Journal of Multivariate Analysis 102 (2011) 768–780

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

# High dimensional data analysis using multivariate generalized spatial quantiles

Nitai D. Mukhopadhyay[a,*], Snigdhansu Chatterjee[b]

[a] *Virginia Commonwealth University, Department of Biostatistics, Richmond VA 23298, United States*

[b] *School of Statistics, University of Minnesota, Minneapolis MN 55455, United States*

ARTICLE INFO

*Article history:*
Received 13 October 2009
Available online 8 December 2010

*AMS subject classifications:*
62H11
62G05

*Keywords:*
Multivariate quantile
Spatial quantile
Projection quantile
Generalized spatial quantile
Multidimensional coverage sets
Multivariate order statistics
Brain imaging
High dimensional data visualization

ABSTRACT

High dimensional data routinely arises in image analysis, genetic experiments, network analysis, and various other research areas. Many such datasets do not correspond to well-studied probability distributions, and in several applications the data-cloud prominently displays non-symmetric and non-convex shape features. We propose using spatial quantiles and their generalizations, in particular, the projection quantile, for describing, analyzing and conducting inference with multivariate data. Minimal assumptions are made about the nature and shape characteristics of the underlying probability distribution, and we do not require the sample size to be as high as the data-dimension. We present theoretical properties of the generalized spatial quantiles, and an algorithm to compute them quickly. Our quantiles may be used to obtain multidimensional confidence or credible regions that are not required to conform to a pre-determined shape. We also propose a new notion of multidimensional order statistics, which may be used to obtain multidimensional outliers. Many of the features revealed using a generalized spatial quantile-based analysis would be missed if the data was shoehorned into a well-known probabilistic configuration.

## 1. Introduction

The use of multivariate Normal distribution, or certain characteristics of multivariate Normal distributions, is routine in statistical data analysis. Prominent among such characteristics are the elliptic shape of the density function concentration regions, convexity and compactness of such concentration ellipsoids, and an overall symmetry of the density function around the location parameter. These characteristics are useful, for example, in describing confidence sets (or in a Bayesian analysis, credible sets), or acceptance regions for hypothesis tests. Sometimes multivariate heavy-tailed, lifetime, or discrete distributions may be put to use, however it is not obvious how to proceed when the properties of the data do not match the characteristics of the chosen family of distributions.

In this paper, we propose to address the issue of how to describe, analyze and conduct inference on datasets where routine assumptions like multivariate Normality may not be viable. Minimal assumptions are made about the nature and shape characteristics of the data-cloud. Also, in view of several recent applications where the dimensions of the observations are extraordinarily high, but the sample size may or may not be high, our methodology does not necessarily require that the sample size be higher than the dimensions of the data.

As an example where routine multivariate data analysis assumptions may not be appropriate, consider the problem of the treatment of Alzheimer's disease using *Deep Brain Stimulation* (DBS). This treatment is conducted by putting DBS electrodes close to the nucleus of the brain, to provide a stimulation deep inside the brain of the patient. The data consists

[*] Corresponding author.
*E-mail addresses:* ndmukhopadhy@vcu.edu (N.D. Mukhopadhyay), chatterjee@stat.umn.edu (S. Chatterjee).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.12.002

of the location of the electrode placement inside the brain, along with measurements on the changes in the neurological patterns of the patients. The measurements on changes of neurological patterns differ from one location to another inside a person's brain, and from one individual to another. The medical interest in this problem lies in obtaining the region of the human brain where the placement of the electrodes and subsequent stimulation results in prominent changes in the neurological patterns. For example, we may want to obtain the region where electrode placement and stimulation results in a 50% or more improvement in cognitive ability. An assumption that such a region is a convex ellipsoid seems tenuous at best given the geometry of the human brain, and medical professionals are generally unwilling to accept such simplistic statistical assumptions.

An example of a statistical application requiring extraction of high dimensional geometrical features is available from microarray gene experimentation. Typically, a large number $n$ $(=O(10^3))$ of genes are observed a number of times $p$ $(=O(10))$. Such studies are often conducted to understand the role of genes in cell-cycle regulation, typically in the context of a disease like cancer where the regular cell-cycle pattern may be altered due to the over or under-expression of a number of genes. In the context of a particular type of cancer, most of the $n$ genes do not participate in the cell-cycle regulation process. In order to understand which genes are "out-of-the-ordinary" in a given context, we need to study the $p$-dimensional profile of each one of the $n$ genes and identify the outlying ones. Standard approaches rely on assumptions like a multivariate Normal distribution pattern, or some characteristic of it, for example, in considering correlation as a sole dependency measure. There is no biological reason to presume that the $p(=10-50)$ dimensional data-cloud formed by the expressions of thousands of genes would correspond to a $p$-dimensional Normal density pattern. We need a method for identifying extraordinary genes, without presuming the data fits into a probabilistic model simply because the model is well understood.

These examples illustrate the need for ways of obtaining and using general multivariate quantiles. Multivariate quantiles and coverage sets are important tools for a number of different problems. They may be used for summarizing multivariate Bayesian and resampling-based inference, for simultaneous hypothesis tests, for evaluation of several competing models for a given data, and several other applications. One of the main roles of multivariate quantiles is to capture the geometry of the data, and hence the dependency among the variables. The listing of coordinate-wise quantiles is uninformative about the joint distribution of the variables, also coordinate-wise quantiles do not retain desirable invariance properties.

The desirable properties for any candidate multivariate quantile include reflecting the shape and other properties of the data, fast and accurate algorithms for computation, and tractable theoretical properties. Moreover, applicability in sparse data in high dimensions should be considered an advantage, since such cases routinely occur in several modern applications. In this paper, we build on the notion of geometric or spatial quantiles presented in [13]. The central idea of Chaudhuri is that multivariate quantiles are indexed by a $p$-dimensional vector of norm between zero and one, where $p$ is the dimension of the observations. This definition naturally includes the classical definition of a quantile for the univariate case, it extends well-studied notions of multivariate medians [16,2,28] to general quantiles, and conforms to the principle adopted in [3,4,18] and others that multivariate quantiles should have both direction and magnitude. By varying the direction, magnitude and the distance metric, we obtain the class of *generalized spatial quantiles*, of which Chaudhuri's quantiles are a special case. Another interesting special case is the *projection quantile*, which stands out in terms of computational ease and theoretical tractability, and is intuitively appealing since it relates to quantiles of one-dimensional projections.

Multivariate quantiles may be used for several purposes, including data description and exploratory analysis, graphical displays, estimation and inference. Some of these tasks may be accomplished by using *data-depth*, which is essentially a center-outward ranking of multivariate data. Data-depths have been studied comprehensively, see [30,22,24,23,31,36] for several seminal developments. The relationship between multivariate quantiles and depth is similar to that of univariate quantiles and ranks, in the sense that depth (or rank) can be computed from quantiles (see [31] and Section 3.3), but depth/rank does not carry as much information as quantiles. Hence, all methodology, theory and applications based on depth are available when quantiles are used as basic quantities. On the other hand, concepts like quantile regression *require* a notion of quantiles (see Section 3.2 for the multivariate version), and may not be satisfactorily obtained using a depth function alone. Moreover, several depth functions do not account for shape features, they may require an unviable amount of computational time, and may not be applicable in high dimensions. Also, since the underlying densities could be posterior or bootstrap densities (and hence conditional on data) in many applications, verification of all technical assumptions relating to data-depth could be problematic.

In Section 2, we first present Chaudhuri's spatial quantiles, then develop projection quantiles, and finally present the generalized spatial quantiles. Properties of the generalized spatial quantiles, some applications, and algorithms to compute them are presented in Section 3. First, in Section 3.1, we obtain a number of theoretical results; on the consistency and asymptotic Normality of the sample generalized spatial quantile, on the consistency of approximating the distribution of the sample generalized spatial quantile using generalized bootstrap, and on a Bahadur-type asymptotic representation. We also establish a one-to-one correspondence between projection quantiles and the unit ball in $\mathbb{R}^p$ where $p$ is the data-dimension, which is a multivariate generalization of the well-known relationship between quantiles and probabilities. In Section 3.2, we propose a method for obtaining credible or confidence regions in dimensions greater than one, when only a data-scatter is available. Such confidence regions are not presumed to conform to a pre-determined feature like symmetry or convexity, and are expected to capture the shape of the data-cloud. We prove that the one-dimensional projections of the projection quantiles-based confidence regions have exact coverage probability, thus illustrating the efficacy of the proposed method. We then discuss, in Section 3.3, the notion of multivariate order statistics, and remark on how they may be used for detecting outliers in high dimensional data and for defining data-depths. Lastly in Section 3, in Section 3.4 we present a coordinate descent algorithm for computing the generalized spatial quantiles, which is especially useful when the sample size is lower than the dimension size.

