MARGINAL REWEIGHTING IN HOMOGENEITY ANALYSIS

JAN DE LEEUW AND VANESSA BEDDO

Abstract. Homogeneity analysis, also known as multiple correspondence analysis (MCA), can easily lead to outliers and horseshoe patterns in the plots for the object scores. In this note we first generalize the notion of homogeneity analysis. We then propose some alternative measures of homogeneity, which lead to algorithms that are computationally very close to MCA. These measures lead to a simple reweighting of the columns of the indicator matrix, and by choosing an appropriate weighting we can control outliers and more evenly fill the space,

1. Introduction

Suppose we have a finite domain Ω of *n objects*, and *m categorical variables* ϕ_j defined on Ω. Variables ϕ_j has a finite range Γ_j of k_j elements. Thus

$$\phi_j : \Omega \rightarrow \Gamma_j.$$

Data can be collected in an $n \times m$ data-frame H, where $h_{ij} \in \Gamma_j$ is defined by $h_{ij} = \phi_j(\omega_i)$. We use K for the sum of the k_j.

Variable ϕ_j defines a partition of the n objects into the k_j subsets

$$\Omega_{j\ell} = \{\omega \in \Omega \mid \phi_j(\omega) = \gamma_{j\ell}\},$$

where $\gamma_{j1} \cdots \gamma_{jk_j}$ are the elements of Γ_j. Clearly there is a total of K such subsets $\Omega_{j\ell}$, with $j = 1, \cdots, m$ and $\ell = 1, \cdots, k_j$.

Date: July 24, 2005.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis.

It is useful to allow for the possibility of missing data, which can be incorporated by allowing ϕ_j to be defined only on a subset Ω_j of Ω. In that case, of course, $\cup_{\ell=1}^{k_j}\Omega_{jl} \subseteq \Omega$ and inclusion maybe strict.

2. Homogeneity

A homogeneity analysis finds a map $\xi : \Omega \rightarrow \mathbb{R}^p$ of all objects in low-dimensional Euclidean space $\mathbb{R}^p$. The map ξ is chosen so that the representations $\xi(\Omega_{j\ell})$ of the K subsets are *homogeneous*, which generally means relatively small and compact. In order words, points corresponding to objects in the same subset should be relatively close to each other. Or, more specifically, the within-subset variation must be small relative to the total variation.

If we have a measure of homogeneity, we can quantify how homogeneous a particular low-dimensional representation of the objects is. And we can construct algorithms to find the most homogeneous representation that is possible. Actually, we will use measures of heterogeneity, and minimize those, but that is just a computational detail.

If we only have a single variable we can always attain perfect homogeneity by mapping each of the K subsets of the partition into a a single point. Thus it is important to remember we compute a compromise solution. by requiring that all variables simultaneously have representations which are as homogeneous as possible, or, more specifically, that the total homogeneity over all variables is maximized.

2.1. **Size of a Point Set.**

There are many possible ways to define the size of a point set in $\mathbb{R}^p$. Some of them are discussed in De Leeuw [2003]. We can use the diameter, the radius of the smallest circle containing all points, the circumference of the convex hull, the length of the minimum spanning tree, the sum of all distances, or the sum of all distances to the Weber point. For this last measure, see De Leeuw and Michailidis [2004]; Michailides and De Leeuw [2004].

In this paper we study a class of quadratic measures of heterogeneity. They have the major advantage that optimizing them generally leads to relatively simple computations. Moreover one specific member of this class is the homogeneity measure used in MCA [De Leeuw and Michailidis, 1999; Michailidis and De Leeuw, 1998]. All measures are based on squared within-set distances.

Suppose $\mathcal{Z}$ is a finite set of τ points in $\mathbb{R}^p$. Consider the following measures of heterogeneity of this point set.

(1) The sum of the squared distances of the τ points to their centroid (i.e. their unweighted average).
(2) The average of the squared distances of the τ points to their centroid.
(3) The sum of the $\binom{\tau}{2}$ squared distances between the τ points.
(4) The average of the $\binom{\tau}{2}$ squared distances between the τ points.

If we write $\sigma(\mathcal{Z})$ for the first measure, which we as the basis for comparison, because it is the one used in MCA. The second one is $\frac{1}{\tau}\sigma(\mathcal{Z})$, the third one is $\tau\sigma(\mathcal{Z})$ and the fourth one is $\frac{1}{\tau-1}\sigma(\mathcal{Z})$ (we set the fourth one equal to zero if $\tau = 1$). The three measures are closely related, because they are all of the form $f(\tau)\sigma(\mathcal{Z})$, for some function f on the natural numbers.

In an actual data analysis we use sums of K heterogeneity measures over all subsets. This means that, no matter which of the three measure we use, we use weighted sums of the σ-heterogeneity of the subsets. The weights depend on the size of the subset. More specifically, the sums are of the form

$$\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} f(d_{j\ell})\sigma(\Omega_{j\ell}),$$

where $d_{j\ell}$ is the number of objects in category ℓ of variable j.

The weighted sum interpretation shows that if we use the sum of the squared distances, with $f(n) = n$, then large point sets are penalized extra in measuring heterogeneity. We especially want large point sets to be homogeneous. If we use the average of the squared distances, with $f(n) = (n-1)^{-1}$, then we especially want small sets to be homogenous, and we do not care much about the larger ones.

More generally, we can look at the family

$$\sigma_f(\mathcal{Z}) = f(\tau)\sigma(\mathcal{Z})$$

which weighs $\sigma(\mathcal{Z})$ with a non-negative function f depending on the number of points. By using increasing powers of τ for instance, we can put more and more emphasis on the larger subsets, by using negative powers we pay more and more attention to the smaller subsets. From a statistical point of view it does make much sense to try to make small subsets homogeneous and ignore the larger ones. That seems tantamount to encouraging chance capitalization. Thus positive powers of τ are a priori more interesting.

3. Indicator Matrices and Profile Frequencies

Define the $n \times K$ indicator super-matrix G, which has as its columns the indicator functions of the subsets $O_{j\ell}$. It can be partiitoned as

$$G = (G_1 \mid \cdots \mid G_m),$$

where each G_j is the $n \times k_j$ indicator matrix of variable j. In each G_j the columns are orthogonal and add up to a column with all elements equal to one (or, if there are missing data, to either zero or one). This is because objects only map into at most one category of a variable.

3.1. **Matrix Expression.** Suppose $g_{j\ell}$ is column ℓ of indicator matrix G_j, and $E_{j\ell}$ is $\mathbf{diag}(g_{j\ell})$.

The sum of the within set squared distances between the points in category ℓ of variable j is

$$\sigma_{j\ell}(X) = \mathbf{tr}\, X'(E_{j\ell} - \frac{1}{d_{j\ell}} g_{j\ell} g'_{j\ell})X$$

We use the weighting function $f(n)$ to define homogeneity. Let

$$E_\star = \sum_{j=1}^m \sum_{\ell=1}^{k_j} f(d_{j\ell})E_{j\ell},$$

$$P_\star = \sum_{j=1}^m \sum_{\ell=1}^{k_j} \frac{f(d_{j\ell})}{d_{j\ell}} g_{j\ell}g_{j\ell}'$$

The f-weighted heterogeneity over variables is

$$\sigma_f(X) = \mathbf{tr}\ X'(E_\star - P_\star)X$$

which we minimize under the condition that $X'E_\star X = I$ and $X'E_\star u = 0$. Or, equivalently, we maximize $\mathbf{tr}\ X'P_\star X$ over all X such that $X'E_\star X = I$ and $X'E_\star u = 0$. This leads to the generalized eigen-problem

$$P_\star X = E_\star X\Lambda^2 \tag{1}$$

which can be solved by standard numerical linear algebra methods.

There are a number of reasons to choose the normalization $X'E_\star X = I$. First, it guarantees that the eigenvalues are between zero and one. This follows from $\sigma_f \geq 0$, which implies $P_\star \lesssim E_\star$ and thus $E_\star^{-1/2}P_\star E_\star^{-1/2} \lesssim I$. Second, as in MCA, the dominant eigenvalue is equal to one and it corresponds with an eigenvector proportional to u, which has all its elements equal one. This follows directly from the fact that $E_\star - P_\star$ is doubly-centered, and consequently has a smallest eigenvalue equal to zero. The other eigenvectors x, with eigenvalues less than one, satisfy $u'E_\star x = 0$.

3.2. **Using the Singular Value Decomposition.**

The matrices $E_\star$ and $P_\star$ are both of order n. This can be unpleasantly large, both in terms of storage and in terms of computing the eigen decomposition.

