Introduction

Bert van der Veen

Department of Mathematical Sciences, NTNU

Welcome! 😄

Intro

See github for all material

Sessions from 14:00 to 20:00 (Monday to Thursday). Sessions will consist of a mix of lectures, in-class discussion, and practical exercises / case studies over Slack and Zoom.

- Monday: Recap and background
- Tuesday: gllvm R-package, classical ordination vs. model-based ordination, unimodal responses
- Wednesday: Ordination with covariates, JSDM, tools, tips and tricks
- Thursday: Beyond `gllvm`

How we will do it

Lectures of about 45 minutes
Practicals of about 60 minutes: datasets and R

- Practical "tasks" serve as guide, not as exhaustive exercise
- Short, live, demonstrations
- Then go and explore yourself (and potentially answer practicals)

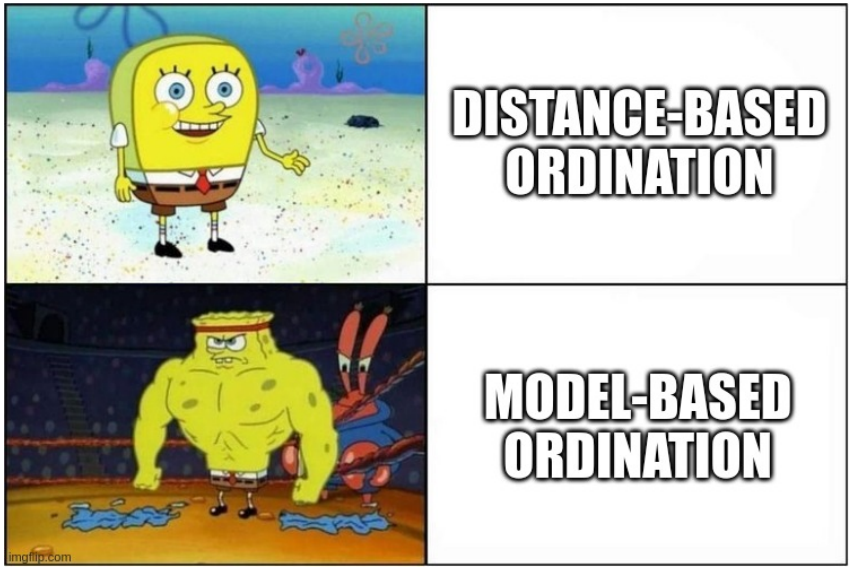
Thursday

1. Other R-packages for GLLVMs/JSDMs
2. A brief look at Hierarchical ordination
3. Own data analysis/questions/suggestions
4. Practical option: Applying a few Machine learning methods for ordination (UMAP, LLE, Laplacian eigenmaps, t-SNE, etc.)

How does that sound?

What I hope you take away

1. The `gllvm` R-package is great!
2. Performing ordination well is hard work
3. Model-based ordination methods do many things better (data properties, diagnostics)
4. One framework for all ordination tools



Detailed outline today

- Generalised linear model (GLM) recap
- Practical: `glm` and `mvabund::manyglm`
- Generalised linear mixed model (GLMM) recap
- Practical: ‘glmmTMB
- Recap classical ordination methods and concepts

15 minute break 15:45-16:00
45 minute break 17:45-18:30

Logistics

All material on github

Please make sure you've downloaded data and updated R/packages

R-packages

- gllvm
- glmmTMB
- mvabund
- HMSC
- sjSDM
- ecopCopula
- labdsv
- umap
- Rdimtools
- vegan
- boral

Some resources on classical ordination

- David Zeneley's website
- Michael Palmer's website
- Numerical ecology
- Numerical ecology with R
- Data analysis in Community and Landscape cology
- Analysis of ecological communities

Resources on model-based ordination

- Some of my other workshop repositories
- gllvm vignette website
- Oxford libraries article
- Warton 2022
- Fahrmeir and Tutz 2001
- Ovaskainen and Abrego
- Bartholomew et al. 2011
- Skrondal and Rabe-Hesketh 2004

Some recommended reading

- Halvorsen (2012)
- Wang et al. (2012)
- Warton et al. (2012)
- Clark et al. (2014)
- Warton et al. (2015)
- Warton et al. (2015)
- Hui et al. (2015)
- Pollock et al. (2015)
- ter Braak and Smilauer (2015)
- Hui et al. (2017)
- Niku et al. (2017)
- Ovaskainen et al. (2017)
- Roberts (2017)
- Warton et al. (2017)
- Niku et al. (2019)
- Niku et al. (2019)
- Roberts (2019)
- Paul (2020)
- Zurell et al. (2020)
- van der Veen et al. (2021)
- Blanchet et al. (2022)
- van der Veen (2022)
- van der Veen et al. (2023)

Resources that cover all kinds of ordination methods

(none)

Motivation

Plant Ecol (2015) 216:669–682
DOI 10.1007/s11258-014-0366-3

Model-based thinking for community ecology

David I. Warton · Scott D. Foster ·
Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

Figure 1: Warton et al. 2015

Motivation

- We need formal, probabilistic, models for community ecology
- That incorporate data properties, rather than transforming our way out of them
- It makes for better/clearer inference
- For nicer for teaching
- Access to tools for testing assumptions
- Overall more flexibility

Why GLLVMs?

GLLVMs are a formal statistical, fully probabilistic, tool for ordination.

1) To step up your multivariate analysis
2) Maybe you want to incorporate random effects
3) To negate the need of distances

Disclaimer

- There will be some equations
- `gllvm` is in active development (some bugs expected, feature requests are very welcome)

What is this presentation about?

- Part I: a brief history of multivariate analysis (ordination)
- Part II: The JSDM
- Part III: Model-based ordination

Multivariate analysis

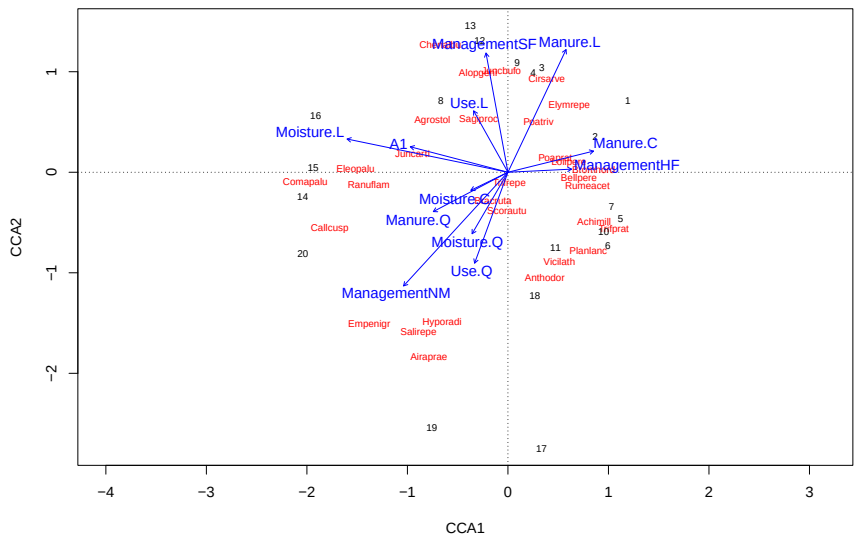


Figure 2: CCA of dune data

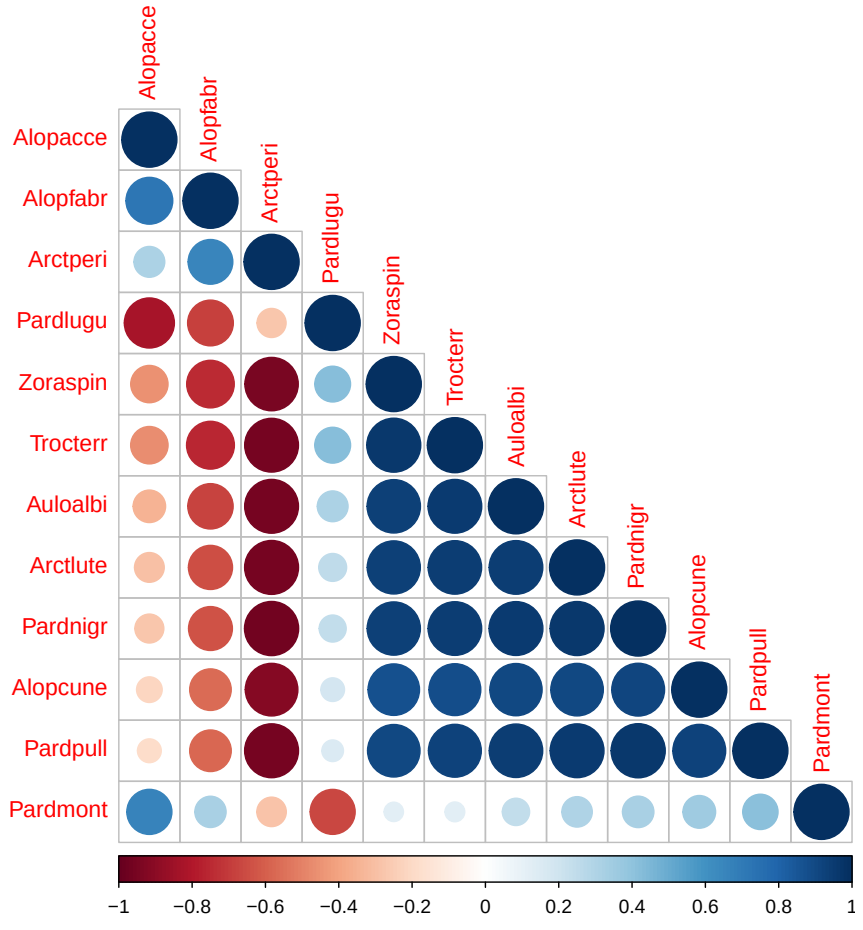


Figure 3: JSDM of spider data

Use of ordination

- One of the most used methods of multivariate analysis
  - but definitely not the only one
- Arranges species and samples in a low-dimensional space
  - or any column and row quantity really
- Summarizes data
- Find underlying structures (gradients/LVs)

A brief history of ordination

- Thoughts from community ecology perspective
- A long history

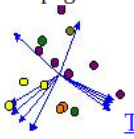
And me:

A brief history of ordination

Milestones in Ordination: a timeline

- 1901 - Pearson develops PCA as a regression technique
- 1927 - Spearman applies factor analysis to psychology
- 1930 - Ramensky uses an informal ordination technique and the term "Ordnung" in ecology
- 1933 - Hotelling develops PCA for understanding the correlation matrix
- 1950 - Curtis and McIntosh employ the "continuum index" approach
- 1952 - Williams uses Correspondence Analysis
- 1954 - Goodall uses the term "ordination" for PCA
- 1957 - Bray-Curtis (Polar) ordination
- 1964 - Kruskal develops NMDS
- 1970's - Whittaker develops theoretical foundations of gradient analysis
- 1973 - Hill revives Correspondence Analysis
- 1976 - Canonical Correlation introduced to ecology
- 1977 - Fasham, Prentice use NMDS
- 1979 - DCA introduced by Hill and Gauch
- 1982 - Gauch's "Multivariate Analysis in Community Ecology"
- 1986 - CCA introduced by ter Braak
- 1986 - Fuzzy set ordination introduced by Roberts
- 1988 - ter Braak and Prentice's "Theory of Gradient Analysis"

This page was created and is maintained by Michael Palmer.



