Other R packages for dimension reduction

Bert van der Veen

Department of Mathematical Sciences, NTNU

Questions so far?

Different packages

I will briefly go through the different model-based ordination packages
I will contrast each package to `gllvm`
Each of these packages really warrants its own presentation

Examples with Bird counts (from CANOCO 5)

```
Y <- read.csv("../data/birdY.csv", header = TRUE, skip = 1, row.names = 1)
Y[is.na(Y)] <- 0;
Y <- Y[,order(colSums(ifelse(Y==0,0,1)),decreasing=TRUE)] #reorder by freq
Y <- as.matrix(Y)
row.names(Y)<-1:nrow(Y) # for VGAM
X <- read.csv("../data/birdX.csv", header = TRUE, skip = 1, row.names = 1)
X[,c(1:3,5:9)] <- scale(X[,c(1:3,5:9)])
X[,-c(1:3,5:9)] <- data.frame(lapply(X[,-c(1:3,5:9)], as.factor))
```

Bayesian Ordination and regression AnaLysis

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 744–750 doi: 10.1111/2041-210X.12514

APPLICATION

BORAL – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R

Francis K.C. Hui*

Mathematical Sciences Institute, The Australian National University, Canberra, ACT 0200, Australia

boral

- The first model-based ordination package for community ecology
- For unconstrained (or residual) ordination (and JSDM)
- Based on JAGS (Plummer, 2012)
- Writes the model to a file, loads it into JAGS, returns results
- Runs on a single MCMC chain

Features

- Covariates
- 4th corner model
- Row intercepts
- Structured LVs
- Natively includes gold-standard residuals

boral

```
This is boral version 2.0.2.
Please note that as of version 2.0, boral will no longer be regularly maintained and updated. However,
if you spot any bugs/typos or have a specific feature requests, please contact the maintainer.
```

boral: code

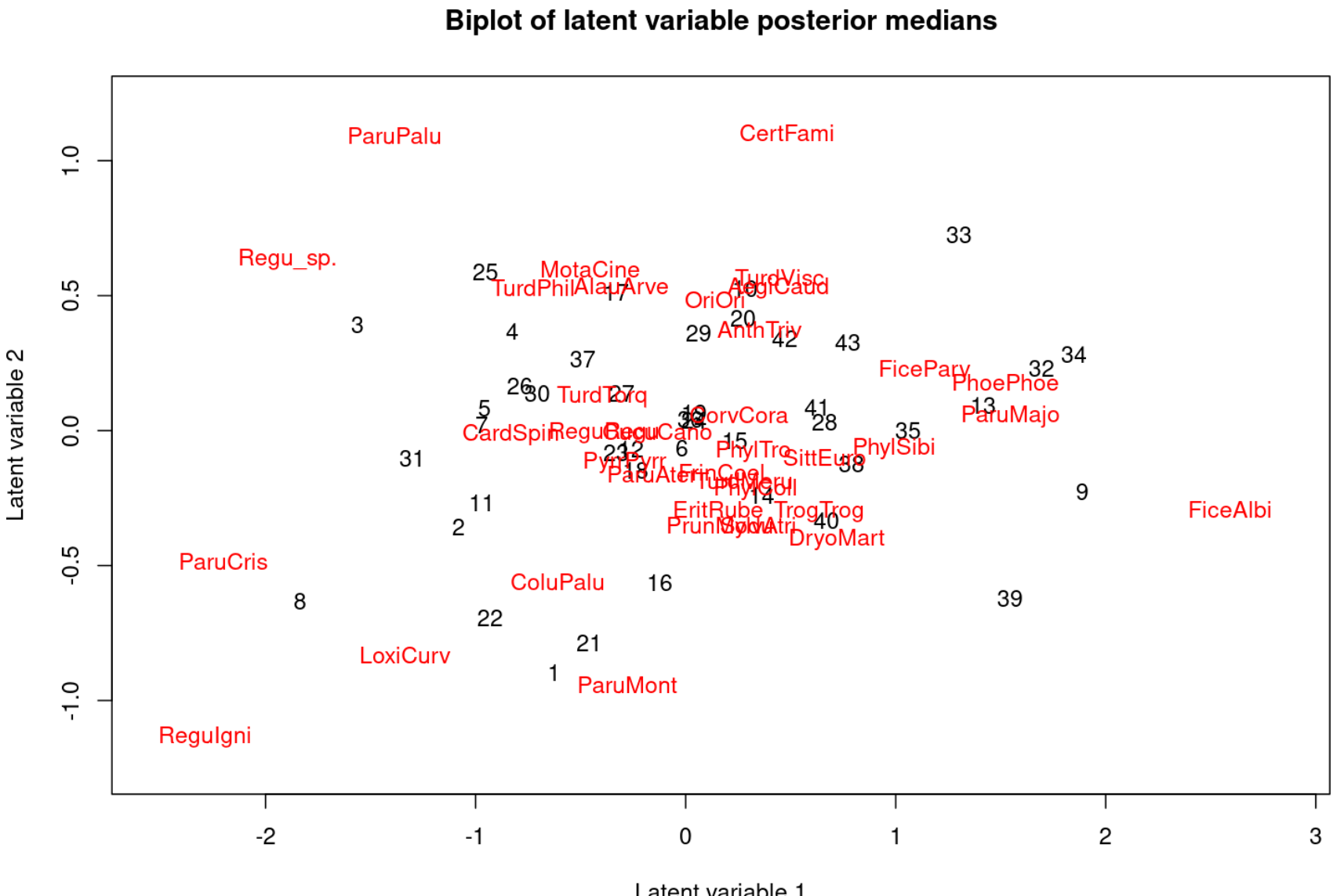
```
model <- boral::boral(Y, X, formula.X = ~ Forest + Altit, lv.control=list(
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 1591
##    Unobserved stochastic nodes: 270
##    Total graph size: 9883
##
## Initializing model
```

```
boral::lvsplot(model)
```

vignette: see paper

boral: plot


Biplot of latent variable posterior medians
Latent variable 2
Latent variable 1
1.0
0.5
0.0
-0.5
-1.0
-2
-1
0
1
2
3
ParuPalu
CertFami
Regu_sp.
MotaCine
TurdVisc
TurdPhil
AlauArve
AegiCaud
OriOri
AnthTriv
FicePary
PhoePhoe
ParuMajo
CardSpin
PhylTro
SittEuro
PhylSibi
ParuCris
ColuPalu
FiceAlbi
DryoMart
LoxiCurv
ParuMont
ReguIgni

boral

Article

Comparison of distance-based and model-based ordinations

David W. Roberts

First published:10 October 2019 | https://doi.org/10.1002/ecy.2908

Corresponding Editor: Helene H. Wagner.