Observe, however, that $P_\star$ is of the form $GD_\star^{-1}G'$, where $D_\star$ is a diagonal matrix with the elements

$$d_{j\ell}^\star = d_{j\ell}/f(d_{j\ell}).$$

It follows that we can alternatively solve the generalized singular value problem

$$GY = E_\star X\Lambda, \tag{2a}$$

$$G'X = D_\star Y\Lambda, \tag{2b}$$

or the (usually smaller) generalized eigenvalue problem

$$C_\star Y = D_\star Y\Lambda^2, \tag{3}$$

where $C_\star = G'E_\star^{-1}G$, which is of order K.

This way of formulating the optimization problem can be used to construct a more efficient algorithm in R, using the singular value decomposition. The code is given in Appendix D.

It also leads directly to generalizations of the *barycentric* or *centroid* principles familiar from MCA. We can rewrite (2) as

$$\begin{aligned} E_\star^{-1}GY &= X\Lambda, \\ D_\star^{-1}G'X &= Y\Lambda, \end{aligned}$$

In the same way (3) gives

$$D_\star^{-1}G'\tilde{X} = Y\Lambda^2,$$

where $\tilde{X} = E_\star^{-1}GY$, and (1) gives

$$E_\star^{-1}G\tilde{Y} = X\Lambda^2,$$

where $\tilde{Y} = D_\star^{-1}G'X$.

The singular value formulation in (2) can also be used to derive *alternating least squares* (or *reciprocal averaging*) algorithms to compute the solution. The algorithm to update the current solution $X^{(s)}$ is

$$\begin{aligned} Y^{(s)} &= D_\star^{-1}G'X^{(s)}, \\ \tilde{X}^{(s+1)} &= E_\star^{-1}GY^{(s)}, \\ X^{(s+1)} &= \mathbf{orth}(\tilde{X}^{(s+1)}), \end{aligned}$$

where **orth** is an operator that orthonormalizes the current X such that $X'E_\star X = I$. As in MCA, this algorithm allows us to only compute the singular vectors we need, and to use the sparsity of G efficiently.

3.3. **Profile Frequencies.** The m variables, with k_j categories each, give a total of $L = \prod_{j=1}^m k_j$ possible *profiles*, where a profile is a binary vector of length K coding a possible combination of categories. Thus the n rows of $G = (G_1 \mid \cdots \mid G_m)$ are profiles. Some profiles may occur more than once in the data and many profiles will not occur at all because usually $L >> n$. Instead of coding the data as an $n \times K$ indicator matrix, we can also code the data as the $L \times K$ indicator matrix of all possible profiles, together with a vector of profile frequencies. In most cases many of these frequencies will be zero. Instead of mapping the n objects into $\mathbb{R}^p$ we now map the L profiles into $\mathbb{R}^p$, but we use a loss function that takes the profile frequencies into account.

Collect the L profile frequencies in a diagonal matrix N. Then the weighted sum of the squared distances of the objects in category ℓ of variable j to the weighted centroid of the category is

$$\sigma_{jl}(X) = \mathbf{tr}\, X'(NE_{jl} - \frac{1}{d_{jl}} N g_{j\ell} g'_{j\ell} N)X$$

where $d_{j\ell} = g'_{jl} N g_{j\ell}$.

By the same reasoning as before, we see that

$$\sigma_f(X) = \mathbf{tr}\, X'(E_\star - P_\star)X,$$

where now

$$E_\star = N\left[\sum_{j=1}^m \sum_{\ell=1}^{k_j} f(d_{j\ell}) E_{j\ell}\right],$$

$$P_\star = N\left[\sum_{j=1}^m \sum_{\ell=1}^{k_j} \frac{f(d_{jl\ell})}{d_{j\ell}} g_{j\ell} g'_{j\ell}\right] N,$$

or

$$E_\star = \left[\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} f(d_{j\ell})\tilde{E}_{j\ell}\right],$$

$$P_\star = \left[\sum_{j=1}^{m}\sum_{\ell=1}^{k_j} \frac{f(d_{jl\ell})}{d_{j\ell}}\tilde{g}_{j\ell}\tilde{g}_{j\ell}'\right],$$

where $\tilde{g}_{j\ell} = Ng_{j\ell}$ and $\tilde{E}_{j\ell} = \mathbf{diag}(\tilde{g}_{j\ell}) = NE_{j\ell}$. Observe that $E_\star$ and $P_\star$ are both of order L, but they have at most n non-zero rows and columns.

3.4. **Pavings.**

With basically the same techniques and definitions we can study *pavings* [De Leeuw, 2003]. A paving is simply a system of any K subsets of a set Ω. The subsets can be used to define the columns of an $n \times K$ binary matrix G, which is now not necessarily an indicator matrix of m categorical variables. The same notions of homogeneity and weighting apply to pavings.

4. Stability

4.1. **Limiting Cases.**

4.2. **Singular Value Decomposition under Rescaling.**

We have demonstrated that our reweighted MCA amounts to computing a singular value decomposition of $E_\star^{-1/2}GD_\star^{-1/2}$.

5. Gauging

5.1. **Large Samples.**

Consider the situation where the rows of G are a random sample from a discrete multivariate distribution. Now $C = G'E_\star^{-1}G$ has elements

$$\underline{c}_{j\ell} = \sum_{i=1}^{n}\frac{1}{\underline{e}_i}\underline{g}_{ij}\underline{g}_{il},$$

and thus

$$\mathbf{E}(\underline{c}_{j\ell}) = \pi_{j\ell}\mathbf{E}(\underline{e}_i \mid \underline{g}_{ij} = 1 \wedge \underline{g}_{il} = 1)$$

5.2. **The Multivariate Normal.**

5.3. **The Guttman Scale.**

6. Examples

6.1. **Multinormal.** Take a sample of size 1000 from a four-variate standard multinormal, with all correlations equal to 0.5. Solutions for $f(n) = (n-1)^{-1}, f(n) = 1, f(n) = n$, and $f(n) = n^2$ are give in Figure 11.

6.2. **GALO.** The second example is the GALO example [Peschar, 1975]. It has data in 1290 sixth graders in the city of Groningen in 1959. Variables are Gender, IQ, Teacher's Advice on secondary education, and SES.

6.3. **Senate.** The third example are twenty votes in the US Senate. For more details we refer to De Leeuw [in press]. The MCA solution is a fuzzy horseshoe with democrats on one side and replublicans on the other. If we weight the smaller marginals more, by using $f(n) = (n-1)^{-1}$, we see that the horseshoe becomes more pronounced. If we go to $f(n) = n$ or $f(n) = n^2$ the horseshoe is broken up, and the data form four different clusters.

6.4. **House.**

6.5. **Summary.** If we compare the solutions, which both give horseshoes [Schriever, 1985; Rijckevorsel, 1987], we see that using $f(n) = n$ either fills up or breaks up the horseshoe. This is clearly because reweighting emphasize homogeneity of the middle categories, and does not pay much attention to the tails. So what we find is like zooming in to the top of the horseshoe.

Appendix A. Differentiating the Singular Value Decomposition

The formulas in this section have been derived various times in many different contexts. See, for instance, Vaccaro [1994]; Papadopulou and Lourakis [2000] for overviews. We give them here for completeness. Our derivation actually presupposes differentiability, which easily follows from the corresponding results for eigenvalues [Kato, 1976]. Given differentiability, we simply compute the differentials.

Suppose X is a matrix of rank r with n rows and m columns. It has a singular value decomposition of the form $X = K\Lambda L'$, or, more precisely,

$$X = \begin{bmatrix} \underset{n\times r}{K_1} & \underset{n\times(n-r)}{K_0} \end{bmatrix} \begin{bmatrix} \underset{r\times r}{\Lambda_{11}} & \underset{r\times(m-r)}{\emptyset} \\ \underset{(n-r)\times r}{\emptyset} & \underset{(n-r)\times(m-r)}{\emptyset} \end{bmatrix} \begin{bmatrix} \underset{r\times m}{L_1'} \\ \underset{(m-r)\times m}{L_0'} \end{bmatrix}.$$

Of course $K'K = KK' = I$ and $L'L = LL' = I$, while Λ_{11} is positive definite.

We now perturb X to $\tilde{X} = X+\Delta_X$ and we compute a first order approximation to the singular vectors and singular values of $\tilde{X}$.

$$\begin{aligned} (X + \Delta_X)(L + \Delta_L) &= (K + \Delta_K)(\Lambda + \Delta_\Lambda), \\ (X + \Delta_X)'(K + \Delta_K) &= (L + \Delta_L)(\Lambda + \Delta_\Lambda), \\ (K + \Delta_K)'(K + \Delta_K) &= I, \\ (L + \Delta_L)'(L + \Delta_L) &= I, \end{aligned}$$

or, retaining only first order terms,

$$\Delta_X L + X\Delta_L = \Delta_K\Lambda + K\Delta_\Lambda, \tag{4a}$$

$$\Delta_X' K + X'\Delta_K = \Delta_L\Lambda + L\Delta_\Lambda, \tag{4b}$$

$$\Delta_K' K + K'\Delta_K = 0, \tag{4c}$$

$$\Delta_L' L + L'\Delta_L = 0. \tag{4d}$$

We can simplify these equations by defining $K_\perp$ to be an $n \times (n - r)$ matrix satisfying $K'K_\perp = 0$ and $K_\perp'K_\perp = I$. $L_\perp$ is is defined in a similar way.

