Model-based vs. classic ordination

Bert van der Veen

Department of Mathematical Sciences, NTNU

Questions so far?

Criteria for a good ordination

(no artefacts)

How I first wanted to approach this

- Theory
- Equations
- Boring

Something slightly more interesting

- There is decades of literature on the performance of ordination methods
- Let's compare to GLLVMs to get a good feeling for how the method behaves
- PCA, CA, DCA, NMDS vs GLLVM all have their deficits that we can check against

1) PCA has a horseshoe
2) CA has an arch
3) DCA has a tongue
4) NMDS has.. something

Name that method

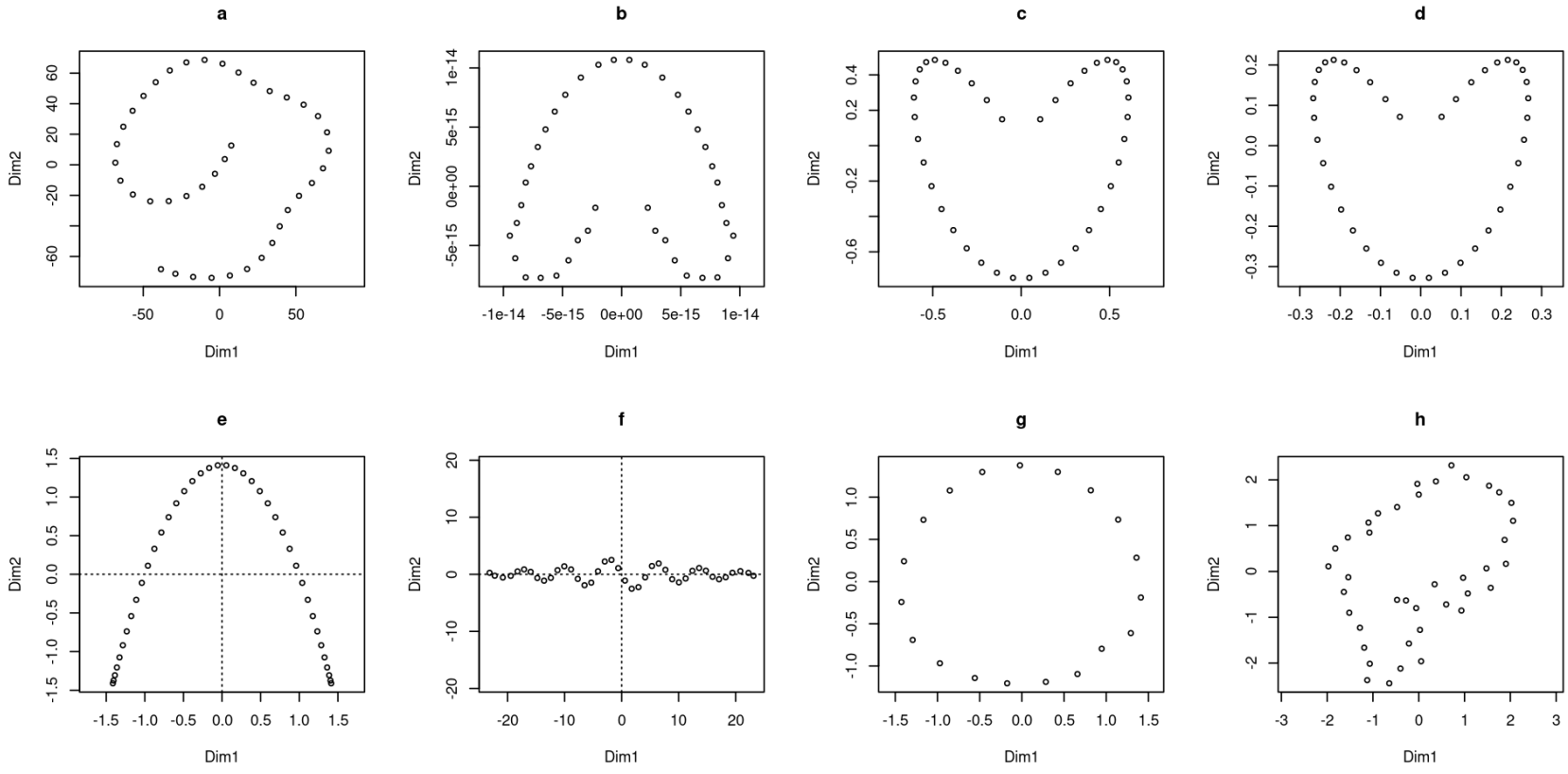


Figure 1: Can you identify the ordination methods?

Name that method: hint

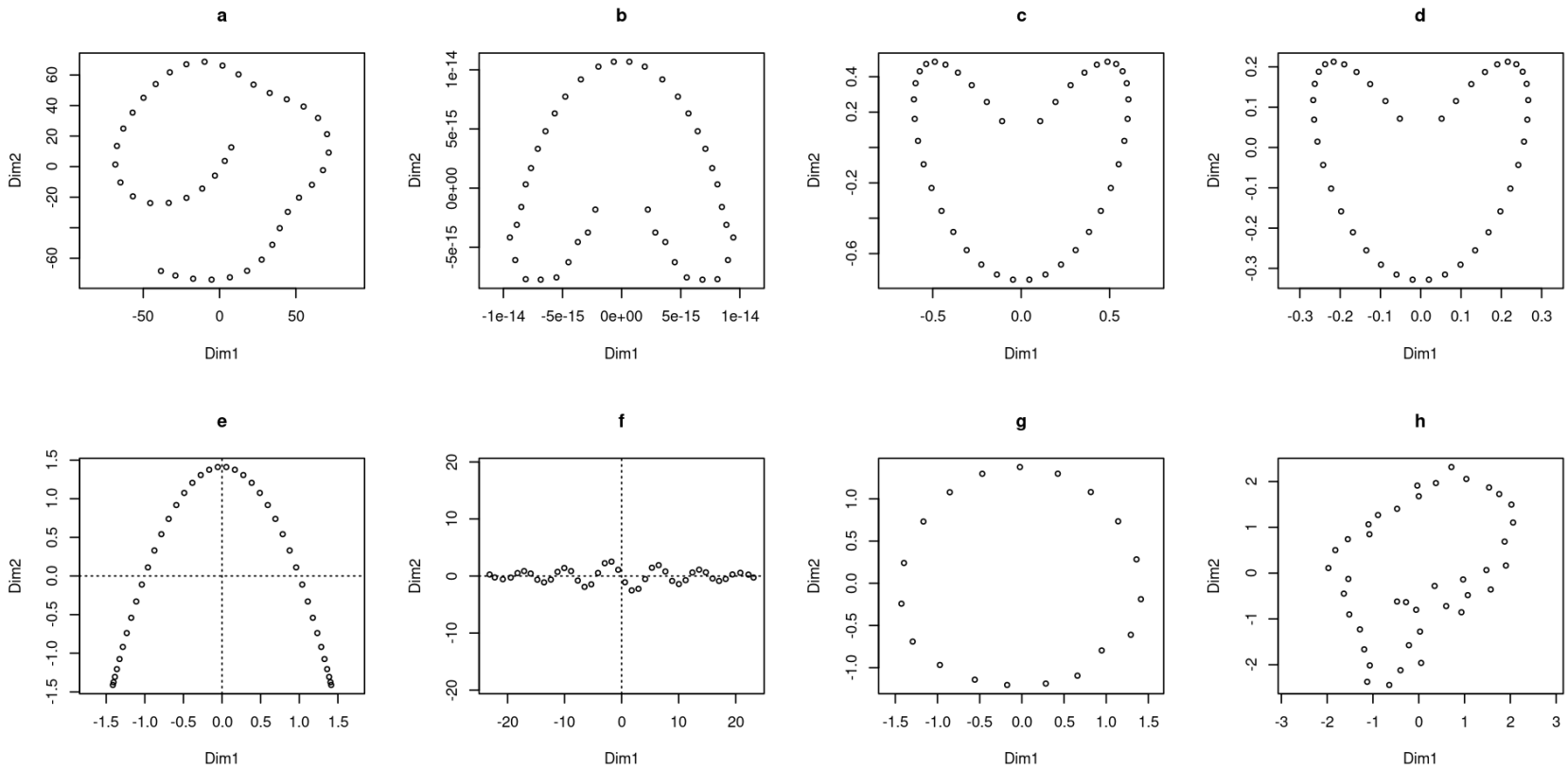


Figure 2: Can you identify the ordination methods?

- NMDS
- DCA

Let's go exploring

So let us use some "notorious" data and explore these

- Podani and Miklos (2002)
- Minchin (1987)

Podani and Miklos (2002)

Ecology, 83(12), 2002, pp. 3331–3343

RESEMBLANCE COEFFICIENTS AND THE HORSESHOE EFFECT IN PRINCIPAL COORDINATES ANALYSIS

J. Podani[1] and I. Miklós

Department of Plant Taxonomy and Ecology, Eötvös University, Pázmány P. s. 1/c H-1117 Budapest, Hungary

Podani and Miklos (2002)

- Four artificial datasets
- With courtesy of Gavin Simpson

1) **Single gradient with unimodal responses**
2) Single gradient with linear responses
3) Single gradient with unimodal responses (from Legendre and Legendre 1998)

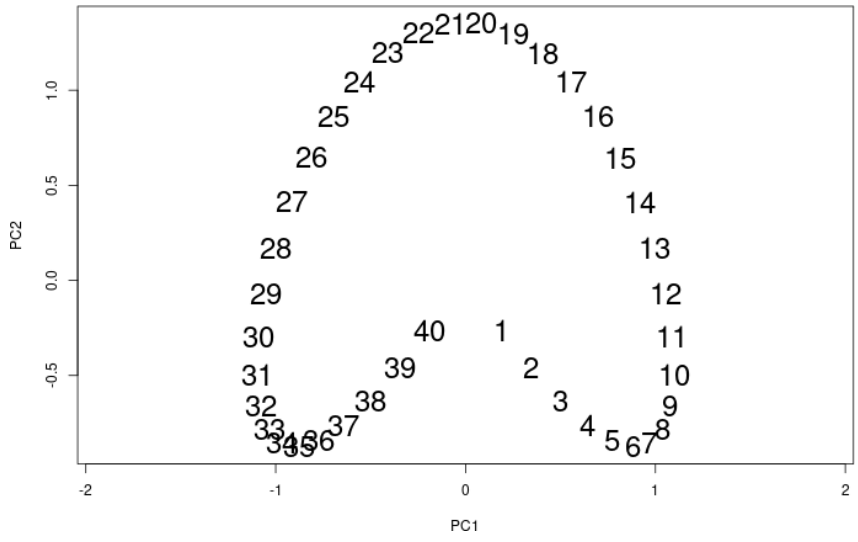
Load the datasets

```
tmp <- tempfile()
download.file("https://github.com/gavinsimpson/random_code/ra
              tmp, method = "wget")
source(tmp)
PM1 <- podani1()
PM2 <- podani2()
PM3 <- podani3()
```