To the ordination web page

Historical Perspective

1901	-	Pearson develops PCA as a regression technique.
1927	-	Spearman applies factor analysis to psychology.
1930	-	Ramensky uses an informal ordination technique and introduces the term 'ordnung' into ecology.
1954	-	D.W. Goodall introduces PCA into ecology and proposes the term 'ordination'.
1970	-	R.H. Whittaker develops theoretical foundations of gradient analysis, especially unimodal species responses and turnover along environmental gradients.
1971	-	K.R. Gabriel develops biplot graphical display.
1973	-	M.O. Hill re-invents correspondence analysis and introduces CA (as 'reciprocal averaging') into ecology.
1986	-	Cajo ter Braak invents canonical correspondence analysis (CCA) and released CANOCO software.
1988	-	Cajo ter Braak and Colin Prentice's "A theory of gradient analysis" (Advances in Ecological Research 18; 271-317) that unifies indirect and direct gradient analysis and highlights the importance of underlying species response models.
1998, 2002	-	Cajo ter Braak and Petr Šmilauer CANOCO 4 & 4.5 software and manual.

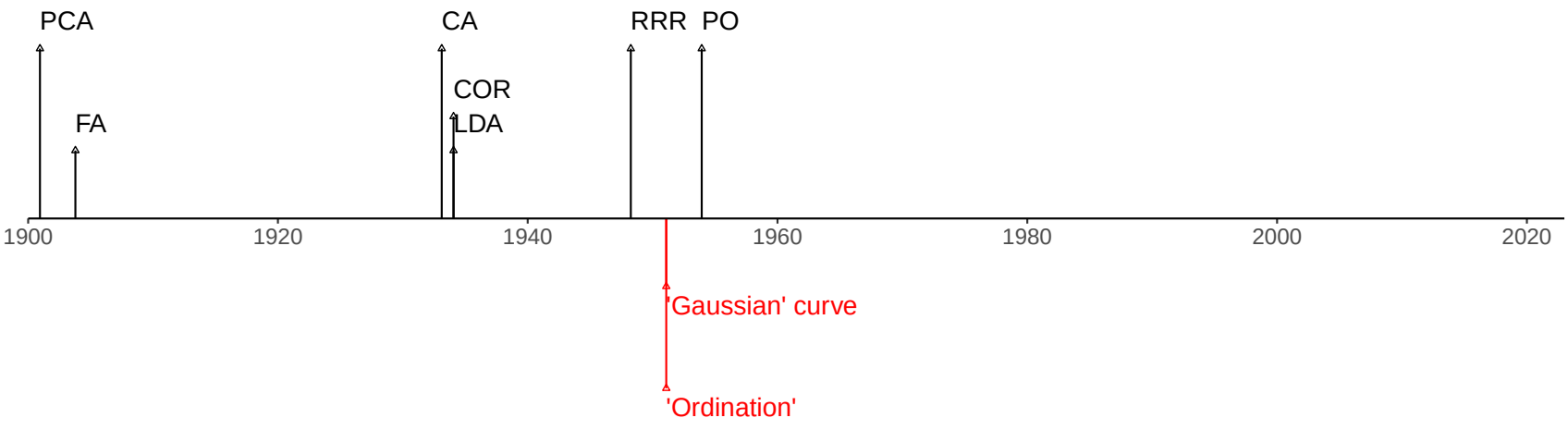
- Inspired by Michael Palmer's and John Birk's
- But in need of a little update

In the beginning, there were PCA and FA

Then, CA was developed

Next, reduced rank regression

Then, there was “Ordination”



Whittaker 1956: Unimodal responses

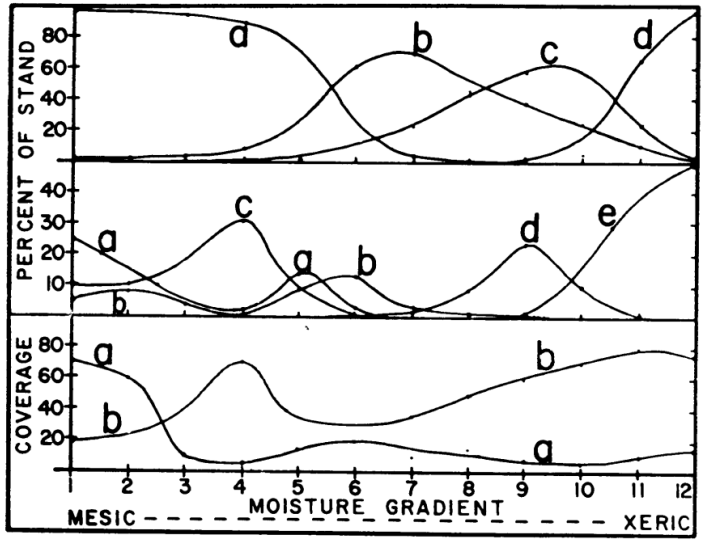
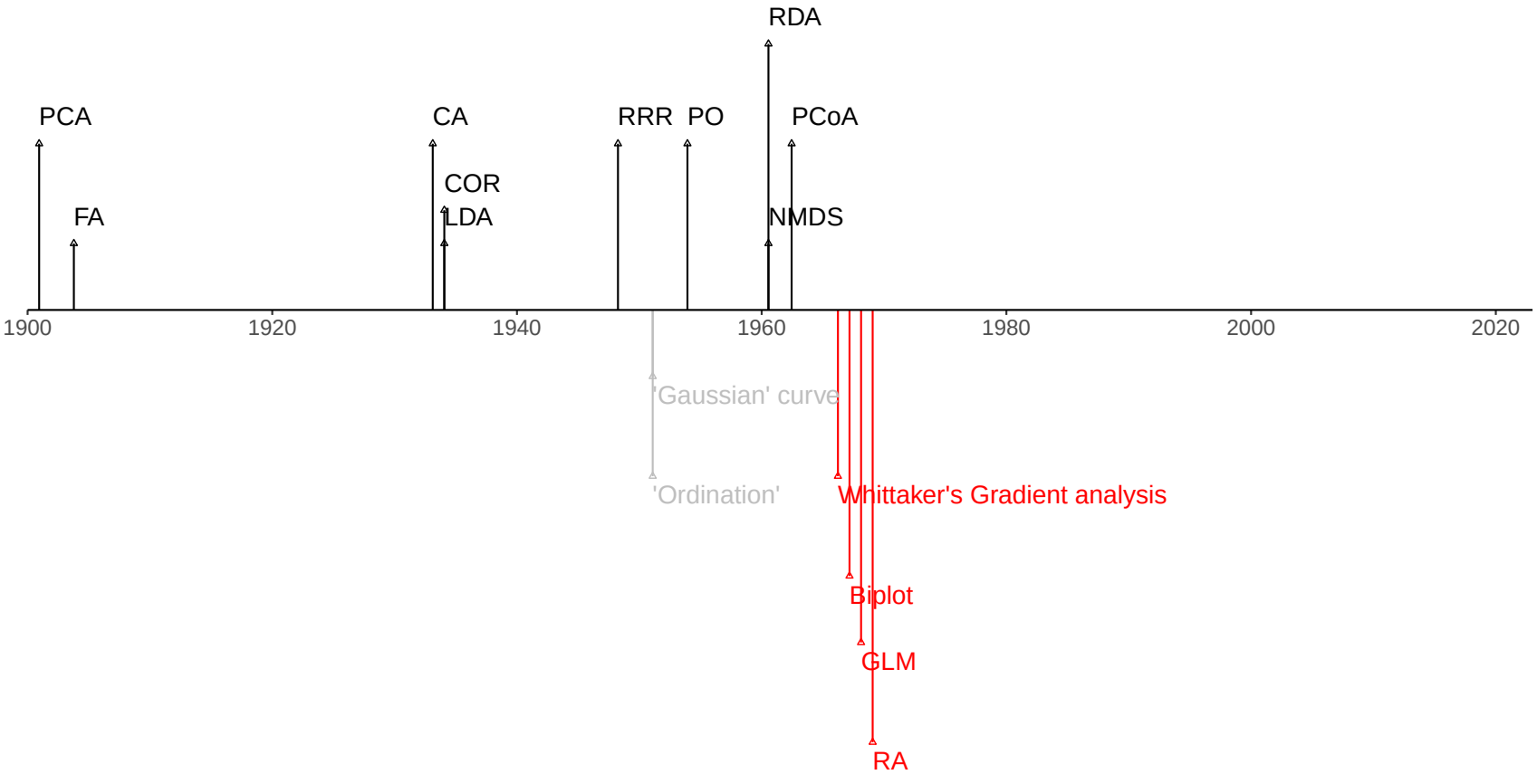
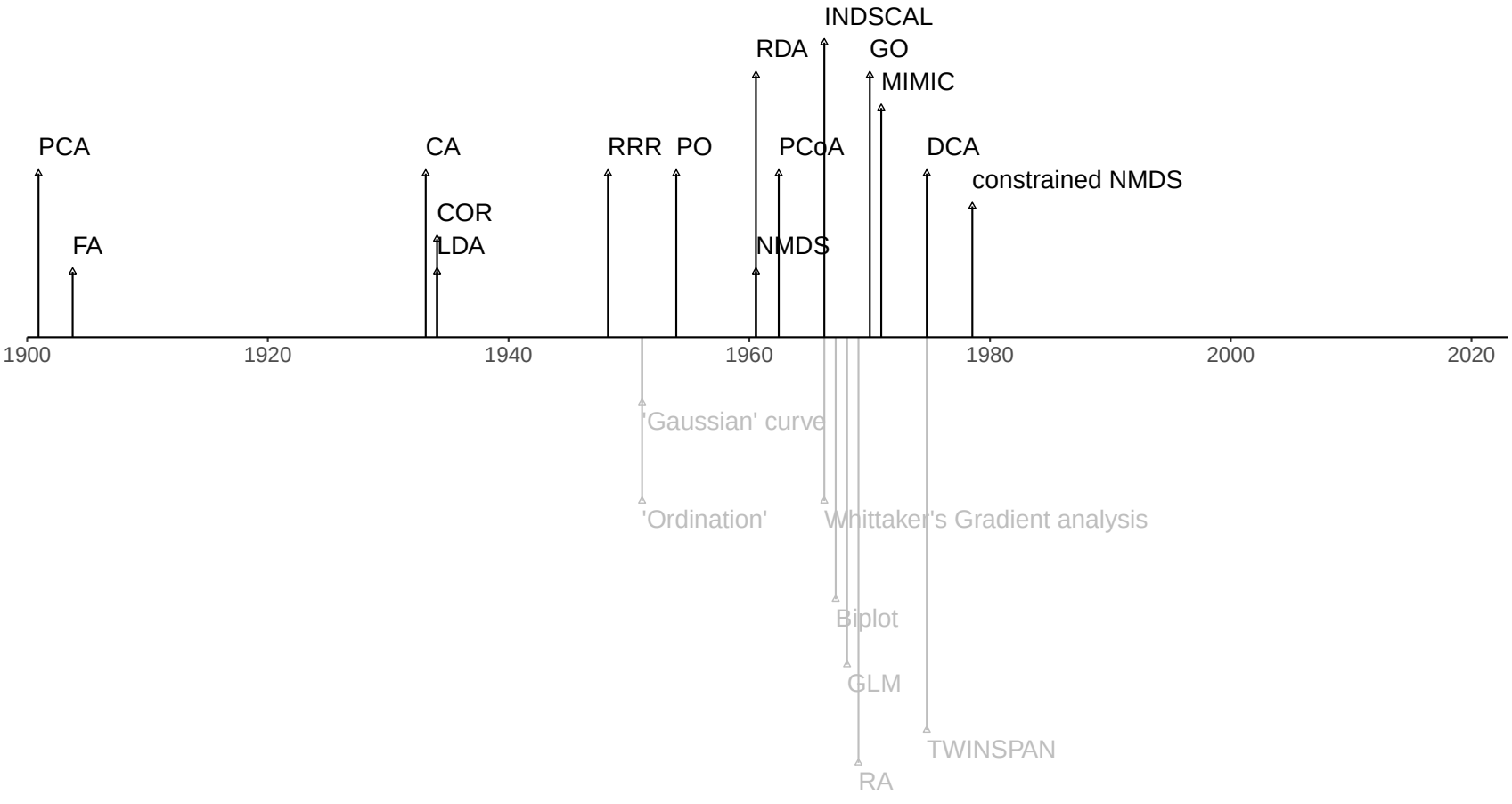