We now write

$$\Delta_K = KA + K_\perp B, \tag{5a}$$

$$\Delta_L = LC + L_\perp D, \tag{5b}$$

where A and C are $r \times r$, B is $(n - r) \times r$, and D is $(m - r) \times r$. We have to solve for A, B, C, and D.

Premultiply (4a) by K' and (4b) by L'. Then

$$K'\Delta_X L + \Lambda C = A\Lambda + \Delta_\Lambda, \tag{6a}$$

$$L'\Delta_X' K + \Lambda A = C\Lambda + \Delta_\Lambda, \tag{6b}$$

while substituting (5a) and (5b) in (4c) and (4d) gives

$$A + A' = 0,$$
$$C + C' = 0,$$

Thus both A and C are anti-symmetric and have a zero diagonal.

Define the partitioned matrix

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} K'\Delta_X L & K'\Delta_X L_\perp \\ K_\perp'\Delta_X L & K_\perp'\Delta_X L_\perp \end{bmatrix}$$

and take the diagonal of either (6a) or (6b) to obtain

$$\Delta_\Lambda = \mathbf{diag}(E).$$

This gives the first order perturbation of the singular values.

Premultiply (6a) by Λ, postmultiply (6b) by Λ, and add the two results. This gives

$$\mathbf{offdiag}(\Lambda E + E'\Lambda) = C\Lambda^2 - \Lambda^2 C.$$

In the same way

$$\mathbf{offdiag}(E\Lambda + \Lambda E') = A\Lambda^2 - \Lambda^2 A.$$

We switch to elementwise notation to give the solutions

$$a_{st} = \frac{\lambda_t e_{st} + \lambda_s e_{ts}}{\lambda_t^2 - \lambda_s^2},$$
$$c_{st} = \frac{\lambda_s e_{st} + \lambda_t e_{ts}}{\lambda_t^2 - \lambda_s^2}$$

Now premultiply (4a) by $K'_\perp$ and (4b) by $L'_\perp$. Then

$$K'_\perp \Delta_X L = G = B\Lambda,$$
$$L'_\perp \Delta'_X K = F' = D\Lambda,$$

and thus

$$b_{st} = \frac{g_{st}}{\lambda_t},$$
$$d_{st} = \frac{f_{ts}}{\lambda_t}.$$

Appendix B. Small Example

This is the coding example used in Gifi [1990, Chapter 2].

a	p	u
b	q	v
a	r	v
a	p	u
b	p	v
c	p	v
a	p	u
a	p	v
c	p	v
a	p	v

a	p	u	3
a	p	v	2
a	p	w	0
a	q	u	0
a	q	v	0
a	q	w	0
a	r	u	0
a	r	v	1
a	r	w	0
b	p	u	0
b	p	v	1
b	p	w	0
b	q	u	0
b	q	v	1
b	q	w	0
b	r	u	0
b	r	v	0
b	r	w	0
c	p	u	0
c	p	v	2
c	p	w	0
c	q	u	0
c	q	v	0
c	q	w	0
c	r	u	0
c	r	v	0
c	r	w	0

a	p	u	3
a	p	v	2
a	r	v	1
b	p	v	1
b	q	v	1
c	p	v	2

Table 1. Data Frame and Profiles Frequencies

a	b	c	p	q	r	u	v	w
1	0	0	1	0	0	1	0	0
0	1	0	0	1	0	0	1	0
1	0	0	0	0	1	0	1	0
1	0	0	1	0	0	1	0	0
0	1	0	1	0	0	0	1	0
0	0	1	1	0	0	0	1	0
1	0	0	1	0	0	1	0	0
1	0	0	1	0	0	0	1	0
0	0	1	1	0	0	0	1	0
1	0	0	1	0	0	0	1	0

