

Methods for Streaming Data with Applications to Multivariate Density Estimation

Jim McDermott, (Bristol-Myers Squibb)

Dennis K.J. Lin, (Penn State University)

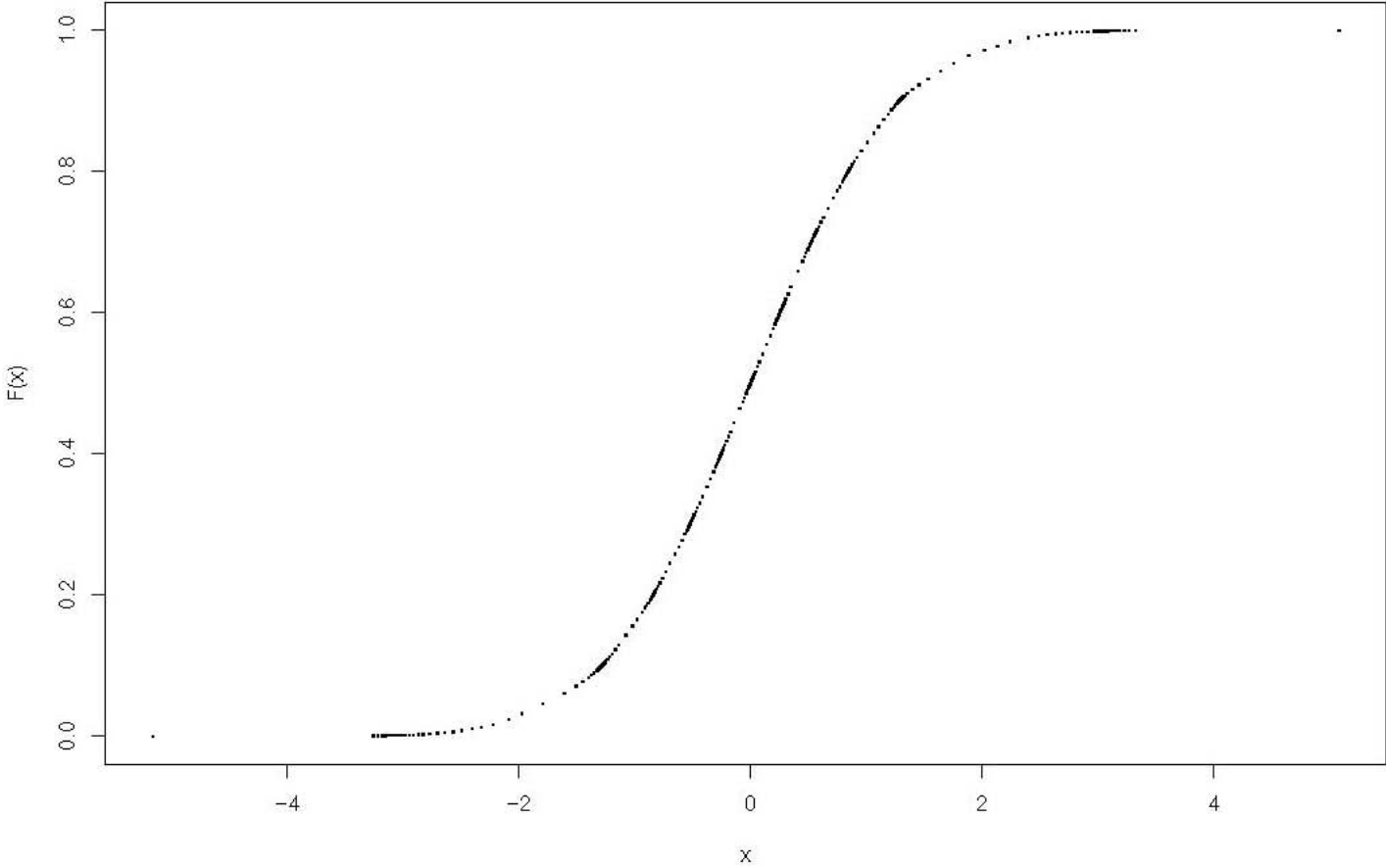
Basic Problem

- How do we keep accurate summaries of a data stream without storing too much of the data and without imposing too much of a computational burden on the system?