Since data-depth measures can accomplish some of the tasks of multivariate quantiles, in Section 4 we first present a simulation example to compare three cases of generalized spatial quantiles and a popular data-depth measure. This simulation example shows that in standard multivariate inferential problems, quantiles and data-depths generally complement and corroborate each other. We then revisit the examples of DBS electrode placement and human cancer cell-cycle regulation that have been briefly introduced above. The advantage of using multivariate quantiles as opposed to data-depth in high dimensions is illustrated in the cell-cycle regulation data. A concluding section collects further remarks, and an Appendix is used for the proofs of some of the theoretical results from Section 3.

## 2. Spatial quantiles

In this section we describe Chaudhuri's quantiles, projection quantiles and generalized spatial quantiles. In this context we also establish some notations that we follow in the rest of this paper.

### 2.1. Chaudhuri's spatial quantiles

In $p$-dimensional Euclidean space $\mathbb{R}^p$, Chaudhuri's spatial quantiles [13] are maps from the open unit ball $\mathcal{B}_p = \{x : \|x\| < 1\}$ to $\mathbb{R}^p$. For any random variable $X \in \mathbb{R}^p$ and every $u \in \mathcal{B}_p$, the $u$th quantile $Q(u)$ is defined as the minimizer of

$$\Psi_u(q) = \mathbb{E}[\|X - q\| + \langle u, X - q\rangle]. \tag{1}$$

The inner product $\langle \cdot, \cdot \rangle$ above is the usual Euclidean inner product, and the norm $\|\cdot\|$ is the usual Euclidean norm. The existence and uniqueness of Chaudhuri's spatial quantiles are discussed in Section 3. If a random sample $X_1, \dots, X_n$ is available, the empirical spatial quantile $Q_n(u)$ imitates the above setup, and is defined as the minimizer of

$$\Psi_{n,u}(q) = \sum_{i=1}^{n} [\|X_i - q\| + \langle u, X_i - q\rangle]. \tag{2}$$

Note that, in the 1-dimensional case, the $\alpha$th sample quantile is traditionally defined as the point below which exactly $\alpha$-proportion of the data falls, for $\alpha \in (0, 1)$. This definition is recovered from (2) using $p = 1$ and $u = 2\alpha - 1 \in (-1, 1)$.

Historically, possibly the earliest example of Chaudhuri's quantiles is Haldane's spatial median [16]. Various properties and applications of Chaudhuri's quantiles are available in [6–9,11].

### 2.2. The projection quantile

Here we present another approach that retains the theme of describing quantiles as function indexed by the unit ball in $\mathbb{R}^p$. Let $U$ denote the unit vector in the direction of $u \in \mathcal{B}^p \setminus \{0\}$,i.e., $U = u/\|u\|$. Let $X_U = \langle X, U\rangle = \|u\|^{-1}\langle X, u\rangle$, thus the projection of the random vector $X \in \mathbb{R}^p$ on the 1-dimensional space spanned by the vector $u \in \mathcal{B}_p$ is $X_U U = \|u\|^{-2}\langle X, u\rangle u$. Let $q_u$ be the $(1 + \|u\|)/2$th quantile of $X_U$, that is, $\mathbb{P}[X_U \le q_u] = (1 + \|u\|)/2$. The $u$th projection quantile is defined as $Q_{\text{proj}}(u) = q_u u/\|u\| = q_u U$.

Thus, the $u$th projection quantile $Q_{\text{proj}}(u)$ is a vector that lies in the subspace spanned by $u$, and has the intuitive appeal of being related to $q_u$. Moreover, it poses no computational burden of any significance, since projecting $X$ on a 1-dimensional subspace is a simple operation. One of the attractive features of quantiles of univariate, continuous distributions is that they are invertible functions of probabilities, that is, there is a one-to-one map between the quantiles and probabilities. In Section 3 we establish the equivalent property for the projection quantile; i.e. the projection quantile is a one-to-one map of the unit ball in $p$-dimensions. There may be several interesting applications developed from this important property, which we will pursue in future.

The use of projections for studying higher dimensional objects is very standard in geometry and statistics. For example, projection pursuit is used extensively in many applications. An early review of projection pursuit may be found in [19], and an overview of applications may be found in [17]. A notion of data depth based on projections has been developed and studied in [36,35,34] and in several other papers. However, we have not been able to trace a reference for the projection quantile, as described in this section.

### 2.3. Generalized spatial quantiles

In this section we present a general approach towards spatial quantiles, which obtains Chaudhuri's spatial quantiles as well as the projection quantiles as special cases. As earlier, define $U$ as the unit vector in the direction of $u \in \mathcal{B}^p \setminus \{0\}$, i.e., $U = u/\|u\|$. Also, for convenience, define $\beta = \|u\|$, thus $u = \beta U$. Let $X_U = \langle X, U\rangle$, $q_U = \langle q, U\rangle$, thus the projections of $X$

and $q$ in the direction of $u$ is $X_U U$ and $q_U U$ respectively. Let $X_{U^\perp} = X - X_U U$, $q_{U^\perp} = q - q_U U$; these are the projections on the space orthogonal to $U$ (or $u$). In particular, we have $\|X - q\|^2 = (X_U - q_U)^2 + \|X_{U^\perp} - q_{U^\perp}\|^2$.

Based on this, for every $\lambda \in \mathbb{R}$, the generalized spatial quantiles $Q(u, \lambda)$ are defined as minimizers of expectation of:

$$\begin{aligned}\Psi_{u,\lambda}(X, q) &= \Psi_{u,\lambda}(X, q_U, q_{U^\perp})\\ &= \|X_U - q_U\|[1 + \lambda(X_U - q_U)^{-2}\|X_{U^\perp} - q_{U^\perp}\|^2]^{1/2} + \beta(X_U - q_U).\end{aligned}$$

Note that for $\lambda = 0$ we get the projection quantile, for $\lambda = 1$ we get Chaudhuri's quantiles.

We may consider another level of generalization here, by replacing the Euclidean norm used in $\Psi_{u,\lambda}(X, q)$ with a $L_k$-norm, for $k \geq 1$. The $L_k$-norm of a vector $x = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is given by $\|x\|_k = \left(\sum_{i=1}^p |x_i|^k\right)^{1/k}$. Thus, the generalized spatial quantiles $Q(u, \lambda, k)$ based on the $L_k$-norm are defined as minimizers of expectation of:

$$\begin{aligned}\Psi_{u,\lambda,k}(X, q) &= \Psi_{u,\lambda,k}(X, q_U, q_{U^\perp})\\ &= \|X_U - q_U\|_k[1 + \lambda(X_U - q_U)^{-k}\|X_{U^\perp} - q_{U^\perp}\|_k^k]^{1/k} + \beta(X_U - q_U).\end{aligned}$$

The notion of a *projection*, and the definitions of $X_U$, $q_U$, $X_{U^\perp}$, $q_{U^\perp}$ based on the Euclidean inner product are retained as earlier. The extension of Chakraborty [5] to Chaudhuri's quantiles is obtained with $\Psi_{u,1,1}(X, q)$. The properties of the quantiles depend on the choice of $k$, but for this paper excepting the occasional remark, we will keep to the use of the Euclidean norm, and not use $k$ as a part of our notation. Note that $\Psi_{u,0,k}(X, q) = \Psi_{u,0}(X, q)$ and the choice of the norm does not matter for projection quantiles. Also, when $u$ is chosen along any Cartesian basis direction $(0, \ldots, 0, 1, 0, \ldots, 0)$, the coordinate-wise quantiles are obtained as a special case of projection quantiles. In applications, certain linear combination of the elements of $X \in \mathbb{R}^p$ may be of interest, for example, certain contrasts or the cross-section mean. Quantiles from the joint distributions of all such interesting linear combinations are easily obtainable by our method. The definition of generalized spatial quantiles effectively imposes the requirement that the quantile of a random variable should reside in its support, and reflect the topological and geometric properties of the support. Hence, quantiles of $p$-dimensional random vectors should be $p$-dimensional, and dependent on the metric and geometry in use.

## 3. Properties, applications and algorithms

### 3.1. Properties of generalized spatial quantiles

We now present a few properties of generalized spatial quantiles. Some of these properties have been discovered earlier for special cases like Chaudhuri's spatial quantiles. Our approach below presents a unified and easily understood framework for every fixed $(u, \lambda) \in \mathbb{R}^p \times \mathbb{R}$, relying on the convexity of $\Psi_{u,\lambda}(X, q_U, q_{U^\perp})$ in $(q_U, q_{U^\perp})$. Our first result is to establish this convexity.

