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

Snigdhansu Chatterjee

School of Statistics, University of Minnesota

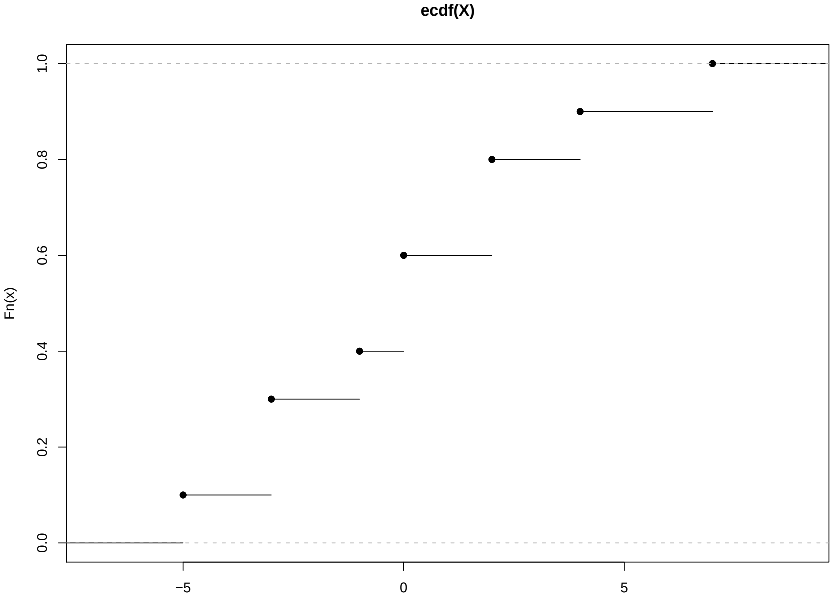
Partially joint work with:
N. Mukhopadhyay, Virginia Commonwealth U.

UMN-CS Workshop, August 16, 2011

Univariate quantiles

- Suppose $X \in \mathbb{R}$ is a random variable.
- For any $\alpha \in (0, 1)$, the α^{th} quantile Q_α is the number below which X is observed with probability α.
- A bit more precisely: $Q_\alpha = \inf\{q : \mathbb{P}[X \leq q] \geq \alpha$.
- If X is continuous, there is a one-to-one relationship between α and Q_α.
- We should not use this co-ordinate-wise for multivariate data, since this ignores all dependency patterns and is statistically inferior.

Plot of cumulative distribution function



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 α^{th} quantile Q_α is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}$.
- **(Alternative notation)** The β^{th} quantile Q_β 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

$$\Psi_u(q) = \mathbb{E}\left[||X - q|| + < u, X - q >\right].$$

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 α^{th} quantile Q_α is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}$.
- **(Alternative notation)** The β^{th} quantile Q_β 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

$$\Psi_u(q) = \mathbb{E}\left[||X - q|| + < u, X - q >\right].$$

- **(Univariate quantiles: an alternative view)** Recall that the *median* is the unique minimizer of $\mathbb{E}|X - q|$.
- **(An extension)** The α^{th} quantile Q_α is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}$.
- **(Alternative notation)** The β^{th} quantile Q_β 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

$$\Psi_u(q) = \mathbb{E}\left[||X - q|| + < u, X - q >\right].$$

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 α^{th} quantile Q_α is the unique minimizer of $\mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}$.
- **(Alternative notation)** The β^{th} quantile Q_β 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

$$\Psi_u(q) = \mathbb{E}\left[||X - q|| + < u, X - q >\right].$$

Chaudhuri's geometric quantiles

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

- 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:

$$\mathbb{E}\left[||X_U - q_U||\left[1 + \lambda(X_U - q_U)^{-2}||X_{U^\perp} - q_{U^\perp}||^2\right]^{1/2} + \beta(X_U - q_U)\right].$$

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.

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.

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.

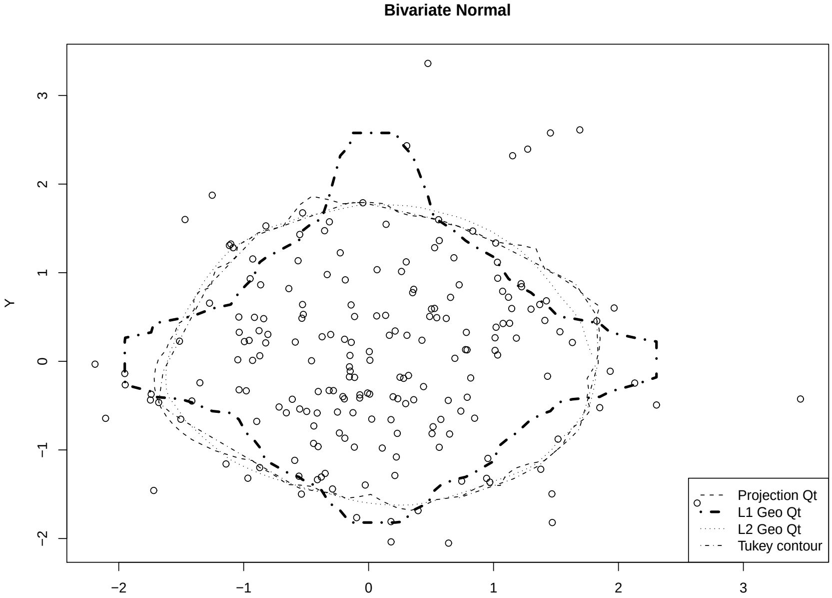
I'm the stat-math guy, remember?

- **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.

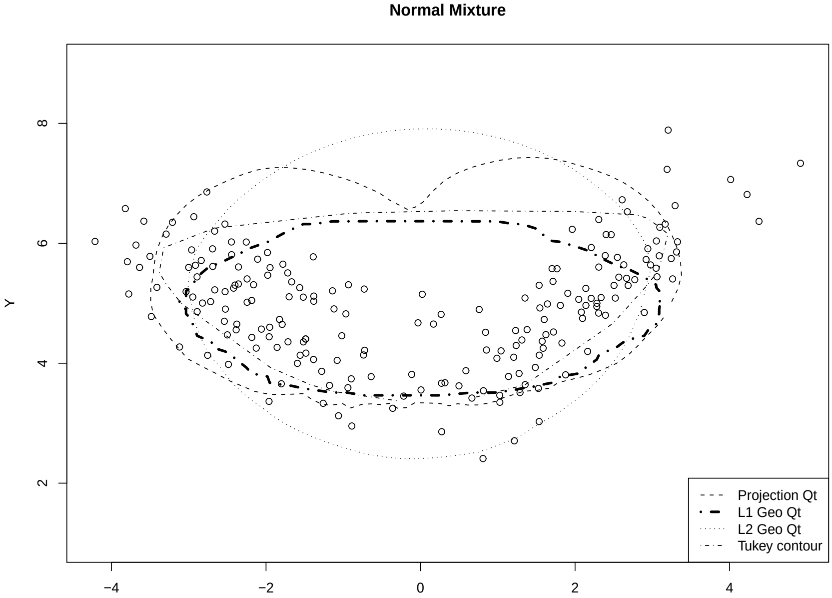
I'm the stat-math guy, remember?

- **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.

Example scatter plot



Example scatter plot



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).