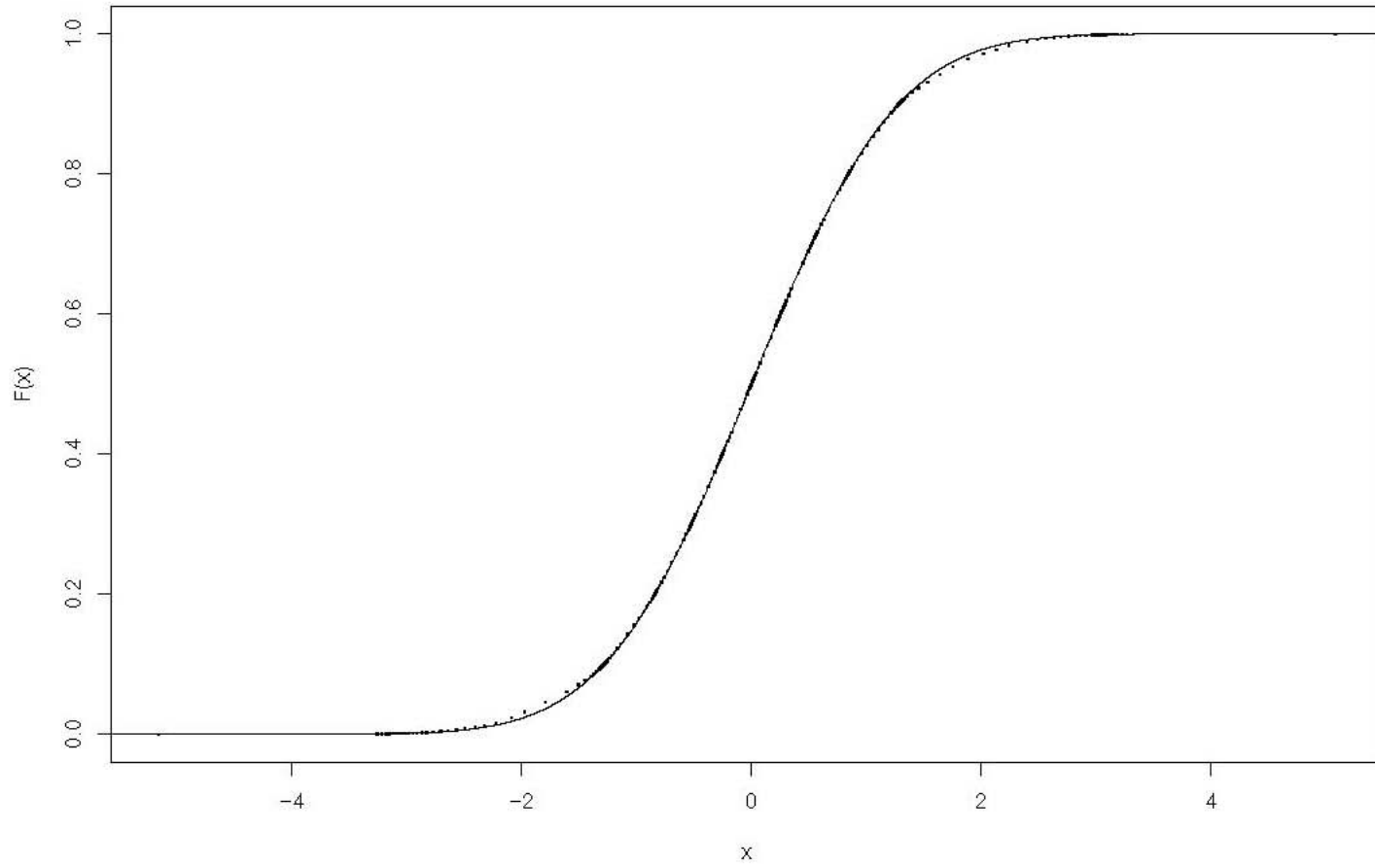
Summaries

- In particular, we would like to have quantile estimates and pictures of the cdf and/or density estimates.
- We have a restriction that we are only allowed to see each data point once.
- Further, we are not allowed to store the entire dataset.

