

# Structure in Data

Often our objective is to identify interesting structures in the data, such as clusters of observations, or relationships among the variables.

Sometimes, the structures allow a reduction in the dimensionality of the data.

Many of the classical methods of multivariate analysis, such as principal components analysis, factor analysis, canonical correlations analysis, and multidimensional scaling, are useful in identifying interesting structure. These methods generally attempt to combine variables in such a way as to preserve information yet reduce the dimension of the dataset.

Dimension reduction generally carries a loss of some information. Whether the lost information is important is the major concern in dimension reduction.

## Finding Groups in Data

Another set of methods for reducing the complexity of a dataset attempts to group observations together, combining observations, as it were.

In the following we will assume that an observation consists of a vector  $x = (x_1, \dots, x_m)$ . In most cases, we will assume that  $x \in \mathbb{R}^m$ .

In statistical analysis, we generally assume that we have  $n$  observations, and we use  $X$  to denote an  $n \times m$  matrix in which the rows correspond to observations.

# Scales of Measurement

In practice, it is common for one or more of the components of  $x$  to be measured on a nominal scale; that is, one or more of the variables represents membership in some particular class. We refer to such variables as “categorical variables”.

Although sometimes it is important to make finer distinctions among types of variables, we often need to make a simple distinction between variables whose values can be modeled by  $\mathbb{R}$  and those whose values essentially indicate membership in some class.

We may represent the observation  $x$  as being composed of these two types, “real” or “numerical”, and “categorical”:

$$x = (x^r, x^c).$$

# Linear Structure and Other Geometric Properties

Numerical data can conveniently be represented as geometric vectors. We can speak of the length of a vector, or of the angle between two vectors, and relate these geometric characteristics to properties of the data. We will begin with definitions of a few basic terms.

The *Euclidean length* or just the *length* of an  $n$ -vector  $x$  is the square root of the sum of the squares of the elements of the vector. We generally denote the Euclidean length of  $x$  as  $\|x\|_2$  or just as  $\|x\|$ :

$$\|x\| = \left( \sum_{i=1}^n x_i^2 \right)^{1/2} .$$

The Euclidean length is a special case of a more general real-valued function of a vector called a “norm”.

# Dot Products and Angles

The *dot product* or *inner product* of two vectors  $x$  and  $y$  that have the same number of elements  $n$  is denoted by  $\langle x, y \rangle$  and is defined by

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

The *angle*  $\theta$  between the vectors  $x$  and  $y$  is defined in terms of the cosine by

$$\cos(\theta) = \frac{\langle x, y \rangle}{\sqrt{\langle x, x \rangle \langle y, y \rangle}}.$$

Linear structures in data are the simplest and often the most interesting. Linear relationships can also be used to approximate other more complicated structures.

# Flats

The set of points  $x$  whose components satisfy a linear equation

$$b_1x_1 + \cdots + b_dx_d = c$$

is called a flat. Such linear structures often occur (approximately) in observational data, leading to a study of the linear regression model,

$$x_d = \beta_0 + \beta_1x_1 + \cdots + \beta_mx_m + \epsilon.$$

A flat through the origin, that is, a set of points whose components satisfy

$$b_1x_1 + \cdots + b_dx_d = 0,$$

is a vector space. Such equations allow simpler transformations, so we often transform regression models into the form

$$x_d - \bar{x}_d = \beta_1(x_1 - \bar{x}_1) + \cdots + \beta_m(x_m - \bar{x}_m) + \epsilon.$$

The data are centered to correspond to this model.

# Linear Transformations

Linear transformations play a major role in analyzing numerical data and identifying structure.

A linear transformation of the vector  $x$  is the vector  $Ax$ , where  $A$  is a matrix with as many columns as the elements of  $x$ . If the number of rows of  $A$  is different, the resulting vector has a dimension different from  $x$ .

# Orthogonal Transformations

An important type of linear transformation is an orthogonal transformation, that is, a transformation in which the matrix of the transformation,  $Q$ , is square and has the property that

$$Q^T Q = I,$$

where  $Q^T$  denotes the transpose of  $Q$ , and  $I$  denotes the identity matrix.

If  $Q$  is orthogonal, for the vector  $x$ , we have

$$\|Qx\| = \|x\|.$$

(This is easily seen by writing  $\|Qx\|$  as  $\sqrt{(Qx)^T Qx}$ , which is  $\sqrt{x^T Q^T Qx}$ .)

Thus, we see that orthogonal transformations preserve Euclidean lengths.

## Orthogonal Transformations

If  $Q$  is orthogonal, for vectors  $x$  and  $y$ , we have

$$\langle Qx, Qy \rangle = (Qx)^T (Qy) = x^T Q^T Qy = x^T y = \langle x, y \rangle,$$

hence,

$$\arccos \left( \frac{\langle Qx, Qy \rangle}{\|Qx\|_2 \|Qy\|_2} \right) = \arccos \left( \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2} \right).$$

Thus, we see that orthogonal transformations preserve angles.

# Gram-Schmidt Orthogonalization

Given two nonnull, linearly independent vectors,  $x_1$  and  $x_2$ , it is easy to form two orthonormal vectors,  $\tilde{x}_1$  and  $\tilde{x}_2$ , that span the same space:

$$\tilde{x}_1 = \frac{x_1}{\|x_1\|_2},$$
$$\tilde{x}_2 = \frac{(x_2 - \tilde{x}_1^T x_2 \tilde{x}_1)}{\|x_2 - \tilde{x}_1^T x_2 \tilde{x}_1\|_2}.$$

These are called *Gram-Schmidt transformations*.

It is easy to confirm by multiplication that  $\tilde{x}_1$  and  $\tilde{x}_2$  are orthonormal.

Further, because they are orthogonal and neither is 0, they must be independent.

## Gram-Schmidt Orthogonalization

The Gram-Schmidt transformations can easily be extended to more than two vectors.

Given a third linearly independent vector,  $x_3$ , a third orthonormal vector,  $\tilde{x}_3$ , would be

$$\tilde{x}_3 = \frac{(x_3 - \tilde{x}_1^T x_3 \tilde{x}_1 - \tilde{x}_2^T x_3 \tilde{x}_2)}{\|x_3 - \tilde{x}_1^T x_3 \tilde{x}_1 - \tilde{x}_2^T x_3 \tilde{x}_2\|_2}.$$

# Geometric Transformations

A set of vectors describes a geometric object.