FIG. 4. Transect of the moisture gradient, 3500-4500 ft. Top—curves for tree classes; a, mesic; b, submesic; c, subxeric; d, xeric. Note expansion of mesic stands, compared with Figs. 2 and 3. Middle—curves for tree species: a, *Tilia heterophylla*; b, *Halesia monticola* (both the preceding are bimodal, with populations on each side of the mode of Tsuga); c, *Tsuga canadensis;* d, *Quercus alba*; e, *Pinus pungens*. Bottom—curves for undergrowth coverages: a, herbs; b, shrubs.

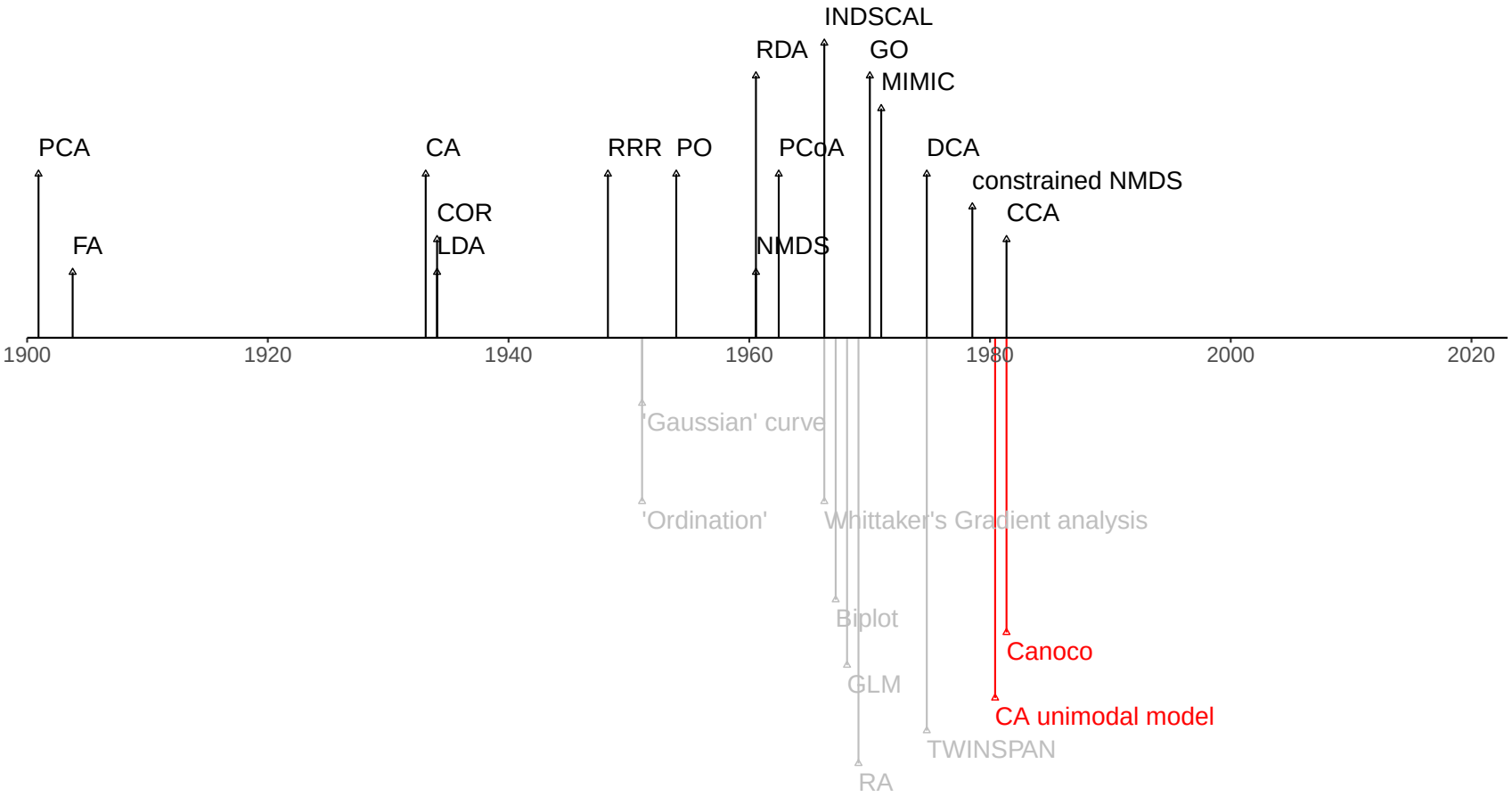
Ordination became popular



Mark Hill's contributions & MIMIC



CTB did his PhD



- The unimodal model continued

Unimodal model (2)

Ecology (1974) **55**: pp. 1382–1390

ORDINATION OF VEGETATION SAMPLES BY GAUSSIAN SPECIES DISTRIBUTIONS[1]

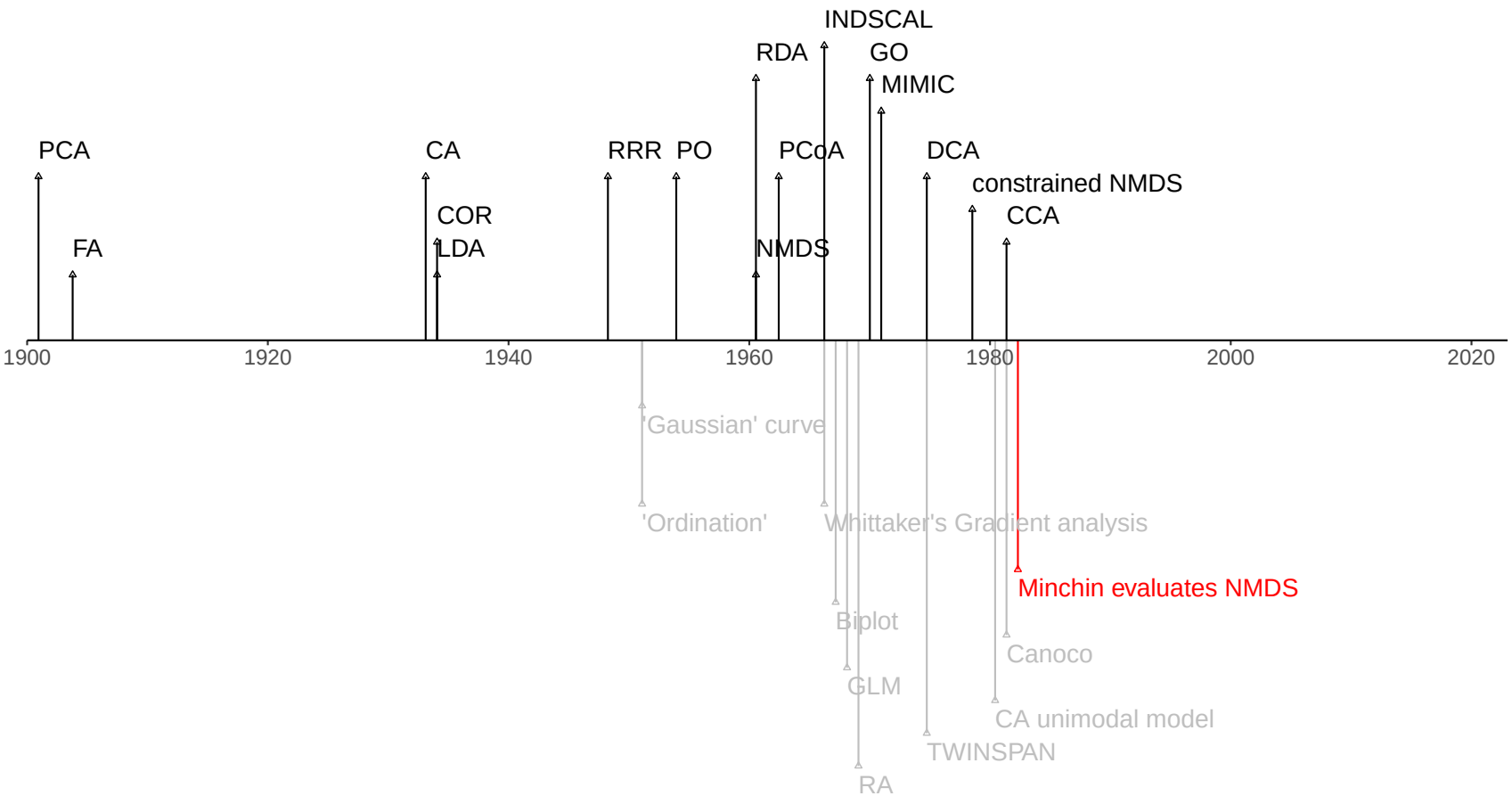
HUGH G. GAUCH, JR.
Ecology and Systematics, Cornell University, Ithaca, New York 14850
GENE B. CHASE[2]
Education, Cornell University, Ithaca, New York 14850
AND
ROBERT H. WHITTAKER
Ecology and Systematics, Cornell University, Ithaca, New York 14850

BIOMETRICS 41, 859–873
December 1985

Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model

Cajo J. F. ter Braak
TNO Institute of Mathematics, Information Processing and Statistics,
P. O. Box 100, 6700 AC Wageningen, The Netherlands

NMDS was shown to be “better”



- The unimodal model
- Development of CCA

Are species response curves symmetric?