Have a look at the data

PM1: PCA

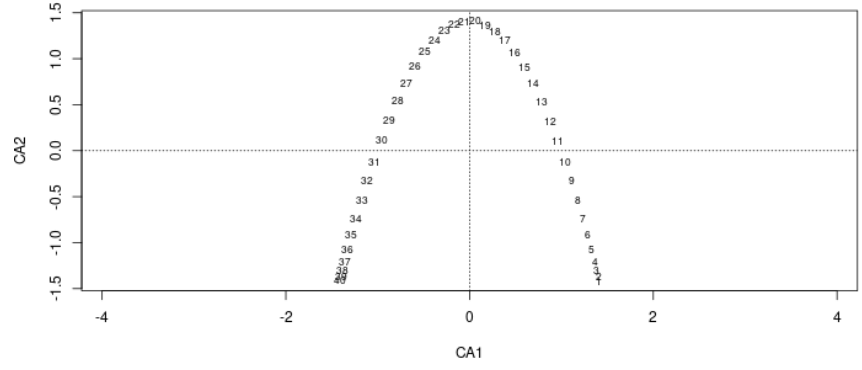
```
PCA <- prcomp(PM1)
vegan::ordiplot(PCA, type = "text", display = "sites", cex = 2)
```

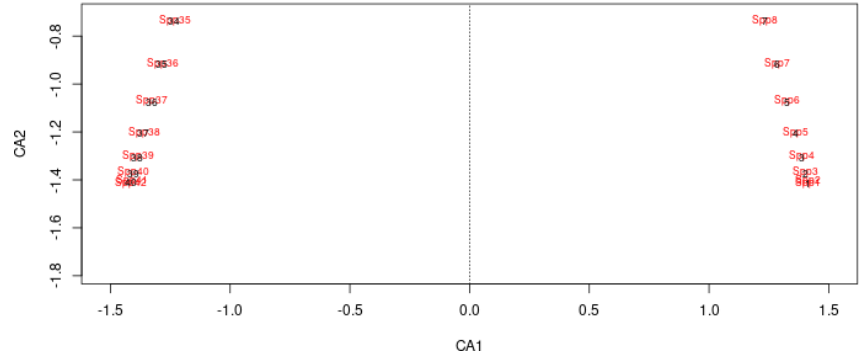


- First dimension is related to quadratic function of the second
- First and last sites are still close together in two-dimensional space
- Quadratic term comes on first dimension because it explains more variation than the linear term

PM1: CA

```
CA <- vegan::cca(PM1)
vegan::ordiplot(CA, type = "text", cex = 5, display = "sites");vegan::ordi
```

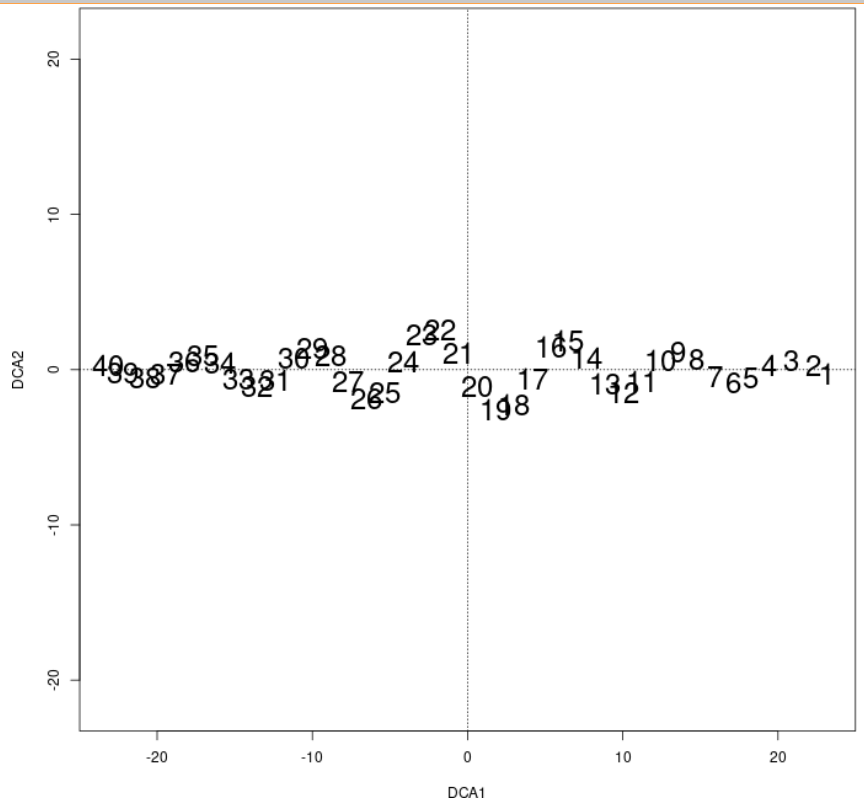




- First dimension is the quadratic function of the second
- No curvature inwards: first and last sites are not actually close together
- Spread of scores smaller on at beginning and end: edge effect

PM1: DCA

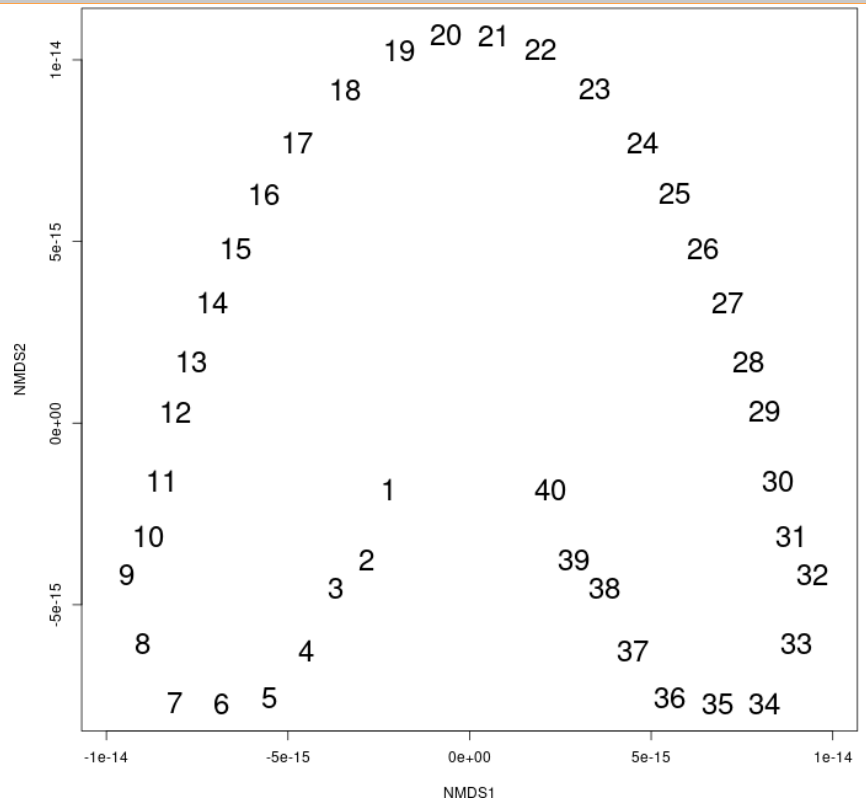
```
DCA <- vegan::decorana(PM1)
vegan::ordiplot(DCA, type = "text", display = "sites", cex = 2)
```



- I have no idea what DCA did here
- The procedure DCA uses is quite criticized
- It "smashes" the arch, and rescales to improve edge issues

PM1: NMDS

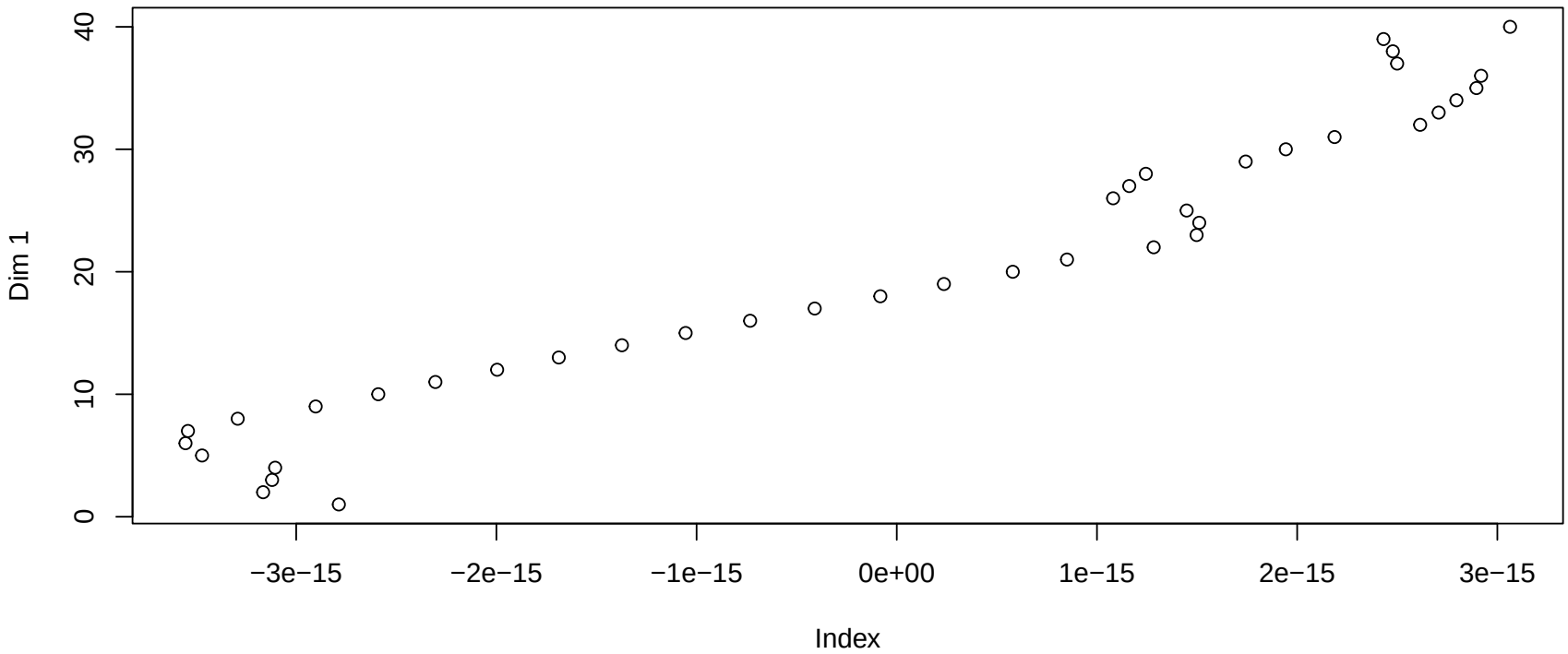
```
NMDS <- vegan::metaMDS(PM1, trace = 0)
vegan::ordiplot(NMDS, type = "text", display = "sites", cex = 2)
```



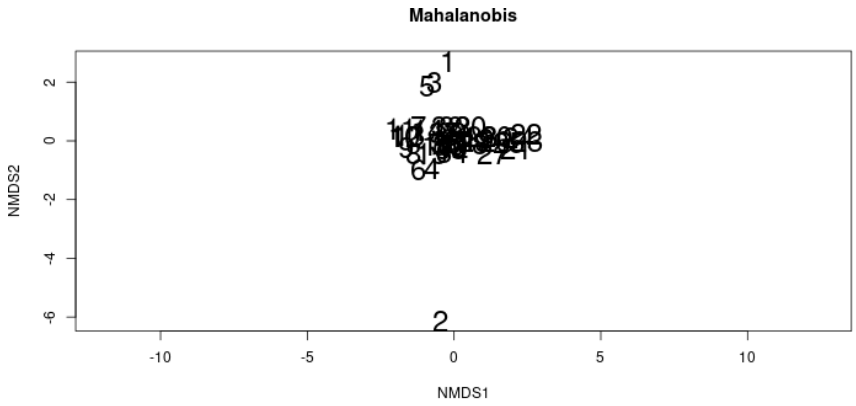
- NMDS also shows a horseshoe
- But note the range of the axes
- By default it uses Bray-Curtis distance