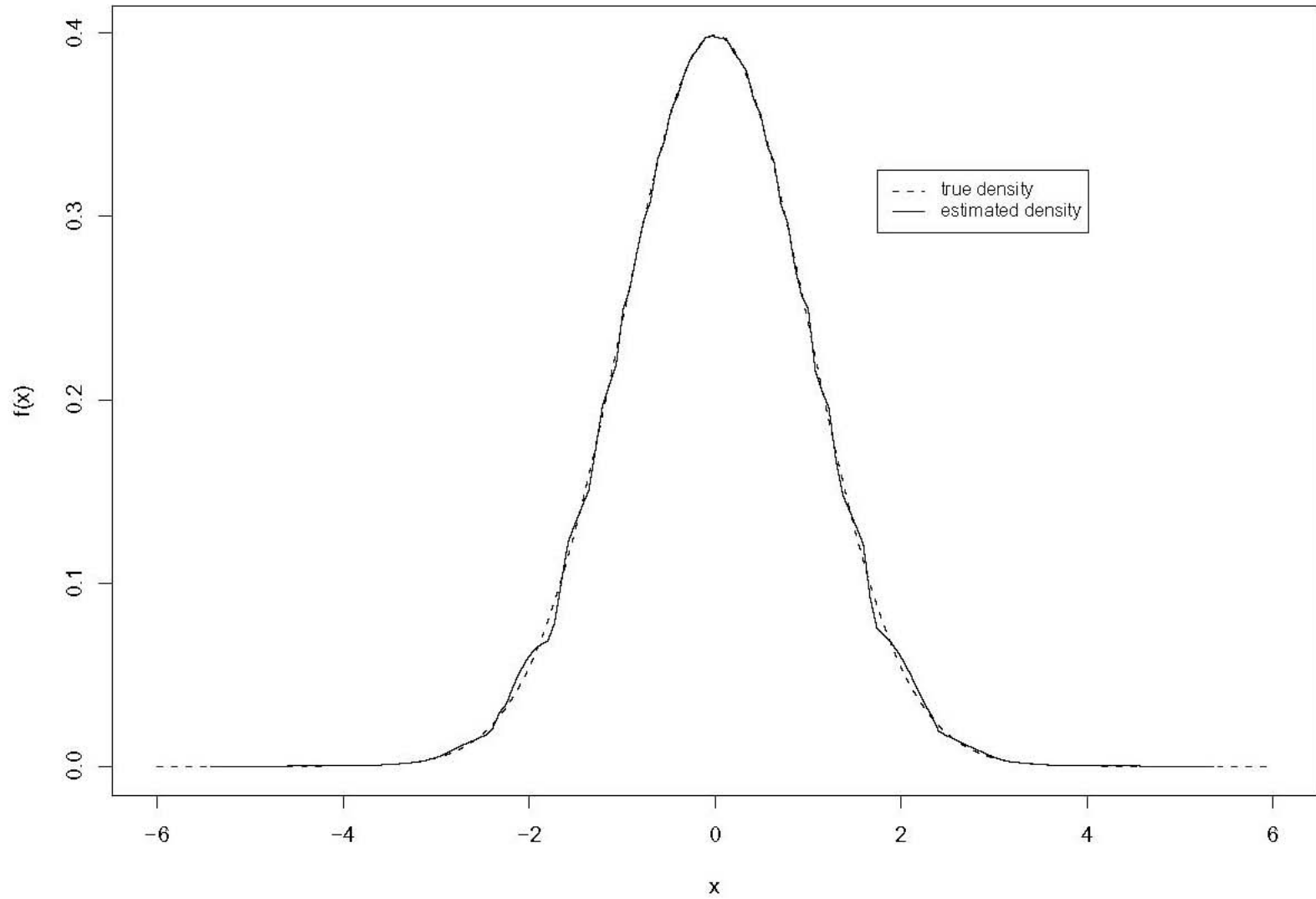
Normal Example with Multiple Quantiles



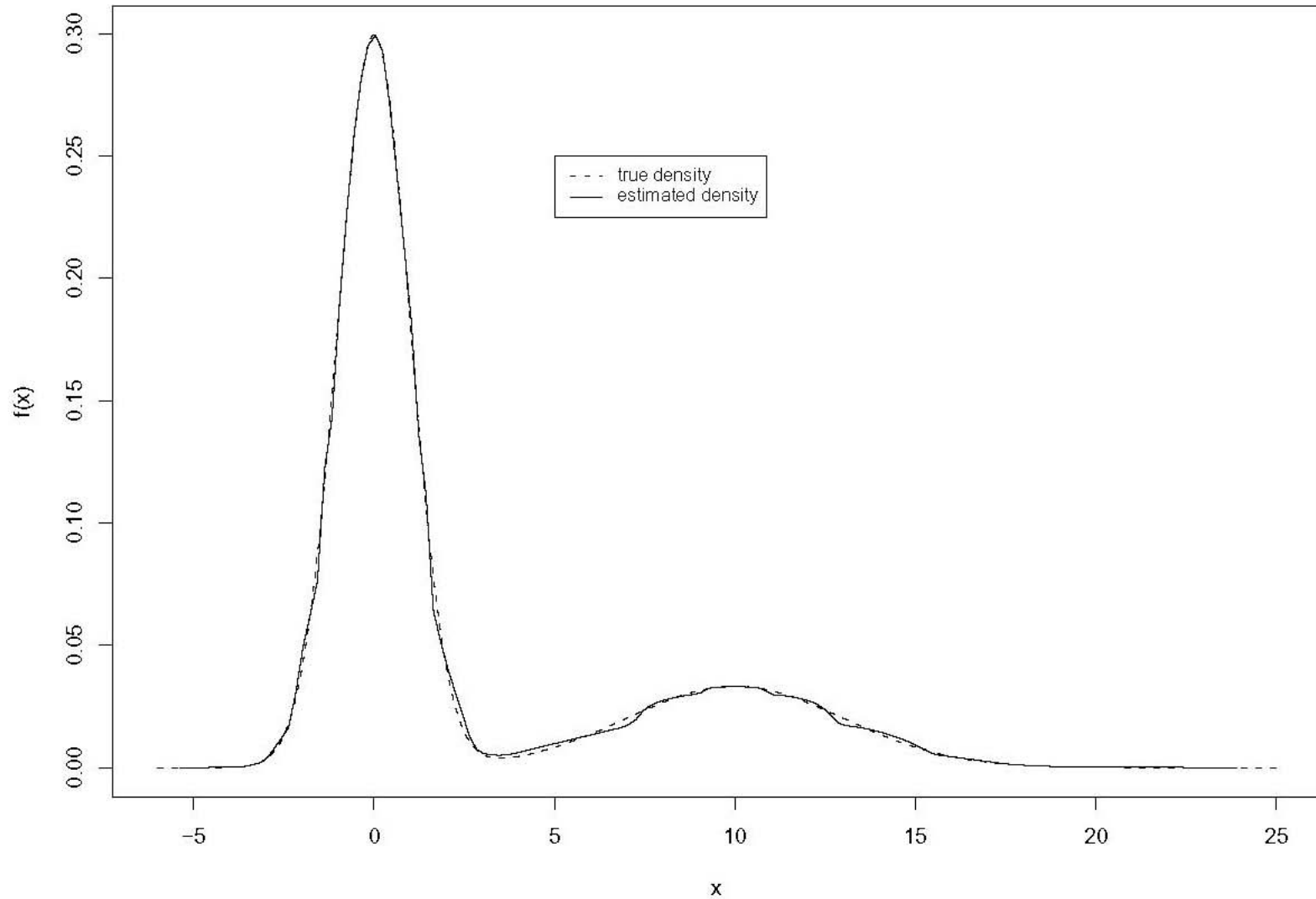
Normal Example with Multiple Quantiles and True CDF