**Proposition 3.1.** *The function*

$$\Psi_{u,\lambda}(X, q_U, q_{U^\perp}) = \|X_U - q_U\|[1 + \lambda(X_U - q_U)^{-2}\|X_{U^\perp} - q_{U^\perp}\|^2]^{1/2} + \beta(X_U - q_U)$$

*is convex in* $(q_U, q_{U^\perp})$*, with the subgradient function*

$$g(X, q_U, q_{U^\perp}) = \begin{pmatrix} -[(X_U - q_U)^2 + \lambda\|X_{U^\perp} - q_{U^\perp}\|^2]^{-1/2}(X_U - q_U) - \beta \\ -\lambda[(X_U - q_U)^2 + \lambda\|X_{U^\perp} - q_{U^\perp}\|^2]^{-1/2}(X_{U^\perp} - q_{U^\perp}) \end{pmatrix}.$$

The proof of this result is easy and hence omitted. We restrict ourselves to such random variables for which $\mathbb{E}\Psi_{u,\lambda}(X, q_U, q_{U^\perp})$ is finite for our choices of $q = q_U U + q_{U^\perp}$. We also assume that the minimizer of $\mathbb{E}\Psi_{u,\lambda}(X, q_U, q_{U^\perp})$, denoted by $q^* = q_U^* U + q_{U^\perp}^*$, which is the population $(u, \lambda)$th quantile, is unique. The conditions of finiteness of the expectation of the population quantile defining function and the uniqueness of the population quantile, are mild and necessary assumptions.

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample. We denote the minimizer of $n^{-1}\sum_{i=1}^n \Psi_{u,\lambda}(X_i, q_U, q_{U^\perp})$, the sample $(u, \lambda)$th quantile, by $q_n = q_{nU} U + q_{nU^\perp}$. Our next set of results relate to the behavior of $q_n$, much of which is characterized by the moments of the subgradient function $g(X, q^*)$ defined in Proposition 3.1.

**Theorem 3.1.** 1. *$q_n \to q^*$ almost surely as $n \to \infty$.*

2. *If $\mathbb{E}\|g(X, q^*)\|^2 < \infty$ and if $\mathbb{E}\Psi_{u,\lambda}(X, q)$ is twice continuously differentiable at $q^*$ with the second derivative $H$ being positive definite, then as $n \to \infty$*

$$n^{1/2}(q_n - q^*) = -n^{-1/2}H^{-1}S_n + o_P(1),$$

*where $S_n = \sum_{i=1}^n g(X_i, q^*)$. This implies, in particular, that $n^{1/2}(q_n - q^*)$ is asymptotically Normal, with asymptotic variance $H^{-1}VH^{-1}$ where $V = \mathrm{Var}\, g(X, q^*)$.*

3. *Under the conditions of the previous item, the generalized bootstrap approximation for the distribution of $n^{1/2}(q_n - q^*)$ is consistent, and resampling may be used for inference.*

4. *In addition to the conditions of the previous item, assume that*

$$\left\| \frac{\partial}{\partial q}\mathbb{E}\Psi_{u,\lambda}(X,q) - \frac{\partial^2}{\partial q^2}\mathbb{E}\Psi_{u,\lambda}(X,q^*)(q-q^*)\right\| = O(\|q-q^*\|^{(3+s)/2}) \quad \text{as } q \to q^*,$$
$$\mathbb{E}\|g(X,q)-g(X,q^*)\|^2 = O(\|q-q^*\|^{1+s}) \quad \text{as } q \to q^*,$$
$$\mathbb{E}\|g(X,q)\|^r < \infty \quad \text{as } q \to q^*,$$

*for some $s \in (0,1)$ and $r > (8+p(1+s))/(1-s)$. Then the following asymptotic Bahadur-type representation holds with probability 1:*

$$n^{1/2}(q_n - q^*) = -n^{-1/2}H^{-1}S_n + O(n^{-(1+s)/4}(\log\ n)^{1/2}(\log\log\ n)^{(1+s)/4})$$

*as $n \to \infty$.*

The above results require considerable algebra in some cases, but are otherwise derivable using the results of Haberman [15], Niemiro [26], and Bose and Chatterjee [1]. We omit the proofs of these to avoid lengthy technical discussions. Our next result is to establish an inverse of the projection quantiles. To simplify notations, we assume that the spatial median is $0 \in \mathbb{R}^p$.

**Theorem 3.2.** *Suppose $X$ is an absolutely continuous random variable in $\mathbb{R}^p$. The projection quantile $Q_{\text{proj}} : \mathcal{B}_p \to \mathbb{R}^p$ defined as $Q_{\text{proj}}(u) = \|u\|^{-1}q_u u$, where $q_u$ is the $(1+\|u\|)/2$-quantile of $X_U = \|u\|^{-1}\langle X, u\rangle$, and the following function*

$$Q_{\text{proj}}^{-1} : \mathbb{R}^p \Rightarrow \mathcal{B}_p \quad \text{defined as}$$
$$Q_{\text{proj}}^{-1}(x) = \frac{x}{\|x\|}(2G_x(\|x\|) - 1)$$
$$\text{where } G_x(\cdot) = P\left(\frac{\langle X, x\rangle}{\|x\|} \le \cdot\right),$$

*are inverse functions of each other, for $u \neq 0$ and $x \neq 0$. The spatial median $0 \in \mathbb{R}^p$ and $u = 0 \in \mathcal{B}^p$ map to each other.*

We prove this result in the Appendix following this paper.

The projection quantile, and the generalized spatial quantile for all choices of $\lambda \neq 1$ are equivariant under location shifts. That is, the quantiles of $Z = a + Y \in \mathbb{R}^p$ for any $a \in \mathbb{R}^p$ are given by the corresponding quantiles of $Y$added to $a$. For Chaudhuri's quantiles, which correspond to $\lambda = 1$, both rotation and location equivariance are obtained. Note however, that when the sample size is considerably large compared to the dimension size, a simple two-step transformation process is adequate to address invariance issues. This is the transformation–retransformation approach proposed by Chakraborty and Chaudhuri [10]. For data in $\mathbb{R}^p$, isolate $p+1$ data points $Y_0, \ldots, Y_p$, and re-center every other observation by subtracting $Y_0$. Then express the re-centered data in terms of a basis given by $\{Y_i - Y_0,\ i = 1, \ldots, p\}$. The results from the statistical analyses performed on the transformed data (excluding the $p+1$ isolated points) can be mapped to the original co-ordinate system by a simple back transformation, and would satisfy all the conditions of affine equivariance.

It is clear that the multivariate projection quantile defined in Section 2.2 shares the same kind of robustness properties as a univariate quantile, and $Q_{\text{proj}}(u)$ has a breakdown value of $(1+\|u\|)/2$th. The robustness properties of the other generalized spatial quantiles are not so apparent. Chakraborty and Chaudhuri [9] have studied the breakdown value of the spatial median.

### *3.2. Spatial confidence sets and quantile regression*

For any choice of $\beta \in (0,1)$ and $\lambda \ge 0$, the set of generalized spatial quantiles $\mathcal{C}_{\beta,\lambda} = \{Q(u,\lambda) : \|u\| \le \beta\} \subseteq \mathbb{R}^p$ is a compact, path connected set, and $\mathcal{C}_{\beta_1,\lambda} \subseteq C_{\beta_2,\lambda}$ if $\beta_1 \le \beta_2$. Since by varying the choice of $\beta$ we can consider an entire range of compact sets from the null set to the support of the random vector under study, we propose to use $\mathcal{C}_{\beta,\lambda}$ as a *generalized spatial confidence set*. Different choices of $\lambda$ correspond to determining the shape of the sets $\mathcal{C}_{\beta,\lambda}$. Later, in Section 4, we show that the choice of the norm also regulates the shape of $\mathcal{C}_{\beta,\lambda}$ to some extent.

A challenging task here is to compute the probability $\mathbb{P}[X \in \mathcal{C}_{\beta,\lambda}]$. Our next result is to show that projection confidence sets achieve the exact coverage probability of $\beta$, for the natural interval resulting from $\mathcal{C}_{\beta,0}$ for any linear combination of the coordinates of $X$.

**Theorem 3.3.** *For every linear combination $c^TX$ with $\|c\| = 1$, consider the interval $\mathcal{B}_{\beta,0} = (-q_{(-c)}, q_c)$ constructed using the projection quantiles corresponding to $-\beta c$ and $\beta c$ for any $\beta \in (0,1)$. This projection quantile based interval has the exact coverage probability of $\beta$.*

Theorem 3.3 is also proved in the Appendix.