Table 2. Indicator Matruces

Appendix C. Figures

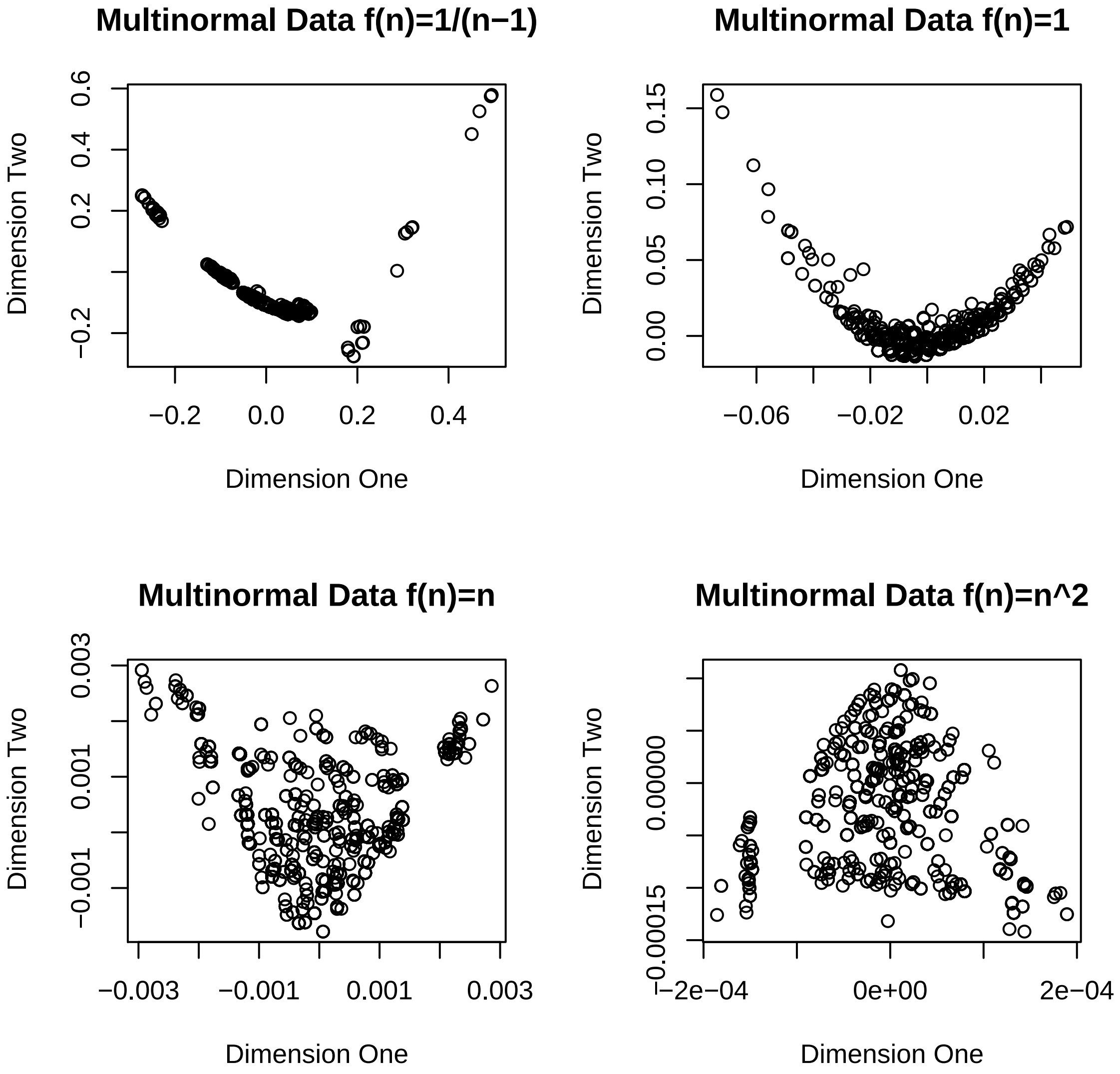


Figure 1. Multinormal Example

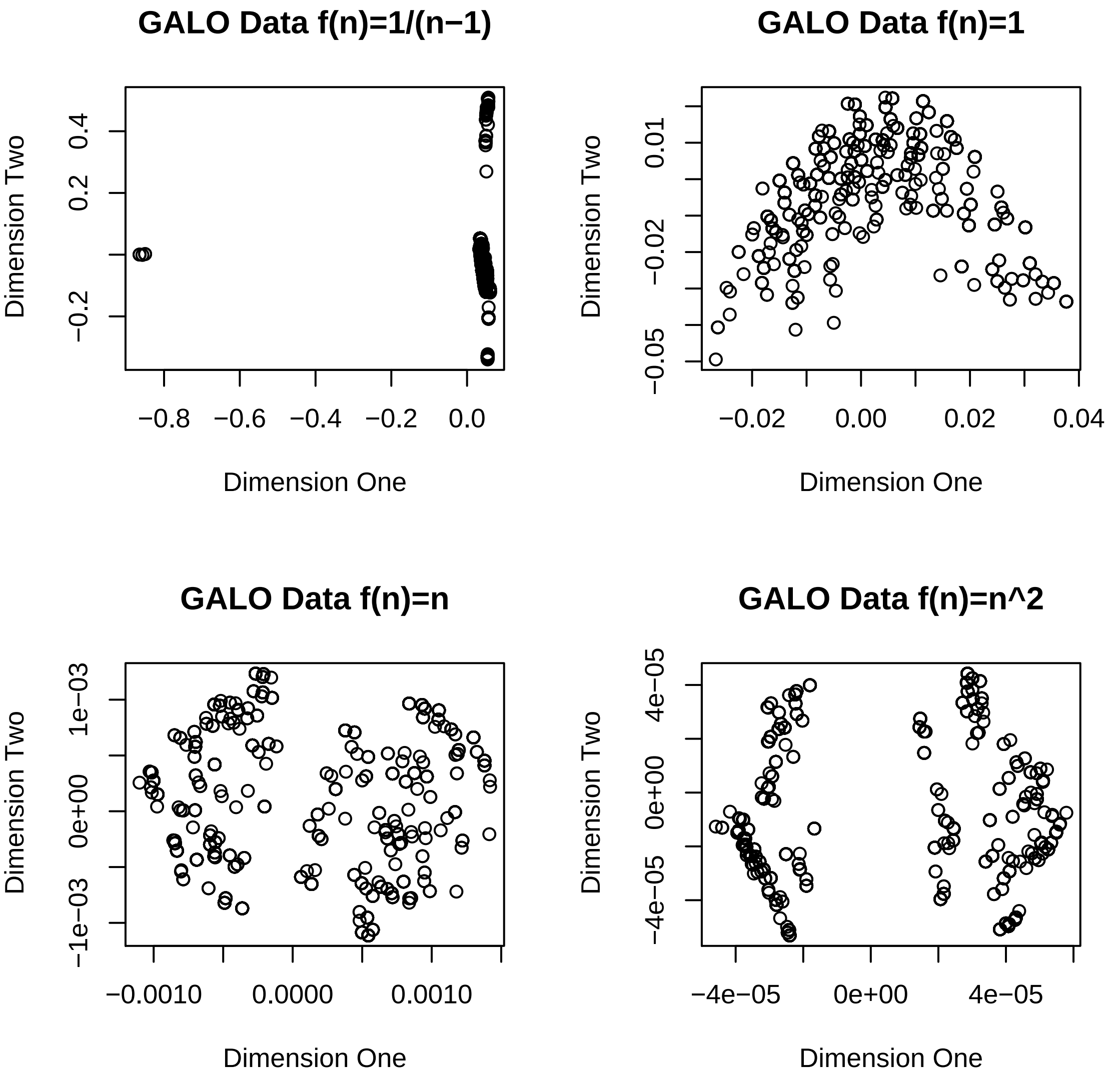


FIGURE 2. GALO Example