Example: Standard Normal – 10,000,000 observations



Example: Mixture of 2 Normals – 10,000,000 observations



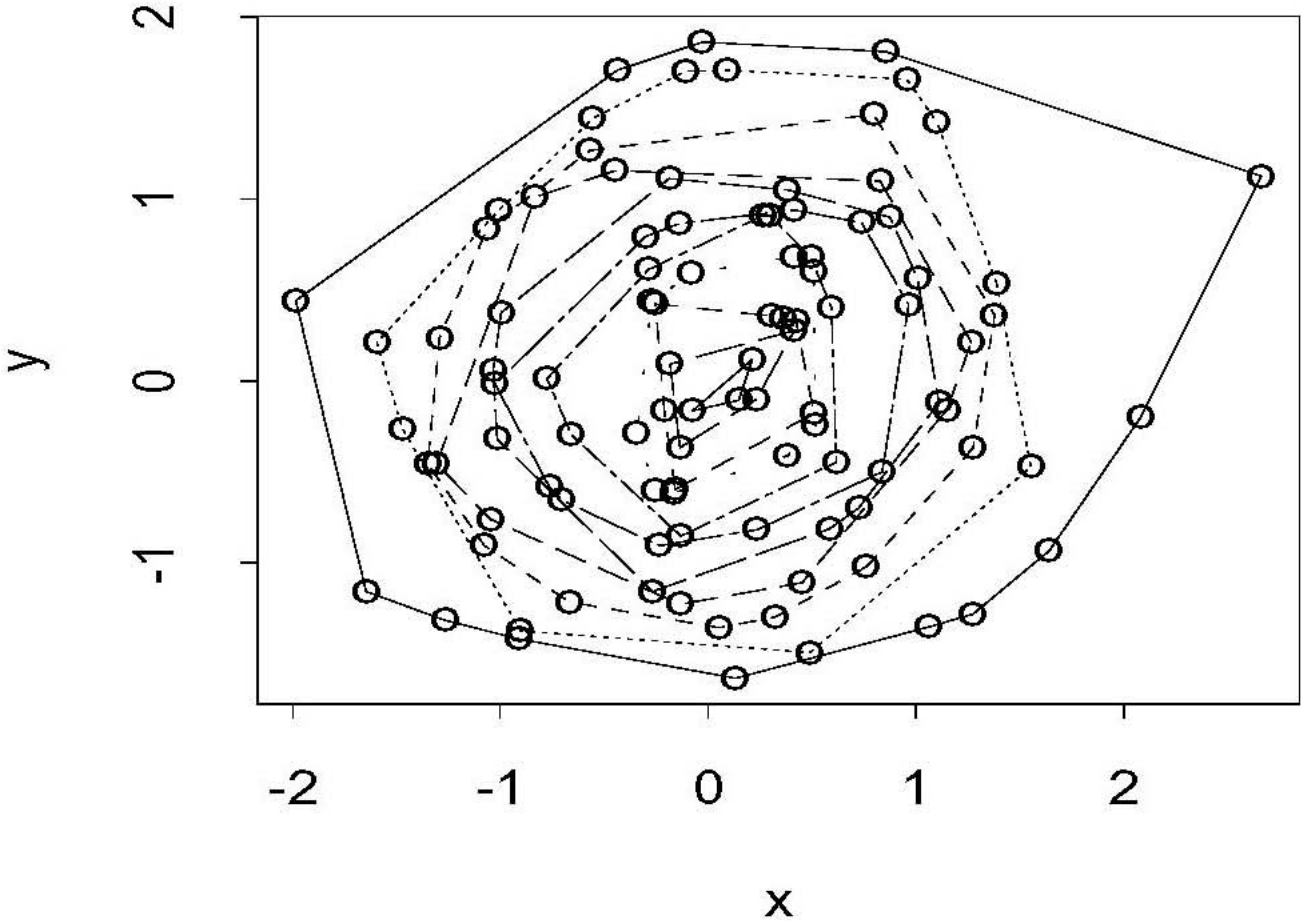
Multivariate Case

- The concept of a quantile is undefined in more than one dimension.
- Instead, we can think of a point's *depth*.
- For example, we can think of the multivariate median as the *most central* point in the dataset.
- There are many different definitions of data depth. We will focus on the convex hull peeling version.

Convex Hulls

- A convex hull in two dimensions is the polygon formed by the outer perimeter of a set of points.
- This definition can be extended to dimensions of three or greater.
- This method is computationally intensive and the entire dataset must be stored.

Convex Hull Peels



Convex Hull Peeling - Multivariate Median

To avoid the storage demands of performing convex hull peeling on a massive dataset, we propose the following modification.