The computation of $\beta$ for which $\mathbb{P}[X \in \mathcal{C}_{\beta,\lambda}] = \alpha$ is achieved for fixed $\alpha \in (0,1)$ and fixed $\lambda$ is an open problem. For $\lambda = 0$ and $p = 2$, if $X$ follows the uniform distribution on the unit square $(0,1) \times (0,1) \subset \mathbb{R}^2$, we have $\beta = 1 - \frac{1}{\pi}(cos^{-1}\sqrt{\alpha} - \sqrt{\alpha(1-\alpha)})$. For $\lambda = 0$, $p = 2$ and $X$ following the bivariate standard Normal distribution with mean zero and identity dispersion matrix, the relation $\beta = 2\Phi(\sqrt{-2\ln(1-\alpha)}) - 1$ holds, where $\Phi(\cdot)$ is the univariate standard Normal cumulative distribution function. For general multivariate data, we adopt a scheme similar to [33], and in order to find a set with $\alpha$-level coverage we choose that value of $\beta$ for which $\alpha$ fraction of the data are inside $\mathcal{C}_{\beta,\lambda}$. Thus, finite-sample coverage properties of our confidence or credible sets are exact.

Multivariate quantiles and coverage sets are important tools for a number of different problems. For example, modern-day Bayesian and resampling-based statistical inference typically involve Monte Carlo sampling from the probability distributions of interest, which are then used to approximate moments, quantiles, credible or confidence regions, and for other statistical purposes. While these inferential tasks are routine when performed for one-dimensional quantities, they can be difficult in higher dimensions. As an illustration, consider a random sample $\mathbf{X} = (X_1, \ldots, X_n)$ from a probability distribution $P_\theta$ for some $\theta \in \Theta \subseteq \mathbb{R}^d$, and suppose $g(\theta)$ is the quantity of interest. In a Bayesian study, a prior probability measure $\pi(\cdot)$ on $\Theta$ is used, then typically a Monte Carlo sample $\theta = (\theta_1, \ldots, \theta_m)$ is generated from the posterior distribution $\pi(\cdot|\mathbf{X})$. Posterior quantiles may then be approximated using the order statistics of $g(\theta_1), \ldots, g(\theta_m)$, if $g(\theta) \in \mathbb{R}$. However, if $g(\theta)$ is two or higher dimensional vector, obtaining its quantiles or a credible set becomes challenging.

Similarly, if $\tilde{g}(\mathbf{X})$ is an estimator of $g(\theta)$, bootstrap-based inference will typically proceed by obtaining the Monte Carlo sample $(\tilde{g}(\mathbf{X}_1^*), \ldots, \tilde{g}(\mathbf{X}_m^*))$, where $\mathbf{X}_i^*$'s are the resamples of $\mathbf{X}$. Then, functionals of the distribution of $\tilde{g}(\mathbf{X})$ can be evaluated empirically in a straightforward way, but if $g(\theta)$ is two or higher dimensional, obtaining its bootstrap-based confidence region is problematic.

One of the motivating factors for empirical likelihood techniques is that bootstrap confidence sets could not be constructed easily in multi-dimensions. Hence, Owen [27] uses the bootstrap only for calibration. Our methods offer a solution to the open problem of constructing multidimensional bootstrap confidence sets, that are different from the depth-based approach advocated by Yeh and Singh (1997).

As an example, consider the data on prey of dippers considered in [27]. There, in Fig. 1, 95% confidence regions, constructed from empirical likelihood and Normal theory, are presented for the bivariate means of (Caddis fly larvae, Stonefly larvae) and (Mayfly larvae, other invertebrates). In Fig. 1, in the top panel we present the bootstrap-based 95% confidence set for the same problem. Notice the lack of convexity for the 95% confidence set for the mean of (Caddis fly larvae, Stonefly larvae), a feature not revealed by the empirical likelihood based region or the Normality-based region. We would like to emphasize that if a convex confidence set is desired, our algorithm can handle that as well with minor changes in the computer code. The extreme variability in the dipper-prey data suggests that median might be better choice of a location parameter to consider, and the bottom panels of Fig. 1 show the 95% confidence sets for the bivariate medians.

We describe *multivariate quantile regression* briefly below. Suppose the $i$th response is the vector $Y_i \in \mathbb{R}^p$, while the $i$th covariate is the matrix $X_i \in \mathbb{R}^p \times \mathbb{R}^d$. Thus, the data consists of $\{(Y_i, X_i) \in \mathbb{R}^p \times (\mathbb{R}^p \times \mathbb{R}^d), i = 1, 2, \ldots, n\}$. Multivariate quantile regression models the $u$th quantile of $Y_i$ as a linear transformation of $X_i$. Adopting notations as earlier of $\beta = \|u\|, U = u/\|u\|$ and for any vector $Z \in \mathbb{R}^p$ that $Z_U = \langle Z, U\rangle U$ and $Z_{U^\perp} = Z - Z_U$, we define the $u$th quantile regression vector $\gamma_u \in \mathbb{R}^d$ as the argument that minimizes

$$\sum_{i=1}^{n}[\|Y_{iU} - q_{iU}\|[1 + \lambda(Y_{iU} - q_{iU})^{-2}\|Y_{iU^\perp} - q_{iU^\perp}\|^2]^{1/2} + \beta(X_{iU} - q_{iU})],$$

where $q_i = X_i\gamma_u$. The *simple multivariate quantile regression* case is obtained when $d = 1$. We obtain the classical univariate quantile regression of Koenker and Bassett [20] as a special case with $p = 1$. Properties of the quantile regression estimator can be derived easily from Section 3.1. The above framework assumes that quantiles of each element of the $p$-dimensional response are dependent on $d$ covariates. This assumption can be dropped and the number of covariates allowed to vary for each element; while the development is easy the algebra is unwieldy.

### 3.3. Multivariate order statistics, data-depth and outliers

We now introduce the notion of an *order statistic* in the context of generalized spatial quantiles. Recall that for a size $n$ sample of real-valued data, the $j$th order statistic is the value below or equal to which $j$ observations fall, and above which $n-j$ observations fall. The elementary transformation $\alpha = j/n \in (0, 1]$ may be used to restate the above notion in terms of the $\alpha$th order statistic, i.e., the value below or equal to which $n\alpha$ observations fall, and above which $n(1-\alpha)$ observations fall. For $\mathbb{R}^p$-valued multivariate data $X_1, \ldots, X_n$, instead of indexing the order statistics by the values $\alpha \in (0, 1]$, we index each observation $X_i$ according to a vector $u_i \in \mathcal{B}_p$, such that $X_i = Q_n(u_i, \lambda)$. That is, for every fixed $\lambda \geq 0$, observation $X_i$ is the $u_i$th order statistic for that value $u_i \in \mathcal{B}_p$ for which it minimizes

$$\sum_{j=1}^{n}[\|X_{jU} - q_U\|[1 + \lambda(X_{jU} - q_U)^{-2}\|X_{jU^\perp} - q_{U^\perp}\|^2]^{1/2} + \beta(X_{jU} - q_U)],$$

where $X_{jU} = \langle X_j, u\rangle/\|u\|$ and $X_{jU^\perp} = X_j - X_{jU}u/\|u\|, j = 1, \ldots, n$. Thus, the order statistics are indexed by directions as well as norms of vectors in the unit sphere in $\mathbb{R}^p$.

For illustration, let us consider the $\lambda = 0$ case corresponding to the projection quantiles. Here, an observation $X_i$ is the $u_i$th order statistic if the sample projection quantile corresponding to $u_i$ is $X_i$ itself. In other words, for projection

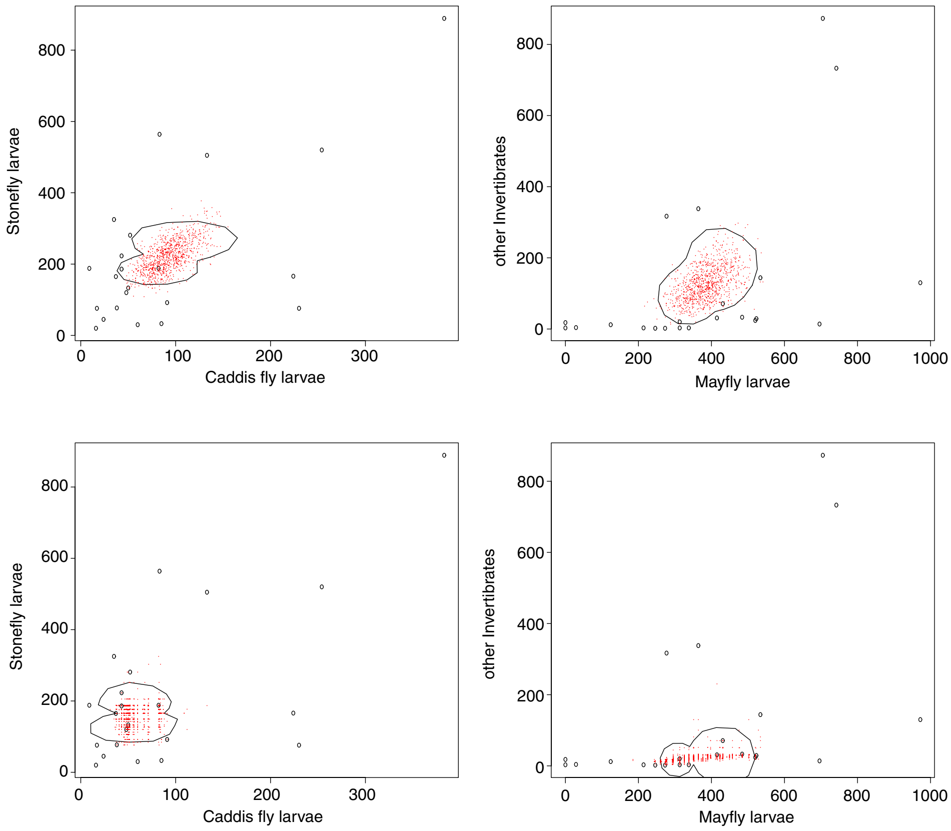

**Fig. 1.** Dipper data with 95% projection confidence interval for mean (top row) and median (bottom row).

quantiles, observations and their order statistic indices correspond to the sample version of Theorem 3.2. Hence, $u_i = \frac{X_i}{\|X_i\|}(2G_{nX_i}(\|X_i\|) - 1)$, where