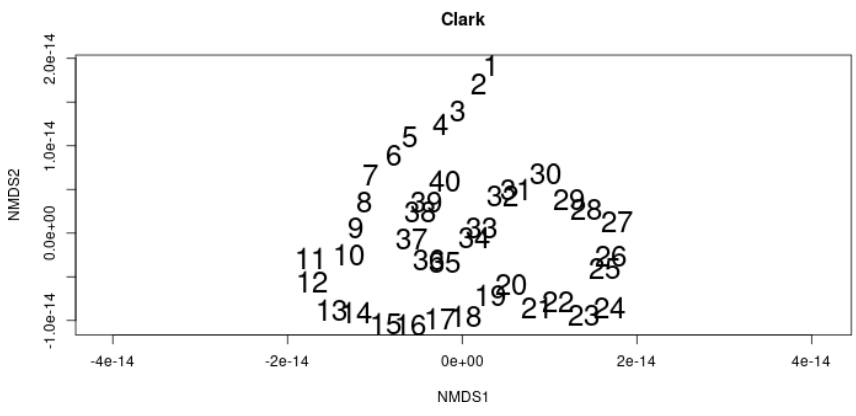
PM1: NMDS

```
NMDS <- vegan::metaMDS(PM1, k =1, trace = 0)
plot(NMDS$points, 1:nrow(PM1), xlab = "Index", ylab = "Dim 1")
```



PM1: NMDS funny plots

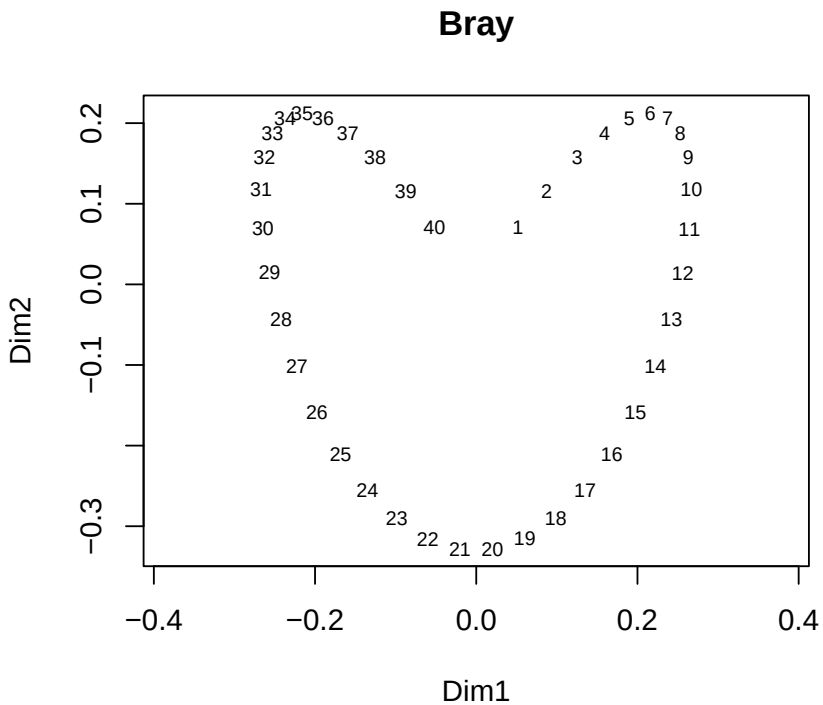


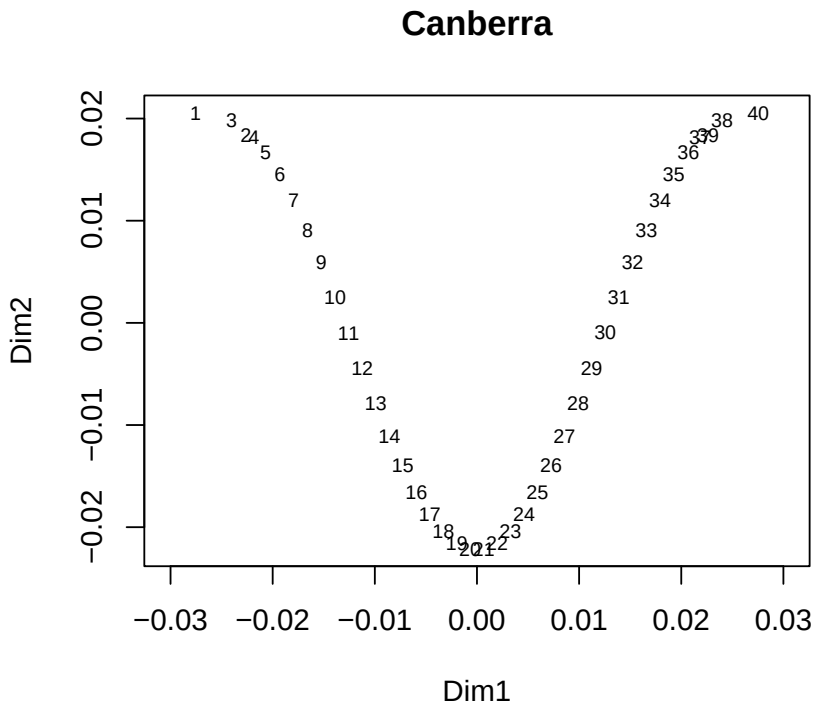


- Doesn't converge
- Not even for one of the axes

PM1: PCoA

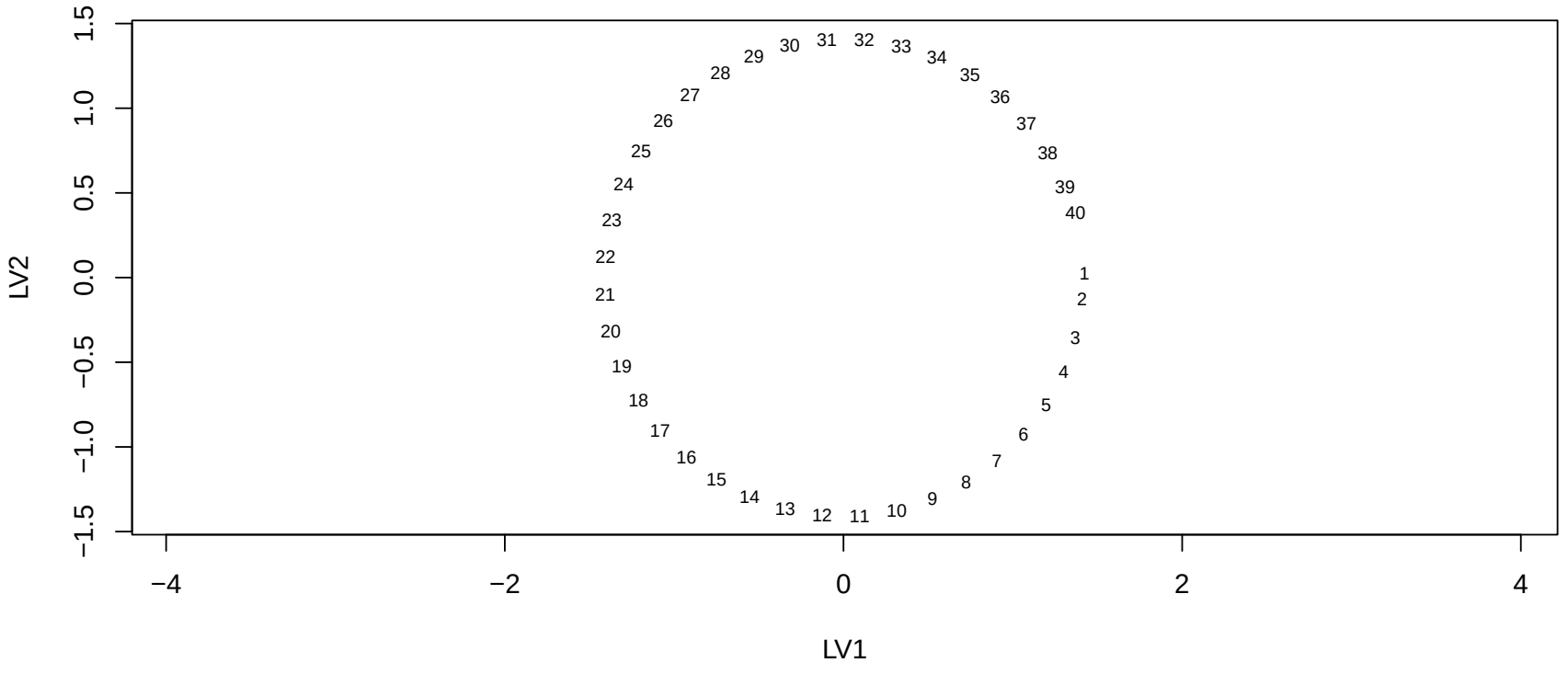
```
PCoA1 <- cmdscale(vegan::vegdist(PM1))
PCoA2 <- cmdscale(vegan::vegdist(PM1, method = "gower"))
vegan::ordiplot(PCoA1, type = "text", main = "Bray")
vegan::ordiplot(PCoA2, type = "text", main = "Canberra")
```





PM1: gllvm

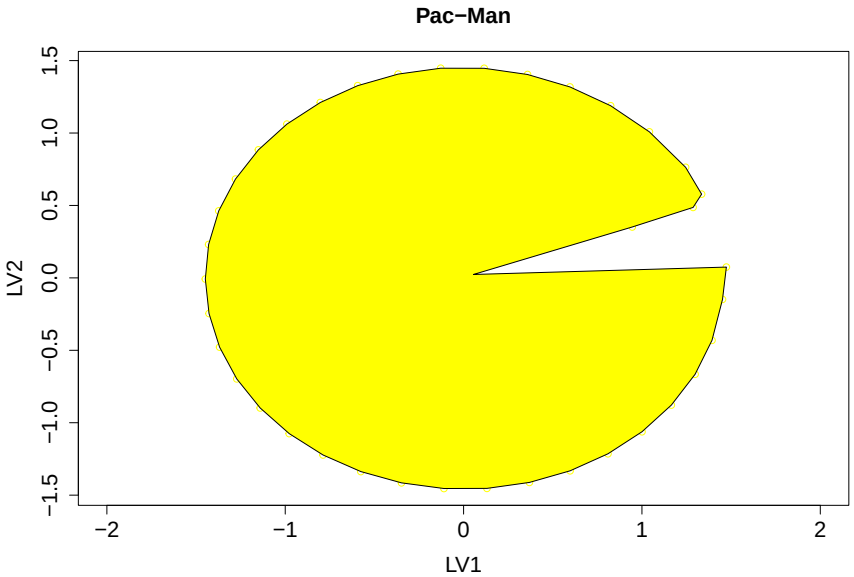
```
uord <- gllvm::gllvm(PM1, num.lv = 2, family = "poisson")
vegan::ordiplot(uord, display = "sites", type = "text") # requires scores.
```

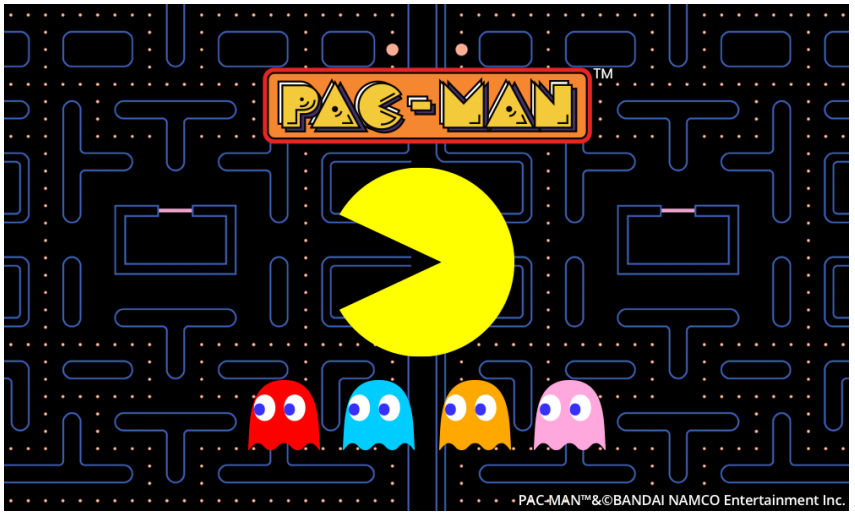


PM1: gllvm

When we flip that around, disconnect the circle, color it yellow, and we get...