Algebraic operations are geometric transformations that rotate, deform, or translate the object.

Although these transformations are often used in the two or three dimensions that correspond to the easily perceived physical space, they have similar applications in higher dimensions.

# Invariance Properties of Geometric Transformations

Important characteristics of these transformations are what they leave *unchanged* (that is, their *invariance properties*).

As we have seen, an orthogonal transformation preserves lengths of vectors and angles between vectors. A transformation that preserves lengths and angles is called an *isometric transformation*. Such a transformation also preserves areas and volumes.

Another isometric transformation is a *translation*, which for a vector  $x$  is just the addition of another vector:

$$\tilde{x} = x + t.$$

# Invariance Properties of Geometric Transformations

A transformation that preserves angles is called an *isotropic transformation*. An example of an isotropic transformation that is not isometric is a uniform scaling or dilation transformation,  $\tilde{x} = ax$ , where  $a$  is a scalar.

The transformation  $\tilde{x} = Ax$ , where  $A$  is a diagonal matrix with not all elements the same, does not preserve angles; it is an *anisotropic scaling*.

Another anisotropic transformation is a *shearing transformation*,  $\tilde{x} = Ax$ , where  $A$  is the same as an identity matrix except for a single row or column that has a one on the diagonal but possibly nonzero elements in the other positions; for example,

$$\begin{bmatrix} 1 & 0 & a_1 \\ 0 & 1 & a_2 \\ 0 & 0 & 1 \end{bmatrix}.$$

# Invariance Properties of Geometric Transformations

Although they do not preserve angles, both anisotropic scaling and shearing transformations preserve parallel lines.

A transformation that preserves parallel lines is called an *affine transformation*.

Preservation of parallel lines is equivalent to preservation of collinearity, so an alternative characterization of an affine transformation is one that preserves collinearity.

# Invariance Properties of Geometric Transformations

More generally, we can combine nontrivial scaling and shearing transformations to see that the transformation  $Ax$  for any nonsingular matrix  $A$  is affine.

It is easy to see that addition of a constant vector to all vectors in a set preserves collinearity within the set, so a more general affine transformation is  $\tilde{x} = Ax + t$  for a nonsingular matrix  $A$  and a vector  $t$ .

All of these transformations are *linear transformations* because they preserve straight lines.

A *projective transformation*, which uses the homogeneous coordinate system of the projective plane, preserves straight lines but does not preserve parallel lines.

These transformations are very useful in computer graphics.

# Rotations

Two major tools in seeking linear structure are rotations and projections of the data matrix  $X$ .

Rotations and projections of the observations are performed by postmultiplication of  $X$  by special matrices.

The simplest rotation of a vector can be thought of as the rotation of a plane defined by two coordinates about the other principal axes. Such a rotation changes two elements of all vectors in that plane and leaves all of the other elements, representing the other coordinates, unchanged.

This rotation can be described in a two-dimensional space defined by the coordinates being changed, without reference to the other coordinates.

# Rotations

Consider the rotation of the vector  $x$  through the angle  $\theta$  into  $\tilde{x}$ . The length is preserved, so we have  $\|\tilde{x}\| = \|x\|$ . We can write

$$\begin{aligned}\tilde{x}_1 &= \|x\| \cos(\phi + \theta), \\ \tilde{x}_2 &= \|x\| \sin(\phi + \theta).\end{aligned}$$

Now, from elementary trigonometry, we know that

$$\begin{aligned}\cos(\phi + \theta) &= \cos \phi \cos \theta - \sin \phi \sin \theta, \\ \sin(\phi + \theta) &= \sin \phi \cos \theta + \cos \phi \sin \theta.\end{aligned}$$

Because  $\cos \phi = x_1/\|x\|$  and  $\sin \phi = x_2/\|x\|$ , we can combine these equations to get

$$\begin{aligned}\tilde{x}_1 &= x_1 \cos \theta - x_2 \sin \theta, \\ \tilde{x}_2 &= x_1 \sin \theta + x_2 \cos \theta.\end{aligned}$$

## Rotations

Putting those results together, we see that multiplying  $x$  by the orthogonal matrix

$$\begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

performs the rotation of  $x$ .

This idea easily extends to the rotation of a plane formed by two coordinates about all of the other (orthogonal) principal axes.

# Rotations

The  $m \times m$  orthogonal matrix

$$Q_{pq}(\theta) = \begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & \ddots & & & & & & & & & \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \cos \theta & 0 & \dots & 0 & \sin \theta & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ & & & & & \ddots & & & & & & \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & -\sin \theta & 0 & \dots & 0 & \cos \theta & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ & & & & & & & & & & \ddots & \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

in which  $p$  and  $q$  denote the rows and columns that differ from the identity, rotates the data vector  $x_i$  through an angle of  $\theta$  in the plane formed by the  $p^{\text{th}}$  and  $q^{\text{th}}$  principal axes of the  $m$ -dimensional cartesian coordinate system.

# Rotations

Rotations can be viewed equivalently as a rotation of the coordinate system in the opposite direction.

The coordinate system remains orthogonal after such a rotation.

In the matrix  $XQ$ , all of the observations (rows) of  $X$  have been rotated through the angle  $\theta$ .

Rotations of the data matrix can reveal structure in the data because they provide alternative views of the data.

There is usually nothing obvious in the data to suggest a particular rotation; however, dynamic rotations coupled with projections that are plotted and viewed as they move are very useful in revealing structure.

## Multiple Rotations

A rotation of any plane can be formed by successive rotations of planes formed by two principal axes.

Furthermore, any orthogonal matrix can be written as the product of a finite number of rotation matrices; that is, any orthogonal transformation can be viewed as a sequence of rotations.

# Projections

Another way of getting useful alternative views of the data is to project the data onto subspaces.

A symmetric idempotent matrix  $P$  *projects* vectors onto the subspace spanned by the rows (or columns) of  $P$ .

Except for the identity matrix, a projection matrix is of less than full rank; hence, it projects a full-rank matrix into a space of lower dimension.

