 1\}$, the $u^{th}$ quantile $Q(u)$ is defined as the minimizer of

$$\psi_u(q) = \mathbb{E}[\|X - q\| + < u, X - q >].$$
Multivariate quantiles

Chaudhuri’s geometric quantiles

For every \( u \in \{ x : \|x\| < 1 \} \), \( Q(u) \) minimizes
\[
E [\|X - q\| + < u, X - q >].
\]

- Define \( U = u/\|u\| \) for \( u \neq 0 \). Define \( \beta = \|u\| \), thus \( u = \beta U \).
- Projection of \( X \) in the direction of \( u \) is \( X_U U \), where \( X_U = < X, U > \). The orthogonal projection is \( X_{U\perp} = X - X_U U \).
- For every \( \lambda \in \mathbb{R} \), the generalized spatial quantiles minimize:
\[
E \left[ \|X_U - q_U\| \left[ 1 + \lambda (X_U - q_U)^{-2} \|X_{U\perp} - q_{U\perp}\|^2 \right]^{1/2} + \beta (X_U - q_U) \right].
\]
  For \( \lambda = 1 \) we get Chaudhuri’s quantiles.
  For \( \lambda = 0 \) we get the projection quantile. Computationally simple, no limitations from sample size and dimension, works for infinite-dimensional observations, plenty of good theoretical properties.
Theorem: (loosely worded)

Sample generalized spatial quantiles are consistent, and asymptotically Gaussian with an intractable asymptotic dispersion parameter.

The generalized bootstrap can be used for inference and obtaining all statistical properties of these quantiles. (Bootstrap works great with parallel processing. Excellent theoretical properties.)

Projection quantiles have a one-to-one relationship like univariate quantiles.

Projection quantiles based confidence sets have exact coverage.
I’m the stat-math guy, remember?

- **Theorem**: (loosely worded)

  Sample generalized spatial quantiles are consistent, and asymptotically Gaussian with an intractable asymptotic dispersion parameter.

  The *generalized bootstrap* can be used for inference and obtaining all statistical properties of these quantiles. (Bootstrap works great with parallel processing. Excellent theoretical properties.)

- Projection quantiles have a one-to-one relationship like univariate quantiles.

- Projection quantiles based confidence sets have exact coverage.
I’m the stat-math guy, remember?

- **Theorem**: (loosely worded)
  - Sample generalized spatial quantiles are consistent, and asymptotically Gaussian with an intractable asymptotic dispersion parameter.
  - The *generalized bootstrap* can be used for inference and obtaining all statistical properties of these quantiles. (Bootstrap works great with parallel processing. Excellent theoretical properties.)

- Projection quantiles have a one-to-one relationship like univariate quantiles.
- Projection quantiles based confidence sets have exact coverage.
Theorem: (loosely worded)

Sample generalized spatial quantiles are consistent, and asymptotically Gaussian with an intractable asymptotic dispersion parameter.

The generalized bootstrap can be used for inference and obtaining all statistical properties of these quantiles. (Bootstrap works great with parallel processing. Excellent theoretical properties.)

Projection quantiles have a one-to-one relationship like univariate quantiles.

Projection quantiles based confidence sets have exact coverage.
Theorem: (loosely worded)
Sample generalized spatial quantiles are consistent, and asymptotically Gaussian with an intractable asymptotic dispersion parameter.
The generalized bootstrap can be used for inference and obtaining all statistical properties of these quantiles. (Bootstrap works great with parallel processing. Excellent theoretical properties.)
Projection quantiles have a one-to-one relationship like univariate quantiles.
Projection quantiles based confidence sets have exact coverage.
Example scatter plot

Bivariate Normal

Ansu Chatterjee (U. Minnesota)
Example scatter plot

Normal Mixture

Ansu Chatterjee (U. Minnesota)
A partial list of applications

- Uncertainty quantification in a variety of ways.
- Robust estimation, inference.
- Less restrictive statistical assumptions needed.
- Heteroscedastic, “local” regression. (Quantile regression is extensively used by economists.)
- Nonlinear, non-smooth projections.
- Modeling of extremes.
- Missing data imputation (missing at random, an alternative to multiple imputation with comparable/better features).