Pac-man!


Pac−Man
LV2
LV1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2


PAC-MAN
PAC-MAN™&©BANDAI NAMCO Entertainment Inc.

Minchin (1987)

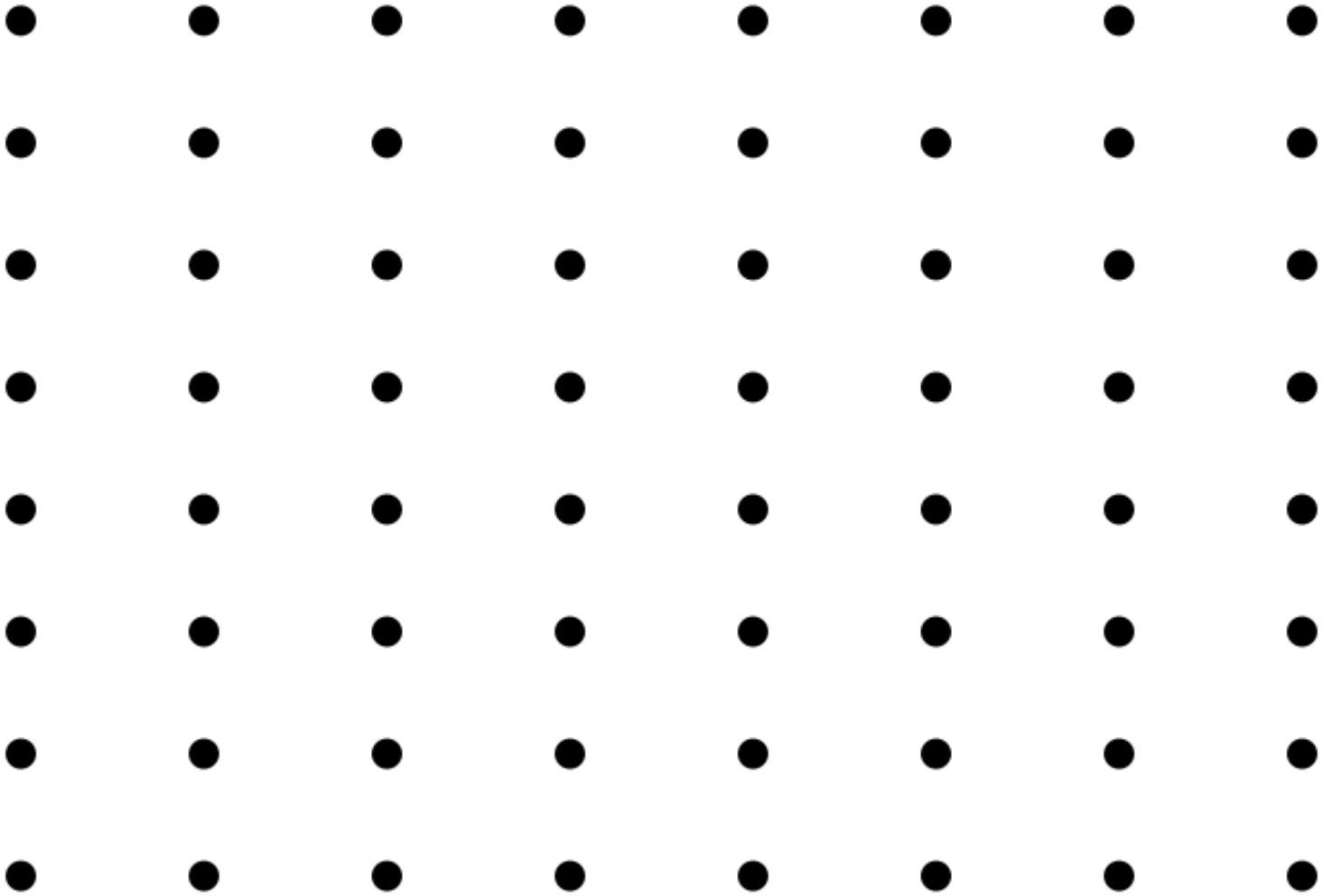
Vegetatio 69: 89–107, 1987

An evaluation of the relative robustness of techniques for ecological ordination

Peter R. Minchin*
CSIRO Division of Water and Land Resources, G.P.O. Box 1666, Canberra, 2601, Australia

- Simulated using the COMPAS software on a lattice
- Skewed and asymmetric response curves
- Found all methods except NMDS to perform poorly

Lattice



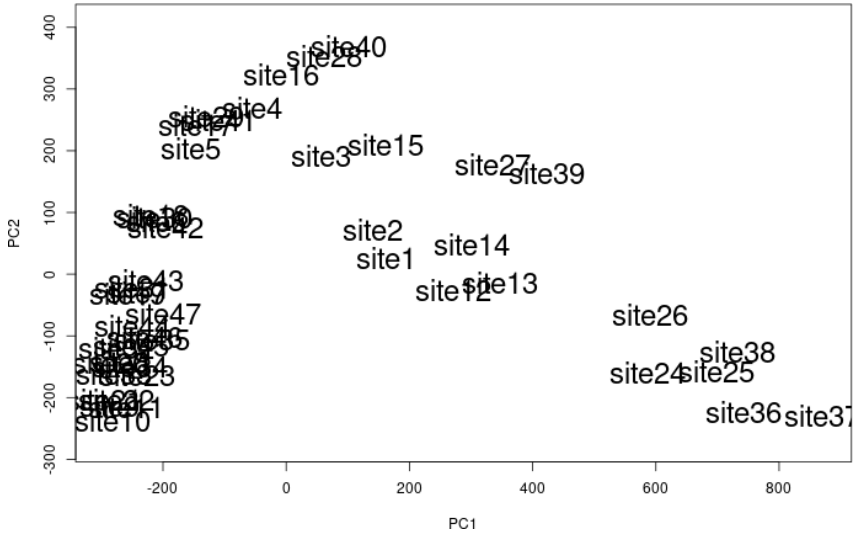
Load the data

```
MC <- read.csv("https://raw.githubusercontent.com/BertvanderV
MC[is.na(MC)] <- 0
```

Have a look at the data

MC: PCA

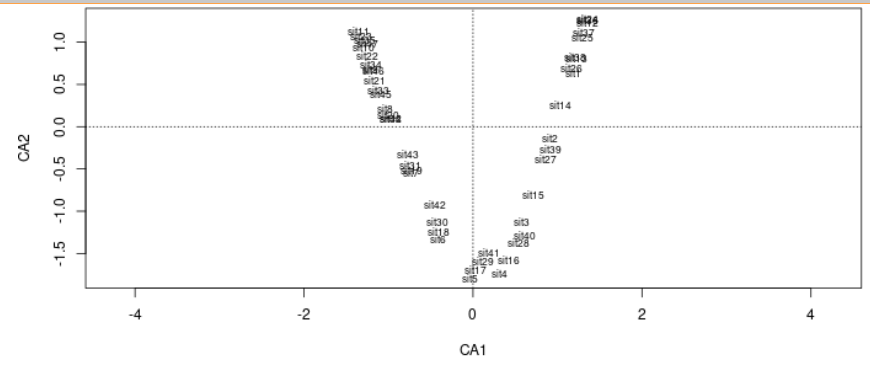
```
PCA <- prcomp(MC)
vegan::ordiplot(PCA, type = "text", display = "sites", cex = 2)
```

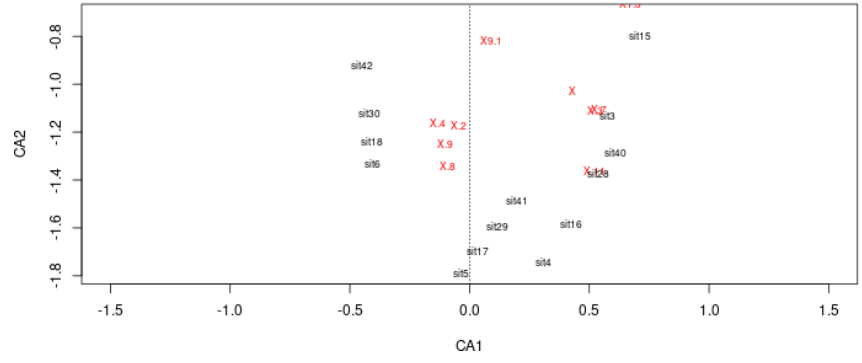


No lattice: PCA is not doing well

MC: CA

```
CA <- vegan::cca(MC)
vegan::ordiplot(CA, type = "text",display = "sites");vegan::ordiplot(CA, t
```

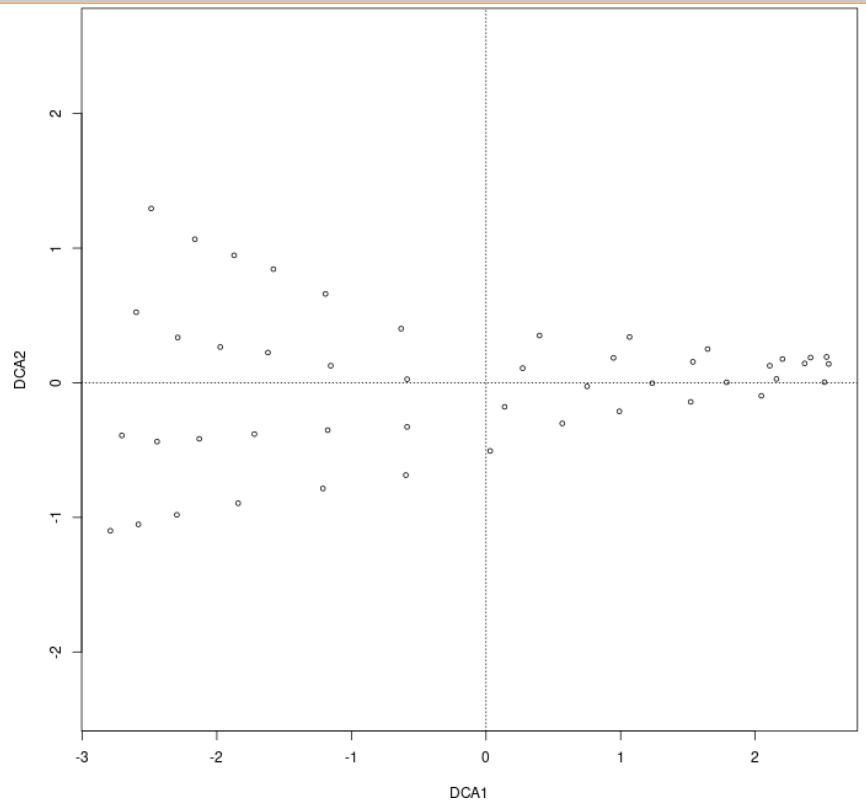




No lattice: CA is not doing well

MC: DCA

```
DCA <- vegan::decorana(MC)
vegan::ordiplot(DCA, type = "text", display = "sites")
```