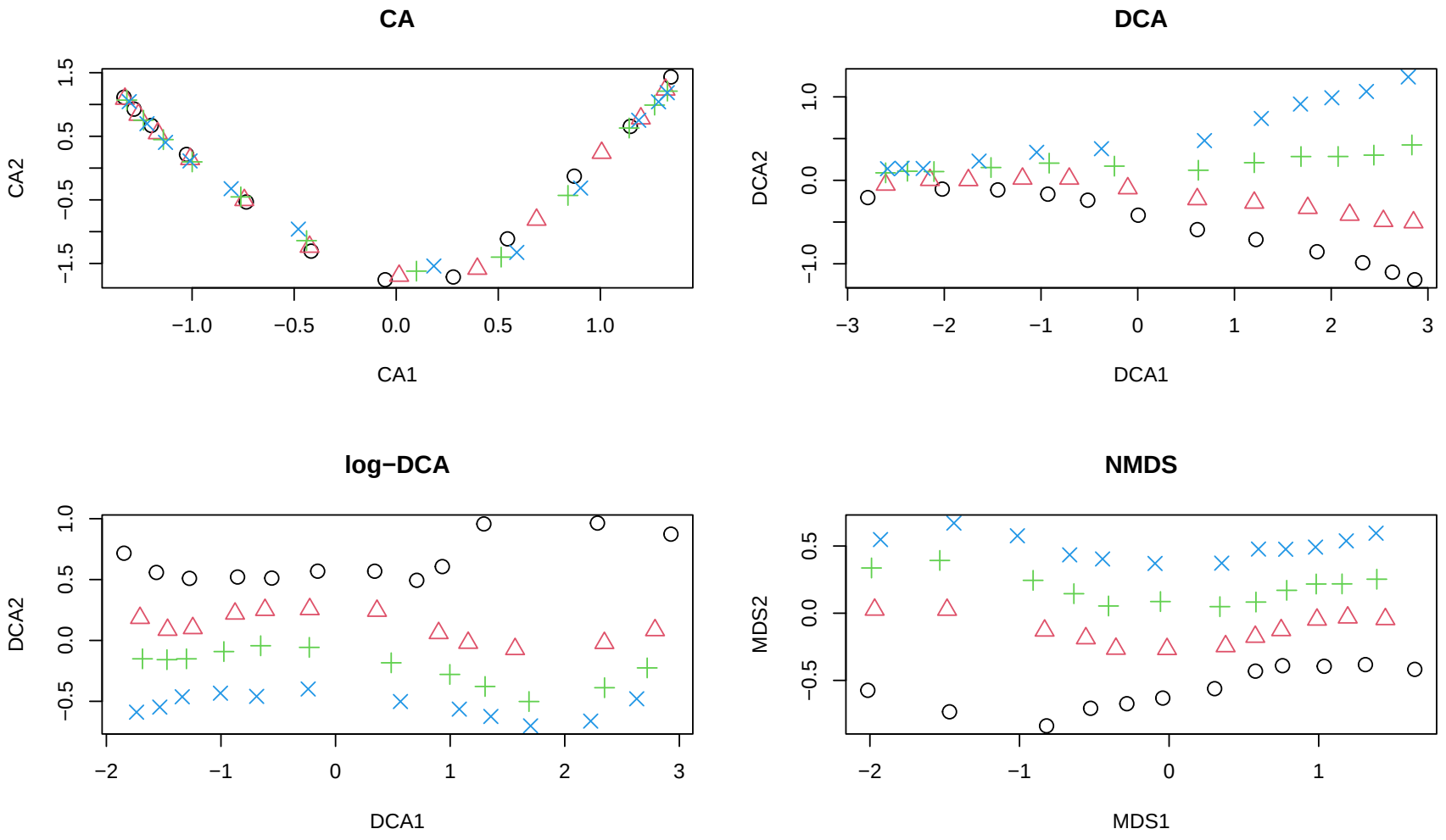
Vegetatio 69: 89–107, 1987

An evaluation of the relative robustness of techniques for ecological ordination

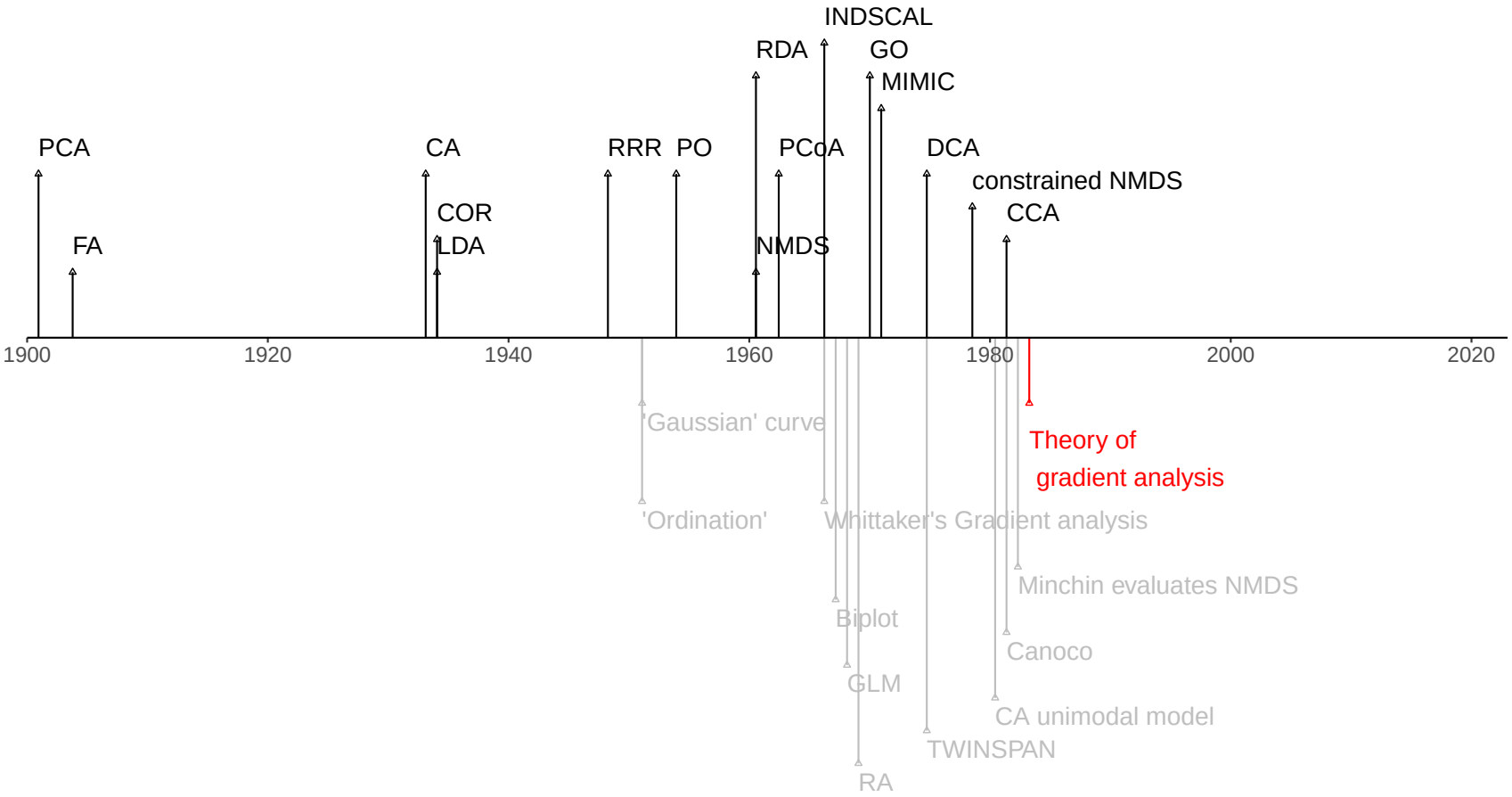
Peter R. Minchin*
CSIRO Division of Water and Land Resources, G.P.O. Box 1666, Canberra, 2601, Australia

- Skewed might be more realistic
- NMDS is “robust”
- Moving away from process-based