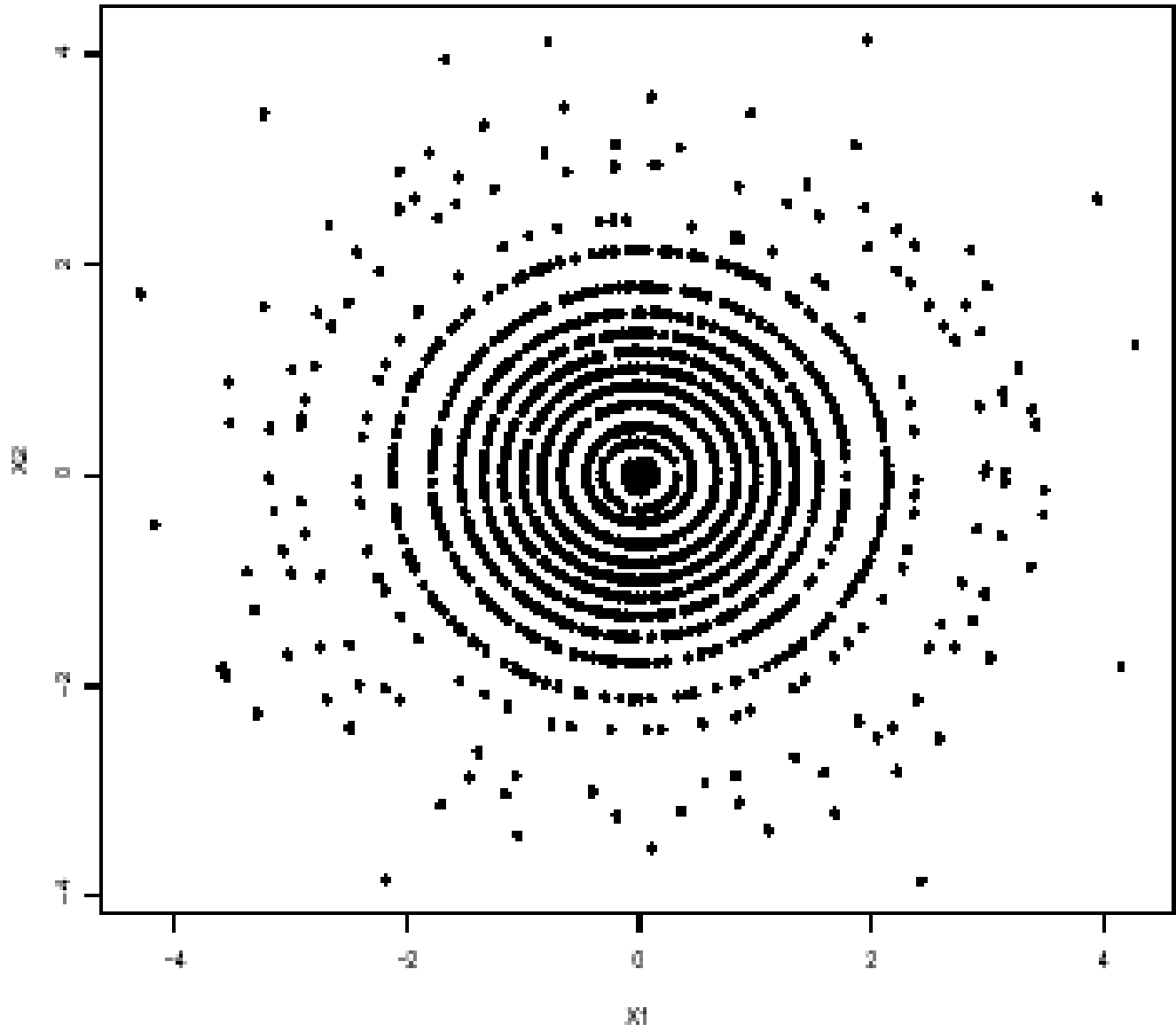
- Take the first 1000 points from the dataset and peel the layers until approximately 500 points are left.
- Add enough points from the remaining dataset to bring the number up to 1000 again.
- Repeat until the dataset is exhausted
- The centroid of the final hull is taken as an estimate of the multivariate median.

This method can be extended to dimensions higher than 2.