boral: MCMC

- MCMC is (kind of) like optimisation, you need to check convergence
- MCMC needs "burn-in", i.e., forget the initial state
- But samples of parameters are stored; so we can expect them
- MCMC needs to mix well (explore whole parameter space)
- The chain is stationary if we have reached a good state
- We can check this visually, or with statistics
- If it has not converged, it needs to be run longer (or your model is poorly formulated)

boral: calc.varpart

Partition variance per species over model terms

```
boral::calc.varpart(model)
```

```
## $varpart.X
##   EritRube   FrinCoel   ParuAter   PrunModu   SylvAtri   TurdMeru   PhylColl
## 0.60780697 0.59865866 0.06173318 0.51924213 0.35864968 0.31753990 0.21357465
##   CucuCano   PyrrPyrr   TurdTorq   ReguRegu    PhylTro   TrogTrog   PhylSibi
## 0.37167046 0.21165534 0.44256406 0.06601324 0.61613997 0.06316656 0.47066437
##   AnthTriv   TurdPhil   ParuMont   CardSpin   PhoePhoe   ColuPalu   FiceAlbi
## 0.68493235 0.11133645 0.10628251 0.12602051 0.27975501 0.09347799 0.04809894
##   ParuMajo   Regu_sp.   CorvCora   SittEuro   LoxiCurv   ParuCris   ReguIgni
## 0.11507590 0.07984541 0.37470580 0.61249621 0.14038543 0.23710468 0.14300492
##   AlauArve   FiceParv   MotaCine     OriOri   AegiCaud   CertFami   DryoMart
## 0.69538115 0.61701777 0.36724482 0.39196857 0.30115936 0.47765464 0.34222404
##   TurdVisc   ParuPalu
## 0.32710083 0.24968084
##
## $varpart.lv
##  EritRube  FrinCoel  ParuAter  PrunModu  SylvAtri  TurdMeru  PhylColl  CucuCano
## 0.3921930 0.4013413 0.9382668 0.4807579 0.6413503 0.6824601 0.7864253 0.6283295
##  PyrrPyrr  TurdTorq  ReguRegu   PhylTro  TrogTrog  PhylSibi  AnthTriv  TurdPhil
## 0.7883447 0.5574359 0.9339868 0.3838600 0.9368334 0.5293356 0.3150676 0.8886635
##  ParuMont  CardSpin  PhoePhoe  ColuPalu  FiceAlbi  ParuMajo  Regu_sp.  CorvCora
## 0.8937175 0.8739795 0.7202450 0.9065220 0.9519011 0.8849241 0.9201546 0.6252942
```

boral

Has a few other helpful functions:

- `get.enviro.cor` and `get.residualcor`
- `predict.boral` and `plot.boral`
- `coefsplot` and `ranefsplot`

boral: compared to gllvm

boral	gllvm
Bayesian	Frequentist
MCMC	Likelihood approximation
Slow	Fast
Correlated LVs	Not yet
Single row effect	Multiple row effects
Stochastic Variable Selection	Adaptive shrinkage?

There is little reason to use boral at this point, except for the SVSS and correlation of LVs.

Hierarchical Modeling of Species Communities

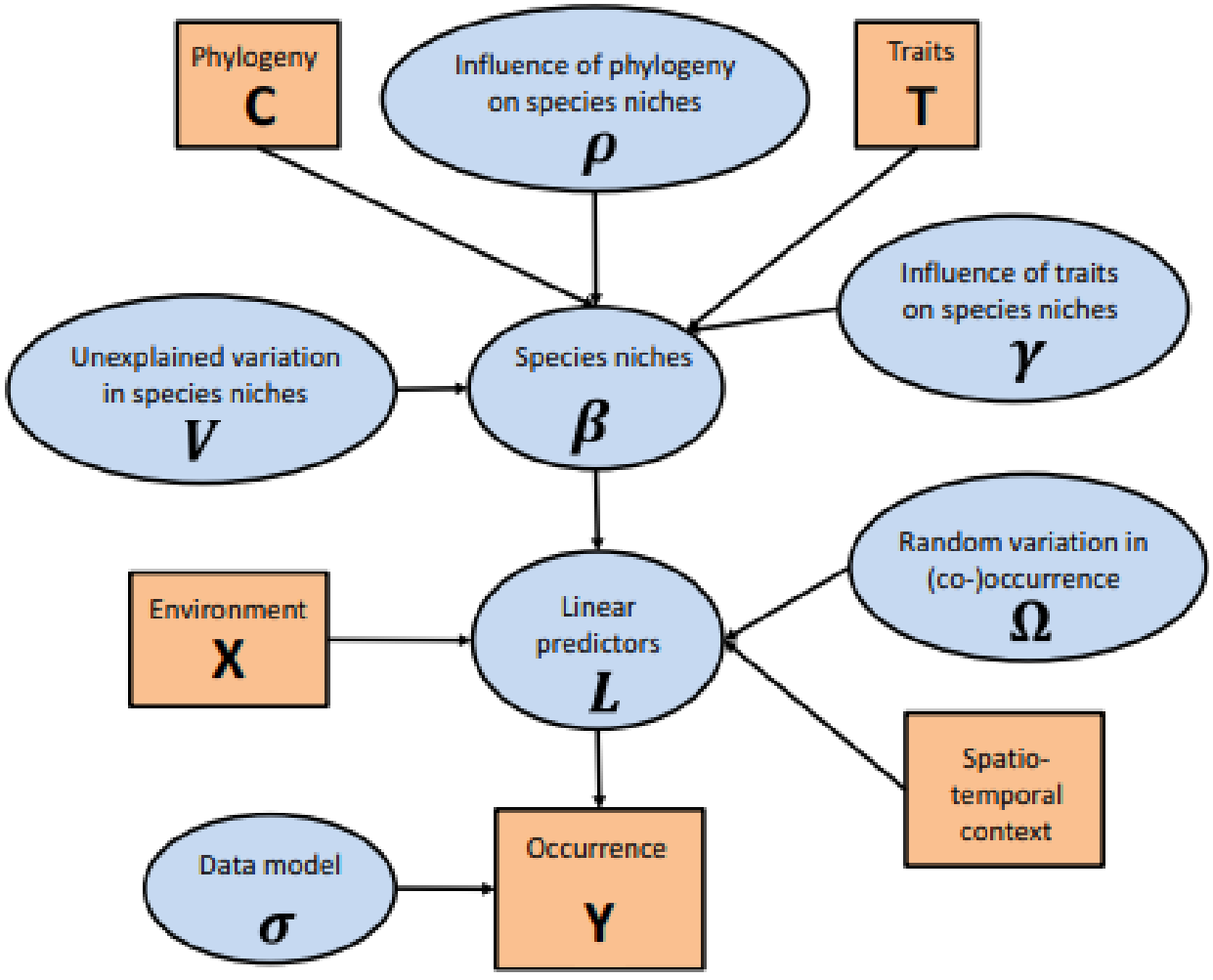
APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Joint species distribution modelling with the R-package HMSC