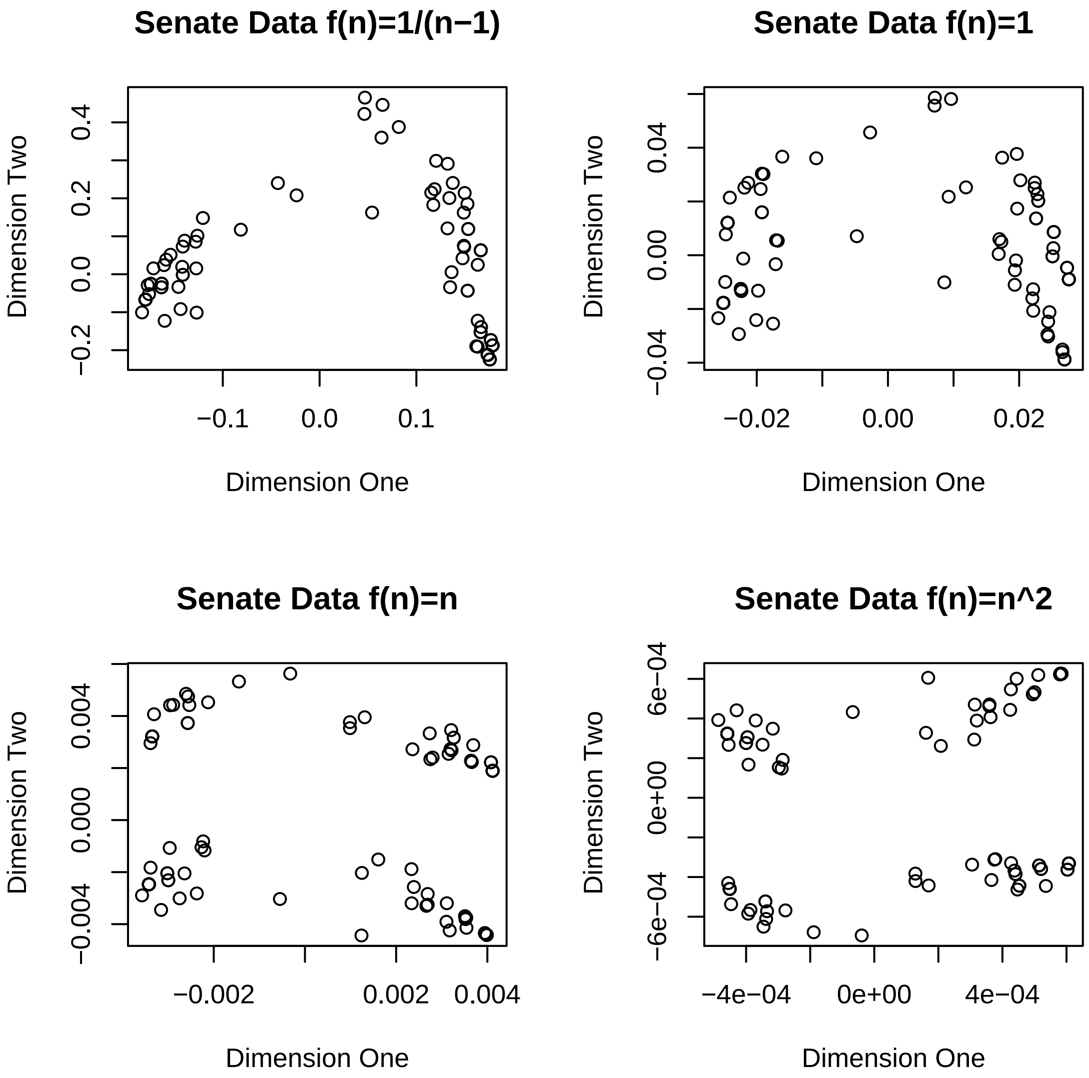


FIGURE 3. Senate Example

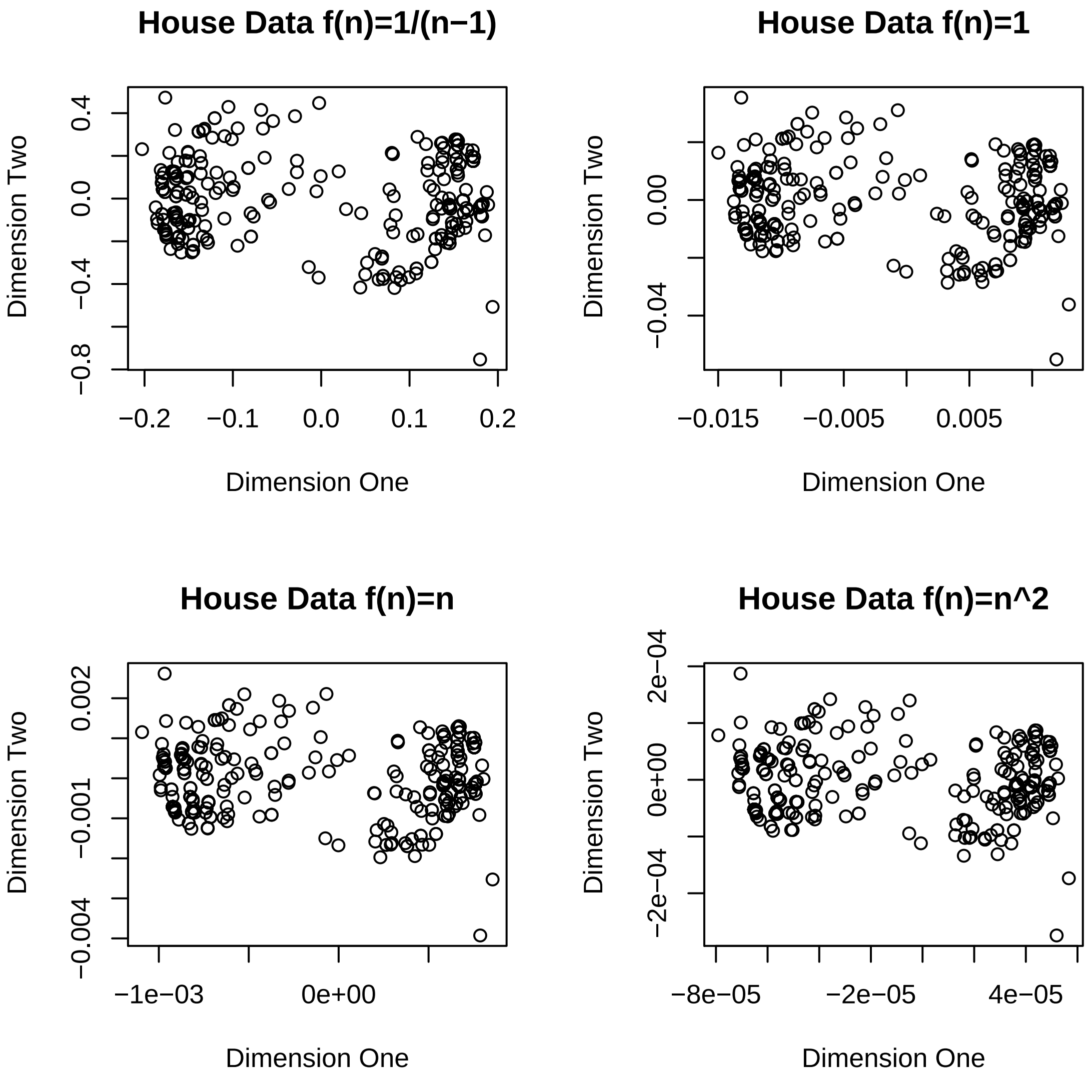
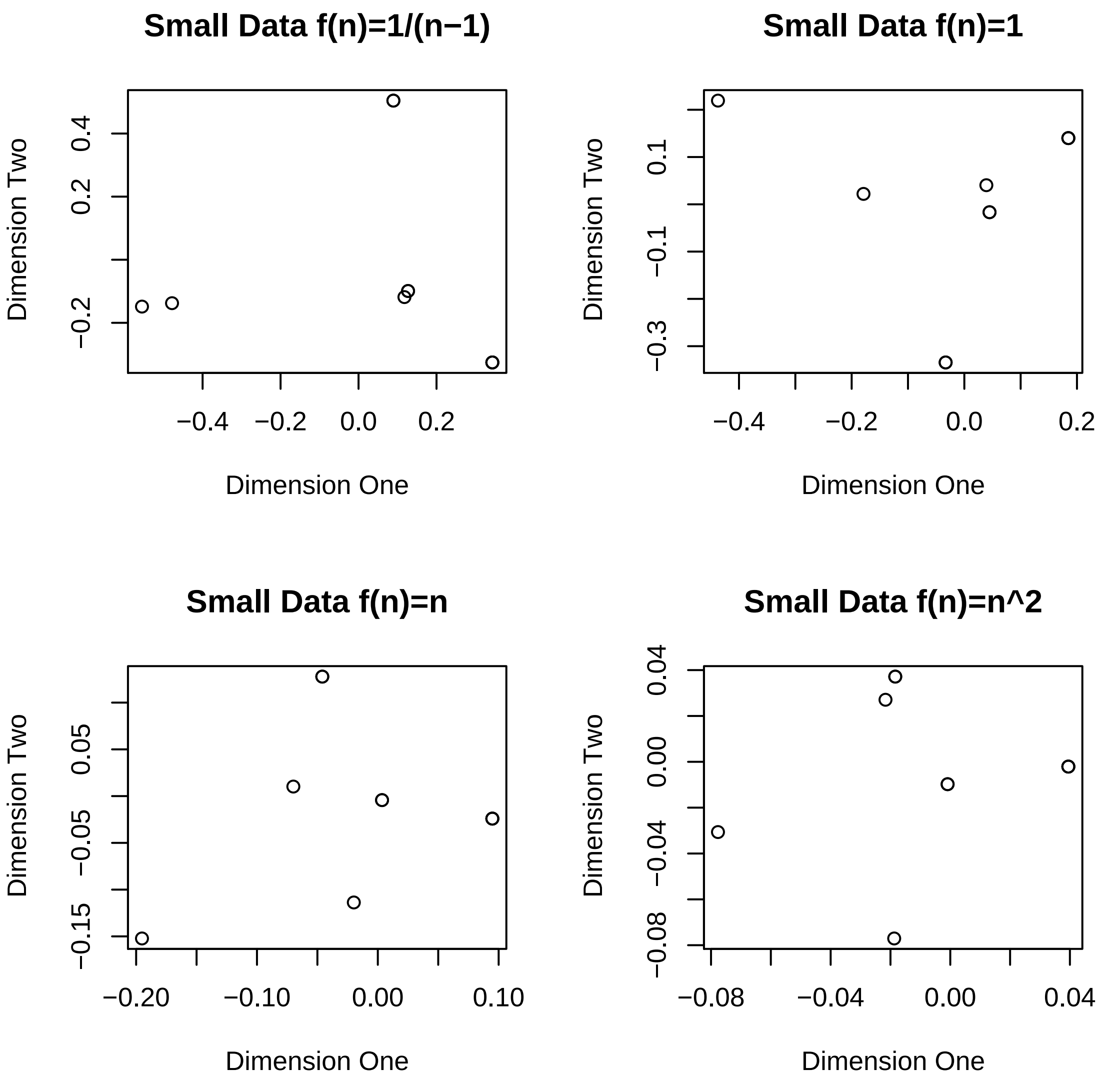


FIGURE 4. House Example



Small Data f(n)=1/(n−1)