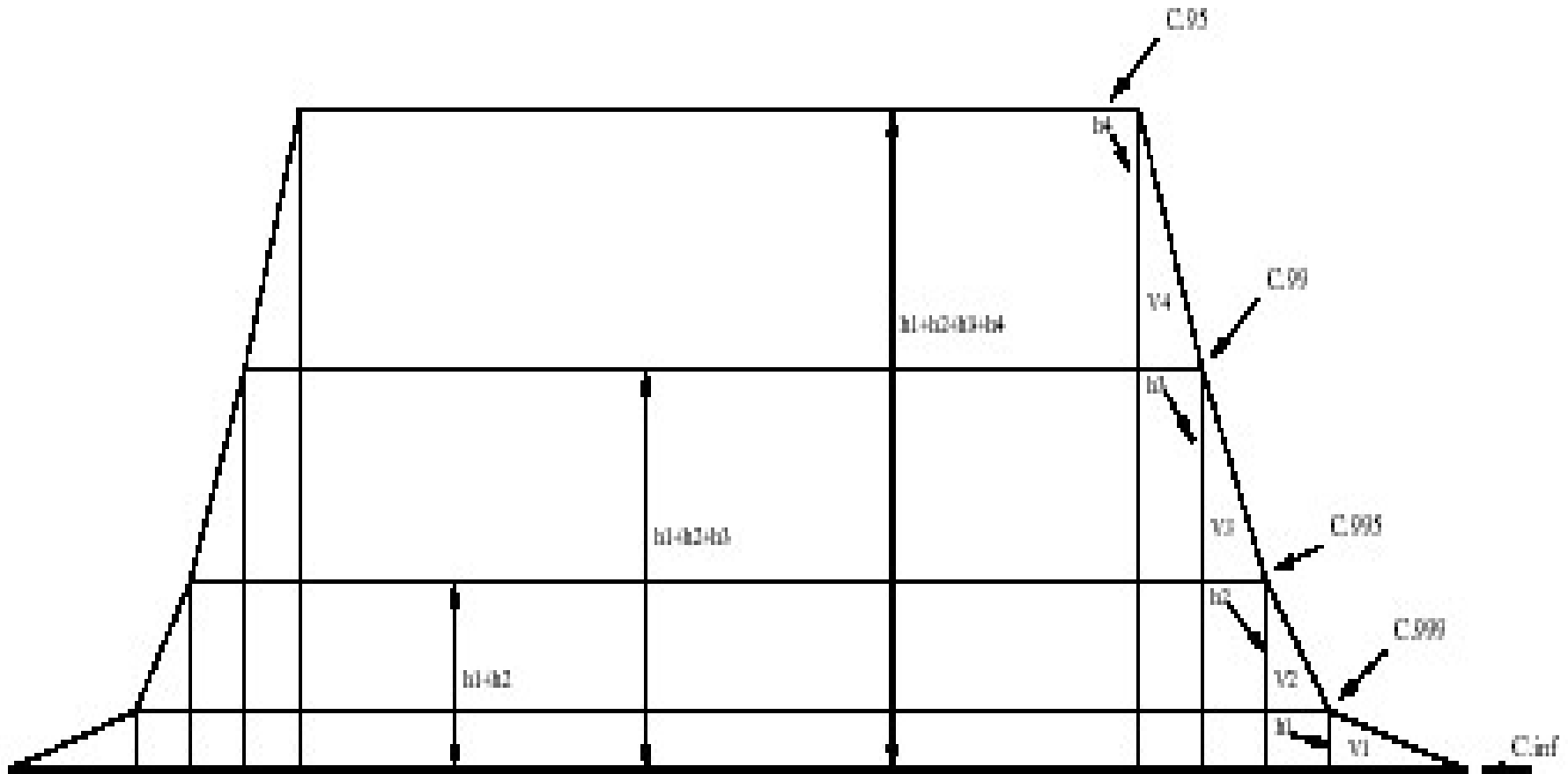
Convex Hull Peeling - Quantile Contours

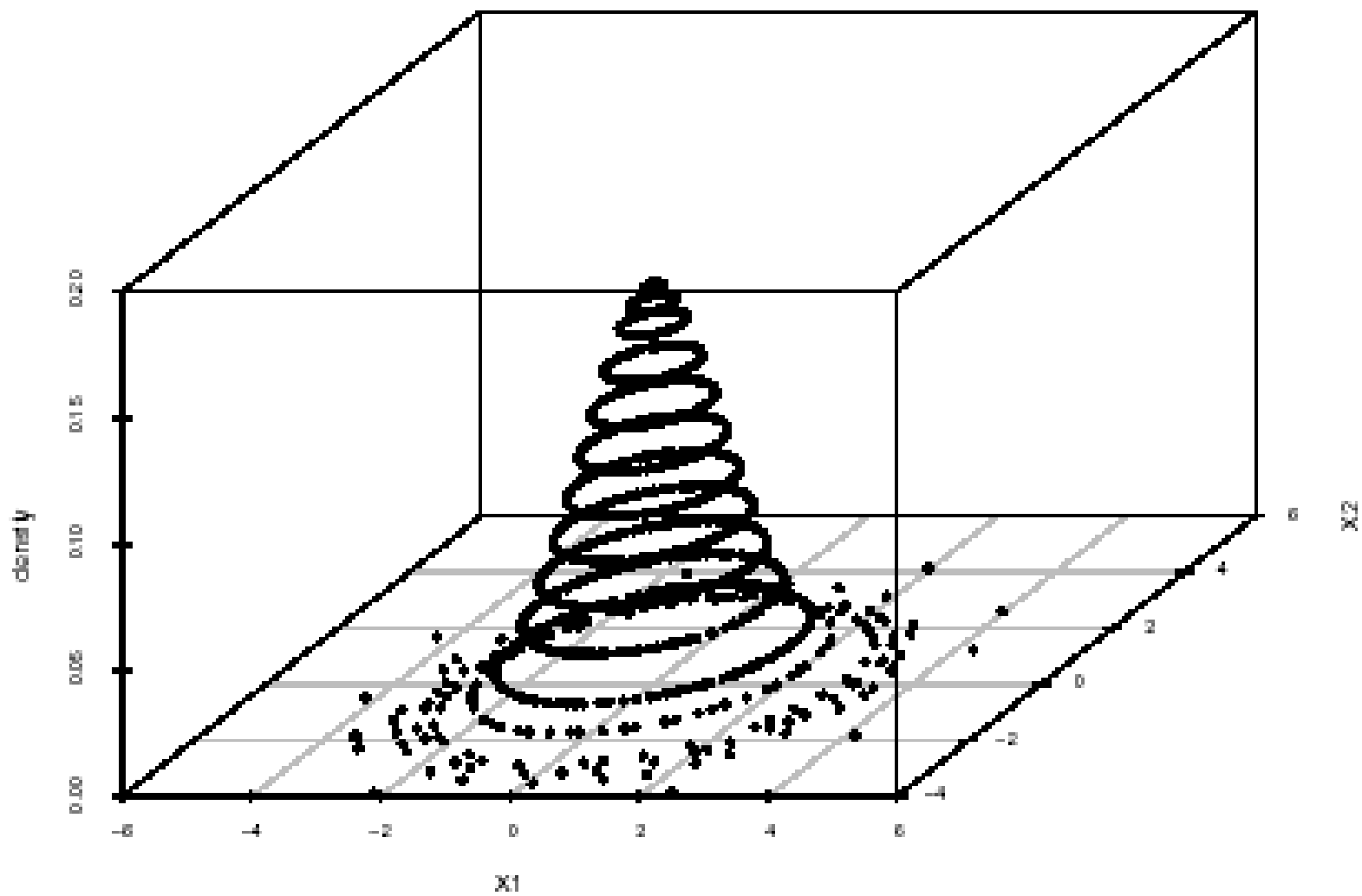
- Take the first 1000 points from the dataset and peel the layers until approximately $1000 \cdot p$ points are left, where $0 < p < 1$.
- Store the points that make up this hull
- Repeat and append the next set of points to the first set
- Repeat until the dataset is exhausted.
- Peel to the center of this remaining set of points until approximately $1/2$ of these points remain.
- This hull is your estimate of the quantile contour.

