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

Major contributors




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 α^{th} quantile Q_α is the number below which X is observed with probability α, *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 $F(X) \sim Uniform(0, 1)$, and there is a one-to-one relationship between α and Q_α.

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 α^{th} quantile Q_α is the (unique) minimizer of

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

Univariate quantiles: an alternative view

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

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

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

Univariate quantiles: an alternative view

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

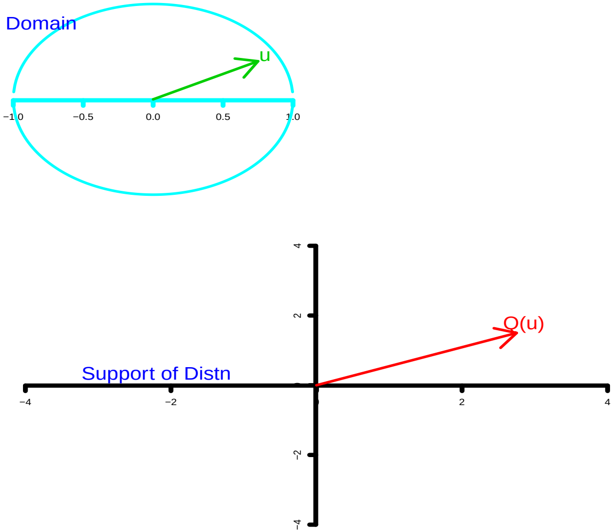
For every $u \in \{x : ||x|| < 1\} \subset \mathbb{R}$, $Q(u)$ minimizes
$\Psi_u(q) = \mathbb{E}\,[||X - q|| + < u, X - q >]$.

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 $u \in \mathcal{B}_p = \{x : ||x|| < 1\}$, and
2. the u-th quantile $Q(u)$ is the minimizer of

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

Bivariate quantiles



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:

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

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.

Example: simulated data plots

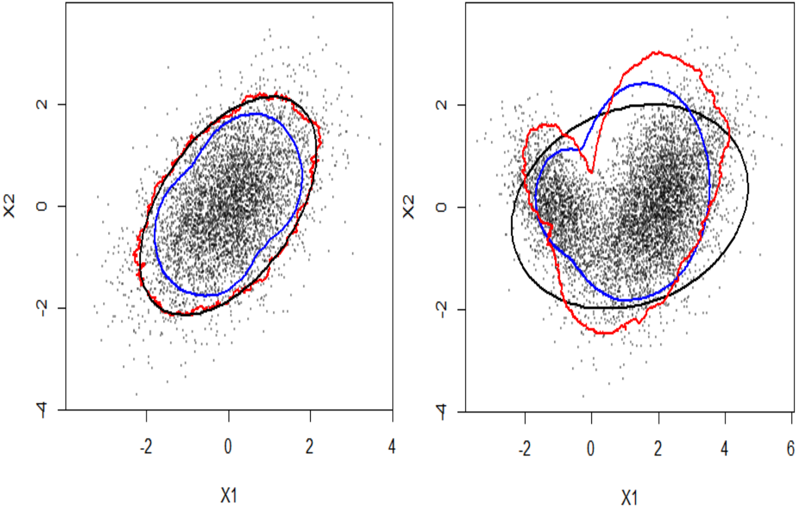


Figure: Simulated data with a few GSQ (covered areas are deliberately different)

Outline

GSQ-depths are great for classification

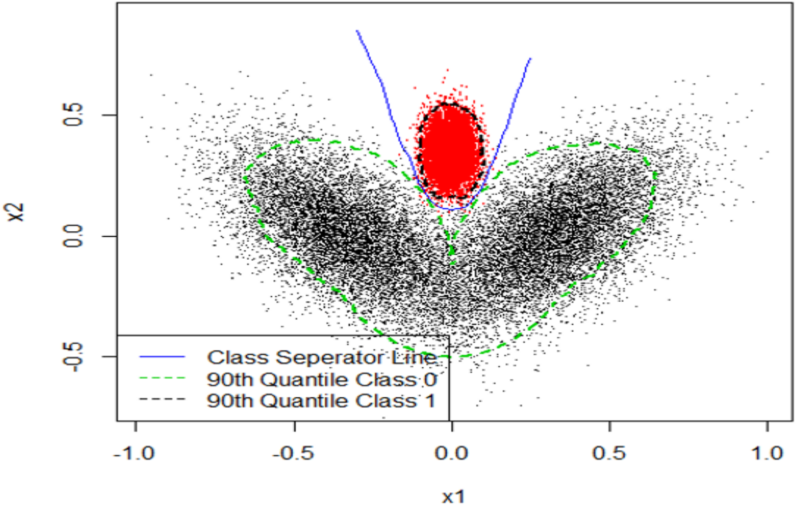


Figure: A simulated 2-class classification problem with GSQ-depth classifier

GSQ-depth based classification: some results

Method	CPU Time	Accuracy
GSQ	3.67	0.925
Random Forest	16714.20	0.895
SVM	966.86	0.842
LDA	0.28	0.74
Logit	0.35	0.69

Table: Arcene classification without feature selection (neural nets did not converge)

