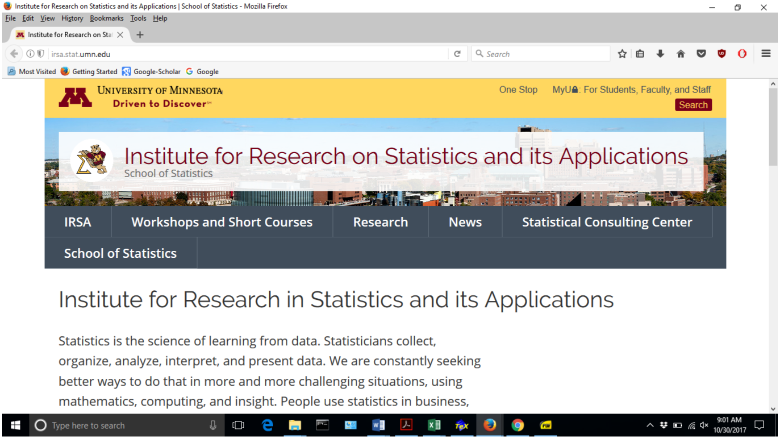
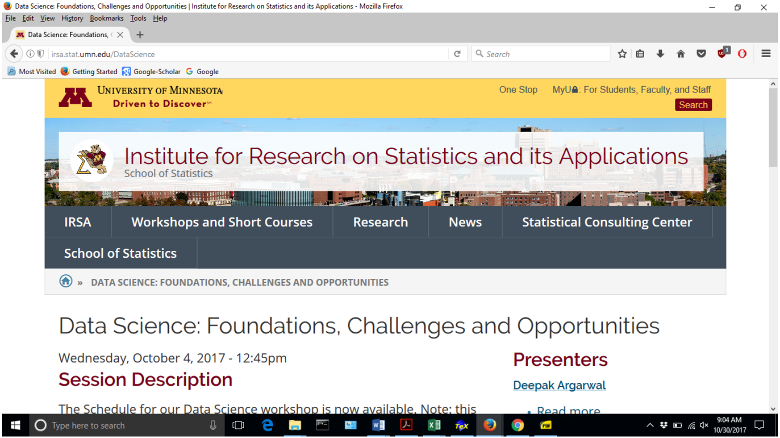
Shameless plug for IRSA

IRSA.stat.umn.edu



Data Science workshop at IRSA

IRSA.stat.umn.edu



Data-geometry and resampling-based inference for selecting predictors for monsoon precipitation

Snigdhansu Chatterjee

School of Statistics, University of Minnesota

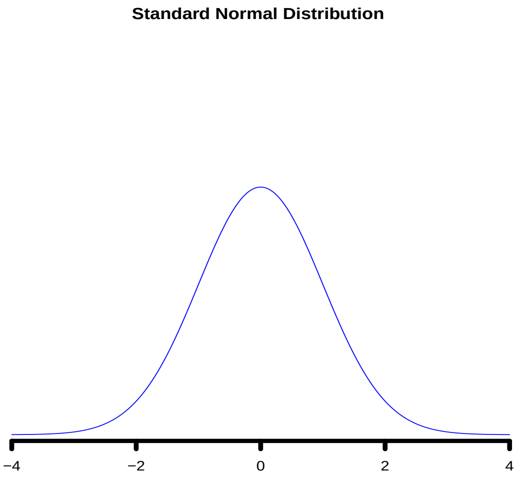
Joint work with Lindsey Dietz, Megan Heyman, Subhabrata (Subho) Majumdar, and Ujjal Mukherjee