Small Data f(n)=1

Small Data f(n)=n

Small Data f(n)=n^2

FIGURE 5. Small Example

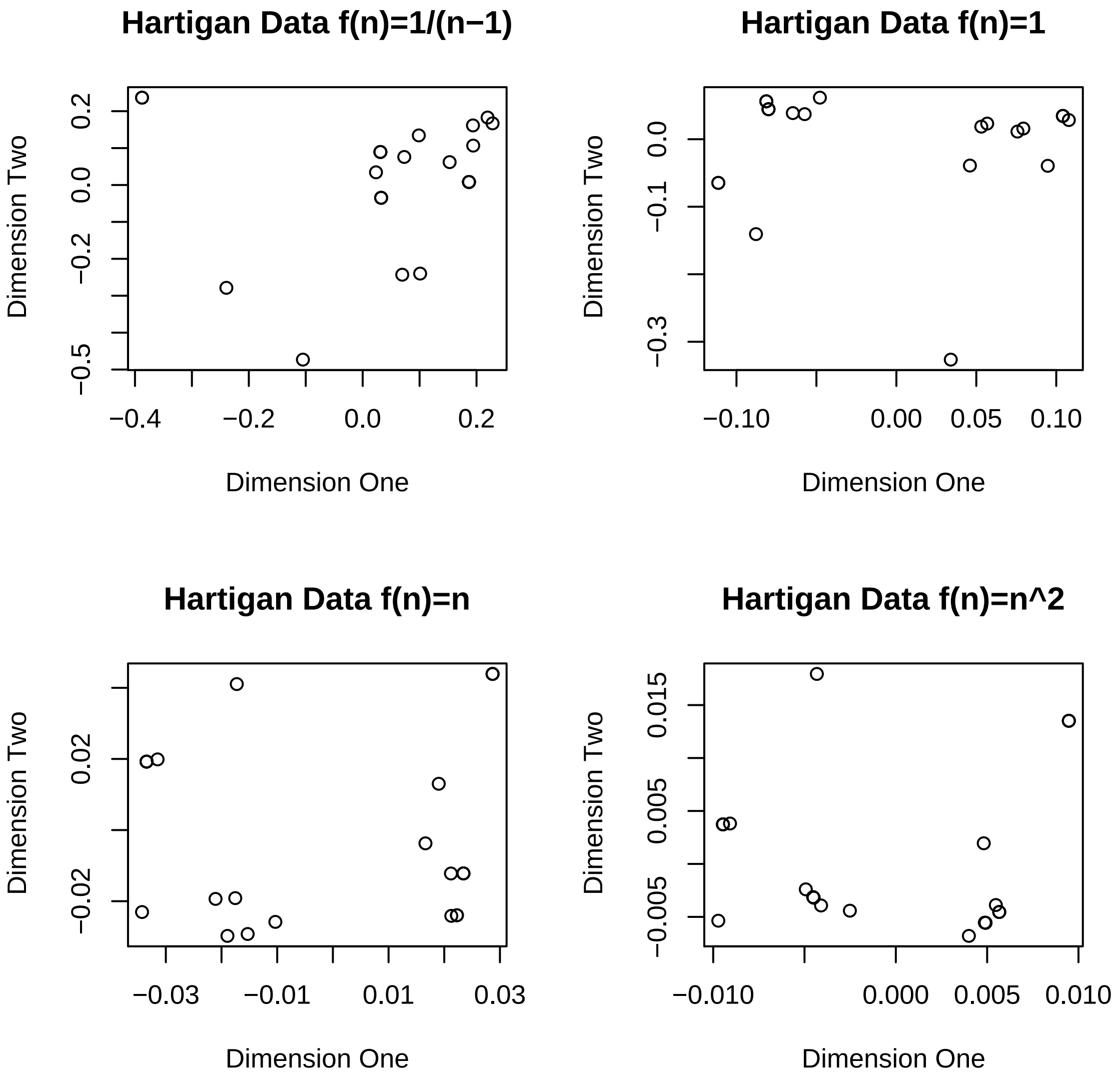


FIGURE 6. Hartigan Example

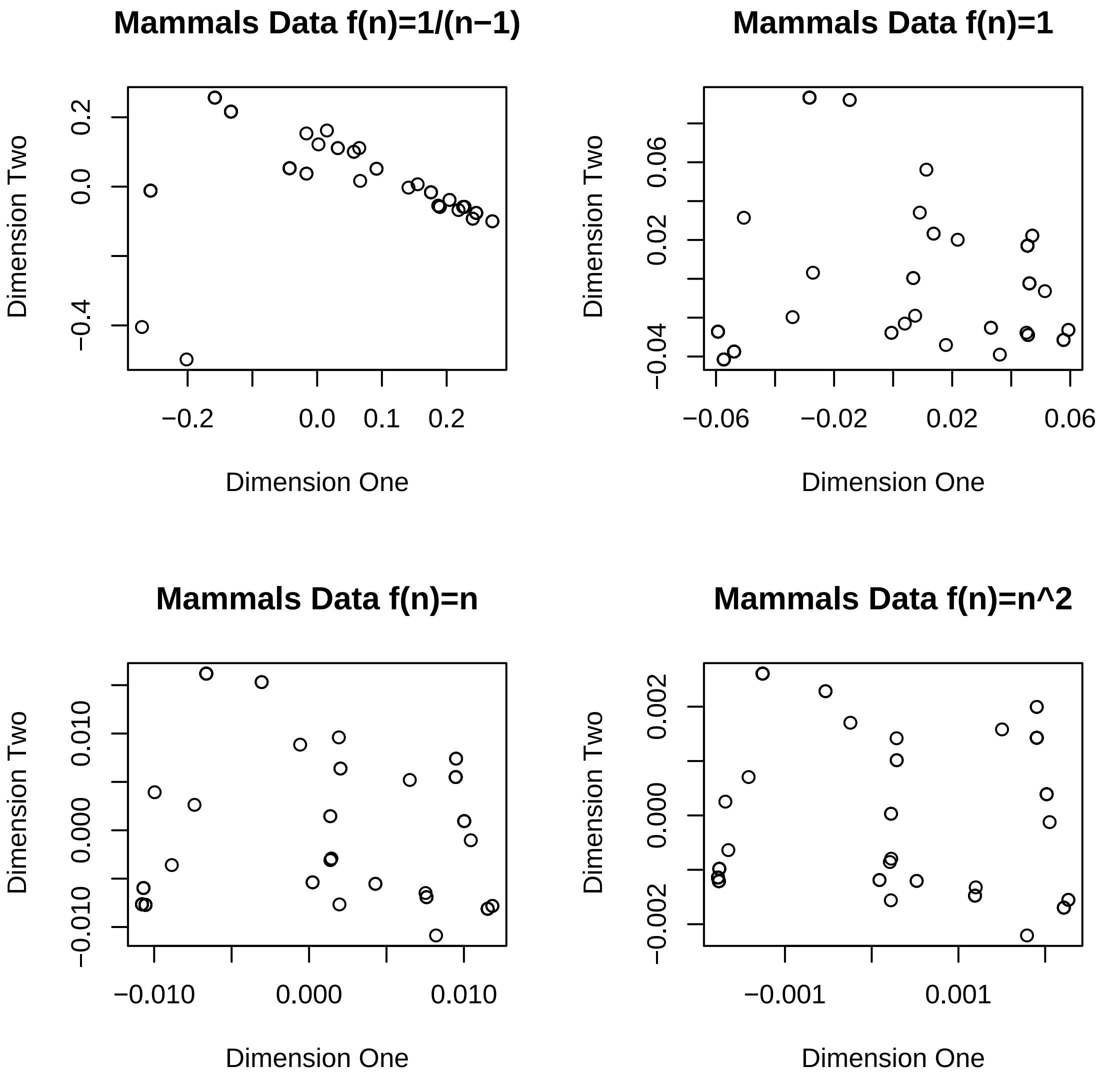


FIGURE 7. Dentition Example

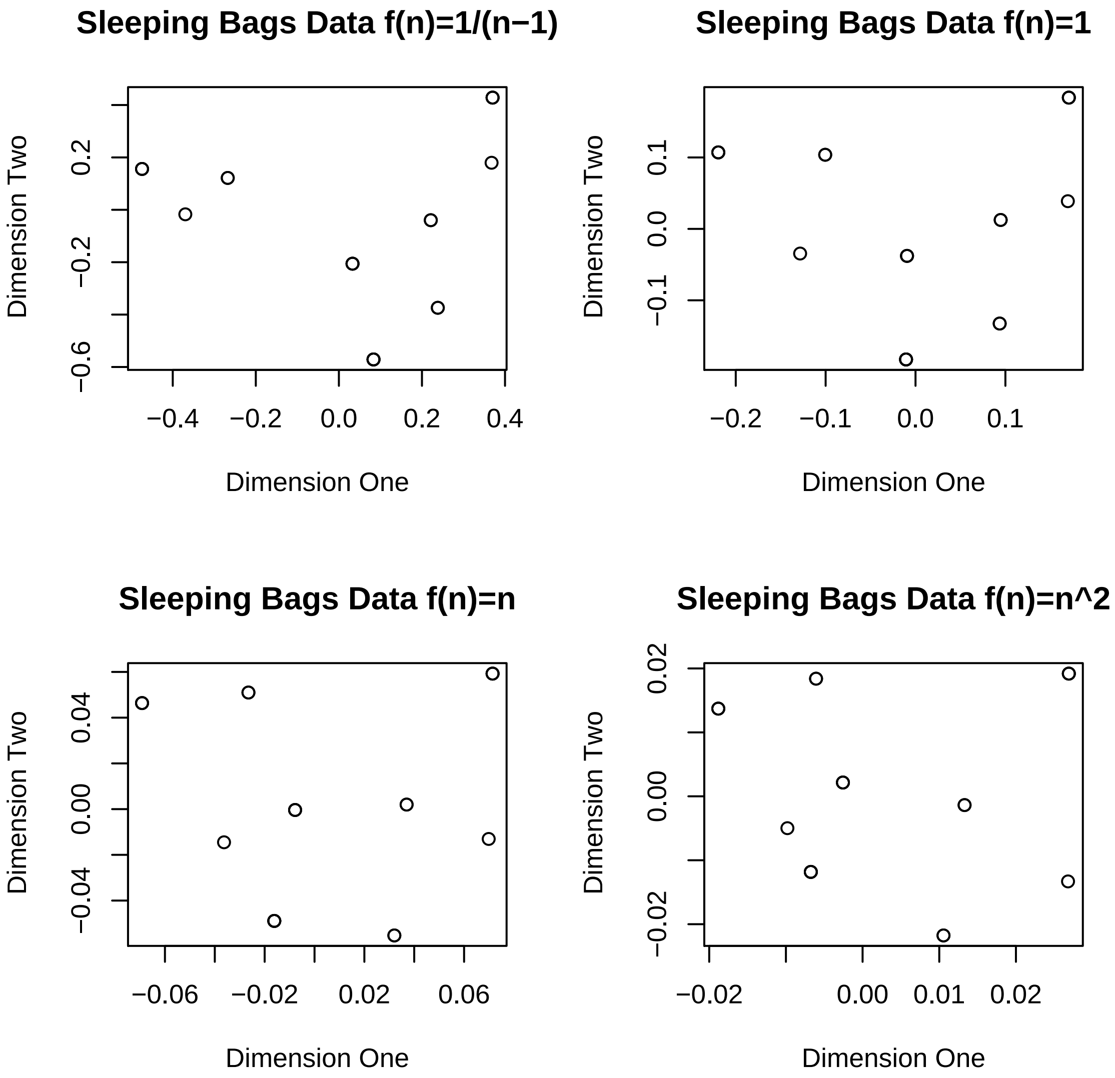


FIGURE 8. Sleeping Bags Example

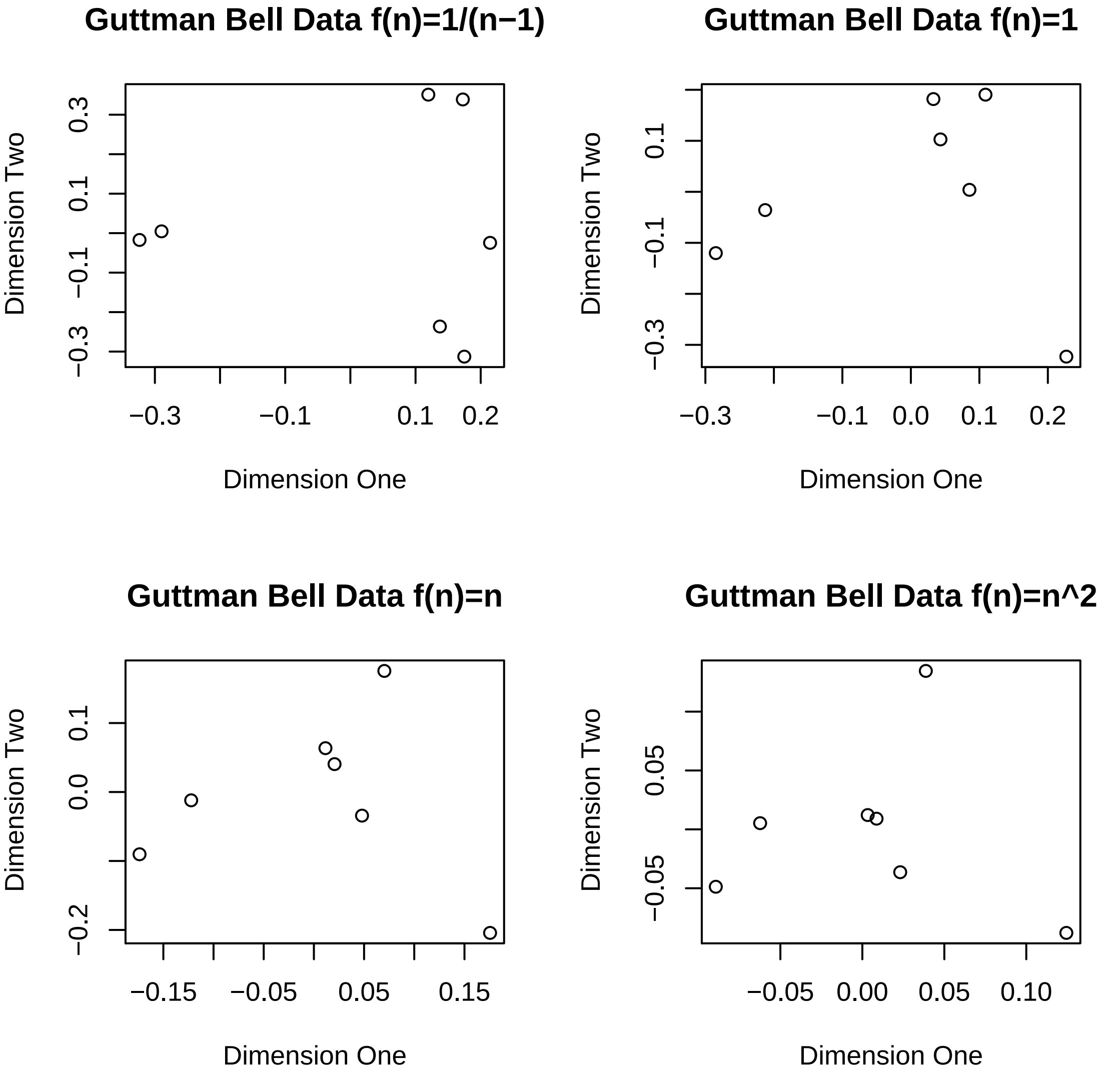


FIGURE 9. Guttman-Bell Example

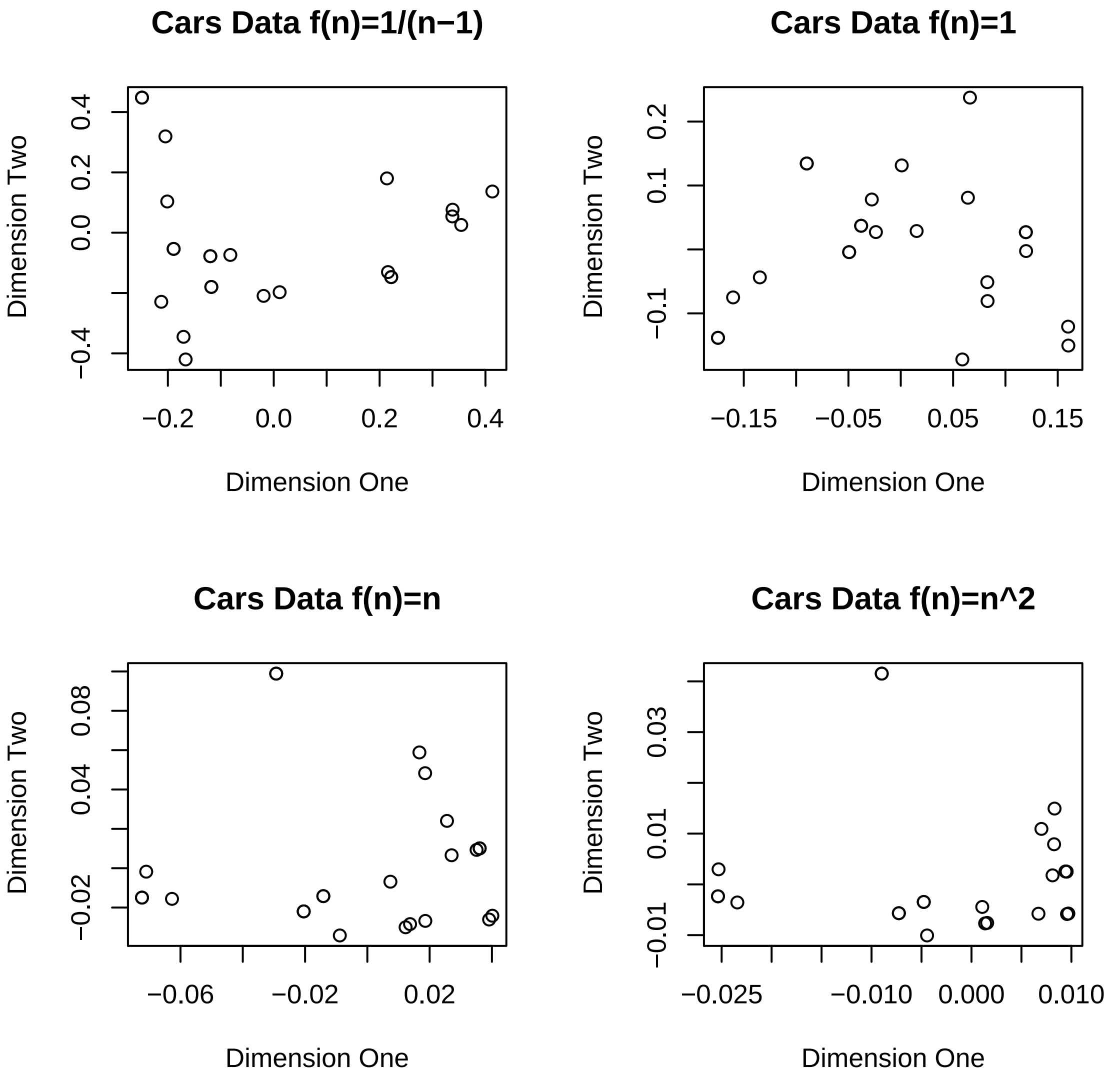


Figure 10. Cars Example

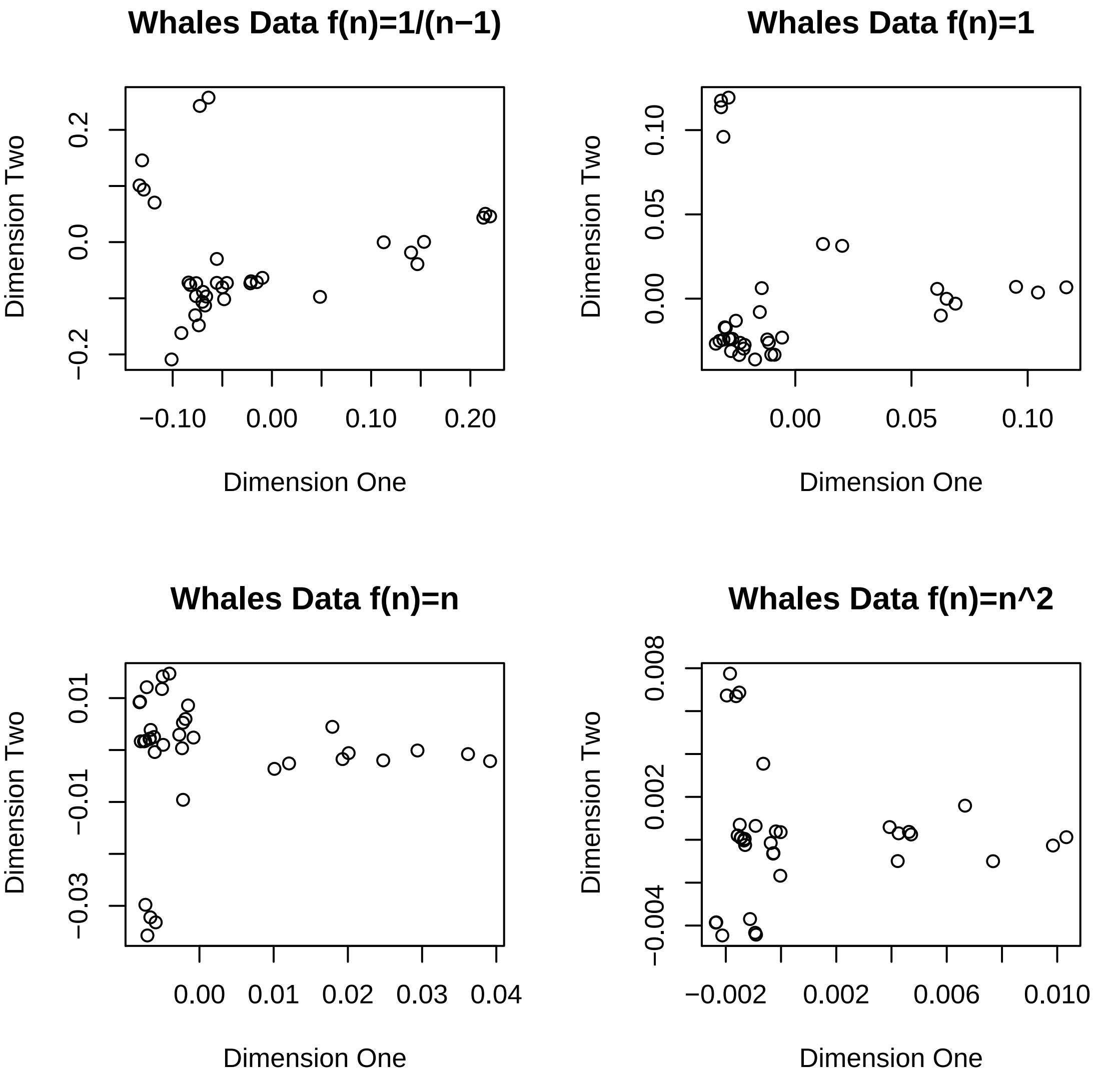


Figure 11. Whales Example

Appendix D. Code

```
# Function to perform a reweighted quadratic
# homogeneity analysis.
# Input a paving matrix tab and a weighting
# function fn. Output a list with singular vectors
# and singular values.

quadHom<-function(tab,fn,pdim=2) {
n<-dim(tab)[1]; m<-dim(tab)[2]; qdim=1+pdim; e<-rep(0,
   n); d<-rep(0,m)
for (i in 1:m) {
        g<-tab[,i]; s<-sum(g)
        d[i]<-s/fn(s)
        e<-e+fn(s)*g
        }
e<-ifelse(e==0,1,e); d<-ifelse(d==0,1,d)
z<-tab/sqrt(outer(e,d)); s<-svd(z,nu=qdim,nv=qdim)
l<-(s$d)[2:qdim]; x<-(s$u/sqrt(e))[,-1]; y<-(s$v/sqrt(
   d))[,-1]
list(x=x,y=y,l=l)
}

# fourPlots is a utility routine for making a two
# by two table of plots for an example, using
# four different reweightings

fourPlots <- function(indmat,dataname) {
s1 <- quadHom(indmat, function(n) if (n<2) 0 else 1/(n
   -1))
s2 <- quadHom(indmat, function(n) 1)
s3 <- quadHom(indmat, function(n) n)
s4 <- quadHom(indmat, function(n) n^2)
m1 <- paste(dataname,"Data␣f(n)=1/(n-1)",sep="␣")
```

```
m2 <- paste(dataname,"Data␣f(n)=n",sep="␣")
m3 <- paste(dataname,"Data␣f(n)=1",sep="␣")