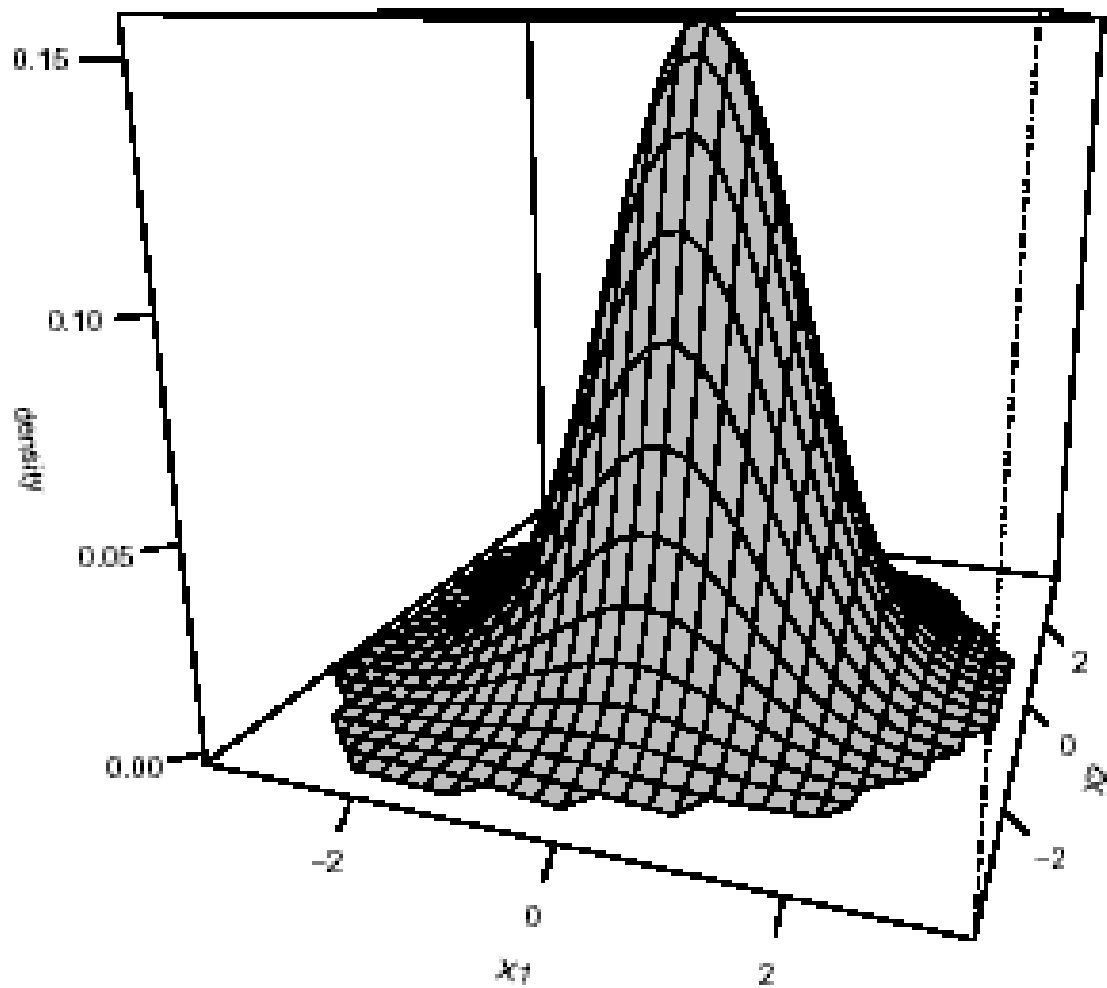
This method can be extended to dimensions higher than 2.