Although we may only know that the rows of the data matrix  $X$  are in  $\mathbb{R}^m$ , the rows of  $XP$  are in the subspace spanned by the rows of  $P$ .

It may be possible to identify relationships and structure in this space of lower dimension that are obscured in the higher-dimensional space.

# Translations

Translations are relatively simple transformations involving the addition of vectors.

Rotations and other geometric transformations such as shearing involve multiplication by an appropriate matrix.

In applications where several geometric transformations are to be made, it would be convenient if translations could also be performed by matrix multiplication. This can be done by using *homogeneous coordinates*.

# Homogeneous Coordinates

Homogeneous coordinates, which form the natural coordinate system for projective geometry, have a very simple relationship to cartesian coordinates.

The point with cartesian coordinates  $(x_1, x_2, \dots, x_d)$  is represented in homogeneous coordinates as  $(x_0^h, x_1^h, x_2^h, \dots, x_d^h)$ , where, for arbitrary  $x_0^h$  not equal to zero,  $x_1^h = x_0^h x_1$ ,  $x_2^h = x_0^h x_2$ , and so on.

Each value of  $x_0^h$  corresponds to a hyperplane in the ordinary cartesian coordinate system.

The special plane  $x_0^h = 0$  does not have a meaning in the cartesian system. It corresponds to a hyperplane at infinity in the projective geometry.

# Homogeneous Coordinates

An advantage of the homogeneous coordinate system is that we can easily perform translations.

We can effect the translation  $\tilde{x} = x + t$  by first representing the point  $x$  as  $(1, x_1, x_2, \dots, x_d)$  and then multiplying by the  $(d+1) \times d$  matrix

$$T = \begin{bmatrix} 1 & 0 & \dots & 0 \\ t_1 & 1 & \dots & 0 \\ & & \dots & \\ t_d & 0 & \dots & 1 \end{bmatrix}.$$

We will use the symbol  $x^h$  to represent the vector of corresponding homogeneous coordinates:

$$x^h = (1, x_1, x_2, \dots, x_d).$$

The translated point can be represented as  $\tilde{x} = Tx^h$ .

# Homogeneous Coordinates

We must be careful to distinguish the point  $x$  from the vector of coordinates that represents the point.

In cartesian coordinates, there is a natural correspondence, and the symbol  $x$  representing a point may also represent the vector  $(x_1, x_2, \dots, x_d)$ .

The vector of homogeneous coordinates of the result  $Tx^h$  corresponds to the vector of cartesian coordinates of  $\tilde{x}$ ,  $(x_1 + t_1, x_2 + t_2, \dots, x_d + t_d)$ .

Homogeneous coordinates are used extensively in computer graphics not only for the ordinary geometric transformations but also for projective transformations, which model visual properties.

# General Transformations of the Coordinate System

Transformations can be thought of either as transforming the data within a fixed coordinate system or as transforming the coordinate system, the coordinate system itself remains essentially a cartesian coordinate system.

Homogeneous coordinates correspond in a simple way to cartesian coordinates.

We can make more general transformations of the coordinate system that can be useful in identifying structure in the data. These include parallel coordinates and polar coordinates.

## **Measures of Similarity and Dissimilarity**

There are many ways of measuring the similarity or dissimilarity between two observations or between two variables.

For numerical data, the most familiar measures of similarity are covariances and correlations.

Dissimilarities in numerical data are generally distances of some type.

## Metrics

The dissimilarity or distance function is often a *metric*, which is a function  $\Delta$  from  $\mathbb{R}^m \times \mathbb{R}^m$  into  $\mathbb{R}$  satisfying the properties

- $\Delta(x_1, x_2) \geq 0$  for all  $x_1, x_2 \in \mathbb{R}^m$ ,
- $\Delta(x_1, x_2) = 0$  if and only if  $x_1 = x_2$ ,
- $\Delta(x_1, x_2) = \Delta(x_2, x_1)$  for all  $x_1, x_2 \in \mathbb{R}^m$ ,
- $\Delta(x_1, x_3) \leq \Delta(x_1, x_2) + \Delta(x_2, x_3)$  for all  $x_1, x_2, x_3 \in \mathbb{R}^m$ .

The last property is called the “triangle inequality”.

## Measures of Similarity and Dissimilarity

Other measures of dissimilarity can often be useful. Nonmetric functions, such as ones allowing ties and that do not obey the triangle inequality, can also be used for defining dissimilarity, especially in applications in which there is some noise or in which there is some subjectivity in the data.

Distance measures defined on a finite set of points,  $x_1, x_2, \dots, x_n$ , may use, instead of the triangle inequality, the “ultrametric” inequality:

$$\Delta(x_i, x_k) \leq \max_j (\Delta(x_i, x_j), \Delta(x_j, x_k)).$$

Ultrametric distances are sometimes used as dissimilarity measures in clustering applications.

## Measures of Similarity and Dissimilarity

Other measures of both similarity and dissimilarity must be used for categorical data or for mixed data (that is, for data consisting of some numerical variables and categorical variables),

$$x = (x^r, x^c).$$

The measures may involve ratings of judges, for example. The measures may not be metrics.

In some cases, it is useful to allow distance measures to be asymmetric. If  $d(x_i, x_j)$  represents the cost of moving from point  $x_i$  to point  $x_j$  it may be the case that  $d(x_i, x_j) \neq d(x_j, x_i)$ . If the distance represents a perceptual difference, it may also be the case that  $d(x_i, x_j) \neq d(x_j, x_i)$ .

## **Similarities: Covariances and Correlations**

Measures of similarity include covariances, correlations, rank correlations, and cosines of the angles between two vectors.

Any measure of dissimilarity, such as the distances can be transformed into a measure of similarity by use of a decreasing function, such as the reciprocal.

For example, whereas the cosine of the angle formed by two vectors can be considered a measure of similarity, the sine can be considered a measure of dissimilarity.

## **Similarities: Covariances and Correlations**

Although we can consider similarities/dissimilarities between either columns (variables) or rows (observations), in our common data structures, we often evaluate covariances and correlations between columns and distances among rows.

We speak of the covariance or the correlation between columns or between variables.

The covariance between a column (variable) and itself is its variance.

