Geometric Statistics for High-Dimensional Data Analysis

Snigdhansu Chatterjee

School of Statistics, University of Minnesota

Joint work with Lindsey Dietz, Megan Heyman, Subhabrata (Subho) Majumdar, and Ujjal Mukherjee

April 25, 2018
**Outline**

**Quantiles: univariate, multivariate**

Geometric quantiles for classification

The Indian Summer Monsoons: GSQ for feature selection

fMRI data: GSQ for spatio-temporal modeling
Univariate quantiles

- 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 \( \alpha \), i.e.
\[
Q_\alpha = \inf\{ q : \mathbb{P} [X \leq q] \geq \alpha \}.
\]

**Theorem**

*If* \( X \) *is (absolutely) continuous with cumulative distribution function* \( F(\cdot) \), *then* \( F(X) \sim \text{Uniform}(0, 1) \), *and there is a one-to-one relationship between* \( \alpha \) *and* \( Q_\alpha \).*
Univariate quantiles: an alternative view

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

- The *median* is the (unique) minimizer of \( \Psi(q) = \mathbb{E}|X - q| \).
- **(An extension)** The \( \alpha^{th} \) quantile \( Q_\alpha \) is the (unique) minimizer of

\[
\Psi(q) = \mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}.
\]
The median is the (unique) minimizer of $\Psi(q) = \mathbb{E}|X - q|$.

(An extension) The $\alpha^{th}$ quantile $Q_\alpha$ is the (unique) minimizer of

$$\Psi(q) = \mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)\}.$$ 

(Alternative notation) Define $u = 2\alpha - 1 \in (-1, 1)$. The $u^{th}$ quantile $Q_u$ is the (unique) minimizer of

$$\Psi(q) = \mathbb{E}\{|X - q| + u(X - q)\}.$$
Univariate quantiles: an alternative view

- The median is the (unique) minimizer of $\psi(q) = \mathbb{E}|X - q|$.

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

- (Alternative notation) Define $u = 2\alpha - 1 \in (-1, 1)$. The $u^{th}$ quantile $Q_u$ is the (unique) minimizer of
  \[\psi(q) = \mathbb{E}\{|X - q| + u(X - q)\} = \mathbb{E}\{\|X - q\| + \langle u, X - q \rangle\}\.\]
The *median* is the (unique) minimizer of $\Psi(q) = \mathbb{E}|X - q|$.

*(An extension)* The $\alpha^{th}$ quantile $Q_\alpha$ is the (unique) minimizer of

$$\Psi(q) = \mathbb{E}\{|X - q| + (2\alpha - 1)(X - q)|\}.$$ 

*(Alternative notation)* Define $u = 2\alpha - 1 \in (-1, 1)$. The $u^{th}$ quantile $Q_u$ is the (unique) minimizer of

$$\Psi(q) = \mathbb{E}\{|X - q| + u(X - q)|\} = \mathbb{E}\{||X - q|| + < u, X - q >\}.$$ 

*Define* quantiles in any inner-product space as minimizers of $\Psi_u(q) = \mathbb{E}\{||X - q|| + < u, X - q >\}$. (Haldane (1948), Chaudhuri (1996).)
Univariate to multivariate quantiles

**Univariate quantiles:**

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

Write \( x = x_u u/\| u \| + x_u \perp \).

Generally, for some \( \lambda \geq 0 \), the *generalized spatial quantile* (GSQ) are:

1. indexed by vectors in the unit ball \( u \in B_p = \{ x : \| x \| < 1 \} \),
   and
2. the \( u \)-th quantile \( Q(u) \) is the minimizer of
   \[
   \Psi_{u\lambda}(q) = \mathbb{E} \left[ |X_u - q_u| \{ 1 + \lambda (X_u - q_u)^{-2} \| X_u \perp - q_u \perp \|^2 \}^{1/2} \right. \\
   \left. + \| u \| (X_u - q_u) \right].
   \]
Bivariate quantiles

Domain

Support of Distn

\[ Q(u) \]
Bahadur representation of generalized spatial quantiles

**Theorem**

The following asymptotic Bahadur-type representation holds with probability 1 for any $u$:

$$n^{1/2} (\hat{Q}(u) - Q(u)) = -n^{-1/2} H^{-1} S_n + O(n^{-(1+s)/4} (\log n)^{1/2} (\log \log n)^{(1+s)/4})$$

as $n \to \infty$.

(Apologies for not including the details.)
Projection quantiles

Generalized spatial quantiles minimize:

$$\Psi_{u\lambda}(q) = \mathbb{E}\left[|X_u - q_u|\{1 + \lambda(X_u - q_u)^{-2}\|X_u - q_u\|_\perp^2\}^{1/2} + \|u\|(X_u - q_u)\right].$$

Set $\lambda = 0$ to get projection quantiles.