Schematic for Multivariate Density Estimation

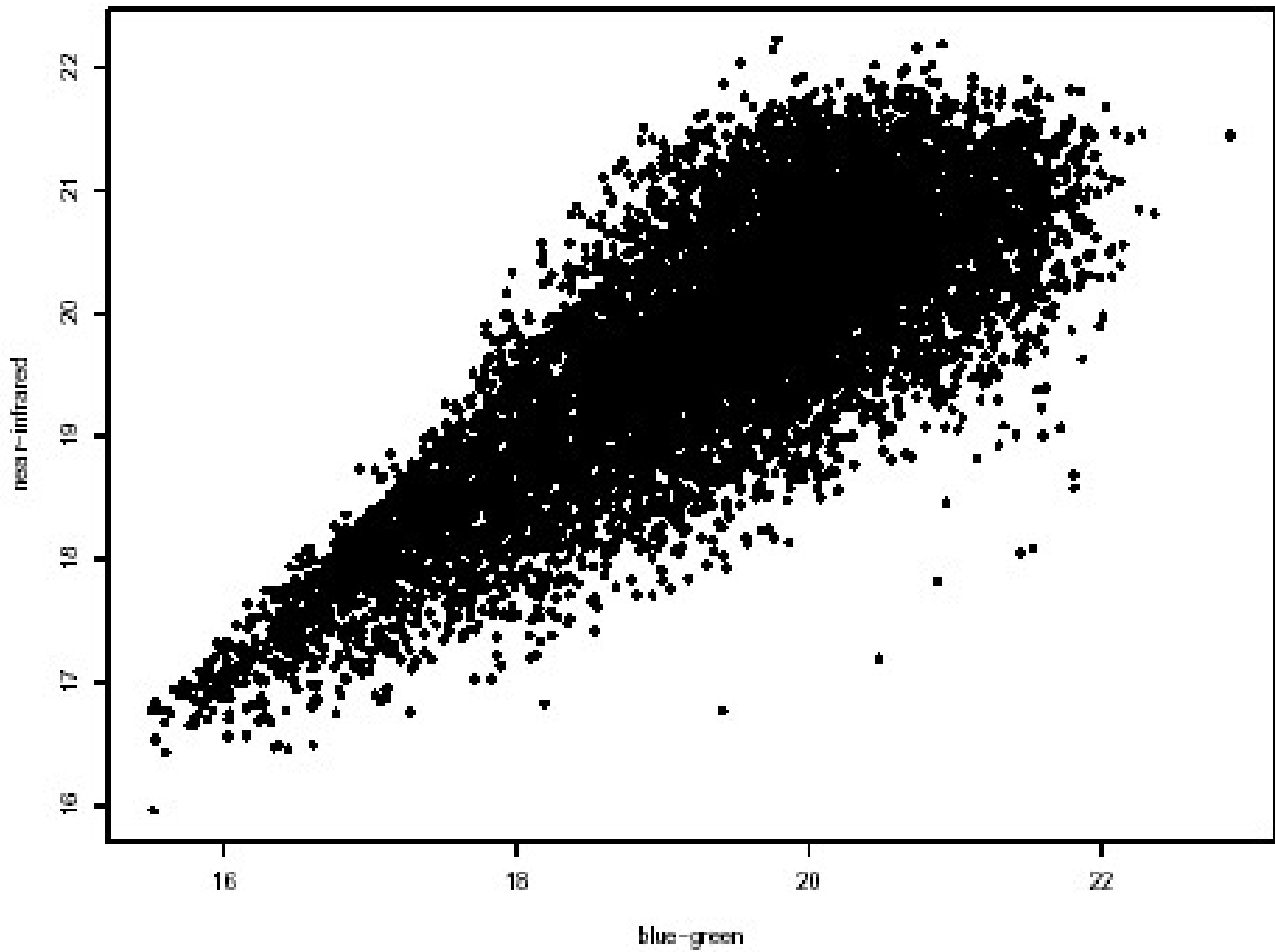




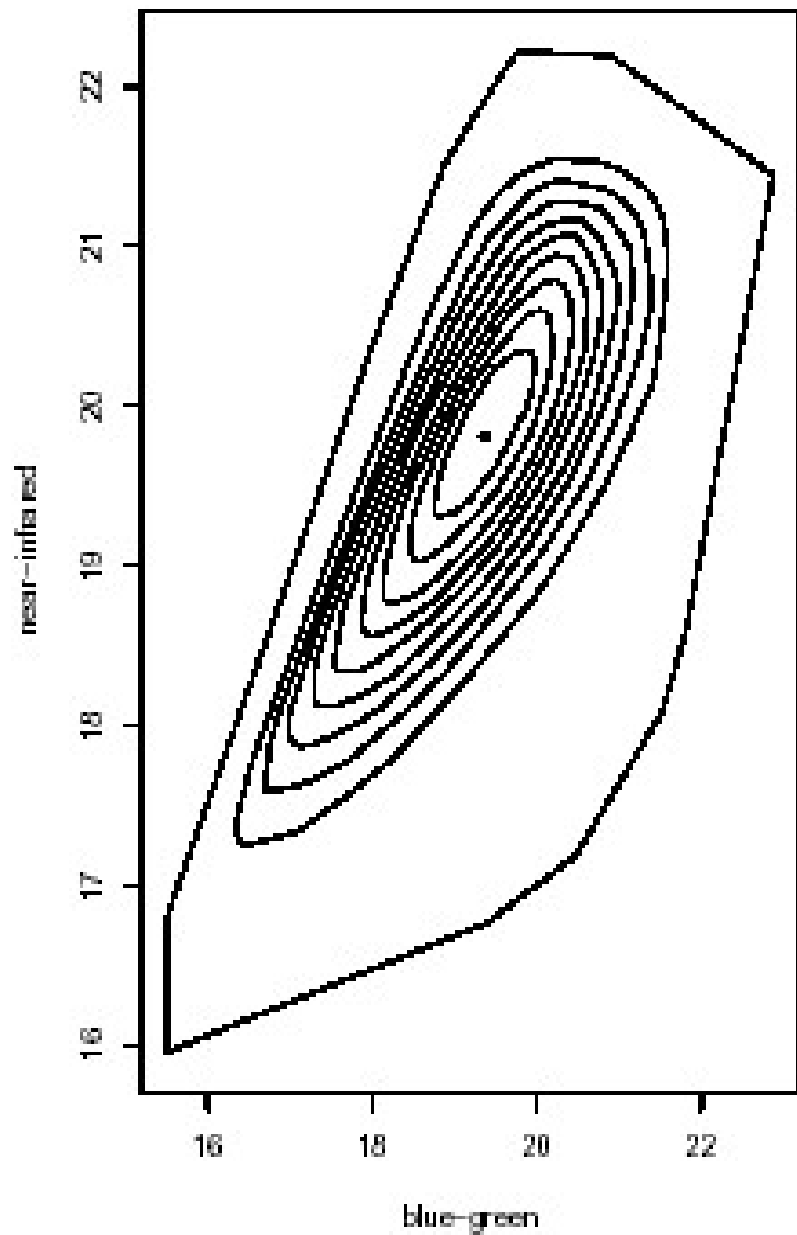


Real Example – Astronomy Data

- We will apply these ideas to a dataset taken from a sky survey.
- The data is bivariate with two different wavelength measurements taken for each object observed.
- For this example, $n = 11,355$.



(a)



(b)