# Covariances

For an  $n \times m$  data matrix  $X$ , we have the  $m \times m$  *variance-covariance matrix* (or just the *covariance matrix*):

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mm} \end{bmatrix},$$

where

$$s_{jk} = s_{kj} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n - 1}.$$

If  $\bar{X}$  is the matrix in which each column consists of the mean of the corresponding column of  $X$ , we see that

$$S = \frac{1}{n - 1} (X - \bar{X})^T (X - \bar{X}).$$

The matrix  $S$  is therefore nonnegative definite. The matrix  $X - \bar{X}$  is called the “centered data matrix”; each column sums to 0.

# Correlations

Assuming none of the variables is constant, the correlation is often a more useful measure because it is scaled by the variances. For an  $n \times m$  data matrix, the  $m \times m$  *correlation matrix* is

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1m} \\ r_{12} & 1 & \cdots & r_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{1m} & r_{2m} & \cdots & 1 \end{bmatrix},$$

where

$$r_{jk} = r_{kj} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}};$$

that is,

$$R = (\text{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{mm}}))^{-1} S (\text{diag}(\sqrt{s_{11}}, \sqrt{s_{22}}, \dots, \sqrt{s_{mm}}))^{-1}.$$

Notice that covariances and correlations are based on the  $L_2$  norm.

# Covariances and Correlations

Because the concepts of covariance and correlation are also used to refer to properties of random variables, we sometimes refer to the quantities that we have defined above as “sample covariance” or “sample correlation” to distinguish them from the “population” quantities of abstract variables.

There are variations of these such as rank correlations and robust covariances.

Rank correlations are computed by first replacing the elements of each column of  $X$  by the ranks of the elements within the column and then computing the correlation as above.

Robust covariances and correlations are computed either by using a different measure than the  $L_2$  norm or by scaling of the covariance matrix based on an expectation taken with respect to a normal (or Gaussian) distribution.

# Similarities When Some Variables Are Categorical

If all of the variables are measured on a scale that can be modeled as a real number, covariances and/or correlations or similar measures are the obvious choice for measuring the similarity between two points,  $x_j$  and  $x_k$ .

If, however, some of the variables are categorical variables, that is, if the generic  $x$  can be represented in the notation introduced earlier,

$$x = (x^r, x^c),$$

a different measure of similarity must be chosen.

Sometimes, the values of the categorical variables represent such different situations that it does not make sense to consider similarities between observations from different groups.

# Similarities When Some Variables Are Categorical

The similarity between

$$x_j = (x_j^r, x_j^c)$$

and

$$x_k = (x_k^r, x_k^c)$$

may be measured by the function

$$s(x_j, x_k) = \frac{\sum_{i=1}^n (x_{ij}^r - \bar{x}_j^r)(x_{ik}^r - \bar{x}_k^r)}{n - 1}, \quad \text{if } x_j^c = x_k^c,$$
$$= 0, \quad \text{otherwise.}$$

# Similarities When Some Variables Are Categorical

Instead of requiring an exact match of the categorical variables, we can allow some degrees of similarity between observations with different values of their categorical variables.

One way would be by using the count of how many variables within  $x_j^C$  and  $x_k^C$  agree.

Such a simple count can be refined to take into account the number of possible values each of the categorical variables can assume. The measure can also be refined by incorporating some measure of the similarity of different classes.

## Similarities among Functional Observations

Interest-bearing financial instruments such as bonds or U.S. Treasury bills have prices that depend on the spot or current interest rate and so-called forward rates at future points in time. (A forward rate at time  $t_1$  for a future time  $t_2$  can be thought of as the value of cash or a riskless security at time  $t_2 > t_1$  discounted back to time  $t_1$ .)

The forward rates depend on, among other things, the investors' perception of future spot or actual rates.

At any point, a set of forward rates together with the spot rate determine the “yield curve” or the “term structure” for a given financial instrument:

$$r(t).$$

Observational data for measuring and comparing term structures consist of functions for a set of securities measured at different time points.

# Similarities among Functional Observations

Another example of observations that are functions are the measurements on various units of individual features of developing organisms taken over time.

For example, the observational unit may be a developing organism, the features may be gene expressions, and the data elements may be measures of these expressions taken at fixed times during the development of the organism.

The observations on feature  $j$  may consist of measurements  $(x_{j1}, x_{j2}, \dots, x_{jm})$  taken at times  $t_1, t_2, \dots, t_m$ . The overall patterns of the measurements may be of interest. The underlying model is a continuous function,

$$x(t).$$

The observation on each feature is a discrete function, evaluated at discrete points in its time domain.

# Similarities among Functional Observations

Consider the three observations

$$x_1 = (1, 2, 1),$$

$$x_2 = (1, 2, 3),$$

and

$$x_3 = (4, 8, 4).$$

Because of the obvious patterns, we may wish to consider  $x_1$  and  $x_3$  more similar than are  $x_1$  and  $x_2$ .

There are several ways to define a similarity measure to capture this kind of relationship.

A very simple one in this case is the relative changes over time. We may first augment the existing data with measures of changes.

## Similarities among Functional Observations

In the example, taking a simplistic approach of just measuring changes and scaling them, and then augmenting the original vectors, we have

$$\tilde{x}_1 = (1, 2, 1, | 1, -\frac{1}{2}),$$

$$\tilde{x}_2 = (1, 2, 3, | 1, \frac{1}{2}),$$

and

$$\tilde{x}_3 = (4, 8, 4, | 1, -\frac{1}{2}).$$

After transforming the data in this way, we may employ some standard similarity measure, possibly one that downweights the first three elements of each observation.

# Similarities among Functional Observations

Another approach is to fit a smoothing curve to each observational vector and then form a new vector by evaluating the smoothing curve at fixed points.

A standard similarity measure would then be applied to the transformed vectors.

There are many issues to consider when comparing curves.

Whereas the data-generating process may follow a model  $x(t)$ , the data are of the form  $x_i(t_{ij})$ .

In the model, the variable  $t$  (usually “time”) may not be measured in an absolute sense, but rather may be measured relative to a different starting point for each observational unit.

## Registration

Even when this shift is taken into consideration two responses that are similar overall may not begin at the same relative time; that is, one observational unit may follow a model  $x(t)$  and another  $x(t + \delta)$ .