- No lattice: DCA is not doing well
- ter Braak and Smilauer (2015) reanalysed with a log-transform

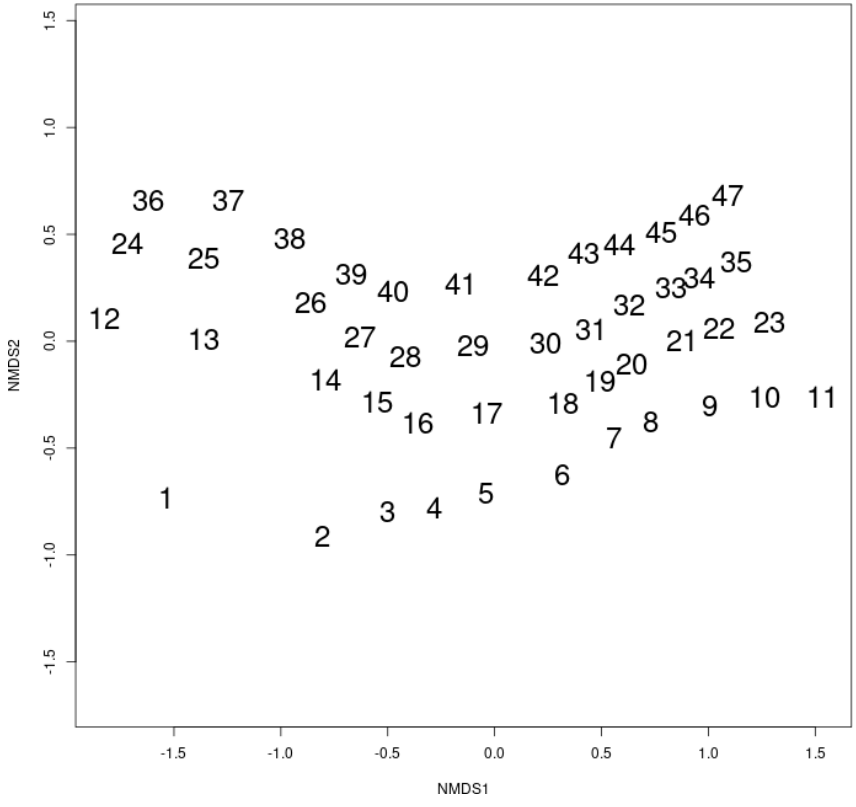
MC: DCA

```
DCA2 <- vegan::decorana(log1p(MC))
vegan::ordiplot(DCA2, type = "text", display = "sites")
```

We start to see a lattice!

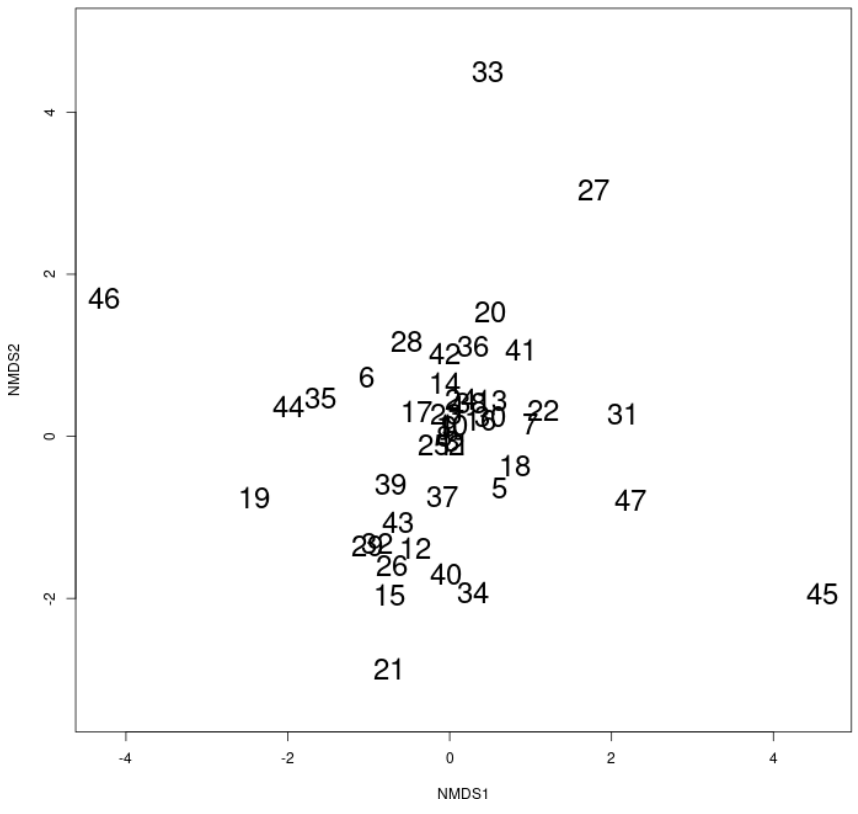
MC: NMDS

```
NMDS <- vegan::metaMDS(MC, trace = 0)
vegan::ordiplot(NMDS, type = "text", display = "sites", cex = 2)
```



- A lattice: NMDS is doing quite well
- NMDS was found to be robust to different response models
- Partly because it condenses to site-level information (no species)

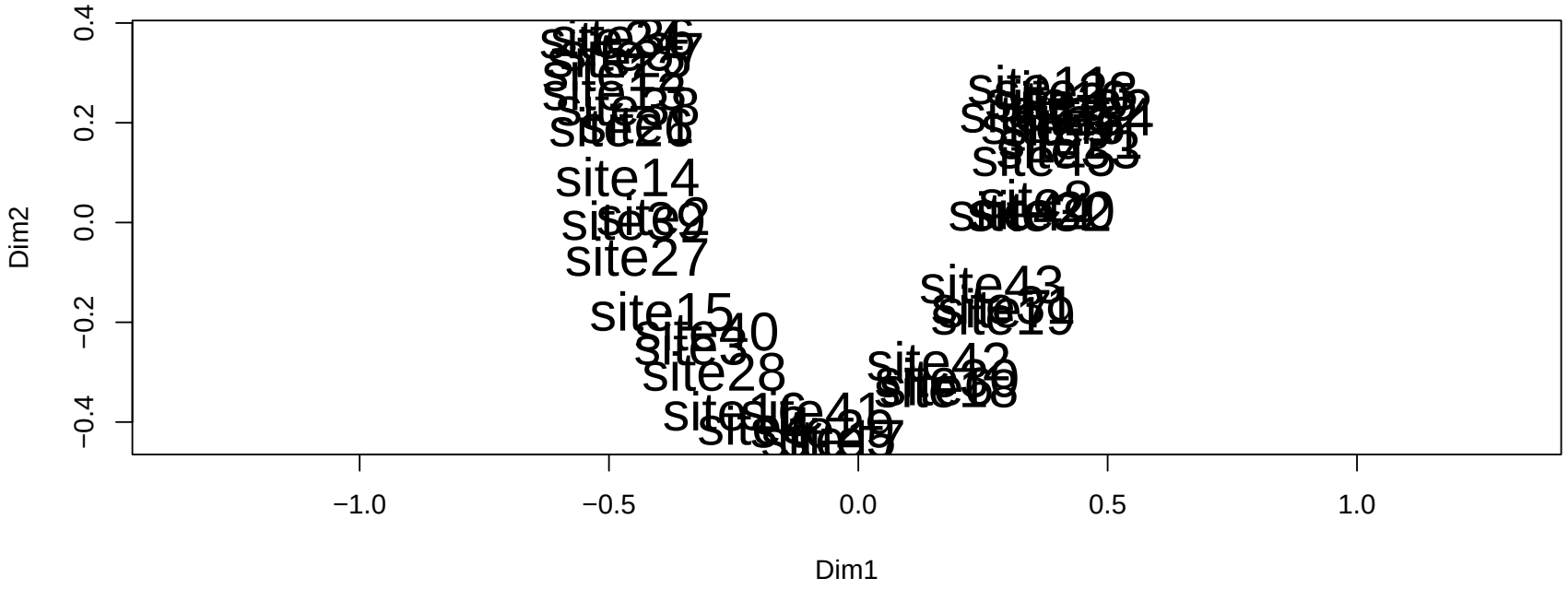
MC: NMDS with Mahalanobis distance



- This looks terrible
- So the used distance measure matters

MC: PCoA

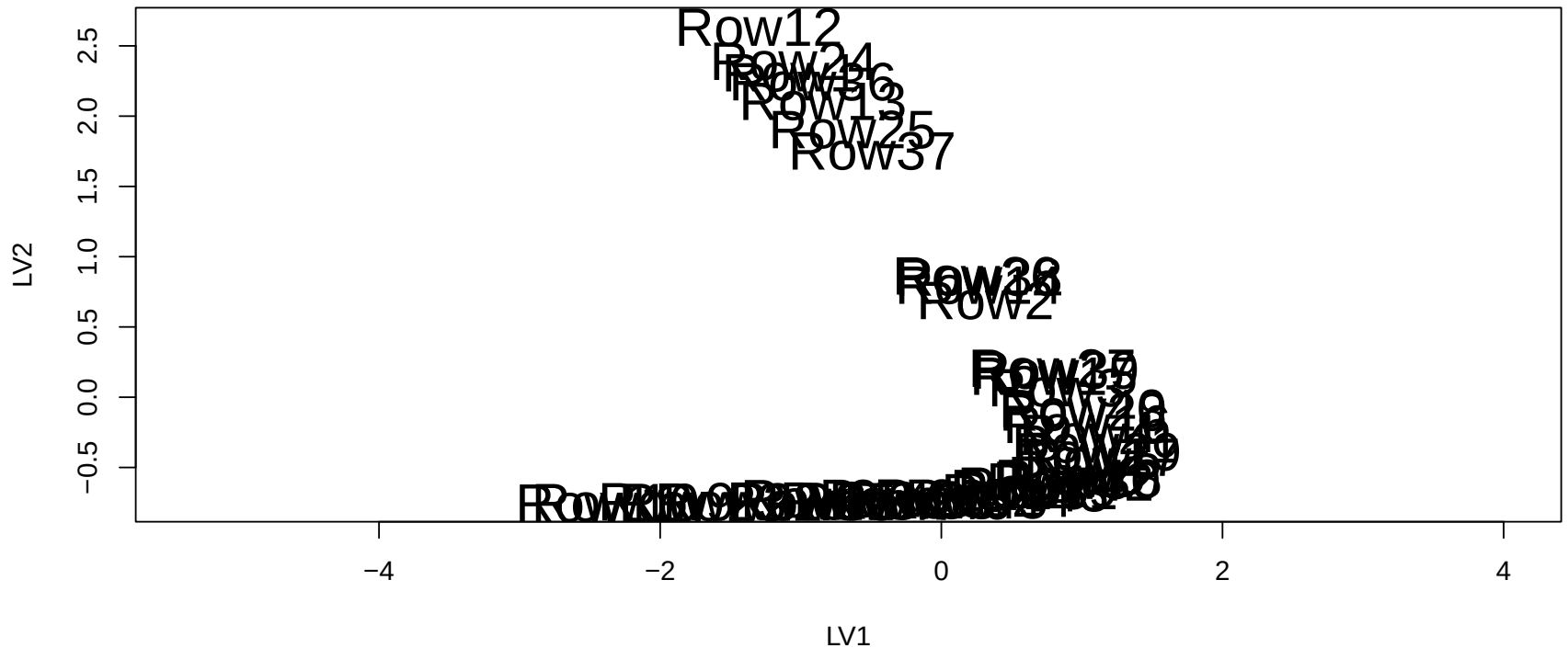
```
vegan::ordiplot(PCoA, type = "text", cex = 2)
```