Gleb Tikhonov[1,2] | Øystein H. Opedal[2,3] | Nerea Abrego[4] | Aleksi Lehikoinen[5] | Melinda M. J. de Jonge[6] | Jari Oksanen[7] | Otso Ovaskainen[2,3]

HMSC


Phylogeny
C
Influence of phylogeny on species niches
ρ
Traits
T
Influence of traits on species niches
γ
Unexplained variation in species niches
V
Species niches
β
Random variation in (co-)occurrence
Ω
Environment
X
Linear predictors
L
Spatio-temporal context
Data model
σ
Occurrence
Y

HMSC

HMSC was introduced by Ovaskainen et al. (2017) but has been expanded a lot since then

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Technological Advances at the Interface Between Ecology and Statistics | Free Access

Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context

Gleb Tikhonov, Nerea Abrego, David Dunson, Otso Ovaskainen

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

APPLICATION | Open Access

Joint species distribution modelling with the R-package HMSC

Gleb Tikhonov, Øystein H. Opedal, Nerea Abrego, Aleksi Lehikoinen, Melinda M. J. de Jonge, Jari Oksanen, Otso Ovaskainen

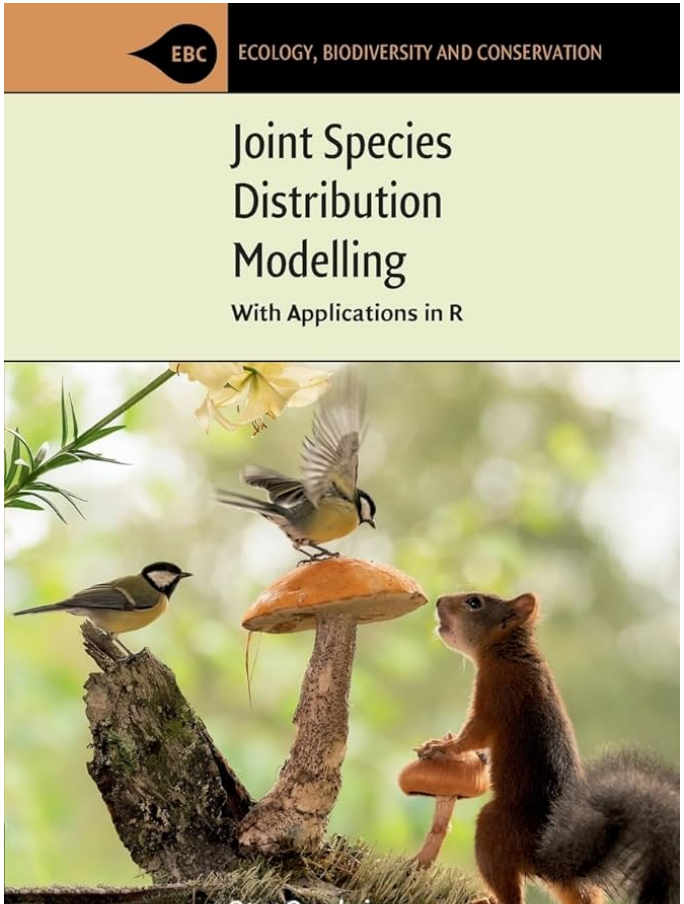
First published: 25 December 2019 | https://doi.org/10.1111/2041-210X.13345 | Citations: 178

ECOLOGICAL SOCIETY OF AMERICA

Statistical Reports | Open Access

Computationally efficient joint species distribution modeling of big spatial data

Gleb Tikhonov, Li Duan, Nerea Abrego, Graeme Newell, Matt White, David Dunson, Otso Ovaskainen



HMSC

- Bayesian; fits with MCMC
- Custom Gibbs samplers
- Flexible package for multispecies hierarchical modeling
- Focuses on prediction and species associations