m4 <- paste(dataname,"Data␣f(n)=n^2",sep="␣")
pdf(paste(dataname,"2X2.pdf",sep=""))
nf <- layout(matrix(c(1,2,3,4),2,2))
plot(s1$x,xlab="Dimension␣One",ylab="Dimension␣Two",
    main=m1)
plot(s3$x,xlab="Dimension␣One",ylab="Dimension␣Two",
    main=m2)
plot(s2$x,xlab="Dimension␣One",ylab="Dimension␣Two",
    main=m3)
plot(s4$x,xlab="Dimension␣One",ylab="Dimension␣Two",
    main=m4)
dev.off()
}

# expand converts a matrix or data-frame to an
# indicator supermatrix and then converts this
# to a data-frame again.
# By default NA becomes zero and constant rows
# and columns  are eliminated.

expand<-function(tab,clean=TRUE,zero=TRUE) {
n<-dim(tab)[1]; m<-dim(tab)[2]; g<-matrix(0,n,0); l
    <-rep("",0)
lab1<-labels(tab)[[1]]; lab2<-labels(tab)[[2]]
for (j in 1:m) {
        y<-as.factor(tab[,j]); h<-levels(y)
        g<-cbind(g,ifelse(outer(y,h,"=="),1,0))
        l<-c(l,paste(lab2[j],"_",h,sep=""))
        }
g<-as.data.frame(g,row.names=lab1)
names(g)<-l
if (zero) g<-ifelse(is.na(g),0,as.matrix(g))
if (clean) {
        g<-g[which(rowSums(g)>0),which(colSums(g)>0)]
        g<-g[,which(colSums(g)<n)]
        }
return(g)
}
```

References

J. De Leeuw. Homogeneity Analysis of Pavings. URL `http://preprints.stat.ucla.edu/389/homPeig.pdf`. August 2003.

J. De Leeuw. Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition. *Computational Statistics and Data Analysis*, in press.

J. De Leeuw and G. Michailidis. Graph Layout Techniques and Multidimensional Data Analysis. In T. Bruss and L. LeCam, editors, *Festschrift for Thomas S. Ferguson*. Institute of Mathematical Statistics, 1999.