To proceed with the analysis of such data, it is necessary to *register* the data (that is, to shift the data to account for such differences in the time).

## Similarities among Functional Observations

More generally, two observational units may follow the same functional process under some unknown transformation of the independent variable:

$$x_1(t) = x_2(h(t)).$$

Unraveling this transformation is a more difficult process of registration.

# Similarities among Functional Observations

We may want to base similarity among observations on some more general relationship satisfied by the observations.

Suppose, for example, that a subset of some bivariate data lies in a circle. This pattern may be of interest, and we may want to consider all of the observations in the subset lying in the circle to be similar to one another and different from observations not lying in the circle.

Many such similarity measures depend on the context (that is, on a subset of variables or observations, not just on the relationship between two variables or two observations).

Similarities defined by a context are of particular use in pattern recognition.

## Similarities between Groups of Variables

We may want to combine variables that have similar values across all observations into a single variable, perhaps a linear combination of some of the original variables.

The general problem of studying linear relationships between two sets of variables is addressed by the method of *canonical correlations*.

## **Dissimilarities: Distances**

There are several ways of measuring dissimilarity. One measure of dissimilarity is distance, and there are several ways of measuring distance.

Some measures of distance between two points are based only on the elements of the vectors defining those two points.

These distances, which are usually defined by a commutative function, are useful in a homogeneous space. Other measures of distance may be based on a structure imposed by a set of observations.

# Dissimilarities: Distances

In a homogeneous space, there are several commonly used measures of distance between two observations.

Most of these are based on some *norm* of the difference between the two numeric vectors representing the observations. For a set of objects  $S$  that has an addition-type operator,  $+_S$ , a corresponding additive identity,  $0_S$ , and a scalar multiplication (that is, a multiplication of the objects by a real (or complex) number), a *norm* is a function,  $\|\cdot\|$ , from  $S$  to  $\mathbb{R}$  that satisfies the following three conditions:

- nonnegativity and mapping of the identity:  
if  $x \neq 0_S$ , then  $\|x\| > 0$ , and  $\|0_S\| = 0$ ;
- relation of scalar multiplication to real multiplication:  
 $\|ax\| = |a|\|x\|$  for real  $a$ ;
- triangle inequality:  
 $\|x +_S y\| \leq \|x\| + \|y\|$ .

A norm of the difference between two vectors is a metric.

# Distances Based on Norms

Some of the commonly used measures of distance between observations of numerical data represented in the vectors  $x_i$  and  $x_k$  are the following:

- Euclidean distance, the root sum of squares of differences:

$$\|x_i - x_k\|_2$$

or

$$\left( \sum_{j=1}^m (x_{ij} - x_{kj})^2 \right)^{1/2}.$$

The Euclidean distance is sometimes called the  $L_2$  norm.

- maximum absolute difference:

$$\|x_i - x_k\|_\infty$$

or

$$\max_j |x_{ij} - x_{kj}|.$$

## More Distances Based on Norms

- Manhattan distance, the sum of absolute differences:

$$\|x_i - x_k\|_1$$

or

$$\sum_{j=1}^m |x_{ij} - x_{kj}|.$$

- Minkowski or  $L_p$  distance:

$$\|x_i - x_k\|_p$$

or

$$\left( \sum_{j=1}^m |x_{ij} - x_{kj}|^p \right)^{1/p}.$$

The  $L_p$  distance is the  $L_p$  norm of the difference in the two vectors. Euclidean distance, maximum difference, and Manhattan distance are special cases, with  $p = 2$ ,  $p \rightarrow \infty$ , and  $p = 1$ , respectively.

## More Distances Based on Norms

- Canberra distance (from Lance and Williams, 1966):

$$\sum_{j=1}^m \frac{|x_{ij} - x_{kj}|}{|x_{ij}| + |x_{kj}|},$$

as long as  $|x_{ij}| + |x_{kj}| \neq 0$ ; otherwise, 0 (sometimes normalized by  $m$  to be between 0 and 1).

- correlation-based distances:

$$f(r_{ik}).$$

The correlation between two vectors  $r_{ik}$  can also be used as a measure of dissimilarity. Values close to 0 indicate small association. The absolute value of the correlation coefficient is a decreasing function in what is intuitively a dissimilarity, so a distance measure based on it,  $f(r_{ik})$ , should be a decreasing function of the absolute value. Two common choices are

$$1 - |r_{ik}| \quad \text{and} \quad 1 - r_{ik}^2.$$

## More Distances Based on Norms

- distances based on angular separation:

$$\frac{x_i^\top x_k}{\|x_i\|_2 \|x_k\|_2}$$

or

$$\frac{\sum_{j=1}^m x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^m x_{ij}^2 \sum_{j=1}^m x_{kj}^2}}.$$

This measure of angular separation is the cosine of the angle; hence, it is a decreasing function in what is intuitively a dissimilarity. Other quantities, such as the sine of the angle, can be used instead. For centered data, the angular separation is the same as the correlation.

## Other Kinds of Distances

For categorical data, other measures of distance must be used. For vectors composed of zeros and ones, for example, there are two useful distance measures:

- Hamming distance: the number of bits that are different in the two vectors;
- binary difference: the proportion of non-zeros that two vectors do not have in common (the number of occurrences of a zero and a one, or a one and a zero divided by the number of times at least one vector has a one).

# Distances Measures

Notice that generally the *distances* are between the *observations*, whereas the *covariances* discussed above are between the *variables*.

The distances are elements of the  $n \times n$  dissimilarity matrix,

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & \cdots & 0 \end{bmatrix}.$$

All of the distance measures discussed above are metrics (in particular, they satisfy  $\Delta(x_1, x_2) = \Delta(x_2, x_1)$  for all  $x_1, x_2 \in \mathbb{R}^m$ ), so any matrix  $D$ , in which the elements correspond to those measures, is symmetric.

# Distances Measures

The measures of distance we have discussed are appropriate in a homogeneous space in which lengths have the same meaning in all directions.

A scaling of the units in any of the cardinal directions (that is, a change of scale in the measurement of a single variable) may change the distances.

In many applications, the variables have different meanings. Because many statistical techniques give preferential attention to variables with larger variance, it is often useful to scale all variables to have the same variance.