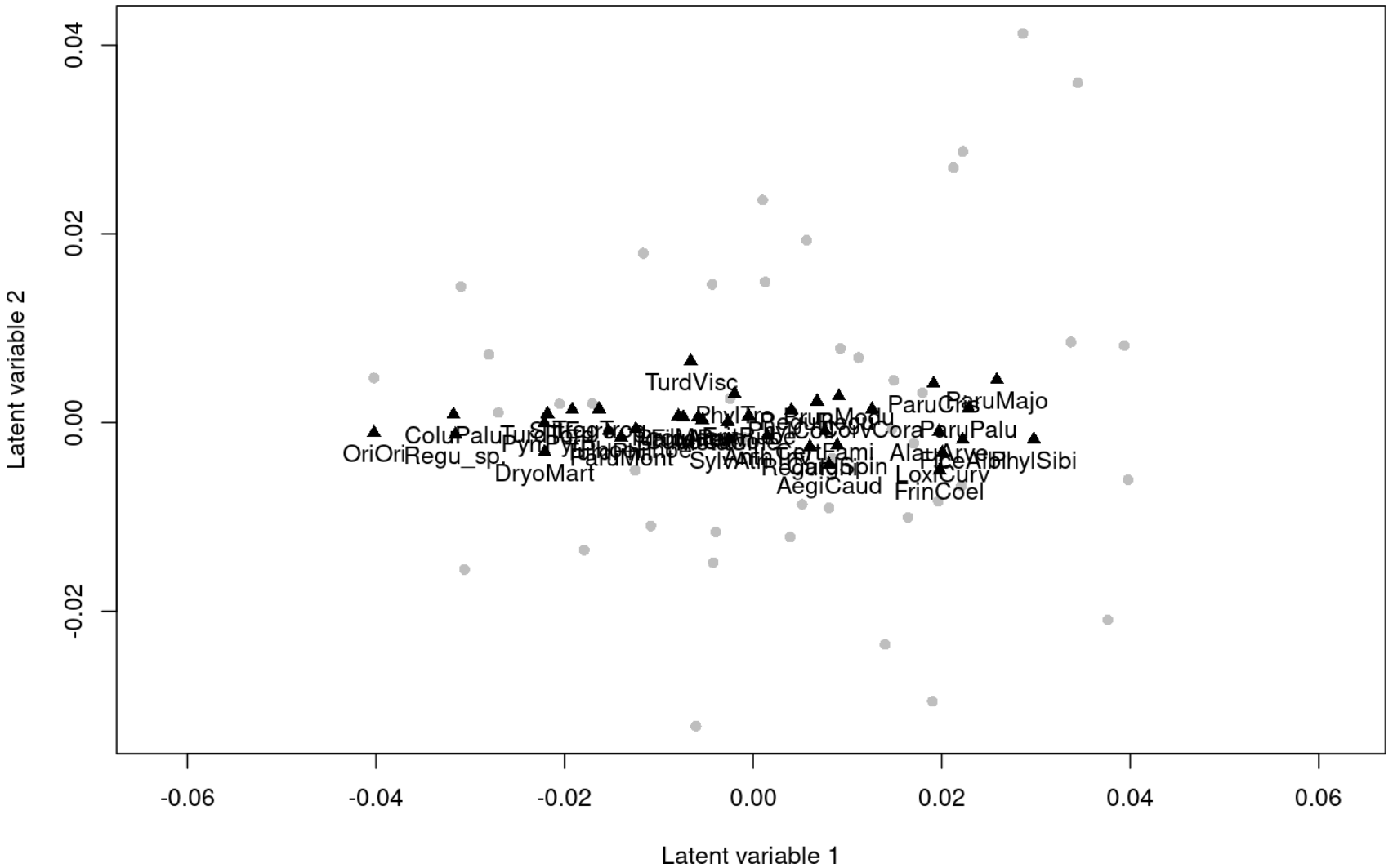
- Phylogenetic effects
- Efficiently implements spatial models with nearest neighbors
- 4th corner model
- Various extra random effects (intercepts and such)
- Effects can be specified at different sampling levels, including sets of LVs
- The "infinite factor model"
- Have a preprint on parallelisation
- Very little support for ordination
- Supports mixed response types

HMSC: code

```
# need to set-up LVs
studyDesign = data.frame(sample=as.factor(1:nrow(Y)))
rL <- Hmsc::HmscRandomLevel(units = studyDesign$sample)
model <- Hmsc::Hmsc(Y, XFormula = ~Forest+Altit, XData= X,
distr = "lognormal poisson", studyDesign = studyDesign,
ranLevels = list(sample = rL))
# Run mcmc
run =  Hmsc::sampleMcmc(model, samples = 1000, nChains = 3,
              transient = 2500)
```

```
## Computing chain 1
## Chain 1, iteration 70 of 3500 (transient)
## Chain 1, iteration 140 of 3500 (transient)
## Chain 1, iteration 210 of 3500 (transient)
## Chain 1, iteration 280 of 3500 (transient)
## Chain 1, iteration 350 of 3500 (transient)
## Chain 1, iteration 420 of 3500 (transient)
```

HMSC: plot


Latent variable 2
0.04
0.02
0.00
-0.02
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
Latent variable 1
TurdVisc
ParuMajo
ParuPalu
OriOri
Regu_sp
DryoMart
AegiCaud
FrinCoel

HMSC

`HMSC`	`gllvm`
Bayesian	Frequentist
MCMC	Likelihood approximation
Slow (but getting quicker)	Fast
normal, Bernoulli, Poisson, lognormal Poisson	Wide range of response types
Effects at different sampling levels	Only at one sampling level
Infinite factor model	Number of LVs fixed a-priori
Efficient spatial implementation	Spatial is a work in progress
Few tools for ordination	many tools for ordination

Ultimately, the focus of these two packages is very different. HMSC focuses on prediction and JSDMs, `gllvm` can do that, but its main focus is different (IMO).

ecoCopula

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Fast model-based ordination with copulas

Gordana C. Popovic[1] | Francis K. C. Hui[2] | David I. Warton[1]

- Employs graphical models for determining species associations
- Requires a secondary model
- Is -very- fast for ordination (faster than NMDS!)
- Can estimate "direct associations" (not as quick)
- Supports mixed response types

ecoCopula: code

```
preModel <- ecoCopula::stackedsdm(Y, formula_X =~1, data = X)
model <- ecoCopula::cord(preModel)
plot(model, biplot=TRUE)
```

vignette

ecoCopula

ecoCopula	gllvm
Frequentist	Frequentist
Gaussian Copula	Likelihood approximation
Faster	Fast
A decent number of distributions	Wide range of response types
Direct species associations	Correlative
None	Many other random effects
Secondary model in parallel	Working on parallel computation
Native residuals	Native residuals
Biplot function	Biplot function
Marginal interpretaiton	Conditional interpretation

ecoCopula has a lot of potential due to its speed, but lacks in support, maintenance, and perhaps some maturity.