Results Minchin dataset

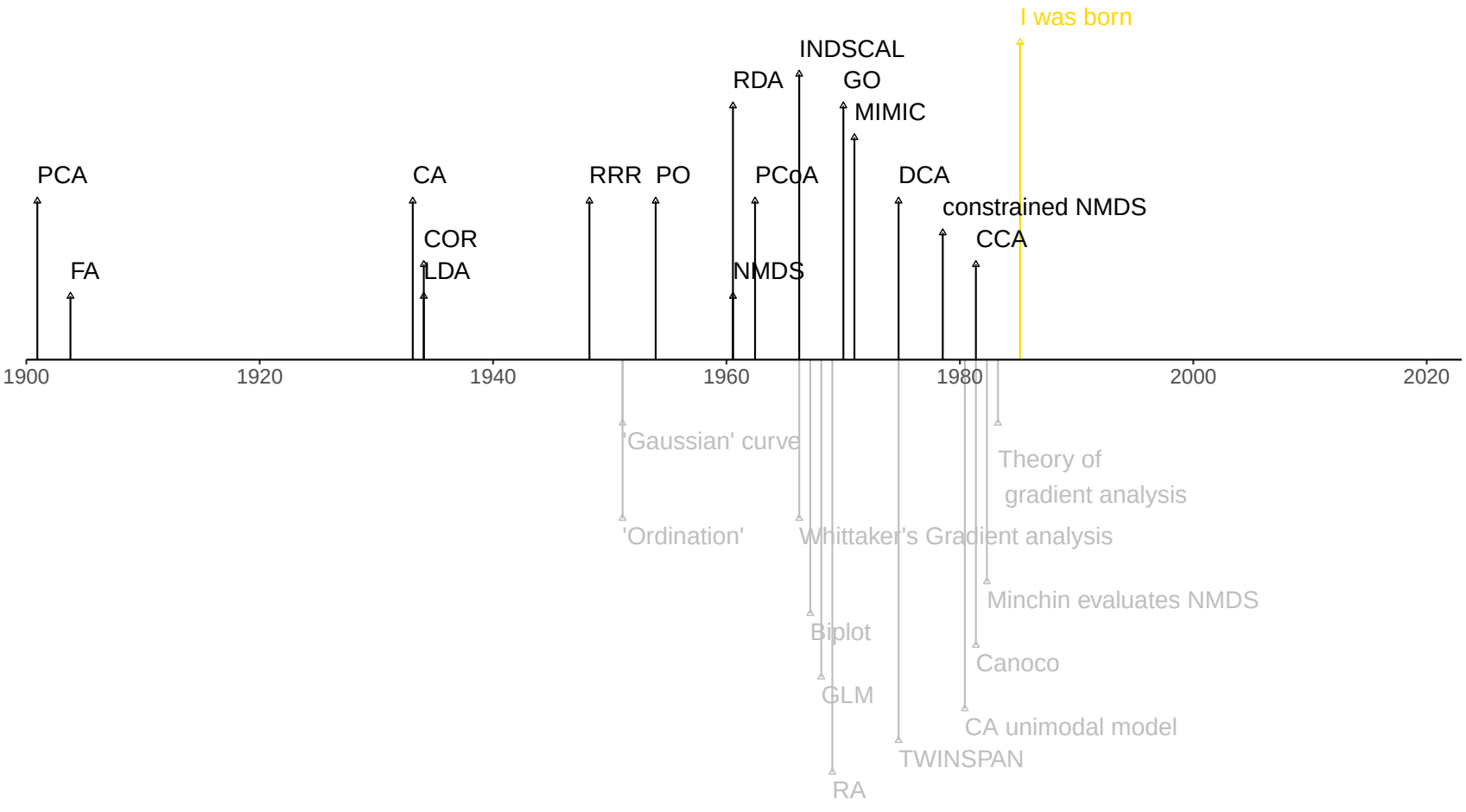


A theory of gradient analysis

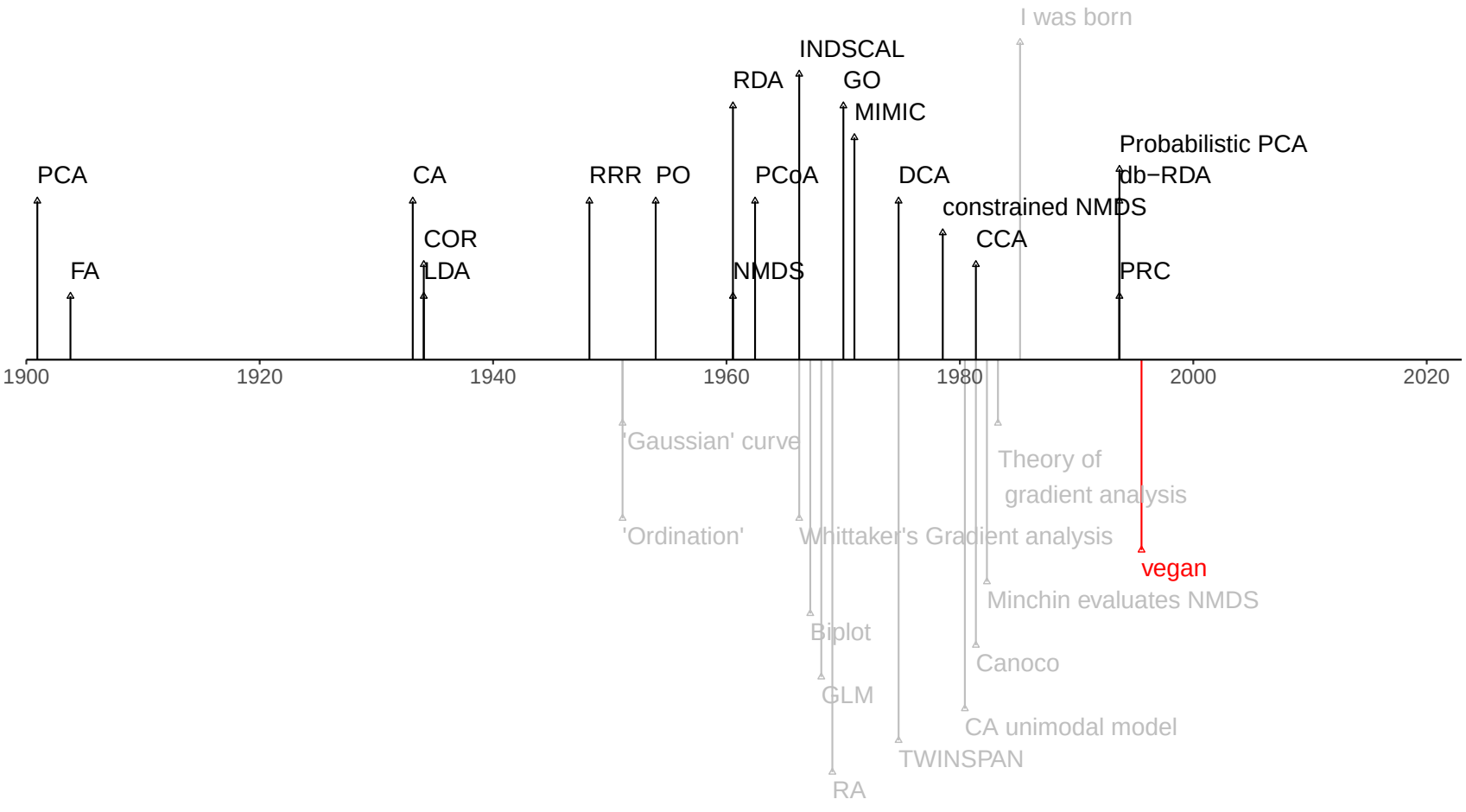


{ "Ter Braak and Prentice (1988) developed a theoretical unification of ordination techniques, hence placing gradient analysis on a firm theoretical foundation." (Michael Palmer) }

I was born

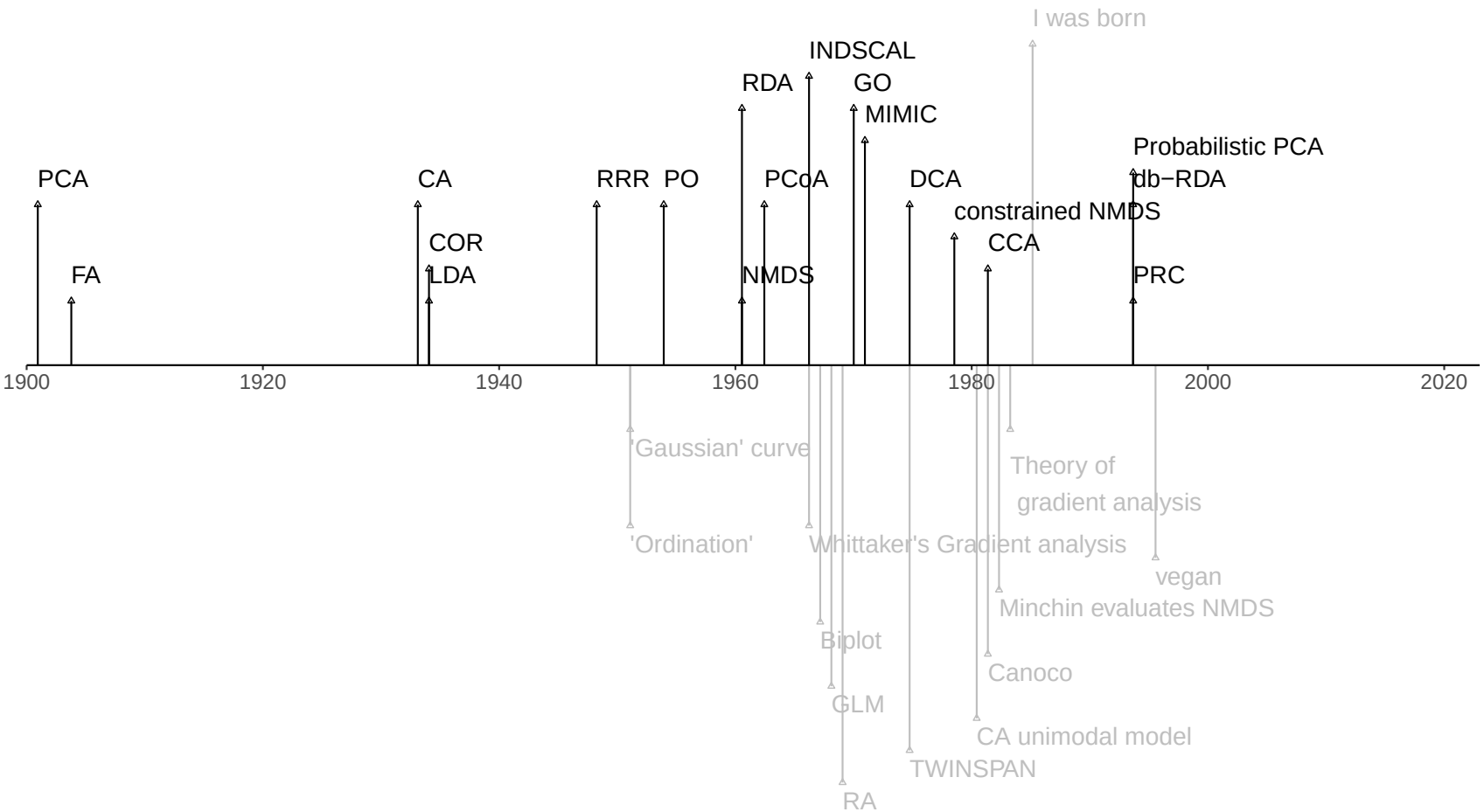

I was born
INDSCAL
RDA
GO
MIMIC
PCA
CA
RRR
PO
PCoA
DCA
constrained NMDS
COR
CCA
FA
LDA
NMDS
1900
1920
1940
1960
1980
2000
2020
'Gaussian' curve
Theory of
gradient analysis
'Ordination'
Whittaker's Gradient analysis
Minchin evaluates NMDS
Biplot
Canoco
GLM
CA unimodal model
TWINSPAN
RA

vegan


I was born
INDSCAL
RDA
GO
MIMIC
Probabilistic PCA
PCA
CA
RRR
PO
PCoA
DCA
db−RDA
constrained NMDS
COR
CCA
FA
LDA
NMDS
PRC
1900
1920
1940
1960
1980
2000
2020
'Gaussian' curve
Theory of
gradient analysis
'Ordination'
Whittaker's Gradient analysis
vegan
Minchin evaluates NMDS
Biplot
Canoco
GLM
CA unimodal model
TWINSPAN
RA