October 30, 2017

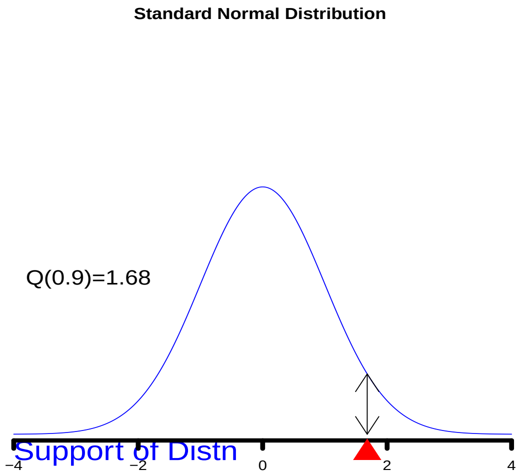
Major contributors



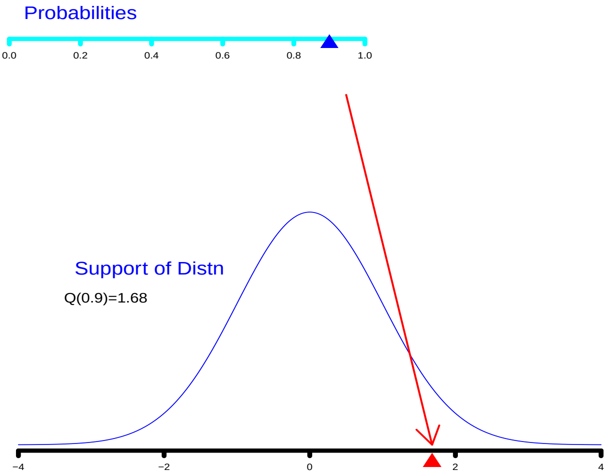

Normal probability density function



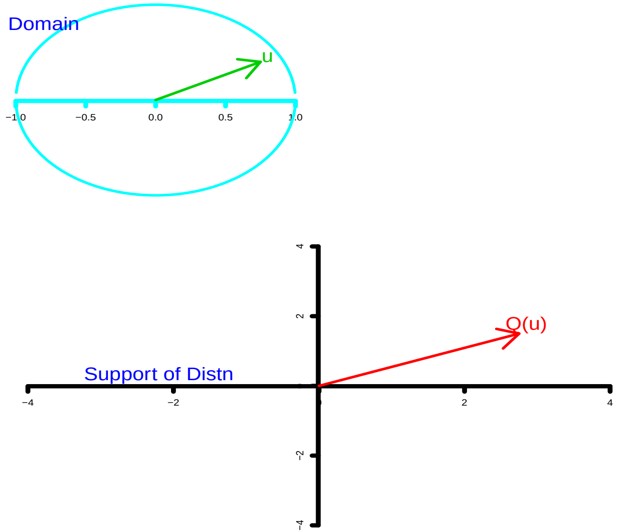
Normal quantile



Univariate quantile mapping



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.)

A few properties

- Computationally can be extremely simple, no limitations from sample size and dimension (high p, low n allowed).
- Confidence sets based on generalized spatial quantiles can have exact coverage.
- Works on infinite-dimensional spaces.
- Some generalized spatial quantiles have a one-to-one relationship with the unit ball, like univariate quantiles.

Example: simulated data plots

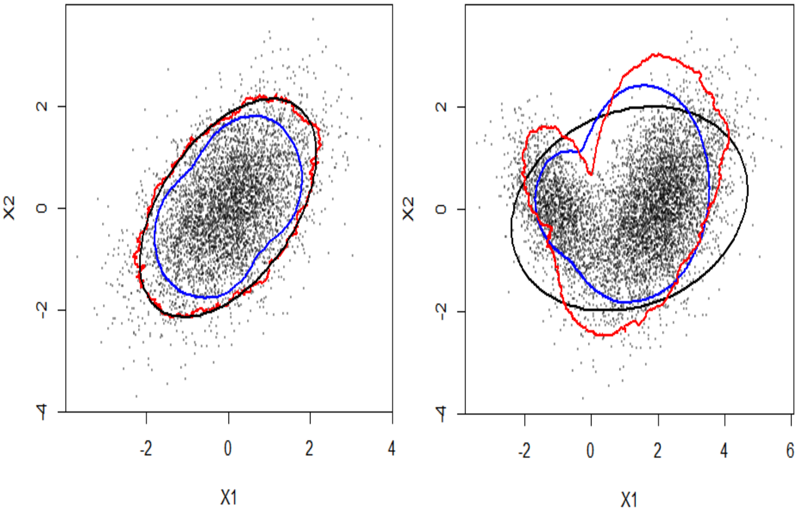


Figure: Simulated data with a few GSQ (covered areas are deliberately different)

What is data-depth?

- Suppose $\mathbb{F}$ is a cumulative distribution function corresponding to the random variable $\mathbf{X} \in \mathbb{R}^p$.
- A *data-depth* is a function of $\mathbb{R}^p$ and measures on $\mathbb{R}^p$ such that *there exists* $\theta \in \mathbb{R}^p$ *such that*

$$D(\theta, \mathbb{F}) \quad \geq \quad D(\theta + t(\mathbf{x} - \theta), \mathbb{F})$$

  for any $\mathbf{x} \in \mathbb{R}^p$ and and any $t \in (0, 1)$.
- Multivariate quantiles naturally yield data-depths.