Vector Generalised Linear and Additive Models

Journal of Statistical Software

January 2010, Volume 32, Issue 10. *http://www.jstatsoft.org/*

The VGAM Package for Categorical Data Analysis

Thomas W. Yee
University of Auckland

Vector Generalised Linera and Additive Models

- Package with a wide range of model types **VGLMs**
- Massive package with a lot of functionality
- An incredible range of response distributions
- Unconstrained and constrained ordination (fixed effects formulation)
- Quadratic and additive ordinations
- Supposed to fit quickly with IWLS
- In my experience, fitting is often difficult (errs often) and can be unstable
- Has some residuals
- Plotting functions are a bit different
- No random effects
- Now (recently) has doubly-constrained ordination!

Centers around `vglm()`, `vgam()`, `rrvglm()`, `cqo()`, `cao()`, `rcim()`

VGAM

The first (model-based) constrained ordination method

Ecological Monographs, 74(4), 2004, pp. 685–701

A NEW TECHNIQUE FOR MAXIMUM-LIKELIHOOD CANONICAL GAUSSIAN ORDINATION

THOMAS W. YEE[1]

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand, and
Department of Statistics and Applied Probability, 6 Science Drive 2, National University of Singapore, Singapore 117546

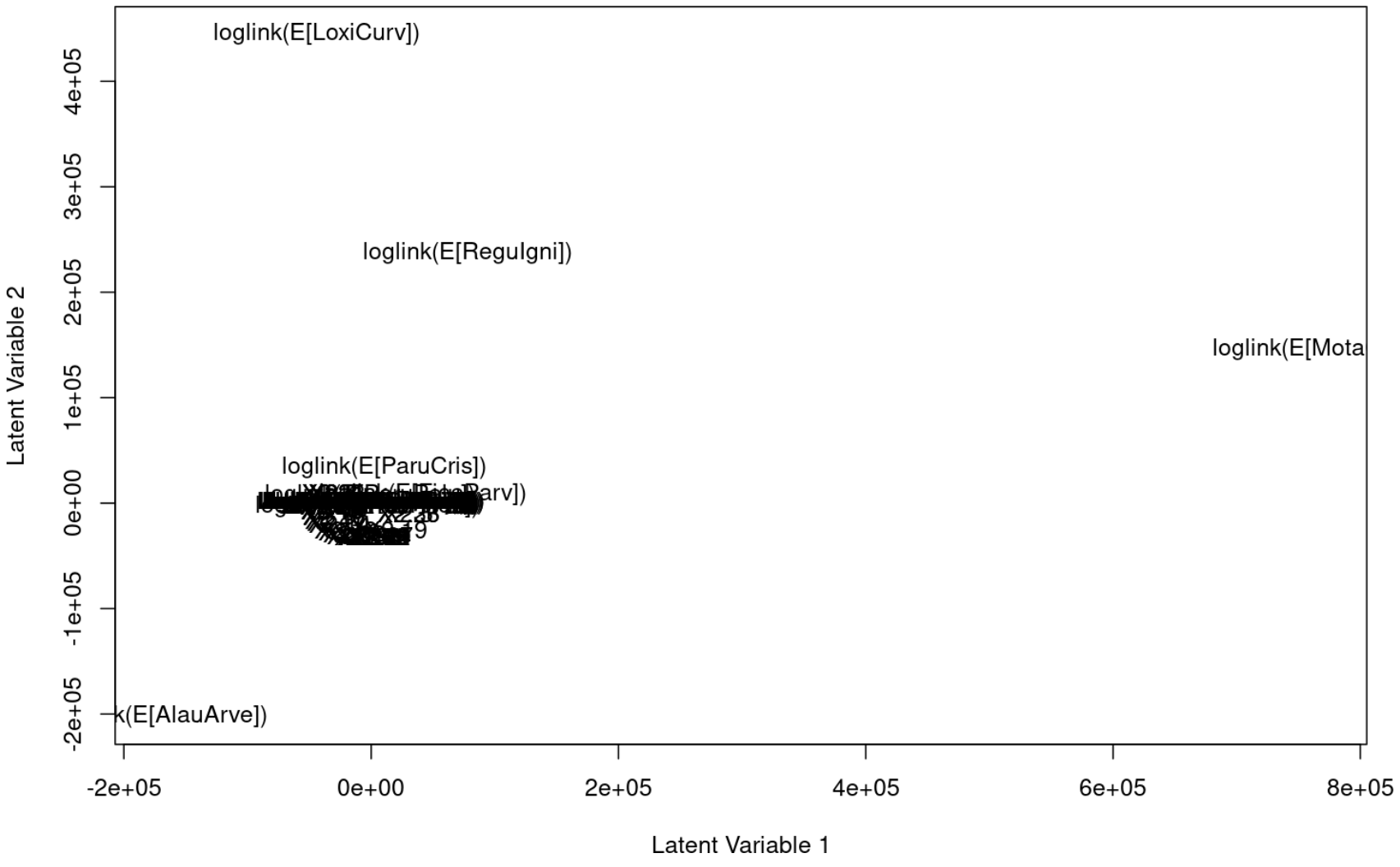
VGAM: code

```
model1 <- VGAM::rcim(Y, Rank = 2, family = VGAM::poissonff)
VGAM::lvplot(model1)
```

```
# Could not get this to work :(
# model2 <- VGAM::rrvglm(Y ~ model.matrix(~.,X[,1:4])[,-1], Rank = 2, fami
```

vignette: see reference card

VGAM: code


loglink(E[LoxiCurv])
loglink(E[ReguIgni])
loglink(E[Mota
loglink(E[ParuCris])
k(E[AlauArve])
Latent Variable 2
-2e+05
-1e+05
0e+00
1e+05
2e+05
3e+05
4e+05
-2e+05
0e+00
2e+05
4e+05
6e+05
8e+05
Latent Variable 1