Things slowed down

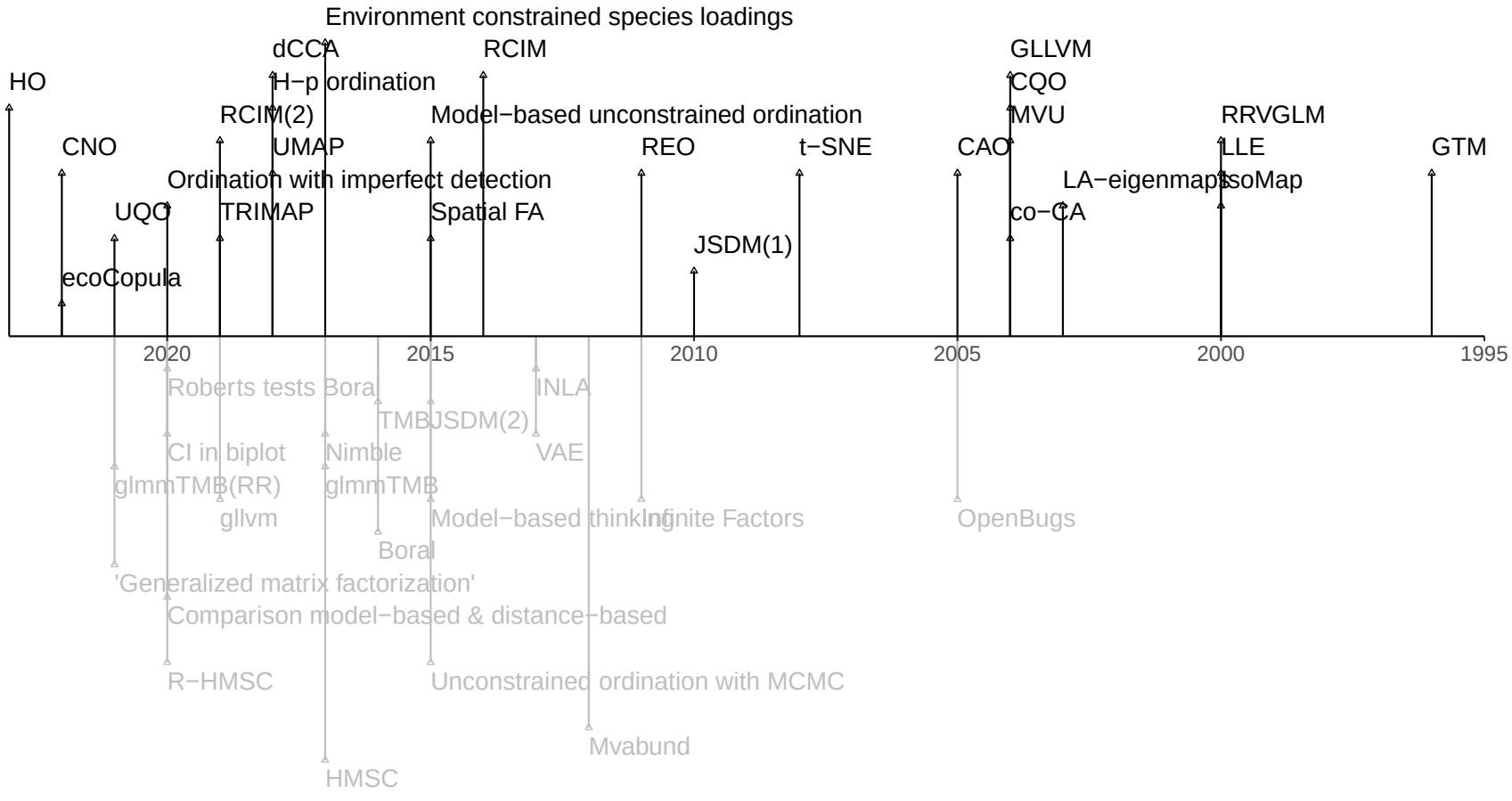
Ordination became “descriptive” and uncool



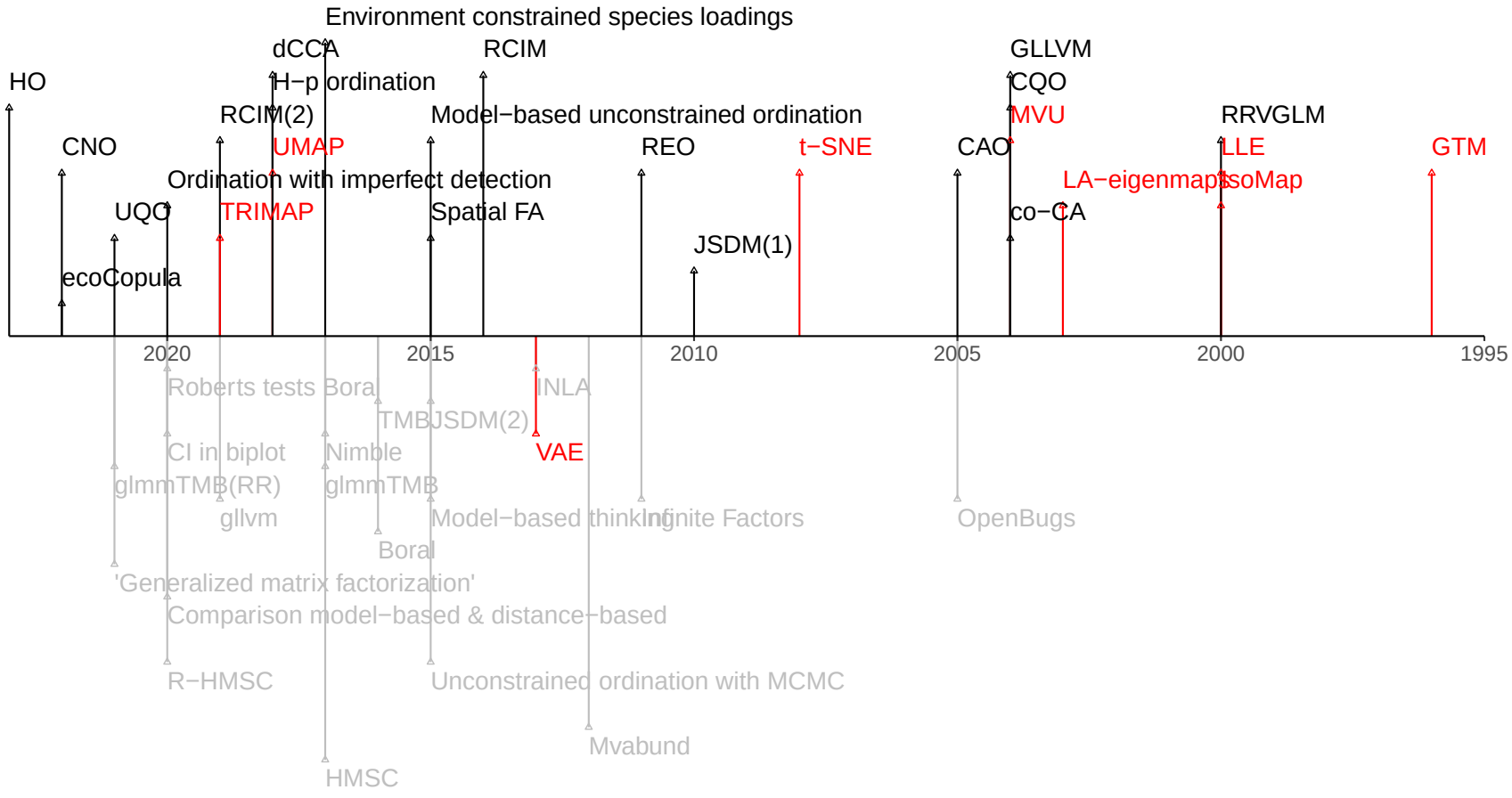
Ordination is well established

- one of the successes currently: so many resources!
  - vegan, anadat
  - many books, websites (Michael Palmer, David Zeleney)
  - I.e., run a search and you will find the answer
- Contemporary ordination methods, not so much (yet)