Example: a depth plot

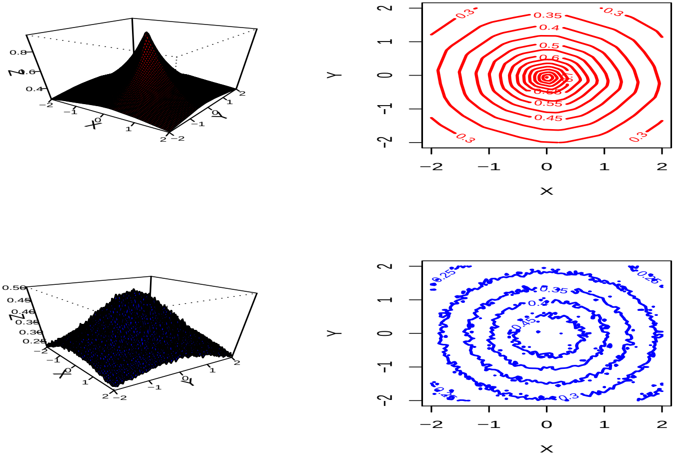


Figure: Perspective and contour plots of projection depth on top, simplicial at bottom.

GSQ-depths are great for classification

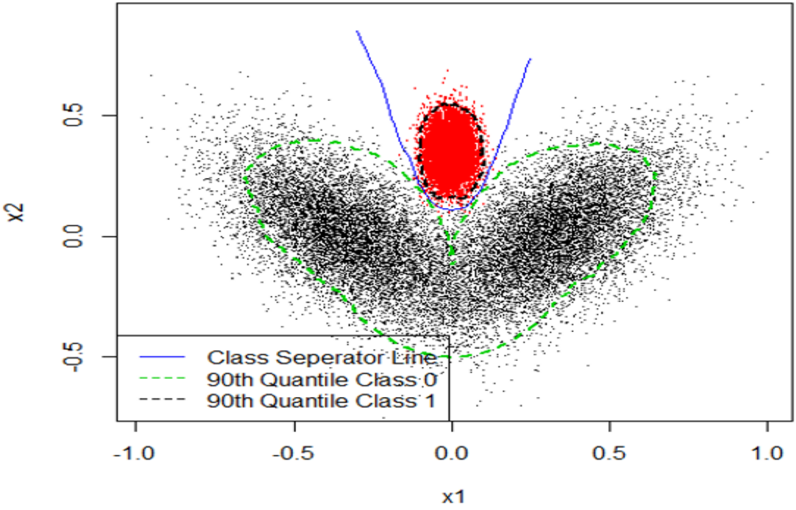


Figure: A simulated 2-class classification problem with GSQ-depth classifier

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)

Simultaneous model selection and inference in LM:

Data: $\{(Y_i, \mathbf{x}_i) \in \mathbb{R} \times \mathbb{R}^p, i = 1, \ldots, n.\}$. We want the best fitting parsimonious linear regression model.

- Fit the largest model with all the p covariates, and get $\hat{\beta}$, with (unknown) sampling distribution F_n.
- Compute $\Delta_0 = \mathbb{E}D(\hat{\beta}, F_n)$, the expected depth of $\hat{\beta}$ with respect to its own distribution.
- For $j = 1, \ldots, p$, define $\hat{\beta}_{(-j)}$ as $\hat{\beta}$, with $\hat{\beta}_j$ *replaced by zero*.
- Compute the *expected data-depth*

$$\Delta_j = \mathbb{E}D(\hat{\beta}_{(-j)}, F_n).$$

- Those variables for which $\Delta_j < \Delta_0$ are the important ones and these collectively form the most wonderful model ever!

Resampling step

Depth-based model selection

Compute and compare

$$\Delta_j = \mathbb{E} D(\hat{\beta}_{(-j)}, F_n).$$

This involves two distributions, and we use resampling to approximate this.

- The *m*-out-of-*n* (*moon*-bootstrap): Get a simple random sample of size m, with replacement, from the data. Assume $m \to \infty$, and $m/n \to 0$ as $n \to \infty$.
- An unusual Bayesian bootstrap: Generate *resampling weights* $\mathbb{W}_1, \ldots, \mathbb{W}_n$ i. i. d. $\sim$ *Gamma*(α, β), such that $\mathbb{E}\mathbb{W}_1 = 1$, $\mathbb{V}\mathbb{W}_1 \to \infty$ as $n \to \infty$. Use $\mathbb{W}_i$ as a weight with the i-th observation.
- Subsampling: Get a simple random sample of size m, without replacement, from the data. Assume $m \to \infty$, and $m/n \to 0$ as $n \to \infty$. (Considerably less efficient.)