- Computationally extremely simple, no limitations from sample size and dimension (high $p$, low $n$ allowed).
- Projection quantiles based confidence sets have exact coverage.
- Works on infinite-dimensional spaces.
Projection quantiles

Theorem

*Projection quantiles have a one-to-one relationship with the unit ball, like univariate quantiles.*
Figure: Simulated data with a few GSQ (covered areas are deliberately different)
Quantiles: univariate, multivariate

**Geometric quantiles for classification**

The Indian Summer Monsoons: GSQ for feature selection

fMRI data: GSQ for spatio-temporal modeling
GSQ-depths are great for classification

**Figure:** A simulated 2-class classification problem with GSQ-depth classifier
<table>
<thead>
<tr>
<th>Method</th>
<th>CPU Time</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSQ</td>
<td>3.67</td>
<td>0.925</td>
</tr>
<tr>
<td>Random Forest</td>
<td>16714.20</td>
<td>0.895</td>
</tr>
<tr>
<td>SVM</td>
<td>966.86</td>
<td>0.842</td>
</tr>
<tr>
<td>LDA</td>
<td>0.28</td>
<td>0.74</td>
</tr>
<tr>
<td>Logit</td>
<td>0.35</td>
<td>0.69</td>
</tr>
</tbody>
</table>

**Table:** Arcene classification without feature selection (neural nets did not converge)
Quantiles: univariate, multivariate

Geometric quantiles for classification

The Indian Summer Monsoons: GSQ for feature selection

fMRI data: GSQ for spatio-temporal modeling
The data on monsoons

Figure: Air from the eastern Indian Ocean (yellow) and air descending over Arabia (blue) converge in the Somali jet. Low pressure at 30S. {Courtesy: UMn Climate Expeditions team.}
<table>
<thead>
<tr>
<th>Variable dropped</th>
<th>( \hat{e}<em>{n}(S</em>{-j}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Tmax</td>
<td>0.1490772</td>
</tr>
<tr>
<td>- X120W</td>
<td>0.2190159</td>
</tr>
<tr>
<td>- ELEVATION</td>
<td>0.2288938</td>
</tr>
<tr>
<td>- X120E</td>
<td>0.2290021</td>
</tr>
<tr>
<td>- ( \Delta TT \text{ Deg Celsius} )</td>
<td>0.2371846</td>
</tr>
<tr>
<td>- X80E</td>
<td>0.2449195</td>
</tr>
<tr>
<td>- LATITUDE</td>
<td>0.2468698</td>
</tr>
<tr>
<td>- TNH</td>
<td>0.2538924</td>
</tr>
<tr>
<td>- Nino34</td>
<td>0.2541503</td>
</tr>
<tr>
<td>- X10W</td>
<td>0.2558397</td>
</tr>
<tr>
<td>- LONGITUDE</td>
<td>0.2563105</td>
</tr>
<tr>
<td>- X100E</td>
<td>0.2565388</td>
</tr>
<tr>
<td>- EAWR</td>
<td>0.2565687</td>
</tr>
<tr>
<td>- X70E</td>
<td>0.2596766</td>
</tr>
<tr>
<td>- ( v _\text{wind}_{850} )</td>
<td>0.2604214</td>
</tr>
<tr>
<td>- X140E</td>
<td>0.2609039</td>
</tr>
<tr>
<td>- X40W</td>
<td>0.261159</td>
</tr>
<tr>
<td>- SolarFlux</td>
<td>0.2624313</td>
</tr>
<tr>
<td>- X160E</td>
<td>0.2626321</td>
</tr>
<tr>
<td>- EPNP</td>
<td>0.2630901</td>
</tr>
<tr>
<td>- TempAnomaly</td>
<td>0.2633658</td>
</tr>
<tr>
<td>- ( u _\text{wind}_{850} )</td>
<td>0.2649837</td>
</tr>
<tr>
<td>- WP</td>
<td>0.2660394</td>
</tr>
<tr>
<td>&lt;none&gt;</td>
<td>0.2663496</td>
</tr>
<tr>
<td>- POL</td>
<td>0.2677756</td>
</tr>
<tr>
<td>- Tmin</td>
<td>0.268231</td>
</tr>
<tr>
<td>- X20E</td>
<td>0.2687891</td>
</tr>
</tbody>
</table>
Figure: Comparing full model rolling predictions with reduced models: (a) Bias across years, (b) MSE across years.
Figure: Comparing full model rolling predictions with reduced models: (c) density plots for 2012, (d) stationwise residuals for 2012
Outline

Quantiles: univariate, multivariate

Geometric quantiles for classification

The Indian Summer Monsoons: GSQ for feature selection

fMRI data: GSQ for spatio-temporal modeling
We consider 19 tests subjects, with 2 kinds of visuals tasks.

Each subject went through 9 runs, where they saw faces or scrambled images, and had to react.

We fit a spatio-temporal model. Temporally, we fit a AR(5) with quadratic drift. Spatially, we consider different layers nearest neighbor voxels.

We measure the degree of spatial dependency in different regions of the brain.

The figures below are for one subject in one run.
Figure: Plot of significant $p$-values at 95% confidence level at the specified cross-sections.
Figure: A smoothed surface obtained from the $p$-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas.
Acknowledgment:

- This research is partially supported by the National Science Foundation (NSF) under grants # DMS-1622483, # DMS-1737918, and by the National Aeronautics and Space Administration (NASA).
- This research is partially supported by the Institute on the Environment (IonE).
Thank you