VGAM

`VGAM`	`gllvm`
Frequentist	Frequentist
ML via IWLS	pproximate marginal likelihood
Fast	Fast
Incredible range of responses	Wide range of response types
Not robust fitting	Relatively robust
No random effects	Many other random effects
UQO, CQO, CAO	UQO, CQO
VGAMs	No smooths
Native residuals	Native residuals
Biplot function	Biplot function

`VGAM` has a lot of potentially useful tools, but I do not find it very usable.

glmmTMB

glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling

by Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Mächler, Benjamin M. Bolker

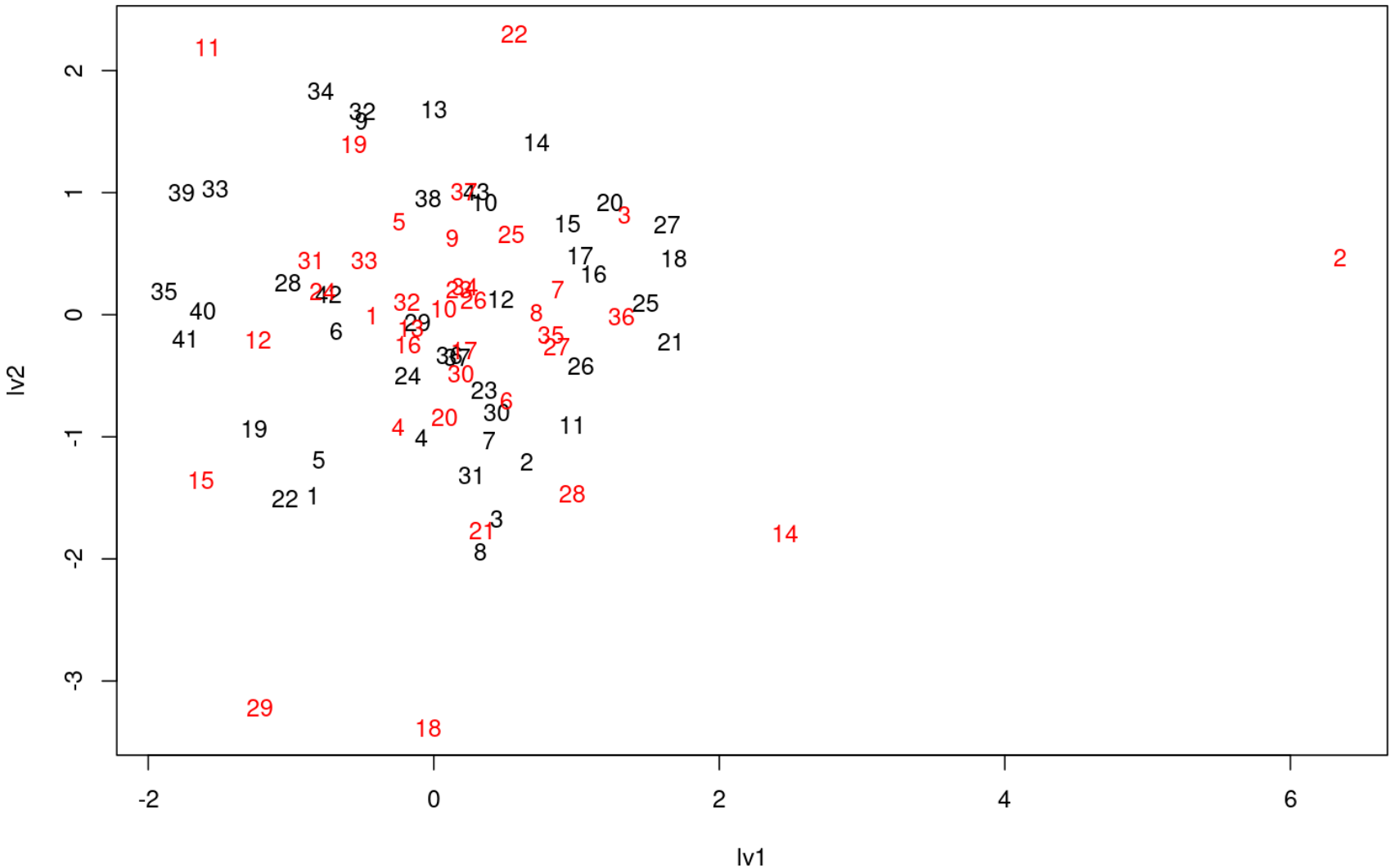
- Kind of similar to `gllvm` in that it uses approximate methods
- Laplace approximation with TMB (state-of-the-art)
- Great usability
- Can include many random-effects
- Unconstrained and constrained ordination (RE formulation)
- Slower than `gllvm`'s VA (I think?)
- Structured random effects (e.g., spatial) soon Phylogenetic
- No other support for ordinations
- Zero-inflated modeling

glmmTMB: code

Note: data needs to be in long format

```
# organize data into long format
tmp <- data.frame(Y)
tmp$id <- 1:nrow(tmp)
glmmDat <- reshape(tmp,
                   idvar = "id",
                   timevar = "col",
                   times =  colnames(Y),
                   varying = list(colnames(Y)),
                   v.names = "y",
                   direction = "long")
model <- glmmTMB::glmmTMB(y ~ col + rr(col + 0|id, d = 2), data = glmmDat,
rrstuf <- extract_rr(model) # see here: https://github.com/glmmTMB/glmmTMB
plot(rbind(rrstuf$b,rrstuf$fl), type = "n");
text(rrstuf$b);text(rrstuf$fl, col = "red")
```

glmmTMB: plot



glmmTMB

glmmTMB	gllvm
Frequentist	Frequentist
Laplace	VA (default) or Laplace
Fast	Fast(er)
Wide range of response types	Wide range of response types
Many (structured) random effects	Many random effects
Can also fit with MCMC	No
Zero-inflated modeling	Work in progress
No residuals	Native residuals
No plotting function	Biplot function
Large community	Small community
Excellent developers	No comment :)

glmmTMB is especially useful if you want user friendliness and many other random effects. Ordination is an afterthought in a package supposed to do many other things (but also still new).