Sometimes, it is more useful to scale the variables so that all have the same range.

## Distances After Transformations

Notice that the angular separation is based on the  $L_2$  norm. A transformation that preserves  $L_2$  distances and angles is called an “isometric transformation”.

If  $Q$  is an orthogonal matrix, the Euclidean distance between  $Qx_i$  and  $Qx_k$  and the angular separation between those two vectors are the same as the distance and angle between  $x_i$  and  $x_k$ .

Hence, an orthogonal matrix is called an isometric matrix because it preserves Euclidean distances and angles.

## Other Dissimilarities Based on Distances

The various distance measures that we have described can be used to define dissimilarities in other ways.

For example, we may define the distance from  $x_j$  to  $x_k$ ,  $d^R(x_j, x_k)$ , as the rank of an ordinary distance  $d_{jk}$  in the set of all distances  $d_{ji}$ .

If  $x_k$  is the point closest to  $x_j$ , then  $d^R(x_j, x_k) = 1$ . This type of dissimilarity depends on the “direction”; that is, in general,

$$d^R(x_j, x_k) \neq d^R(x_k, x_j).$$

A distance measure such as  $d^R(\cdot, \cdot)$  is dependent on the neighboring points, or the “context”.

## Other Dissimilarities Based on Distances

If we think of the distance between two points as the cost or effort required to get from one point to another, the distance measure often may not be symmetric. (It is therefore not a metric.)

Common examples in which distances measured this way are not symmetric arise in anisotropic media under the influence of a force field (say, electrical or gravitational) or in fluids with a flow.

# Dissimilarities in Anisometric Coordinate Systems: Sphering Data

If the elements of the observation vectors represent measurements made on different scales, it is usually best to scale the variables so that all have the same variance or else have the same range.

A scaling of the data matrix  $X$  so that all columns have a variance of 1 is achieved by postmultiplication by a diagonal matrix whose elements are the square roots of the diagonal elements of the sample covariance matrix,  $S$ .

If this is applied to the centered data, we have the “standardized” data matrix:

$$X_S = (X - \bar{X}) \text{diag}(\sqrt{s_{ii}}).$$

# Dissimilarities in Anisometric Coordinate Systems: Sphering Data

If there are relationships among the variables whose observations comprise the columns of  $X$ , and if there are more rows than columns (that is,  $n > m$ ), it may be appropriate to perform an oblique scaling,

$$X_W = (X - \bar{X})H,$$

where  $H$  is the Cholesky factor of  $S^{-1}$ ; that is,

$$\begin{aligned} H^T H &= (n - 1)((X - \bar{X})^T (X - \bar{X}))^{-1} \\ &= S^{-1}. \end{aligned}$$

(If the matrix  $S$  is not of full rank, the generalized inverse is used in place of the inverse. In any case, the matrix is nonnegative definite, so the decomposition exists.) The matrix  $X_W$  is a *centered and sphered* matrix. It is sometimes called a *white* matrix.

The matrix is orthonormal; that is,  $X_W^T X_W = I$ .

# Dissimilarities in Anisometric Coordinate Systems: Sphering Data

In general, a structure may be imposed on the space by the sample covariance matrix  $S$ .

A very useful measure of the distance between two vectors is the *Mahalanobis distance*. The Mahalanobis distance between the  $i^{\text{th}}$  and  $k^{\text{th}}$  observations,  $x_i$  and  $x_k$  (the  $i^{\text{th}}$  and  $k^{\text{th}}$  rows of  $X$ ) is

$$(x_i - x_k)^{\top} S^{-1} (x_i - x_k).$$

Notice that the Mahalanobis distance is the squared Euclidean distance after using  $S$  to scale the data. It is the squared Euclidean distance between rows in the  $X_S$  matrix above.

## Asymmetric Dissimilarities

There are other types of distance. Certain paths from one point to another can be specified.

The distance can be thought of as the cost of getting from one node on a graph to another node.

Although distances are usually considered to be symmetric (that is, the distance from point  $x_i$  to point  $x_k$  is the same as the distance from point  $x_k$  to point  $x_i$ ), a more general measure may take into account fluid flow or elevation gradients, so the dissimilarity matrix would not be symmetric.

# Directional Dissimilarities

Another type of data that presents interesting variations for measuring dissimilarities or similarities is directional data, or circular data (that is, data that contain a directional component).

The angular separation measures this, of course, but often in directional data, one of the data elements is a plane angle. As the size of the angle increases, ultimately it comes close to a measure of 0.

A simple example is data measured in polar coordinates. When one of the data elements is an angle, the component of the overall distance between two observations  $i$  and  $j$  attributable to their angles,  $\theta_i$  and  $\theta_j$ , could be taken as

$$d_{ij}^d = 1 - \cos(\theta_i - \theta_j).$$

The directional component must be combined additively with a component due to Euclidean-like distances,  $d_{ij}^r$ .

## Directional Dissimilarities

In polar coordinates, the radial component is already a distance, so  $d_{ij}^r$  may just be taken as the absolute value of the difference in the radial components  $r_i$  and  $r_j$ .

The overall distance  $d_{ij}$  may be formed from  $d_{ij}^d$  and  $d_{ij}^r$  in various ways that weight the radial distance and the angle differently.

There are many examples, such as wind direction in meteorology or climatology, in which directional data arise.

# Properties of Dissimilarities

A dissimilarity measure based on a metric conforms generally to our intuitive ideas of distance. The norm of the difference between two vectors is a metric, that is, if

$$\Delta(x_1, x_2) = \|x_1 - x_2\|,$$

then  $\Delta(x_1, x_2)$  is a metric.

Distance measures such as the  $L_p$  distance and the special cases of Euclidean distance, maximum difference, and Manhattan distance, which are based on norms of the difference between two vectors, have useful properties, such as satisfying the triangle inequality:

$$d_{ik} \leq d_{ij} + d_{jk}.$$

There are many different measures that may be useful in different applications.

## **Dissimilarities between Groups of Observations**

In clustering applications, we need to measure distances between groups of observations.

We are faced with two decisions. First, we must choose the distance metric to use, and then the points in the two groups between which we measure the distance.

Any of the distance measures discussed above could be used.