$$G_{nx}(z) = n^{-1}\sum_{j=1}^{n} I_{\left\{\frac{\langle X_j, x\rangle}{\|x\|} \le z\right\}}, \quad z \in \mathbb{R}.$$

There are several applications of the above notion of a multivariate order statistic. Define the direction of an order statistic $U_i = u_i/\|u_i\|$ and its norm $\beta_i = \|u_i\|$ for reference. First, all the usage of one-dimensional order statistics and ranks, and similar other univariate summarizations of the data may be carried out for multivariate data by associating $\beta_i$ with $X_i$ (and ignoring $U_i$). A new notion of data-depth may be developed, with a function of $\beta_i$ being the depth of $X_i$. Such depths may be used to define another confidence set for multivariate random variables, extending the work of Yeh and Singh [33]. The $\beta_i$'s may also be used for outlier detection. The $U_i$'s are *directional data*, and can be used for testing whether the data shows spherical symmetry, for example. Tests for multivariate Normality may also be devised using the $U_i$'s and the $\beta_i$'s. Robust analysis of multivariate data, including robust estimation and inference, may be carried out using the above notion of order statistics. These applications will be pursued in future research.

### 3.4. Fast computing of generalized spatial quantiles

The computation of projection quantiles $Q_{\text{proj}}(\cdot)$ is immediate, and does not require the sample size to be greater than the dimension of the data. However, for arbitrary generalized spatial quantiles, a Newton–Raphson type algorithm may be used when $n > p$, and for the case of $n \le p$ an exhaustive grid-search needs to be carried out for exact computation. Neither alternative is attractive, or viable in high dimensions, hence we present below a coordinate descent algorithm to approximate any generalized spatial quantile, which is applicable regardless of the relationship between $n$ and $p$, or the choice of $\lambda$ and $\beta$. Recall that generalized spatial quantiles are obtained by minimizing $\sum_{i=1}^{n} \Psi_{u,\lambda}(X_i, q)$. Our coordinate descent algorithm iterates the following steps till convergence:

1. Start with a tentative minimizer $q^{(0)} = (q_1^{(0)}, \ldots, q_p^{(0)})^T$ of $\sum_{i=1}^{n} \Psi_{u,\lambda}(X_i, q)$. The projection quantiles $Q_{\text{proj}}(u)$ may be used for this initial value.

**Table 1**
Expectation and standard deviation of the relative approximation error (E(Rel. Err) and (SD(Rel. Err))) *scaled by a factor of* $10^5$, and the number of iterations (E(iter) and SD(iter)) of the coordinate-wise updating algorithm for normal data for dimensions 2, 4, 6 and 8 and $\lambda = 0.5, 1, 1.5$.

| | Dimension | 2 | 4 | 6 | 8 |
|---|---|---|---|---|---|
| $\lambda = 0.5$ | E(Rel.Err) | 0.92 | 0.74 | 0.70 | 0.66 |
| | SD(Rel. Err) | 1.17 | 0.31 | 0.28 | 0.30 |
| | E(Iter) | 5.70 | 5.03 | 4.95 | 4.98 |
| | SD(Iter) | 1.10 | 0.26 | 0.22 | 0.14 |
| $\lambda = 1.0$ | E(Rel.Err) | 0.93 | 0.64 | 0.52 | 0.36 |
| | SD(Rel.Err) | 2.01 | 0.29 | 0.25 | 0.22 |
| | E(Iter) | 5.39 | 4.97 | 4.88 | 4.81 |
| | SD(Iter) | 1.14 | 0.17 | 0.33 | 0.39 |
| $\lambda = 1.5$ | E(Rel.Err) | 0.62 | 0.56 | 0.36 | 0.30 |
| | SD(Rel. Err) | 0.41 | 0.24 | 0.22 | 0.16 |
| | E(Iter) | 5.33 | 4.91 | 4.73 | 4.56 |
| | SD(Iter) | 0.77 | 0.29 | 0.45 | 0.50 |

2. For each coordinate $i \in \{1, \ldots, p\}$, sequentially consider $\sum_{i=1}^{n} \Psi_{u,\lambda}(X_i, q)$ to be a function of $q_i$ *only* for minimization, and obtain $q_i^{(1)}$ as its minimizer, for $i = 1, \ldots, p$.
3. At the end of the above step, a new vector $q^{(1)} = (q_1^{(1)}, \ldots, q_p^{(1)})^T$ is obtained. Convergence is achieved if the distance between $q^{(1)}$ and $q^{(0)}$ is small, otherwise the above steps are repeated with $q^{(1)}$ in place of $q^{(0)}$.

To check the performance of the above computation method, we implemented it for multivariate Normal data. We simulated 200 multivariate normal random numbers in dimensions ranging between 2 and 8, and computed the correct generalized spatial quantile directly using multidimensional optimization using "nlm" function in the software `R, version 2.11.1` and also using the above algorithm. We computed the approximation error of our algorithm and the number of iterations it takes to converge. We defined an iteration as one revision of all the coordinates of the quantile, that is, for Step 2 above being implemented for all $i = 1, \ldots, p$. The *relative error in approximation* is defined as the Euclidean norm of the difference between the generalized spatial quantile and the approximation obtained by the above algorithm, divided by the norm of the generalized spatial quantile. We use $u = (0, \ldots, 0, 0.8)$ for this simulation. The results reported below are not affected by our choice of $u$, since the coordinate descent methodology is invariant to the choice of $u$. We considered $\lambda = 0.5, 1, 1.5$.

This exercise is repeated 100 times, and the average (E(Rel. Err)) and the standard deviation (SD(Rel. Err)) of the relative error in approximation scaled up by $10^5$, and the average (E(Iter)) and standard deviation (SD(Iter)) of the number of iterations required are reported in Table 1. Note that approximation errors are $O(10^{-5})$ in about 5 iteration steps, thus the above algorithm performs excellently. Also, the number of iterations required does not increase with the dimension. However, since each iteration in $p$ dimensions involves $p$ implementations of Step 2, the actual number of optimizations carried out increases linearly with dimension.

## 4. Simulation examples and applications

We divide this section in three parts. In the first part, we compare three generalized spatial quantiles, and the halfspace depth due to [30] with four bivariate densities. We compare the volumes of 80% coverage sets from each of these four methods, as well as their shape features.

In the second part, we present our projection quantile-based analysis of the DBS electrode placement experiment. We report the 90% confidence set for the region of the human brain where cognitive ability improvement of 50% or more have been reported. This image clearly shows an asymmetric, non-convex figure, which is in close correspondence of the geometry of the human brain, and in accordance with the medical knowledge relating to Alzheimer's disease.

In the third part, we use projection quantiles-based order statistics and Tukey's depth on a microarray experiment, to identify genes that display extraordinary behavior in human cancer cell-cycle regulation. This example illustrates that prohibitive computational requirements for data-depth, and the inherent features of high-dimensional data, result in too many points having discretized, low depth values, which results in poor quality inference.