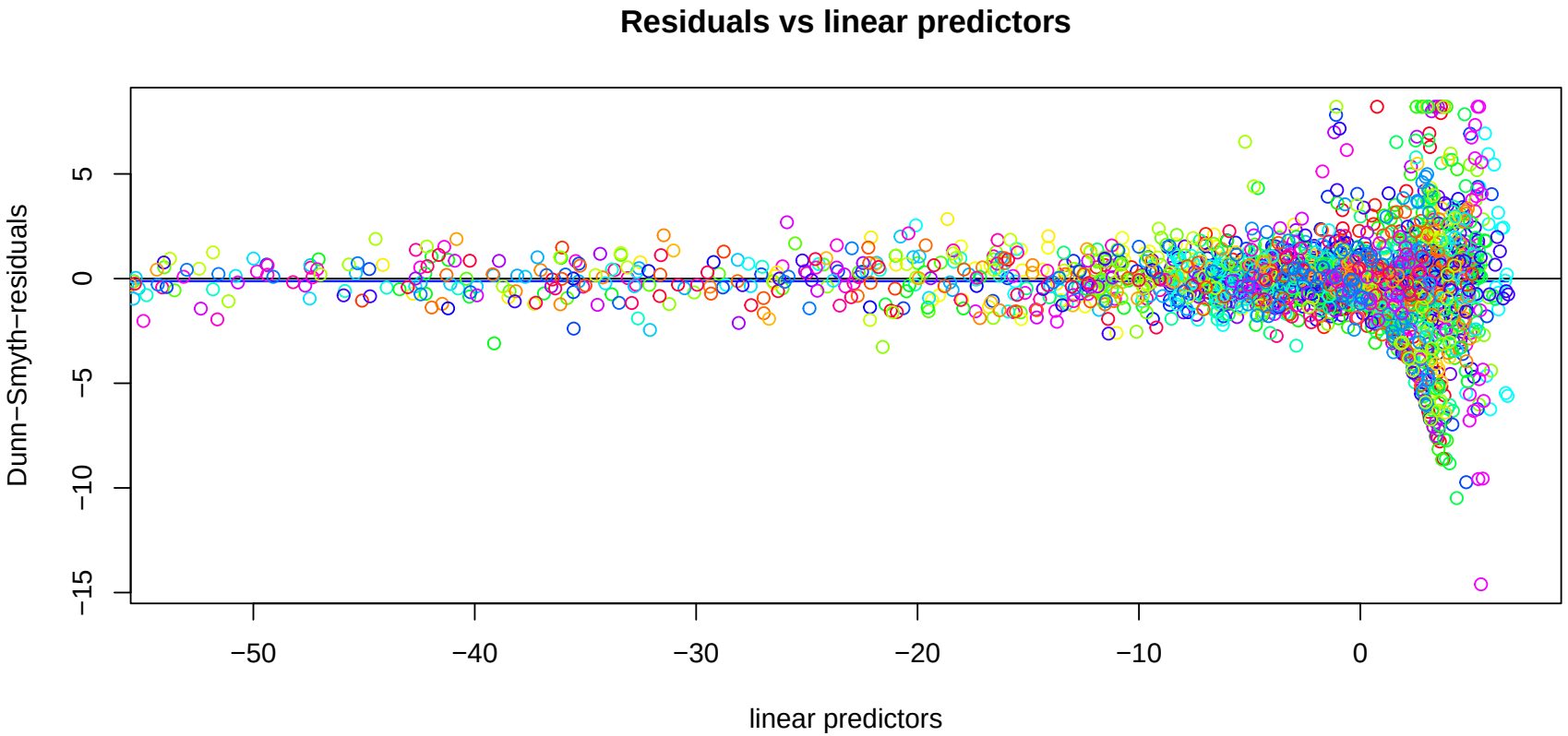
PCoA fails with different distance measures.

MC: gllvm (Poisson)

```
uord <- gllvm::gllvm(MC, num.lv = 2, family = "poisson")
vegan::ordiplot(uord, display = "sites", type = "text", cex = 2) # require
```



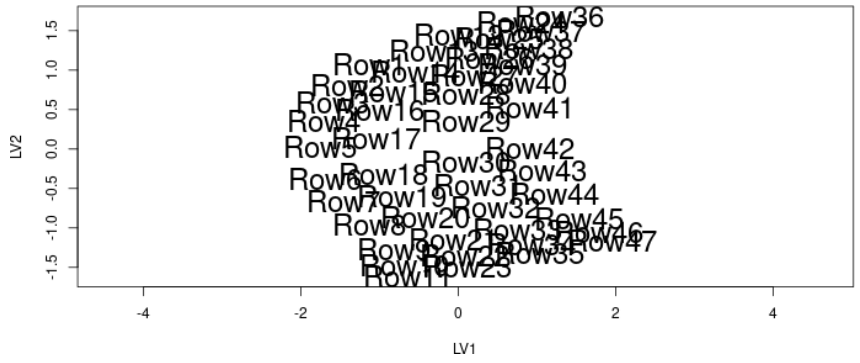
MC: gllvm diagnostics (Poisson)



Clearly, a Poisson distribution will not cut it. At least we can detect a bad ordination.

MC: gllvm (NB)

```
uord <- gllvm::gllvm(MC, num.lv = 2, family = "negative.binomial", row.eff
vegan::ordiplot(uord, display = "sites", type = "text", cex = 2) # require
```



- A lattice: gllvm is doing quite well
- We did have to use a few tricks (NB + row-effects)
- GLLVMs non-linearly relate the ordination space to the data

A real dataset

PeerJ

Multivariate ordination identifies vegetation types associated with spider conservation in brassica crops

Hafiz Sohaib Ahmed Saqib[1,2], Minsheng You[1,2,3] and Geoff M. Gurr[1,2,3,4]

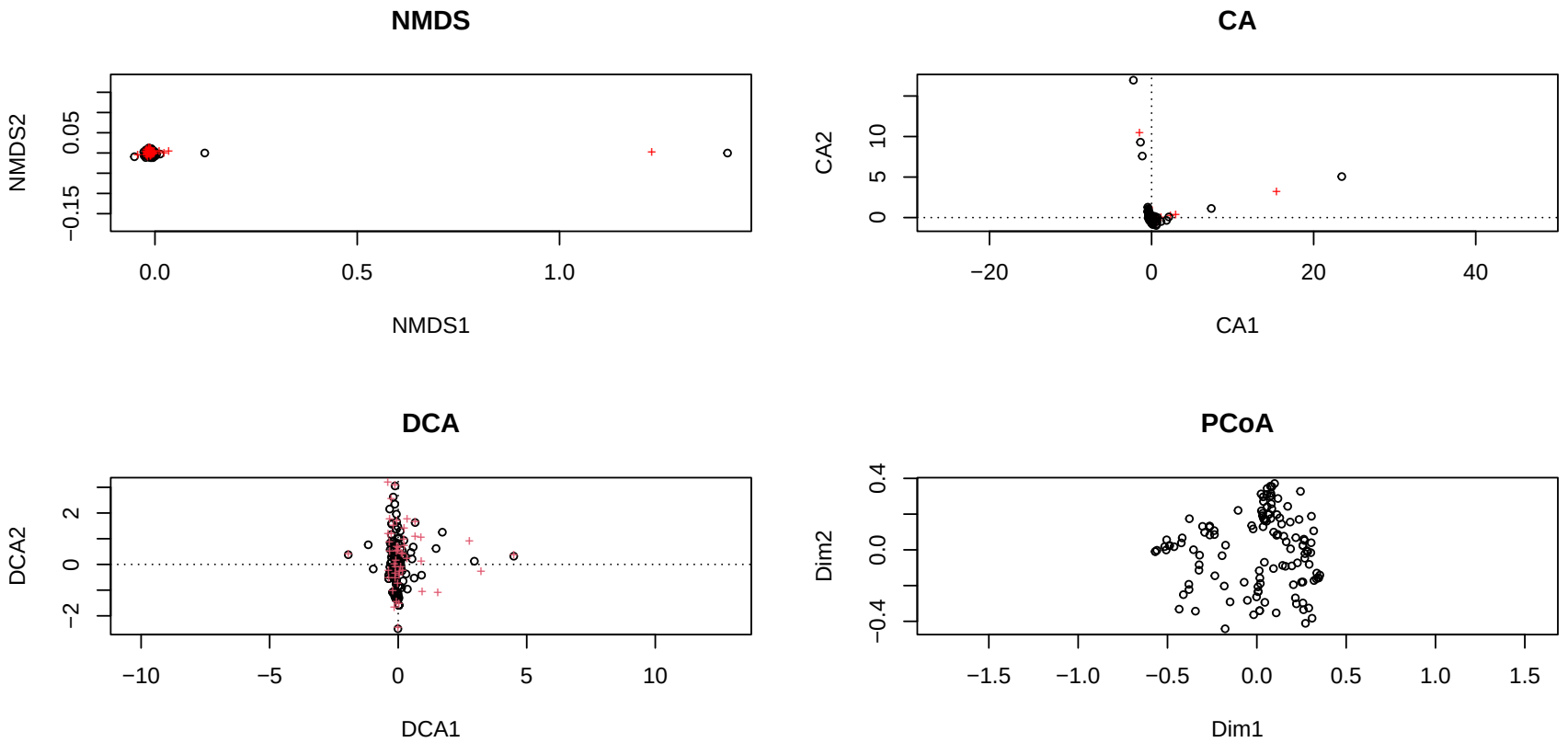
[1] State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, Fujian Agriculture and Forestry University, Fuzhou, China
[2] Institute of Applied Ecology, Fujian Agriculture and Forestry University, Fuzhou, China
[3] Fujian-Taiwan Joint Centre for Ecological Control of Crop Pests, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China
[4] Graham Centre for Agricultural Innovation, Charles Sturt University, Orange, New South Wales, Australia

- Abundance of spiders in **brassica** crops
- 3 sites in China
- Sampled spiders at 25-29 points in 50x50m grids
- Grids were in crops and adjacent vegetation

Difficulty: many zeros

Classical ordinations of SQ data

```
## species scores not available
```

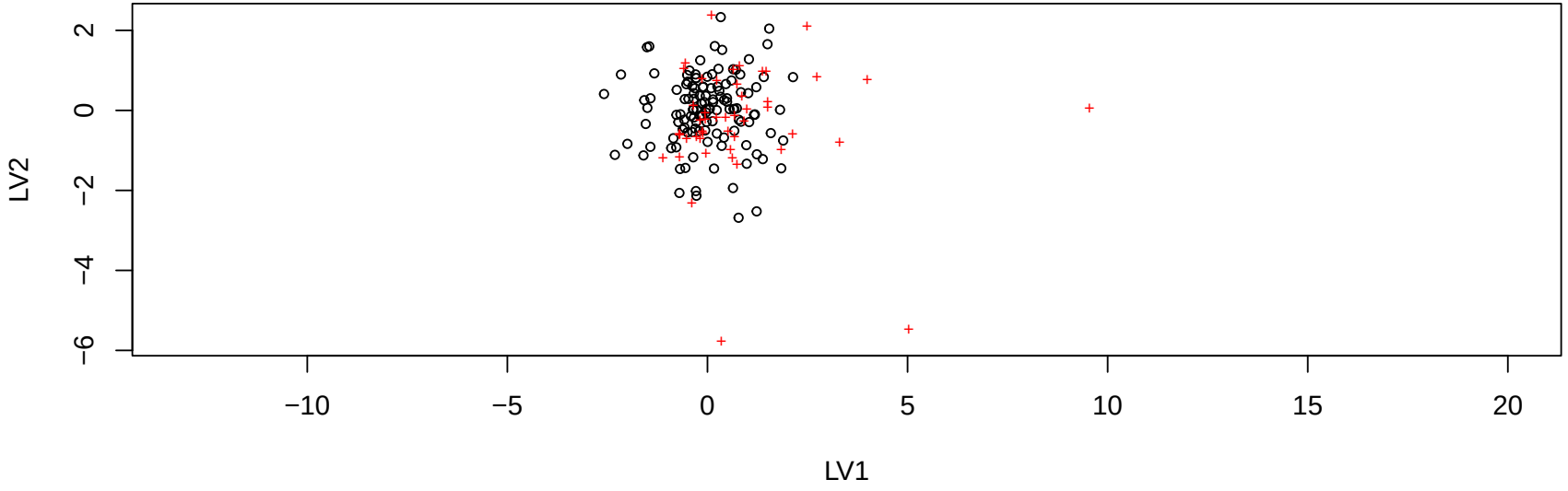


But, what makes for a good ordination method?