Outline

The data on monsoons

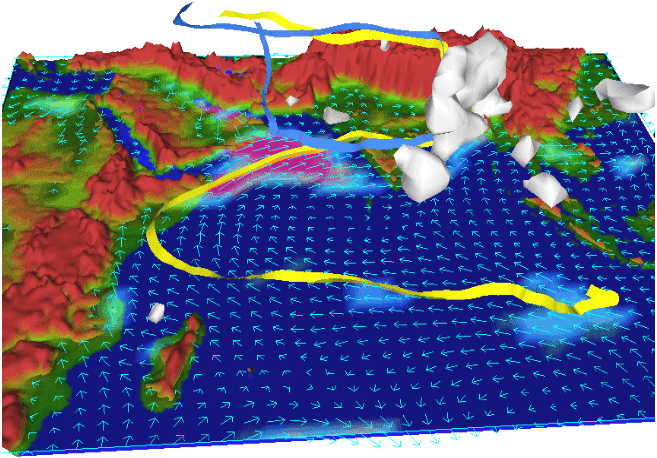
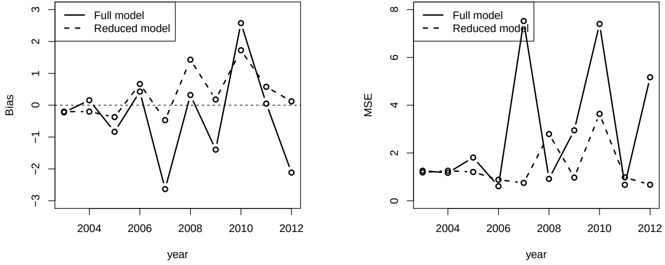


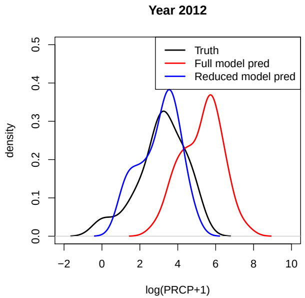
Figure: Air from the eastern Indian Ocean (yellow) and air descending over Arabia (blue) converge in the Somali jet. Low pressure at 30S. {*Courtesy: UMn Climate Expeditions team.*}

Variable dropped	$\hat{e}_n(\mathcal{S}_{-j})$
- Tmax	0.1490772
- X120W	0.2190159
- ELEVATION	0.2288938
- X120E	0.2290021
- ΔTT_Deg_Celsius	0.2371846
- X80E	0.2449195
- LATITUDE	0.2468698
- TNH	0.2538924
- Nino34	0.2541503
- X10W	0.2558397
- LONGITUDE	0.2563105
- X100E	0.2565388
- EAWR	0.2565687
- X70E	0.2596766
- v_wind_850	0.2604214
- X140E	0.2609039
- X40W	0.261159
- SolarFlux	0.2624313
- X160E	0.2626321
- EPNP	0.2630901
- TempAnomaly	0.2633658
- u_wind_850	0.2649837
- WP	0.2660394
$<$none$>$	0.2663496
- POL	0.2677756
- Tmin	0.268231
- X20E	0.2687891

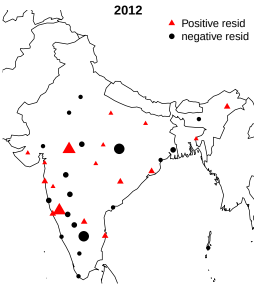


(a) (b)

Figure: Comparing full model rolling predictions with reduced models: (a) Bias across years, (b) MSE across years.



(c)



(d)

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

A brief outline

- 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.

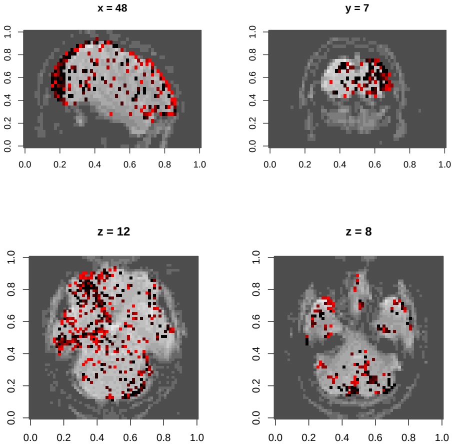


Figure: Plot of significant p-values at 95% confidence level at the specified cross-sections.

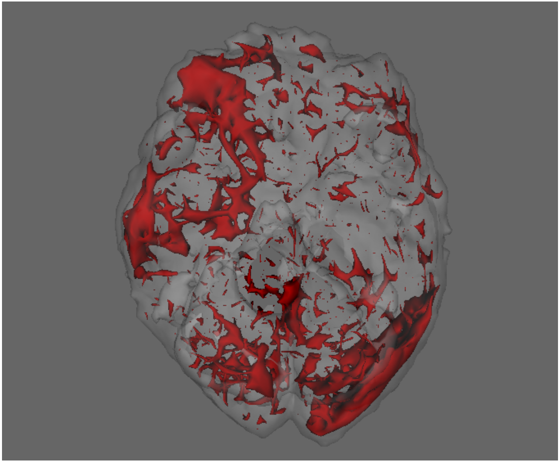


Figure: A smoothed surface obtained from the *p*-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas

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