Generalized Matrix Factorization

Generalized Matrix Factorization: efficient algorithms for fitting generalized linear latent variable models to large data arrays

Łukasz Kidziński LUKASZ.KIDZINSKI@STANFORD.EDU
Department of Bioengineering
Stanford University
Stanford, CA 94305, USA

Francis K.C. Hui FRANCIS.HUI@ANU.EDU.AU
Research School of Finance, Actuarial Studies and Statistics
The Australian National University
Canberra, ACT 2601, Australia

David I. Warton DAVID.WARTON@UNSW.EDU.AU
School of Mathematics and Statistics
and Evolution & Ecology Research Centre
The University of New South Wales
Sydney, NSW 2052, Australia

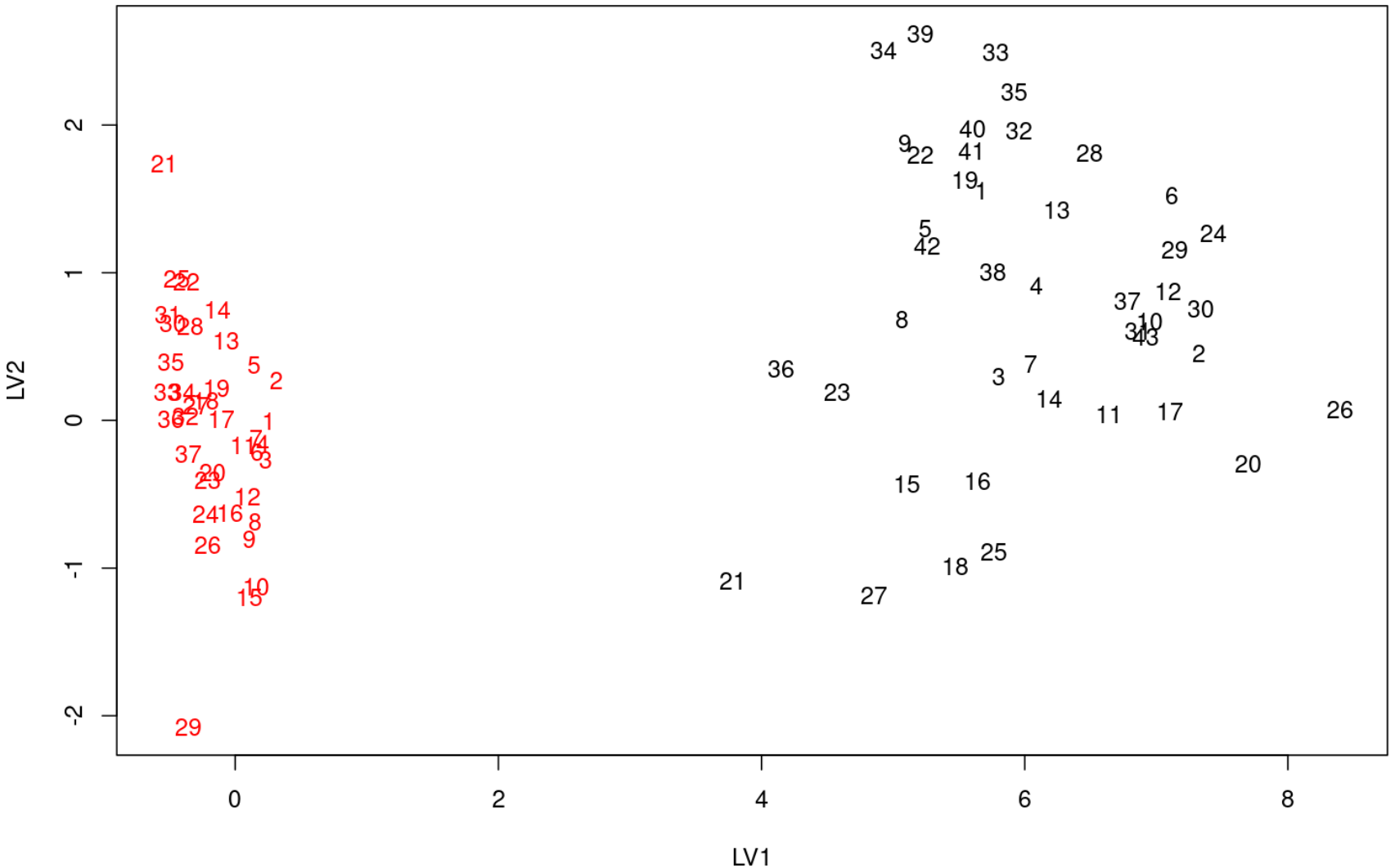
Trevor Hastie HASTIE@STANFORD.EDU
Department of Statistics and Biomedical Data Science
Stanford University
Stanford, CA 94305, USA

- Very quick; fits by penalized likelihood
- Unconstrained or residual ordination only
- No extra random-effects
- Can be unstable due to the approximation
- Stale package not on CRAN

gmf: code

```
# devtools::install_github("kidzik/gmf")
model <- gmf::gmf(Y, family = poisson(), p = 2)
plot(rbind(model$u,model$v), type = "n", xlab="LV1", ylab="LV2")
text(model$u)
text(model$v, col="red")
```

gmf: plot


LV2
LV1
2
1
0
-1
-2
0
2
4
6
8
21
29
39
34
33
35
40
32
9
22
41
28
19
1
6
13
5
42
24
29
38
4
12
37
30
8
10
43
2
36
7
3
23
14
11
17
26
20
15
16
25
18
21
27

gmf

gmf	gllvm
Frequentist	Frequentist
Penalized likelihood	VA or LA approximation
Fast(er)	Fast
A few response types	Wide range of response types
Fitting is fine	Relatively robust
No random effects	Many other random effects

A skeleton of a package, not very useful at this point.