### 4.1. Comparative inference with quantiles and depth

Data from four bivariate density functions are used for our simulation experiment on comparison of coverage sets obtained by different quantile-based and depth-based methods. The density functions are: (1) the standard bivariate Normal distribution, with the marginals being standard normal and with zero correlation between the two variables, (2) an even mixture of two bivariate Normal components, with the means being (−2, 5) and (2, 5), all variances equal to one and the two correlations being −0.75 and 0.75:

$$0.5N\left(\begin{pmatrix}-2\\5\end{pmatrix}, \begin{pmatrix}1 & -0.75\\-0.75 & 1\end{pmatrix}\right) + 0.5N\left(\begin{pmatrix}2\\5\end{pmatrix}, \begin{pmatrix}1 & 0.75\\0.75 & 1\end{pmatrix}\right),$$

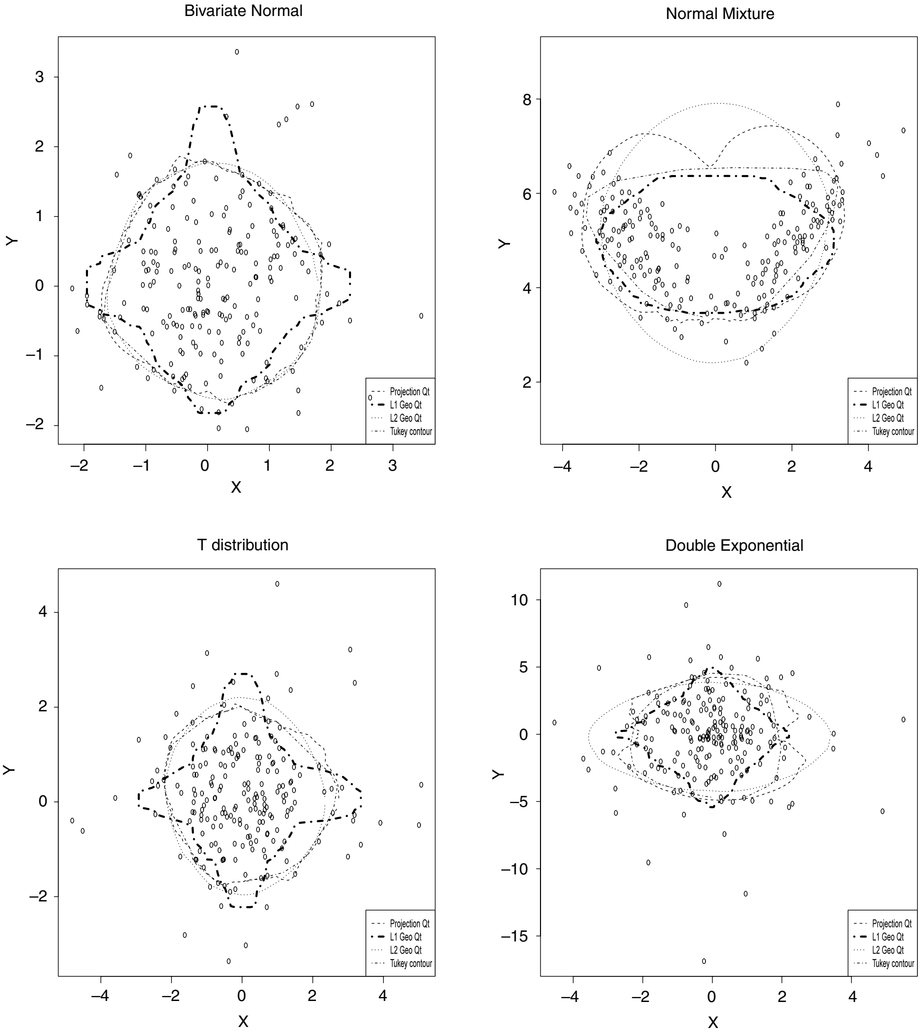

**Fig. 2.** Generalized spatial quantiles in all directions corresponding to 90% coverage for four distributions.

(3) a standard bivariate-$T$ distribution with 5 and 10 degrees of freedom for $X$ and $Y$ coordinates, and (4) a standard bivariate double exponential.

We generate a sample of size $n = 200$ from each of these distributions, and compute the 80% coverage regions obtained by using the projection quantile, generalized spatial quantile using the $L_1$-norm and $\lambda = 1$, and Chaudhuri's quantiles, which correspond to the $L_2$-norm and $\lambda = 1$. We also use the package `depth` in R to obtain the 80% coverage region by Tukey's depth, according to the principle of Yeh and Singh [33].

Table 2 provides the coverage and volume of the regions enclosed by 80% coverage sets in the four distributions. Notice that for all the methods the volumes differ across distributions, but are very similar to each other for the bivariate Normal and Student's-$t$ distributions. The shape characteristics in the mixture Normal and the double exponential distributions create some difference in the volumes. However, note from Fig. 2 the difference in shape features of the four coverage sets. The projection quantile method captures the approximate shape of the data-cloud in all the cases and can be non-convex, while the Tukey depth-based and Chaudhuri's quantile-based sets are always near-ellipsoid convex sets. Based on volume alone, the $L_1$-norm based generalized spatial quantile method seems best, with the projection quantile method being a close second.

### *4.2. Analysis of the DBS electrode placement data*

The data for this experiment consists of the locations in the brain where the DBS electrodes have been placed, and a binary variable indicating whether more than 50% improvement in cognitive ability has resulted from the brain stimulation. The locations of the DBS electrode placement are given with respect to a common coordinate system defined by the anterior commissure (AC) and posterior commissure (PC) planes and the midline. The medical interest in this experiment centers

**Table 2**
Volume of the 80% coverage sets in four simulated populations.

| Algorithm | Biv. nor. | Mix nor. | $T$ dist | Double exp. |
|---|---|---|---|---|
| Projection | 9.12 | 22.77 | 14.00 | 36.35 |
| $L_1$ Geometric | 9.30 | 16.52 | 15.18 | 36.30 |
| $L_2$ Geometric | 8.94 | 25.18 | 14.86 | 42.59 |
| Tukey depth | 9.14 | 18.34 | 14.66 | 48.46 |

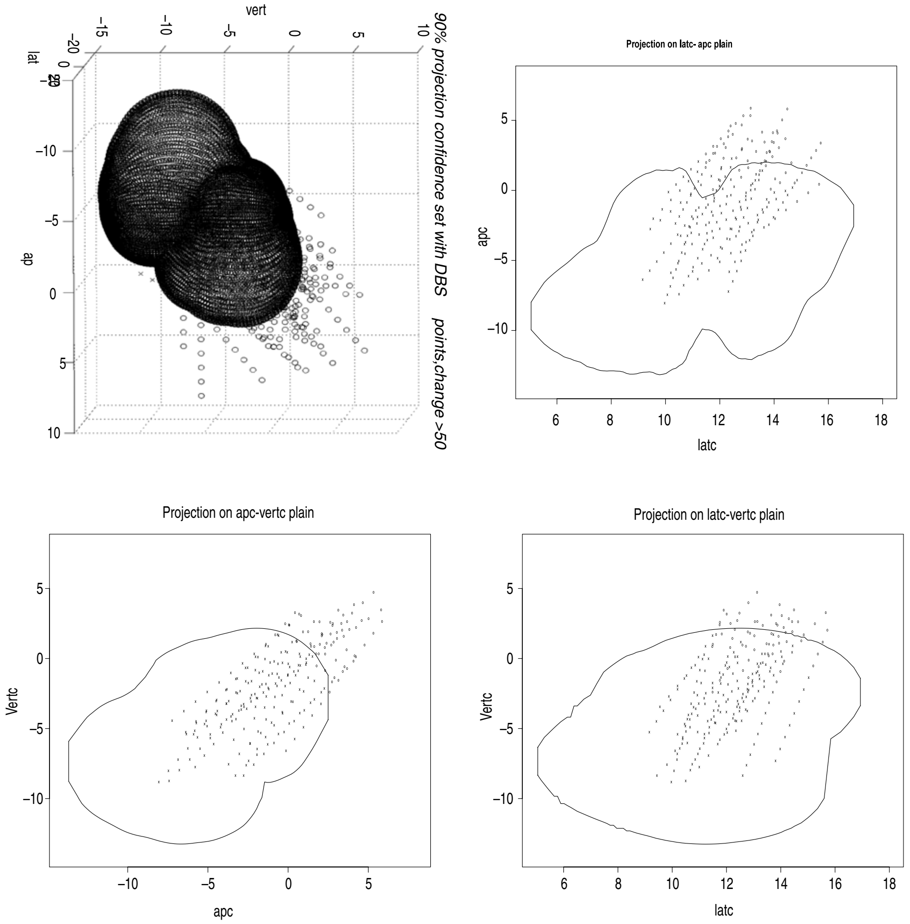

**Fig. 3.** 90% 3D confidence Set of the final DBS location for patients with 50% or more improvement in cognitive ability and projection of the set on lat-ap, ap-vert and lat-vert plains. Also the path of the electrodes are shown.

around the efficacy of the DBS electrode-based treatment for long term improvement in the cognitive ability of a patient suffering from Alzheimer's disease. It is thought that some of the region surrounding the nucleus of the brain should be stimulated for long term improvement; however, the shape or the size of this region is unknown. Our goal here is to map the region of the brain where 90% or more success (defined as $>50\%$ improvement in cognitive ability) has been reported. We obtain the region using projection quantiles, by varying $\beta$ such that 90% coverage is achieved, as described in Section 3.2. This region is displayed in Fig. 3, which also includes the trajectory of the insertion path of some of the electrodes. We also present three two-dimensional cross-sectional plots in the same figure, for greater clarity. Note that the shape of the 90% confidence region in Fig. 3 is irregular, and is neither convex nor symmetric about a point or a line or a plane. However, it closely imitates the shape of the nucleus of the human brain. The region of the brain thus identified from the data using projection quantiles is in agreement with the opinion of scientists and physicians studying Alzheimer's disease; however, biological knowledge about the human brain is still scant.