## **Dissimilarities between Groups of Observations**

Once a distance measure is chosen, the distance between two groups can be defined in several ways, such as the following;

- the distance between a central point, such as the mean or median, in one cluster and the corresponding central point in the other cluster;
- the minimum distance between a point in one cluster and a point in the other cluster;
- the largest distance between a point in one cluster and a point in the other cluster;
- the average of the distances between the points in one cluster and the points in the other cluster.

## **Dissimilarities between Groups of Observations**

The average of all of the pairwise point distances is the most common type of measure used in some applications.

This type of measure is widely used in genetics, where the distance between two populations is based on the differences in frequencies of chromosome arrangements (for example, Prevosti's distance) or on DNA matches or agreement of other categorical variables (for example, Sanghvi's distance).

## Effects of Transformations of the Data

In the course of an analysis of data, it is very common to apply various transformations to the data. These transformations may involve various operations on real numbers, such as scaling a variable (multiplication), summing all values of a variable (addition), and so on.

Do these kinds of operations have an effect on the results of the data analysis?

Do they change the relative values of such things as measures of similarity and dissimilarity?

# Effects of Transformations of the Data:

## Example

Consider a very simple case in which a variable represents length, for example. The actual data are measurements such as 0.11 meters, 0.093 meters, and so on. These values are recorded simply as the real numbers 0.11, 0.093, and so on.

In analyzing the data, we may perform certain operations (summing the data, squaring the data, and so on) in which we merely assume that the data behave as real numbers. (Notice that 0.11 is a real number but 0.11 meters is not a real number — 0.11 meters is a more complicated object.)

After noting the range of values of the observations, we may decide that millimeters would be better units of measurement than meters. The values of the variable are then scaled by 1,000. Does this affect any data analysis we may do?

## Example Continued

Although, as a result of scaling, the mean goes from approximately  $\mu$  (for some value  $\mu$ ) to  $1,000\mu$ , and the variance goes from  $\sigma^2$  (for some value  $\sigma$ ) to  $1,000,000\sigma^2$ , the scaling certainly should not affect any analysis that involves that variable alone.

Suppose, however, that another variable in the dataset is also length and that typical values of that variable are 1,100 meters, 930 meters, and so on.

For this variable, a more appropriate unit of measure may be kilometers.

## Example Continued

To change the unit of measurement results in dividing the data values by 1,000.

The differential effects on the mean and variance are similar to the previous effects when the units were changed from meters to millimeters; the effects on the means and on the variances differ by a factor of 1,000.

Again, the scaling certainly should not affect any analysis that involves that variable alone.

This scaling, however, does affect the relative values of measures of similarity and dissimilarity.

## Example Continued

Consider, for example, the Euclidean distance between two observations,  $x_1 = (x_{11}, x_{12})$  and  $x_2 = (x_{21}, x_{22})$ . The squared distance prior to the scaling is

$$(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2.$$

Following the scaling, it is

$$10^6(x_{11} - x_{21})^2 + 10^{-6}(x_{12} - x_{22})^2.$$

The net effect depends on the relative distances between  $x_1$  and  $x_2$  as measured by their separate components.

As we mention above, an orthogonal transformation preserves Euclidean distances and angular separations; that is, it is an isometric transformation. An orthogonal transformation also preserves measures of similarity based on the  $L_2$  norm. An orthogonal transformation, however, does not preserve other measures of similarity or distance.

# Outlying Observations and Collinear Variables

Many methods of data analysis may be overly affected by observations that lie at some distance from the other observations, or by observations that lie along some lower dimensional manifold.

Using a least squares criterion for locating the center of a set of observations, for example, can result in a “central point” that is outside of the convex hull of all of the data except for just one observation.

As an extreme case, consider the mean of 100 univariate observations, all between 0 and 1 except for one outlying observation at 100.

The mean of this set of data is larger than 99% of the data.

## Outlying Observations

An outlier may result in one row and column in the dissimilarity matrix  $D$  having very large values compared to the other values in the dissimilarity matrix. This is especially true of dissimilarities based on the  $L_2$  norm.

Dissimilarities based on other norms, such as the  $L_1$  norm, may not be as greatly affected by an outlier.

Methods of data analysis that are not as strongly affected by outlying observations are said to be “robust”. (There are various technical definitions of robustness, which we will not consider here.)

## Outlying Observations

The variance-covariance matrix  $S$ , because it is based on squares of distances from unweighted means, may be strongly affected by outliers.

A robust alternative is

$$S_R = (s_{Rjk}),$$

where the  $s_{Rjk}$  are robust alternatives to the  $s_{jk}$ . There are various ways of defining the  $s_{Rjk}$ .

# Outlying Observations

In general, the  $s_{Rjk}$  are formed by choosing weights for the individual observations to decrease the effect of outlying points; for example,

$$s_{Rjk} = \frac{\sum_{i=1}^n w_i^2 (x_{ij} - \bar{x}_{Rj})(x_{ik} - \bar{x}_{Rk})}{\sum_{i=1}^n w_i^2 - 1},$$

where

$$\bar{x}_{Rj} = \frac{\sum_{i=1}^n w_i x_{ij}}{\sum_{i=1}^n w_i},$$

for a given function  $\omega$ ,

$$w_i = \omega(d_i)/d_i,$$

and

$$d_i = (x_i - \bar{x}_R)^T S_R^{-1} (x_i - \bar{x}_R).$$

(In this last expression,  $x_i$  represents the  $m$ -vector of the  $i^{\text{th}}$  observation, and  $\bar{x}_R$  represents the  $m$ -vector of the weighted means.)

## Outlying Observations

The function  $\omega$  is designed to downweight outlying observations. One possibility, for given constants  $b_1$  and  $b_2$ , is

$$\begin{aligned}\omega(d) &= d && \text{if } d \leq d_0 \\ &= d_0 e^{-\frac{1}{2}(d-d_0)^2/b_2^2} && \text{if } d > d_0,\end{aligned}$$

where  $d_0 = \sqrt{m} + b_1/\sqrt{2}$ .

## Collinear Variables

A problem of a different type arises when the variables are highly correlated.

In this case, the covariance matrix  $S$  and the correlation matrix  $R$ , which are based on the  $L_2$  norm, are both ill-conditioned.