J. De Leeuw and G. Michailidis. Weber Correspondence Analysis: The One-dimensional Case. *Journal of Computational and Graphical Statistics*, 13:946–953, 2004.

A. Gifi. *Nonlinear multivariate analysis*. Wiley, Chichester, England, 1990.

T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, Heidelberg, New York, second edition, 1976.

G. Michailides and J. De Leeuw. Homogeneity Analysis Using Absolute Deviations. *Computational Statistics and Data Analysis*, 48:587–603, 2004.

G. Michailidis and J. De Leeuw. The Gifi system for Descriptive Multivariate Analysis. *Statistical Science*, 13:307–336, 1998.

T. Papadopulou and M.I.A. Lourakis. Estimating the Jacobian of the Singular Value Decomposition: Theory and Applications. In D. Vernon, editor, *Proceedings of the 6th European Conference on Computer Vision*, pages 554–570. Springer, 2000.

J. L. Peschar. *School, Milieu, Beroep*. Tjeek Willink, Groningen, The Netherlands, 1975.

J.L.A. Van Rijckevorsel. *The Application of Fuzzy Coding and Horseshoes in Multiple Correspondence Analysis*. PhD thesis, University of Leiden, The Netherlands, 1987. Also published in 1987 by DSWO-Press, Leiden, The Netherlands.

B.F. Schriever. *Order Dependence*. PhD thesis, University of Amsterdam, The Netherlands, 1985. Also published in 1985 by CWI, Amsterdam, The Netherlands.

R.J. Vaccaro. A Second-order Perturbation Expansion for the SVD. *SIAM Journal on Matrix Analysis and Applications*, 15:661–671, 1994.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu

E-mail address, Vanessa Beddo: beddo@stat.ucla.edu

URL, Vanessa Beddo: http://www.stat.ucla.edu/~beddo