Step 1): A poisson gllvm

We have abundances, so we (usually) start at fitting a Poisson model

```
model<-gllvm::gllvm(Y, num.lv=2, family = "poisson", row.eff = ~(1|sites),
vegan::ordiplot(model) # requires scores.gllvm
```



An additional random effect is included to account for replication within sites

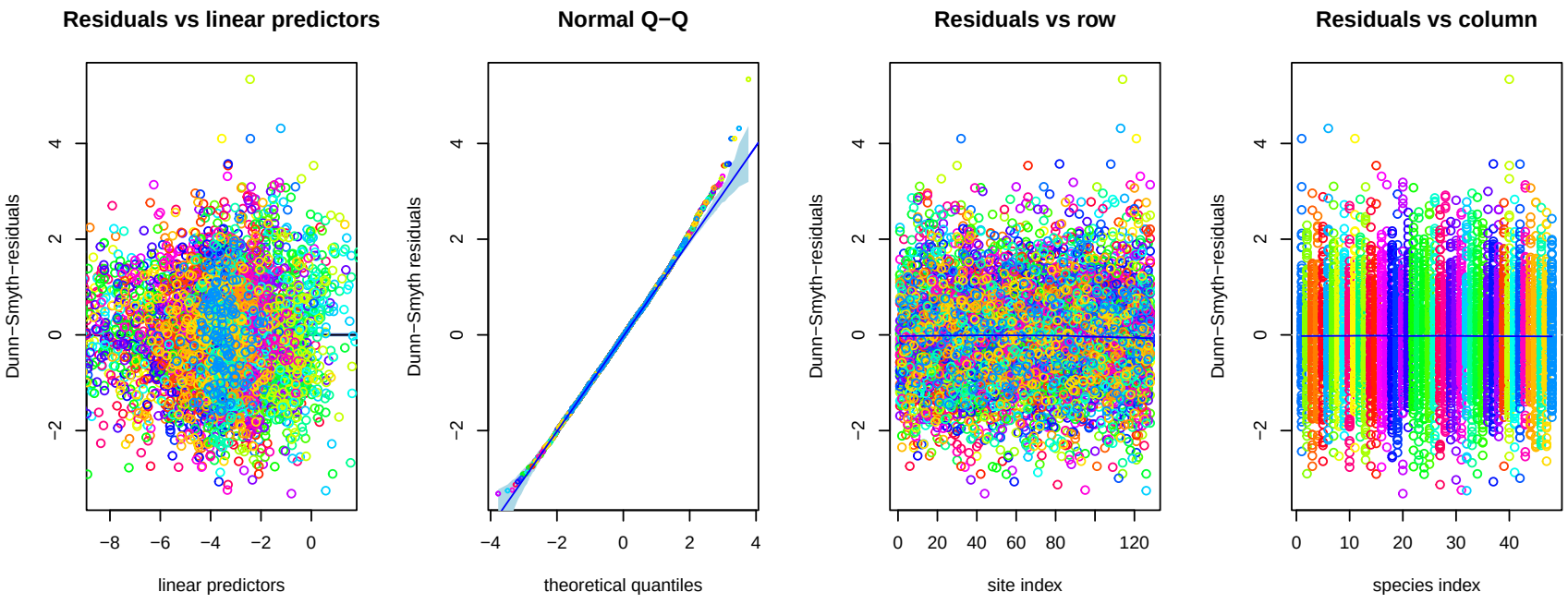
Evaluating fit

In GLLVMs we can quantitatively assess if we have a decent model

1) We look at the likelihood (e.g., with information criteria)
2) We check residuals (to see if we have accounted for all data properties)

Step 2): residuals

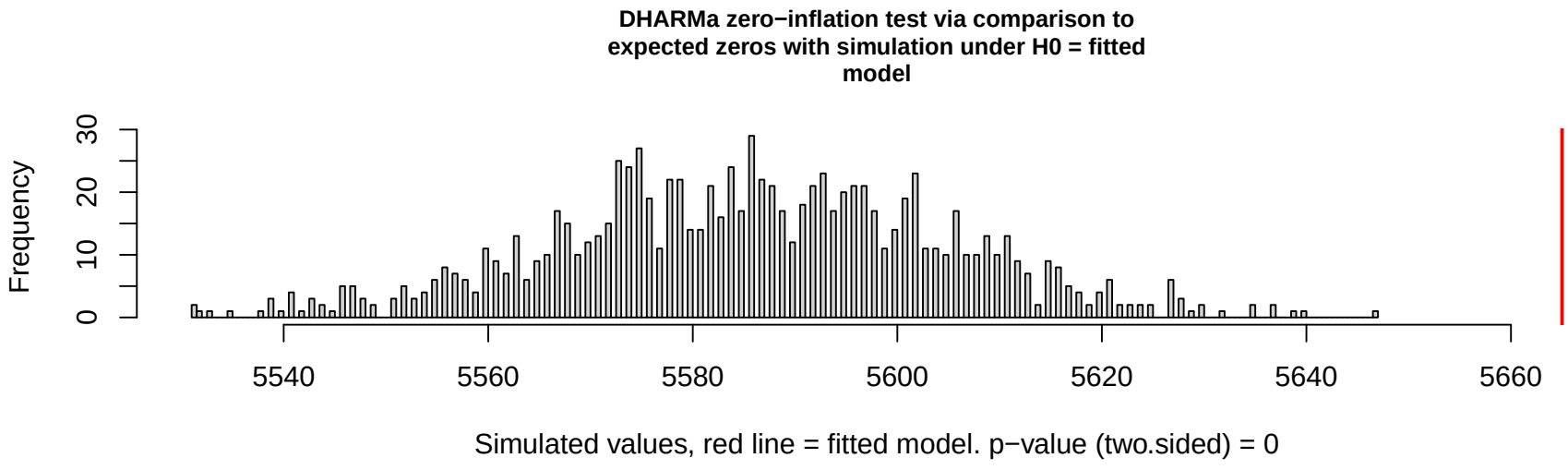
```
plot(model, which = 1:4)
```



We see some deviation in the tail. Maybe we should adjust our model?

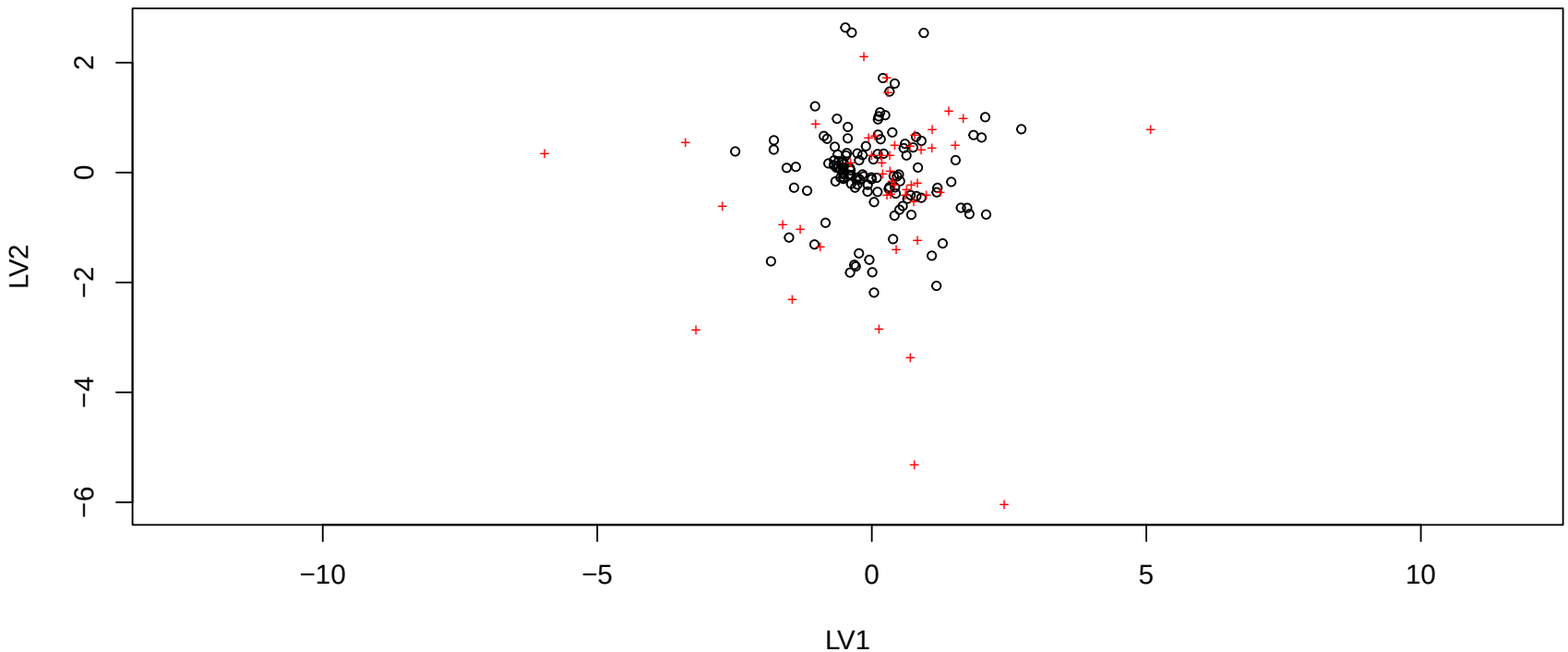
Step 3): checking for zero-inflation

```
sim <- do.call("cbind", replicate(1000, c(as.matrix(gllvm::simulate(model,
dharma <- DHARMa::createDHARMa(simulatedResponse=sim, observedResponse=as.
DHARMa::testZeroInflation(dharma)
```