The ranking transformation results in a correlation matrix that is better conditioned.

# Multidimensional Scaling: Determining Observations that Yield a Given Distance Matrix

Given an  $n \times n$  distance matrix such as  $D$ , could we reconstruct an  $n \times m$  data matrix  $X$  that yields  $D$  for some metric  $\Delta(\cdot, \cdot)$ ?

The question, of course, is constrained by  $m$  (that is, by the number of variables).

The problem is to determine the elements of rows of  $X$  such that

$$\begin{aligned}\tilde{d}_{ij} &= \Delta(x_i, x_j) \\ &\approx d_{ij}.\end{aligned}$$

This is called *multidimensional scaling*.

# Multidimensional Scaling

The approximation problem can be stated precisely as an optimization problem to minimize

$$\frac{\sum_i \sum_j f(\tilde{d}_{ij} - d_{ij})}{\sum_i \sum_j f(d_{ij})},$$

where  $f(\cdot)$  is some function that is positive for nonzero arguments and is monotone increasing in the absolute value of the argument, and  $f(0) = 0$ . An obvious choice is  $f(t) = t^2$ .

If the distances in  $D$  do not arise from a metric, the discussed ways of transforming the dissimilarities so that the least squares approach would still work.

The larger the value of  $m$ , of course, the closer the  $\tilde{d}_{ij}$  will be to the  $d_{ij}$ . If  $m \ll n$  and the approximations are good, significant data reduction is achieved.

# Data Mining

It is now common to search through datasets and compute summary statistics from various items that may indicate relationships that were not previously recognized.

The individual items or the relationships among them may not have been of primary interest when the data were originally collected.

This process of prowling through the data is sometimes called *data mining* or *knowledge discovery in databases* (KDD).

(The names come and go with current fads; there is very little of substance indicated by use of different names.)

# Data Mining

The objective is to discover characteristics of the data that may not be expected based on the existing theory. In the language of the database literature, the specific goals of data mining are:

- classification of observations;
- linkage analysis;
- deviation detection;

and finally

- predictive modeling.

# Data Mining

Data mining is exploratory data analysis (EDA) applied to large datasets. An objective of an exploratory analysis is often to generate hypotheses, and exploratory analyses are generally followed by more formal confirmatory procedures.

The explorations in massive datasets must be performed without much human intervention.

Searching algorithms need to have some means of learning and adaptively improving. This will be a major area of research for some time.

Predictive modeling uses inductive reasoning rather than the more common deductive reasoning, which is much easier to automate.

# Data Mining

In the statistical classification of observations, the dataset is partitioned recursively.

The partitioning results in a classification tree, which is a decision tree, each node of which represents a partition of the dataset.

The decision at each node is generally based on the values of a single variable at a time, as in the two most commonly used procedures, CART and C4.5, or its successors, See5 and C5.0.

CART can also build nodes based on linear combinations of the variables. This is sometimes called “oblique partitioning” because the partitions are not parallel to the axes representing the individual variables.

Seeking good linear combinations of variables on which to build oblique partitions is a much more computationally intensive procedure than just using single variables.

# Data Mining

Linkage analysis is often the most important activity of data mining. In linkage analysis, relationships among different variables are discovered and analyzed. This step follows partitioning and is the interpretation of the partitions that were formed.

It is also important to identify data that do not fit the patterns that are discovered.

The deviation of some subsets of the data often makes it difficult to develop models for the remainder of the data.

There are several commercially available software packages that implement data mining, usually of datasets in some standard format.

# Computational Feasibility

Data must be stored, transported, sorted, searched, and otherwise rearranged, and computations must be performed on it.

The size of the dataset largely determines whether these actions are feasible. Huber (1994, 1996) proposed a classification of datasets by the number of bytes required to store them. Huber described as “tiny” those requiring on the order of  $10^2$  bytes; as “small” those requiring on the order of  $10^4$  bytes; as “medium” those requiring on the order of  $10^6$  bytes (one megabyte); as “large”,  $10^8$  bytes; “huge”,  $10^{10}$  bytes (10 gigabytes); and as “massive”,  $10^{12}$  bytes (one terabyte). (“Tera” in Greek means “monster”.)

This log scale of two orders of magnitude is useful to give a perspective on what can be done with data. Online or out-of-core algorithms are generally necessary for processing massive datasets.

# Computational Feasibility

For processing massive datasets, the order of computations is a key measure of feasibility.

We can quickly determine that a process whose computations are  $O(n^2)$  cannot be reasonably contemplated for massive ( $10^{12}$  bytes) datasets.

If computations can be performed at a rate of  $10^{12}$  per second (teraflop), it would take over three years to complete the computations. (A rough order of magnitude for quick “year” computations is  $\pi \times 10^7$  seconds equals approximately one year.)

A process whose computations are  $O(n \log n)$  could be completed in 230 milliseconds for a massive dataset. This remarkable difference in time required for  $O(n^2)$  and  $O(n \log n)$  processes is the reason that the fast Fourier transform (FFT) algorithm was such an important advance.

# Computational Feasibility

Exponential orders can make operations even on tiny ( $10^2$  bytes) datasets infeasible. A process whose computations require time of  $O(2^n)$  may not be completed in four centuries.

Sometimes, it is appropriate to reduce the size of the dataset by forming groups of data. “Bins” can be defined, usually as nonoverlapping intervals covering  $\mathbb{R}^d$ , and the number of observations falling into each bin can be determined.

This process is linear in the number of observations. The amount of information loss, of course, depends on the sizes of the bins.

Binning of data has long been used for reducing the size of a dataset, and earlier books on statistical analysis usually had major sections dealing with “grouped data”.

## Computational Feasibility

Another way of reducing the size of a dataset is by sampling.

This must be done with some care, and often, in fact, sampling is not a good idea.

Sampling is likely to miss the unusual observations, and it is precisely these outlying observations that are most likely to yield new information.

# Computational Feasibility

Advances in computer hardware continue to expand what is computationally feasible.

It is interesting to note, however, that the order of computations is determined by the problem to be solved and the algorithm to be used, not by the hardware.

Advances in algorithm design have reduced the order of computations for many standard problems, while advances in hardware have not changed the order of the computations. Hardware advances change the constant in the order of time.