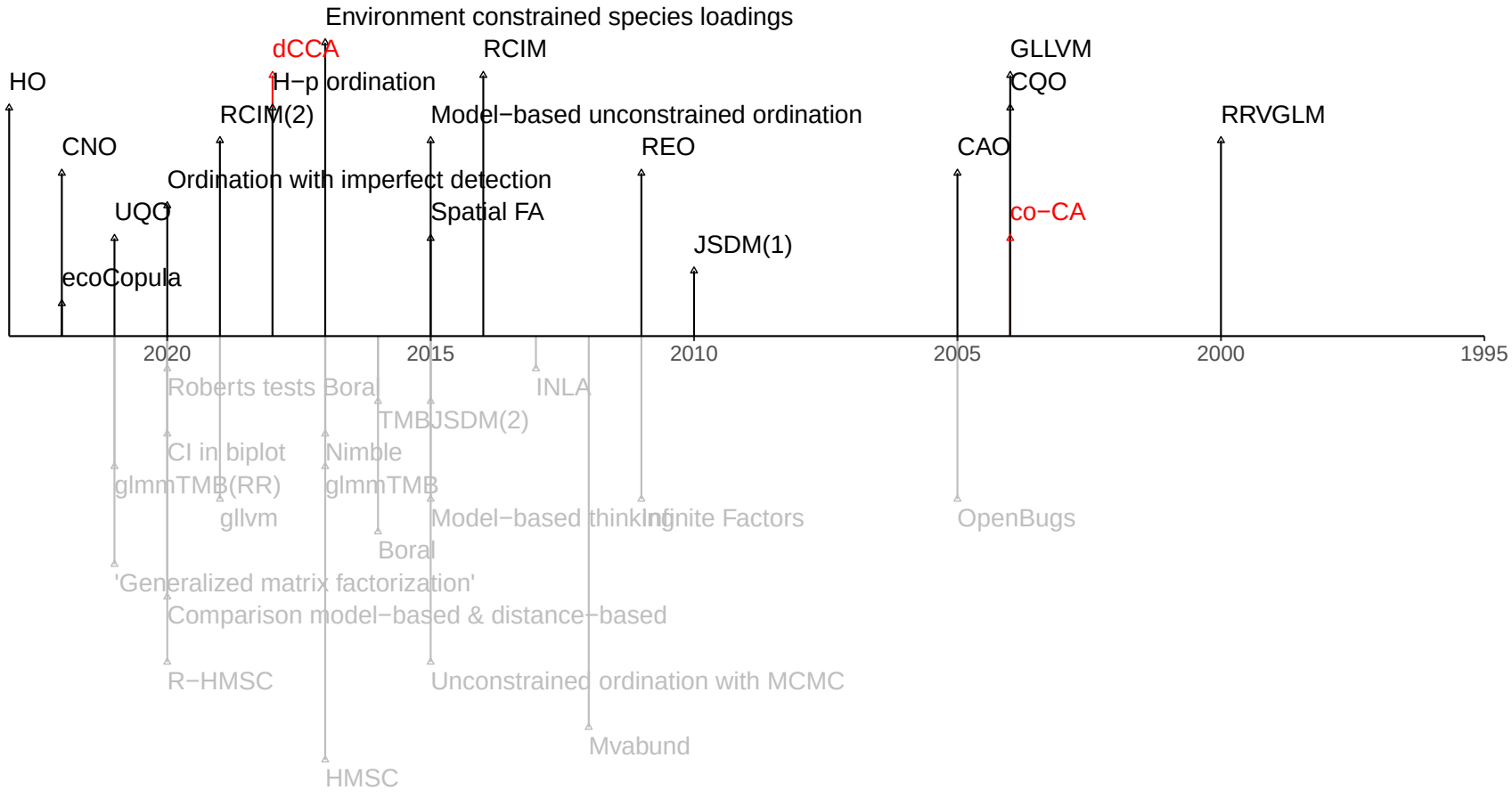
Updating the ordination timeline



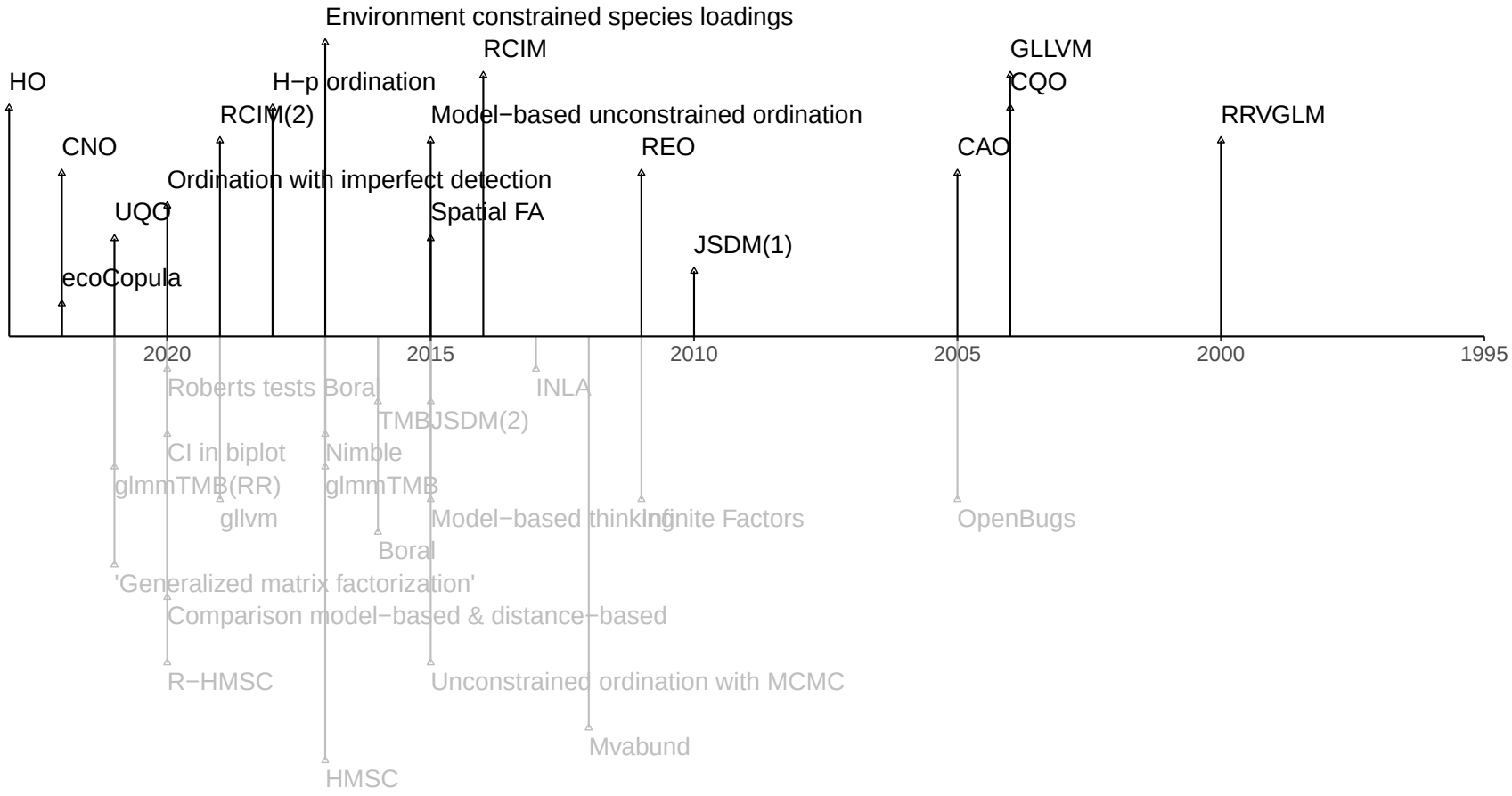
Ditching the machine learning developments



Eigenanalysis-based methods



Timeline model-based ordination



The 3 schools of ordination

1) "Simple-method"
2) "Algorithm-based" (also called "distance-based")
3) "Model-based"

The 3 schools of ordination

1) "Simple-method"
2) "Algorithm-based"
3) "Model-based"

Principal Component Analysis
(Detrended) Correspondence Analysis
Non-metric multidimensional scaling
UMAP
t-SNE
Row-column interaction models
Graphical models (copula)
Generalized Linear Latent Variable Models

Algorithm-based methods deviated from process-based thinking
Model-based methods *reinforce* process-based thinking

Algorithm-based methods deviated from process-based thinking
Model-based methods *reinforce* process-based thinking

Model-based ordination tries to capture an ecological process.
Algorithm-based ordination makes pretty pictures.

Joint Species Distribution Modeling

- First suggested by Ovaskainen et al. 2010
- Named by Pollock et al. 2014

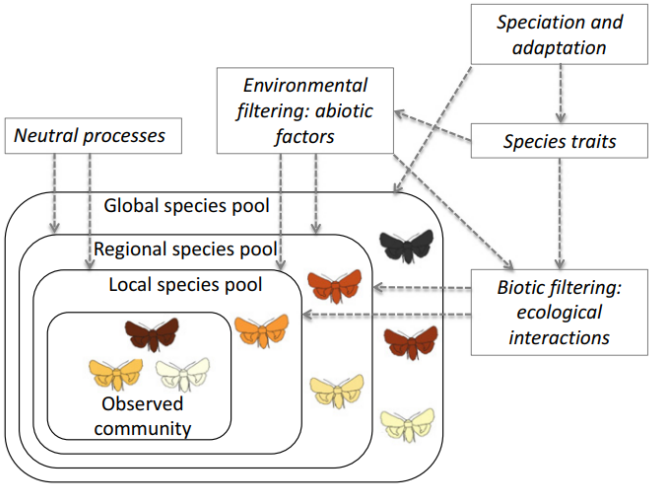


Figure 1 A conceptual diagram of the assembly processes influencing ecological communities at different spatiotemporal scales. The composition and dynamics of local, regional and global communities are influenced by the combined effects of environmental filters, biotic interactions and neutral processes. The responses of the species to these factors depend on their traits, which are ultimately shaped by evolutionary history and therefore constrained by phylogenetic relationships.

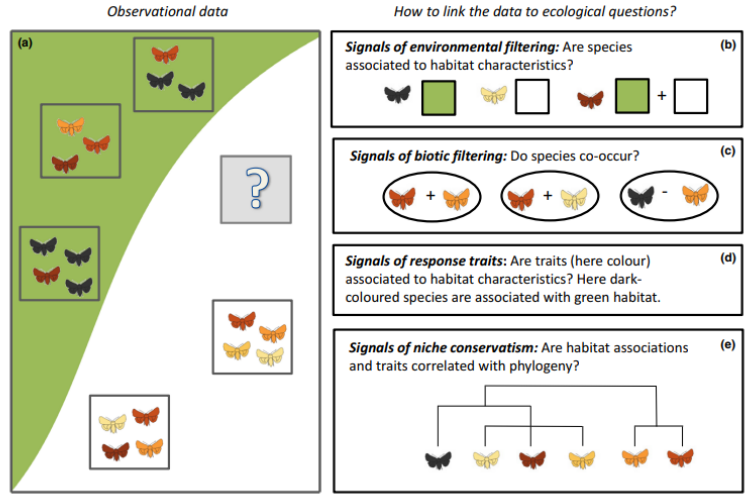


Figure 2 A conceptual illustration of some key questions in community ecology. The green and white colours represent differences in the environmental

Figure 4: Figures from Ovaskainen et al. 2017

Joint Species Distribution Modeling (2)

- A niche process generates data
- Distributions are driven by 1) environment and 2) interactions
- Statistically, interaction = correlation of species
- So we model that:

$$g\{\mathbb{E}(y_{ij}|\boldsymbol{\epsilon}_i)\} = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \qquad (1)$$

Can be fit using standard mixed-effects modeling software:

```
glmer(abundance~species+x:species+(0+species|sites))
```

Problem: Number of parameters grows quadratically

"Fun" ecological inference

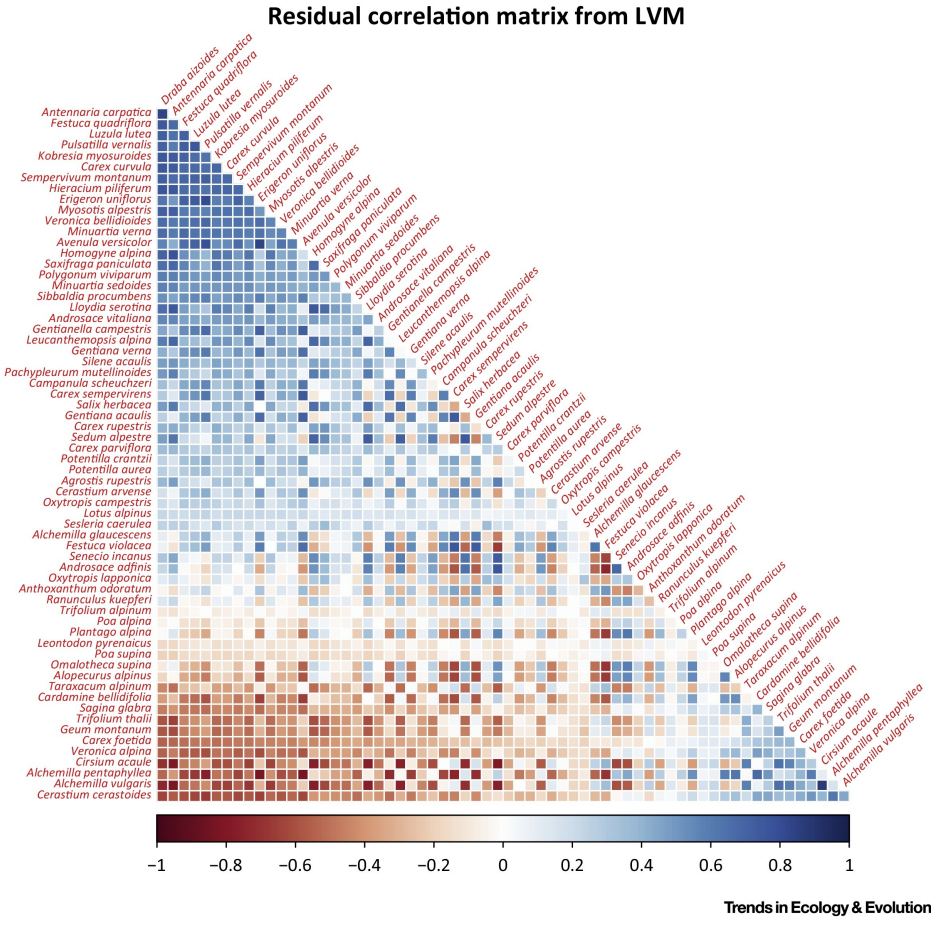


Figure 5: Warton et al. 2015 Alpine plants

Enter: Generalized Linear Latent Variable Models

- GLLVM for short
- Adds "factor analytic" structure to Σ
- $\epsilon_{ij} = \mathbf{u}_i^\top \boldsymbol{\theta}_j$
  - i.e. $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Theta}\boldsymbol{\Theta}^\top)$
- Faster and fewer parameters:
  - Number of parameter doesn't grow so fast

$$\boldsymbol{\Sigma} = \begin{bmatrix} \theta_{11} & 0 & 0 \\ \theta_{12} & \theta_{22} & 0 \\ \vdots & \ddots & \vdots \\ \theta_{1j} & \cdots & \theta_{dj} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1j} \\ 0 & \theta_{22} & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{dj} \end{bmatrix} \tag{2}$$

Huh, this is ordination

- More dimensions, better estimation of $\boldsymbol{\Sigma}$
- Each dimension is a "latent variable" so ordination!