### 4.3. Gene behavior in cell-cycle regulation experimentation

We consider the data on the human cancer cell (HeLa S3) cycle data, available at http://genome-www.stanford.edu/Human-CellCycle/Hela for this part of our analysis. In this particular data, [32] identify 1134 genes out of a total of 42 920 as

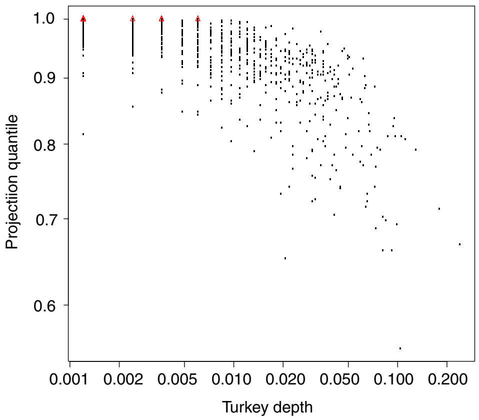

**Fig. 4.** Projection quantile and Tukey depth measure of each periodic gene in cell cycle data. The top 35 genes are indicated by triangles.

periodic, or cell-cycle regulators, based on a periodicity analysis of the marginal distributional behavior of each gene. Gene-network causality and related dependence across pairs of genes has been reported in [25]. Several other studies report other low-dimensional patterns of genes in this or similar datasets, for example, through the computation of various kinds of correlations between gene-pairs.

Here we are interested in identifying those genes that stand out, compared to the overall data cloud of gene expressions, and thereby are of interest in understanding the cell-cycle regulatory mechanism. A parametric distribution for the underlying population of gene expressions is not easy to express, use as a statistical model, verify in practice, or justify on biological grounds. We use ranking based on the projection quantiles to identify those genes that correspond to extreme quantiles. Also, a ranking based on Tukey's depth is obtained.

Here we report our results on experiment 1 of the `Hela S3` cell cycle data, which has $p = 11$ time points over which the expressions of the different genes have been obtained. Spellman et al. [29] showed that of the 15 536 genes studies in this experiment, $n = 828$ are periodic, and are candidates for a possible role in cell-cycle regulation. We ignore the genes that have not been identified as periodic, since they lack relevance in the biological process. For each gene $g$ among the 828 periodic genes, we compute its projection order statistic $u_g$; i.e. the vector $u_g \in \mathcal{B}_{11}$ such that the sample projection quantile with respect to $u_g$ is the gene expression $g$. Details of this method have been discussed in Section 3.3.

The genes that have the highest $\beta$ values are more significant. We set $\beta \geq 0.9999677$ as a cutoff point, based on computations for the projection quantile confidence region of 90% coverage for the standard Normal distribution in $\mathbb{R}^{11}$. However, the data clearly does not fit such a distributional pattern, and only 35 of the genes obtain $\beta \geq 0.9999677$. Tukey's depth for the $n = 828$ gene profiles in $\mathbb{R}^{11}$ and their $\beta$ values are presented in Fig. 4. The 35 identified genes with $\beta$ values above the cutoff are identified with triangles. Some of these 35 genes are also among those with the least Tukey's depth, but there are some genes with higher depth. However, notice that there are 179 genes among the 828 have the same minimum Tukey's depth. It is extremely unlikely biologically that 179 out of 828 genes would be influential in cell-cycle regulation, thus depth-based inference seems to be greatly influenced by false positives. The high number of genes with very low depth show that care must be taken with depth-based inference in high dimensions. Note that computing depths precisely in high dimension is virtually an impossibility; computing just the Tukey's median takes $O(n^{p-1})$ expected time [12]. The potential lack of precision in approximate depth computation may also lead to misleading results.

We may compare this list of genes with those of Li et al. [21], where 20 genes have been studied, that are thought to be associated with human cell-cycle regulatory pattern [14]. Eighteen of these genes are part of our set of 828 genes, and three of these are obtained among the 35 most significant genes that were significant according to our projection order statistic based analysis. These genes are PCNA, PLK and CDC20. A match of three out of eighteen possible genes serves as a strong reinforcement of the utility of our approach. Significantly, the Tukey depth-based approach fails to place PCNA among its large list of 179 genes with lowest depth, although it identifies correctly PLK and CDC20 as significant genes.

## 5. Discussion

The analysis of high dimensional data is a challenging area of research. The traditional approach is to model the data in a probabilistic framework that is often just convenient for the statistician, and/or to replace the high dimensional open problems with lower dimensional ones. Our approach of using generalized spatial quantiles for summarization, estimation and inference is one possible avenue, which neither requires arbitrary probabilistic assumptions nor takes recourse to reduce the data reducing to lower dimensional structures. We have aimed to build on earlier attempts at using geometric quantiles and related methods, and have tried to integrate several approaches towards a quantile-based analysis of multivariate data that have a commonality between them.

The concept of the generalized spatial quantile provides a common platform showing the connection between the projection quantile and the geometric quantile. Interpretation and applicability of the λ parameter is under study at the moment. The computation of the generalized spatial quantile, by means of the iterative algorithm has been demonstrated to work well in examples. Further research is under way to understand the effect of different choices of λ and the effect of outliers on these quantiles and breakdown properties.

There are unique challenges in analyzing multimodal data in high dimensions, and in this paper we have not addressed the issue of multimodality or mixture distributions, other than in a small way in Section 4.1. Depending on the application at hand, one option is to use a classification or clustering step, followed by computing multivariate quantiles in each cluster separately. Projection quantiles may turn out to be particularly useful, since they do not require any condition linking cluster size with data dimensions.

## Acknowledgments

The second authors research is partially supported by grants from the University of Minnesota, and from the National Science Foundation of the USA.

## Appendix

For any $u \in \mathscr{B}_p \setminus \{0\}$, let us adopt the notation $F_{X_U}$ for the (absolutely continuous) distribution function of $X_U$. The following result is useful for proving Theorem 3.2.

**Lemma A.1.** *Under the conditions of Theorem 3.2, for every $x \in \mathbb{R}^p \setminus \{0\}$, $q_{Q_{\text{proj}}^{-1}(x)} = \|x\|$.*

**Proof of Lemma A.1.** For $x \in \mathbb{R}^p \setminus \{0\}$, let $\tilde{x} = x/\|x\|$. Hence, $G_x(\cdot) = P\left(\frac{\langle X, x\rangle}{\|x\|} \leq \cdot\right) = F_{X_{\tilde{x}}}(\cdot)$ in our adopted notation. Note that $\|x\| > 0$, and since the spatial median is zero, we have that $G_X(\|x\|) > 1/2$. Hence, for $x \in \mathbb{R}^p \setminus \{0\}$, we have $\|Q_{\text{proj}}^{-1}(x)\| = (2G_x(\|x\|) - 1)$. Thus, $Q_{\text{proj}}^{-1}(x)/\|Q_{\text{proj}}^{-1}(x)\| = x/\|x\|$, and hence $Q_X = \langle X, Q_{\text{proj}}^{-1}(u)/\|Q_{\text{proj}}^{-1}(u)\|\rangle = \|x\|^{-1}\langle X, x\rangle = X_{\tilde{x}}$.

Thus we have

$$\begin{aligned} q_{Q_{\text{proj}}^{-1}(x)} &= (1+ \parallel Q_{\text{proj}}^{-1}(x))/2\text{-quantile of } Q_X \\ &= (1 + 2G_x(\|x\|) - 1)/2\text{-quantile of } Q_X \\ &= G_x(\|x\|)\text{-quantile of } X_{\tilde{x}} \\ &= F_{X_{\tilde{x}}}(\|x\|)\text{-quantile of } X_{\tilde{x}}, \end{aligned}$$

where, recall, the (absolutely continuous) distribution of $X_{\tilde{x}}$ is $F_{X_{\tilde{x}}}$.

Thus, the result is proved. □

**Proof of Theorem 3.2.** We show that for any $x \in \mathbb{R}^p \setminus \{0\}$, $Q_{\text{proj}}(Q_{\text{proj}}^{-1}(x)) = x$, and for every $u \in \mathscr{B}_p \setminus \{0\}$, $Q_{\text{proj}}^{-1}(Q_{\text{proj}}(u)) = u$.