The data on monsoons

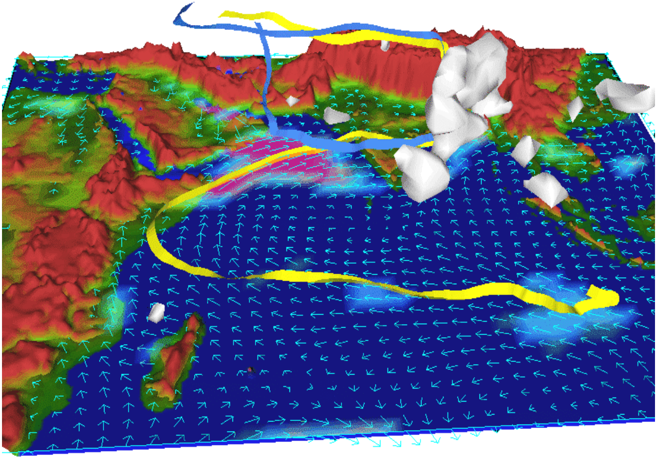
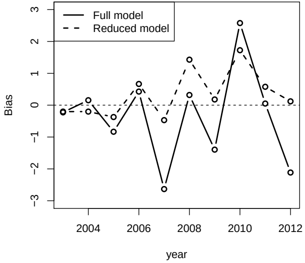
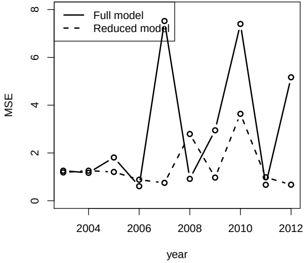


Figure: Air from the eastern Indian Ocean (yellow) and air descending over Arabia (blue) converge in the Somali jet. Low pressure at 30S. {*Courtesy: UMn Climate Expeditions team.*}

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
- EA	0.2690791
- u_wind_200	0.2692731
- u_wind_600	0.2695297
- SCA	0.2700276
- DMI	0.2700579

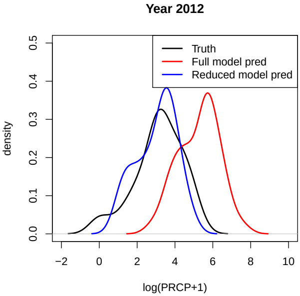


(a)

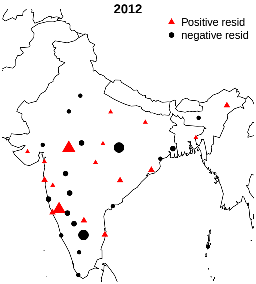


(b)

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

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.

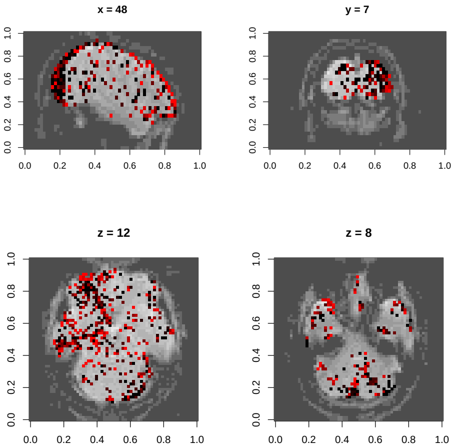


Figure: Plot of significant p-values at 95% confidence level at the specified cross-sections.

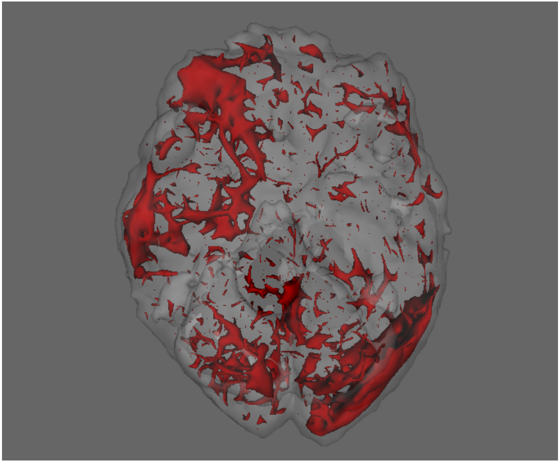


Figure: A smoothed surface obtained from the *p*-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas

Acknowledgment:

- This research is partially supported by the National Science Foundation (NSF) under grants # IIS-1029711, # 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

An example

Example

Data: $\{(Y_i, x_{i1}, x_{i2}), i = 1, \ldots, n$.
True model: $Y_i = 5x_{i1} + e_i$, $e_i iidN(0, 1)$.
Candidate models:

$$\begin{aligned}
\mathcal{M}_1 : & Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \\
\mathcal{M}_2 : & Y_i = \beta_1 x_{i1} + e_i, \\
\mathcal{M}_3 : & Y_i = \beta_2 x_{i2} + e_i, \\
\mathcal{M}_4 : & Y_i = e_i.
\end{aligned}$$

Example: model selection