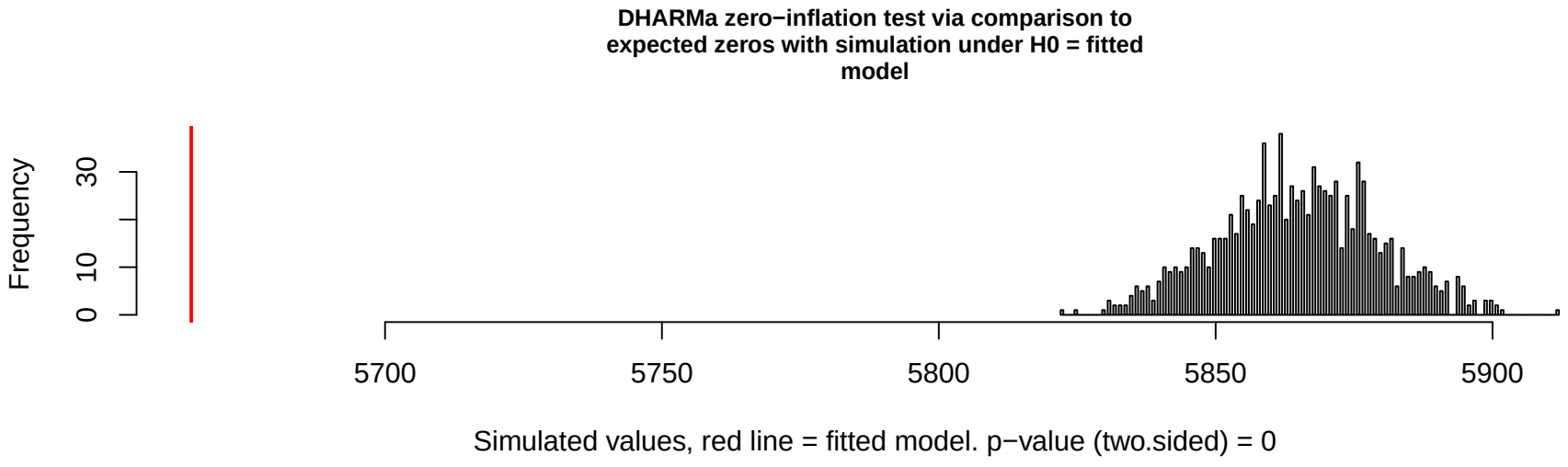


Step 4): Fit a zero-inflated model (or alternatively, NB)

```
model <- update(model, family = "ZIP")
vegan::ordiplot(model)
```



Step 5): Checking for zero-inflation again



Looks like now we are predicting too many zeros. Oh well.

Classical ordination methods

- Form a loosely connected set of methods
- That makes teaching them challenging
- Are not "state-of-the-art" anymore
- Which would be fine, if they did not have deficits
- Still a very useful set of methods (because they are so easy to use)

In contrast, GLLVMs form a state-of-the-art framework that extends GL(M)Ms

Model-based ordination

Suggested to use Generalized Linear Latent Variable Models for unconstrained ordination

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

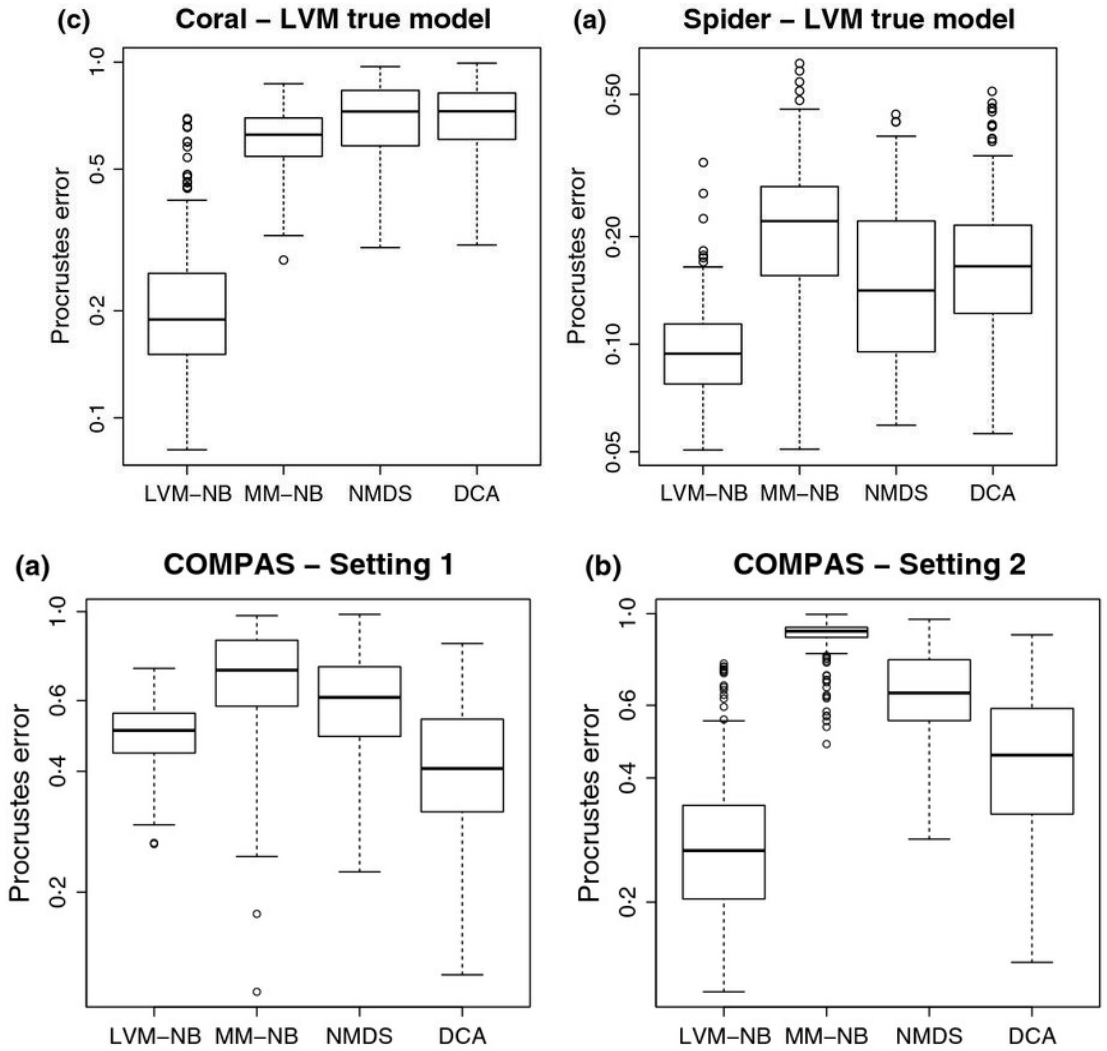
Special Feature: New Opportunities at the Interface Between Ecology and Statistics | Free Access

Model-based approaches to unconstrained ordination

Francis K.C. Hui, Sara Taskinen, Shirley Pledger, Scott D. Foster, David I. Warton

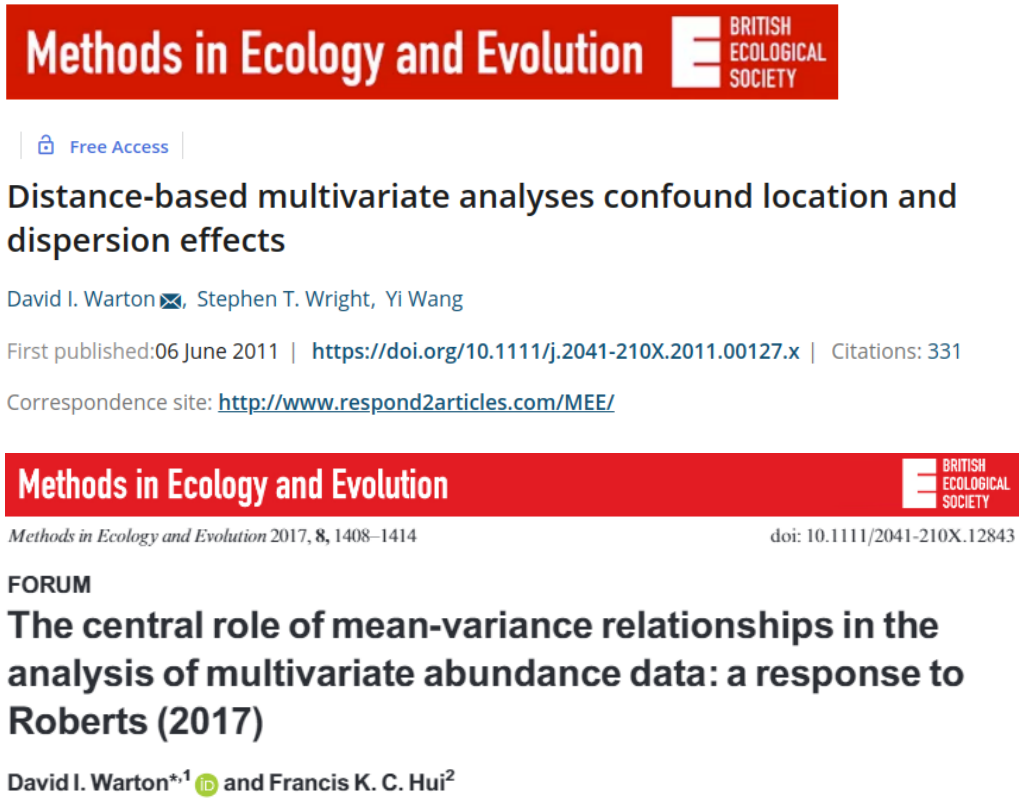
First published:23 July 2014 | https://doi.org/10.1111/2041-210X.12236 | Citations: 57

Model-based unconstrained ordination



- GLVMs perform similarly (or better) than classical methods

Distance-based ordination has issues



Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Free Access

Distance-based multivariate analyses confound location and dispersion effects

David I. Warton, Stephen T. Wright, Yi Wang

First published:06 June 2011 | https://doi.org/10.1111/j.2041-210X.2011.00127.x | Citations: 331

Correspondence site: http://www.respond2articles.com/MEE/

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, **8**, 1408–1414 doi: 10.1111/2041-210X.12843

FORUM

The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017)

David I. Warton*,[1] and Francis K. C. Hui[2]

In essence: not accounting for the properties of your data gives a bad ordination

(spoiler: yes you should)

Received: 3 October 2019 | Revised: 9 December 2019 | Accepted: 20 December 2019

DOI: 10.1002/ece3.6059

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

Should ecologists prefer model- over distance-based multivariate methods?

Jonathan F. Jupke | Ralf B. Schäfer

- Concludes that both frameworks have their place

What makes for a good ordination method?

Michael Palmer:

The "Ideal" ordination method does not exist, but if it did it would possess the following qualities.

1) It recovers gradients without distortion.
2) If clusters exist in nature, this should be obvious in the ordination.
3) It does not produce clusters which do not exist.
4) It gives the same result every time for a given data set.
5) There is a unique solution.
6) Ecological similarity is related to proximity in ordination space.
7) Scaling of axes is related to beta diversity.
8) The method is not sensitive to noise.
9) "Signal" and "Noise" are easily separated.
10) You do not need to pre-specify number of axes.
11) The solution is the same, no matter how many dimensions one chooses to look at.
12) Unless by choice, all sites/stands/quadrats are treated equally.
13) The solution does not take much computer time.
14) The method is robust: it works well for short and for long gradients, for low and high noise, for sparse and full matrices, for big and for small data sets, for species-rich and species-poor systems.
15) For the mathematician: elegant.
16) For the ecologist: available, inexpensive, and easy to understand.

What makes for a good ordination method?

Gauch (1982):

Three criteria are basic for ordination techniques.

(1) Effective (realistic in assumptions, suitably convey information)
(2) Robust (to real data)
(3) Practical (computer time)

Conclusion

- I would say that GLLVMs pass almost all of these
- They are effective, robust, and (mostly) practical (ok, it is a work in progress)
- They usually do better than classical methods
- Above all, we can see when they do not perform well