We start with the first identity. Note that for any $x \in \mathbb{R}^p \setminus \{0\}$,

$$\begin{aligned} Q_{\text{proj}}(Q_{\text{proj}}^{-1}(x)) &= \frac{Q_{\text{proj}}^{-1}(x)}{\|Q_{\text{proj}}^{-1}(x)\|} q_{Q_{\text{proj}}^{-1}(x)} \\ &= \frac{x}{\|x\|} q_{Q_{\text{proj}}^{-1}(x)}. \end{aligned}$$

Use Lemma A.1 to establish that this is equal to $x$.

For the other identity, for any $u \in \mathscr{B}_p \setminus \{0\}$, note that $\|Q_{\text{proj}}(u)\| = q_u = F_{X_U}^{-1}((1 + \|u\|)/2)$. Thus we have $\|u\| = 2F_{X_U}(\|Q_{\text{proj}}(u)\|) - 1$. Also note that $\tilde{Q}_X = \langle X, Q_{\text{proj}}(u)/\|Q_{\text{proj}}(u)\|\rangle = \|u\|^{-1}\langle X, u\rangle = X_U$. Thus, $G_{Q_{\text{proj}}(u)}(\cdot) = \mathbb{P}[\tilde{Q}_X \leq \cdot] = \mathbb{P}[X_U \leq \cdot] = F_{X_U}(\cdot)$, and hence $2G_{Q_{\text{proj}}(u)}(\|Q_{\text{proj}}(u)\|) - 1 = 2F_{X_U}(\|Q_{\text{proj}}(u)\|) - 1 = \|u\|$.

Hence

$$\begin{aligned} Q_{\text{proj}}^{-1}(Q_{\text{proj}}(u)) &= \frac{Q_{\text{proj}}(u)}{\|Q_{\text{proj}}(u)\|}(2G_{Q_{\text{proj}}(u)}(\|Q_{\text{proj}}(u)\|) - 1) \\ &= \frac{u}{\|u\|}\|u\| \\ &= u. \quad \square \end{aligned}$$

**Proof of Theorem 3.3.** Note that if $-u$ is the diametrically opposite vector of $u$, we have $X_{(-u)} = \{-\langle X, U\rangle\}(-U)$ and thus $X_{(-U)} = -\langle X, U\rangle = -X_U$.

We assume that $c \in \{x \in \mathbb{R}^p : \|x\| = 1\}$. Note that $c^T X = \langle X, c\rangle \sim F_{C_X}(\cdot)$ by our notation. Along the line $\{\langle c, x\rangle : x \in \mathbb{R}^p\}$, the set $\mathscr{B}_{\beta,0}$ carves out the interval $(-q_{(-c)}, q_c)$, and we have $\mathbb{P}[c^T X \leq q_c] = (1 + \beta)/2$ and

$$\mathbb{P}[-c^T X \le q_{(-c)}] = (1+\beta)/2,$$
$$\Leftrightarrow \mathbb{P}[c^T X < -q_{(-c)}] = 1 - (1+\beta)/2 = (1-\beta)/2.$$

Thus, $\mathbb{P}[-q_{(-c)} \le c^T X \le q_c] = \beta$. □

## References

[1] A. Bose, S. Chatterjee, Generalized bootstrap for estimators of minimizers of convex functionals, J. Statist. Plann. Inference 117 (2003) 225–239.

[2] B.M. Brown, Statistical uses of the spatial median, J. R. Stat. Soc. Ser. B Stat. Methodol. 45 (1) (1983) 25–30.

[3] B.M. Brown, T.P. Hettmansperger, Affine invariant rank methods in the bivariate location model, J. R. Stat. Soc. Ser. B Stat. Methodol. 49 (3) (1987) 301–310.

[4] B.M. Brown, T.P. Hettmansperger, An affine invariant bivariate version of the sign test, J. R. Stat. Soc. Ser. B Stat. Methodol. 51 (1) (1989) 117–125.

[5] B. Chakraborty, On affine equivariant multivariate quantiles, Ann. Inst. Statist. Math. 53 (2) (2001) 380–403.

[6] B. Chakraborty, P. Chaudhuri, On an adaptive transformation–retransformation estimate of multivariate location, J. R. Stat. Soc. Ser. B Stat. Methodol. 60 (1) (1998) 145–157.

[7] B. Chakraborty, P. Chaudhuri, On multivariate rank regression, in: $L_1$-Statistical Procedures and Related Topics, Neuchâtel, 1997, in: IMS Lecture Notes Monogr. Ser., vol. 31, Inst. Math. Statist., Hayward, CA, 1997, pp. 399–414.

[8] B. Chakraborty, P. Chaudhuri, On affine invariant sign and rank tests in one- and two-sample multivariate problems, in: Multivariate Analysis, Design of Experiments, and Survey Sampling, in: Statist. Textbooks Monogr., vol. 159, Dekker, New York, 1999, pp. 499–522.

[9] B. Chakraborty, P. Chaudhuri, A note on the robustness of multivariate medians, Statist. Probab. Lett. 45 (3) (1999) 269–276.

[10] B. Chakraborty, P. Chaudhuri, On a transformation and re-transformation technique for constructing an affine equivariant multivariate median, Proc. Amer. Math. Soc. 124 (8) (1996) 2539–2547.

[11] B. Chakraborty, P. Chaudhuri, H. Oja, Operating transformation retransformation on spatial median and angle test, Statist. Sinica 8 (1998) 767–784.

[12] T.M. Chan, An optimal randomized algorithm for maximum Tukey depth, in: Proc. 5th ACM-SIAM Symposium on Discrete Algorithms, 2004, pp. 423–429.

[13] P. Chaudhuri, On a geometric notion of quantiles for multivariate data, J. Amer. Statist. Assoc. 91 (434) (1996) 862–872.

[14] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[15] S.J. Haberman, Concavity and estimation, Ann. Statist. 17 (1989) 1631–1661.

[16] J.B.S. Haldane, Note on the median of a multivariate distribution, Biometrika 35 (1948) 414–415.

[17] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, 2009.

[18] T.P. Hettmansperger, J. Nyblom, H. Oja, Affine invariant multivariate one-sample sign tests, J. R. Stat. Soc. Ser. B Stat. Methodol. 56 (1) (1994) 221–234.

[19] Peter J. Huber, Projection pursuit. With discussion, Ann. Statist. 13 (2) (1985) 435–525.

[20] R. Koenker, G. Bassett Jr., Regression quantiles, Econometrica 46 (1) (1978) 33–50.

[21] X. Li, et al., Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling, BMC Bioinformatics 7 (2006) 26.

[22] R.Y. Liu, On a notion of data depth based on random simplices, Ann. Statist. 18 (1990) 405–414.

[23] R.Y. Liu, J.M. Parelius, K. Singh, Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion), Ann. Statist. 27 (1999) 783–858.

[24] R.Y. Liu, K. Singh, A quality index based on data depth and multivariate rank tests, J. Amer. Statist. Assoc. 88 (1993) 252–260.

[25] N. Mukhopadhyay, S.B. Chatterjee, Causality and pathway search in microarray time series experiment, Bioinformatics 23 (2007) 442–449.

[26] W. Niemiro, Asymptotics for $M$-estimators defined by convex minimization, Ann. Statist. 20 (1992) 1514–1533.

[27] Art B. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (2) (1988) 237–249.

[28] C.G. Small, A survey of multidimensional medians, Int. Stat. Rev. 58 (1990) 263–277.

[29] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.

[30] J.W. Tukey, Mathematics and picturing data, in: R.D. James (Ed.), Proceedings of the International Congress on Mathematics, in: Canadian Math. Congress, vol. 2, 1975, pp. 523–531.

[31] Y. Vardi, C.-H. Zhang, The multivariate $L_1$-median and associated data depth, Proc. Natl. Acad. Sci. 97 (4) (2000) 1423–1426.

[32] M.L. Whitfield, G. Sherlock, A. Saldanha, J. Murray, C.A. Ball, K. Alexander, J. Matese, C.M. Perou, M. Hurt, P. Brown, D. Botstein, Identification of genes periodically expressed in the human cell cycle and their expression in tumors, Mol. Biol. Cell 13 (2002) 1977–2000.

[33] A. Yeh, K. Singh, Balanced confidence regions based on Tukey's depth and the bootstrap, J. Roy. Statist. Soc. Ser. B 59 (3) (1997) 639–652.

[34] Yijun Zuo, Multidimensional trimming based on projection depth, Ann. Statist. 34 (5) (2006) 2211–2251.

[35] Yijun Zuo, Hengjian Cui, Dennis Young, Influence function and maximum bias of projection depth based estimators, Ann. Statist. 32 (1) (2004) 189–218.

[36] Y. Zuo, R. Serfling, General notions of statistical depth function, Ann. Statist. 28 (2) (2000) 461–482.