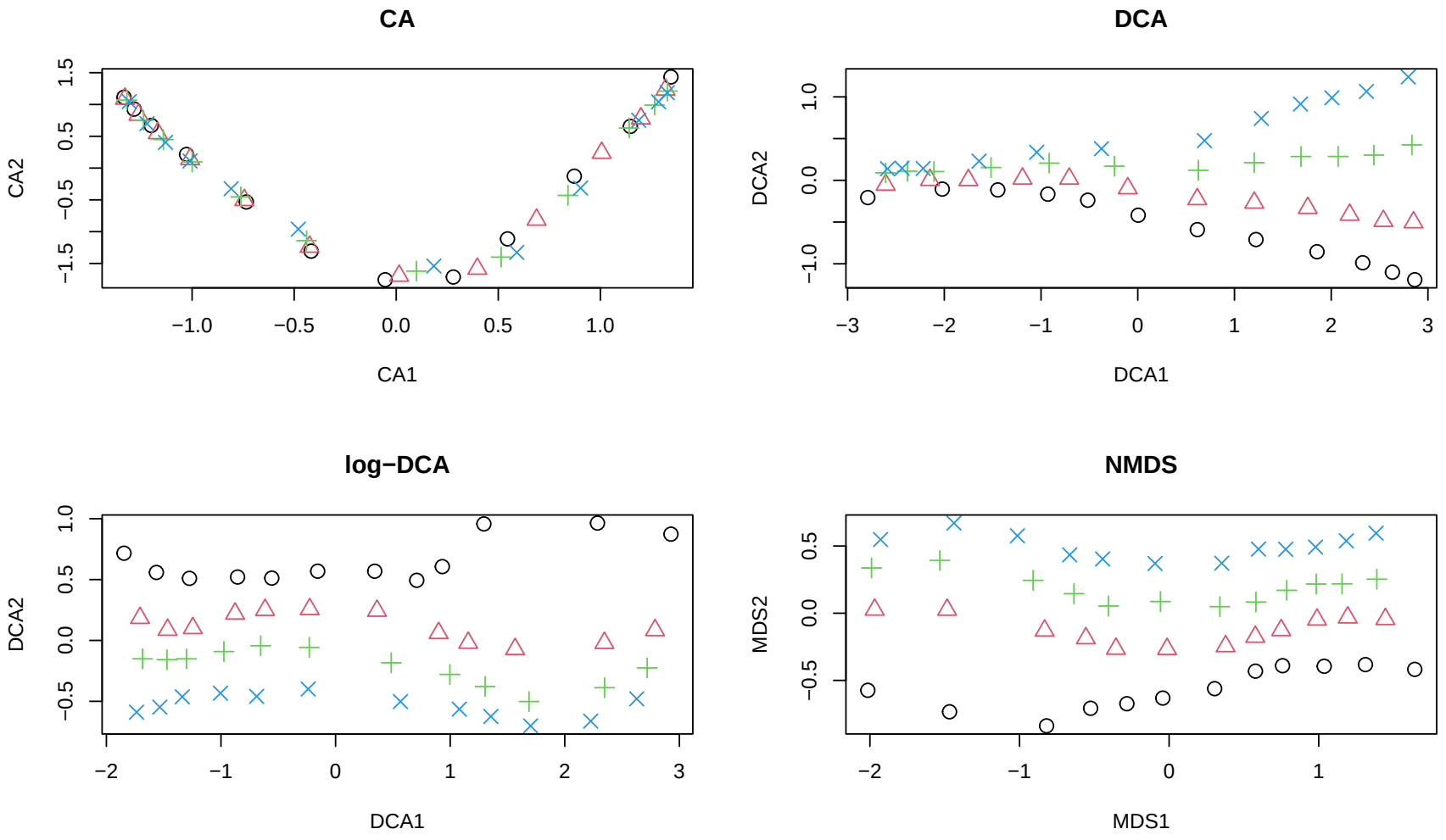
Ecological Monographs, 81(4), 2011, pp. 635–663

Random-effects ordination: describing and predicting multivariate correlations and co-occurrences

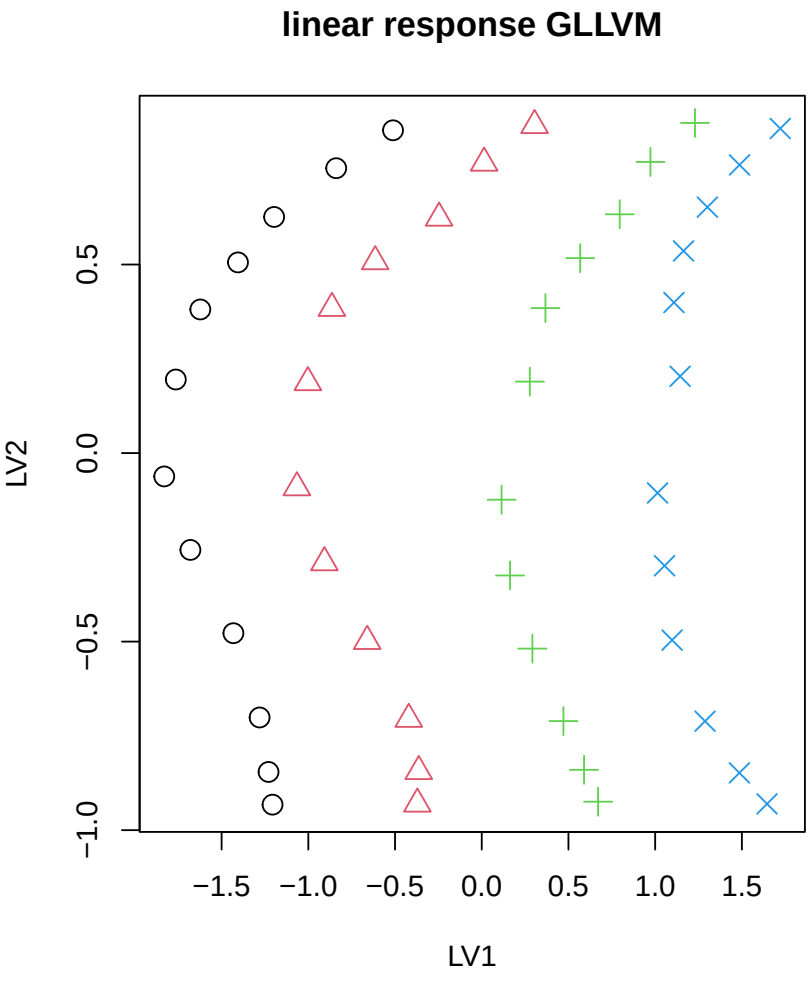
Steven C. Walker[1] and Donald A. Jackson

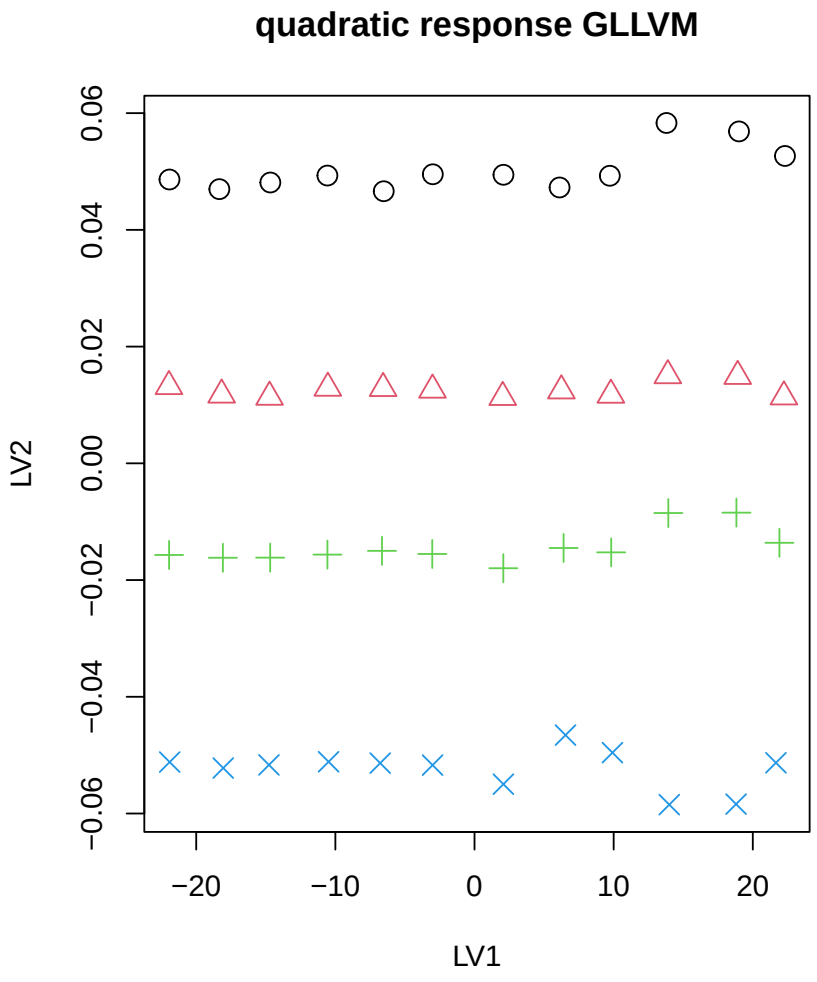
Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario M5S 3G5 Canada

Recall: Results Minchin dataset



Results Minchin dataset: gllvm





Why model-based ordination

1) Model-based methods outperform Algorithm-based methods

2) Model-based methods are conceptually superior

3) Model-based methods are more flexible

goal: use one analysis that can answer all questions in a community ecological study

Model-based ordination

- More flexible & statistically appropriate
- Now straightforward available
- Very much in active development
- Has a steep learning curve & computational difficulties

Allows us to do all kinds of fun things

- Residual diagnostics
- Species-specific effecs from constrained ordination
- Random-effects
- Flexible models
- Etc.

Classical ordination ≈ model-based

- PCA ≈ Gaussian linear GLLVM
- CA ≈ Poisson/Bernoulli/Gamma linear GLLVM
- CCA ≈ constrained GLLVM

But model-based ordination is much more flexible than that

Challenges

- Complex model
  - products of random-effects
  - identifiability
- No analytical solution (need approximate methods)
- Computationally intensive (need fast fitting)
- Non-convex objective function (need robust optimization)
- Usability needs a lot of thought

Towards the end

- I think model-based ordination is the way
- Developments are very active and we need researchers to push the frontier
  - ask new questions
  - show that existing methods are insufficient
  - apply new advances
  - update the course material
- Computation is again an active topic.
- Learning curve can be/is steep
  - ordination should now be taught differently. The model-based angle

Ordination is **so much more** than "a descriptive method"
It is a framework for complex (reduced rank) multivariate statistical modeling


IF YOU HAVE
A QUESTION... ASK
YOU WON'T BE
MAULED, PROMISE
imgflip.com