RCM

A unified framework for unconstrained and constrained ordination of microbiome read count data

Stijn Hawinkel[1]*, **Frederiek-Maarten Kerckhof**[2], **Luc Bijnens**[3,4], **Olivier Thas**[1,4,5]

1 Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium, **2** Center for Microbial Ecology and Technology, Ghent University, Ghent, Belgium, **3** Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson, Beerse, Belgium, **4** Center for Statistics, Hasselt University, Hasselt, Belgium, **5** National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia

* stijn.hawinkel@ugent.be

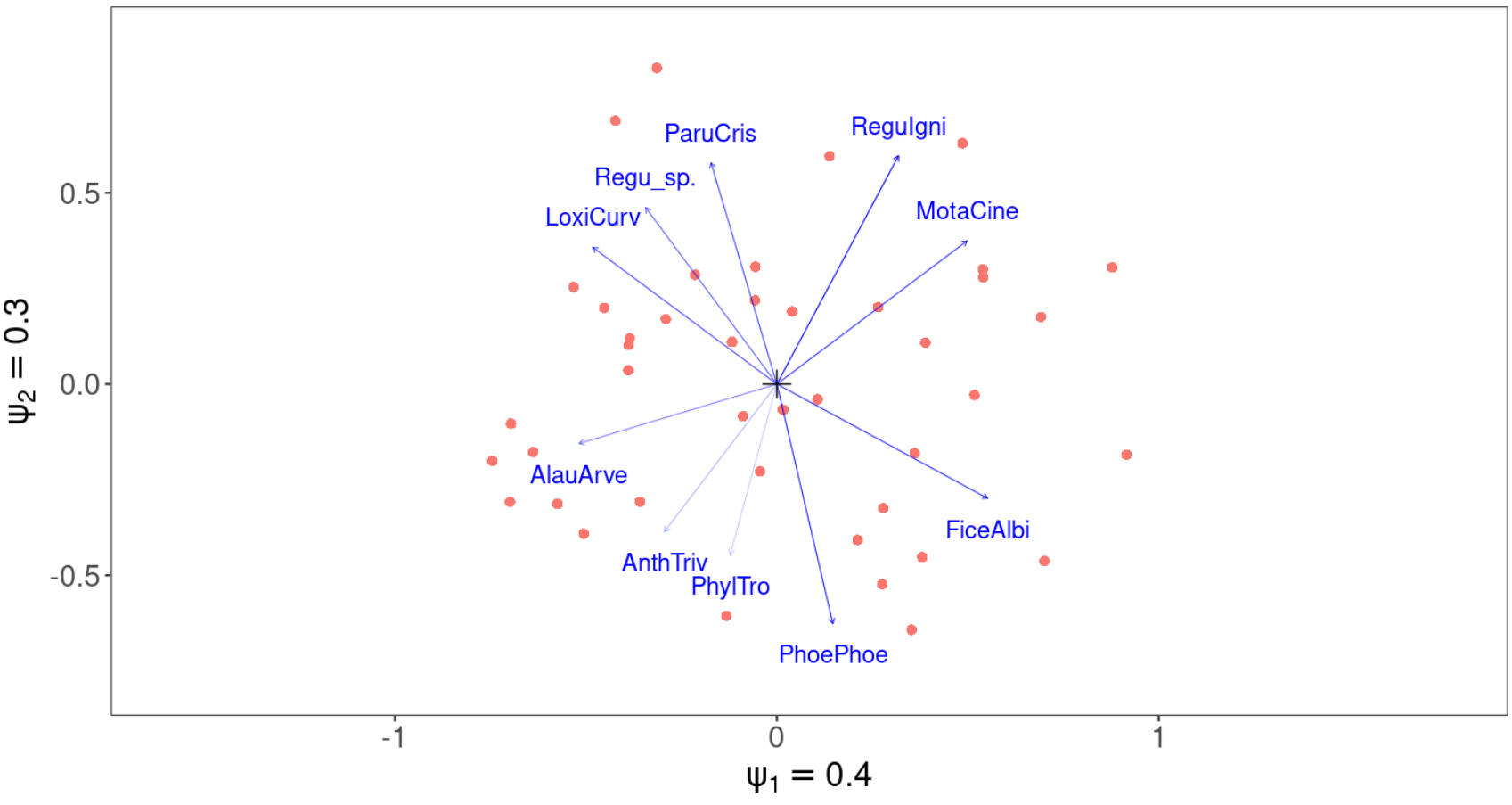
- Does both unconstrained and unconstrained ordination
- Even additive constrained ordination
- All based on fixed effects formulations, no random effects
- Only the negative binomial distribution
- Not a "true" statistical model (according to the authors)
- Permanova functionality
- Residual plots

RCM: code

```
# devtools::install_github("CenterForStatistics-UGent/RCM")
model <- RCM::RCM(Y, k = 2)
plot(model)
```

vignette

RCM: plot


ParuCris
ReguIgni
Regu_sp.
LoxiCurv
MotaCine
AlauArve
FiceAlbi
AnthTriv
PhylTro
PhoePhoe
0.5
0.0
-0.5
-1
0
1
$\psi_2 = 0.3$
$\psi_1 = 0.4$

RCM

`RCM`	`gllvm`
Frequentist	Frequentist
Maximum likelihood	Approximate marginal likelihood
Fast	Fast
Only NB	Wide range of response types
UO, CO, CQO, CAO	UO, CO, CQO
No random effects	Many other random effects

`RCM` seems good at what it does, but functionality is limited.

Experimental: latent INLA

README

Bob O'Hara 8/29/2021

LatentINLA Readme

This package is made to play around with latent variable models in INLA.

To Do

- look at speeding it up
- are initial values going to help with speed or convergence?
- Try a constrained model
- Try a spatial model, making the site scores spatial rather than iid.
- plant some trees to offset the atmospheric warning created by running these models

R-CMD-check passing | CRAN not published | lifecycle experimental | codecov unknown

- Relatively fast; Bayesian with Laplace approximation
- Unconstrained, concurrent ~~, and hierarchical ordination~~
- Fast fitting of large spatial effects
- **Very much a work in progress**

latent INLA: code

don't try, this crashed :(

```
# devtools::install_github("oharar/LatentINLA")
model <- LatentINLA::FitGLLVM(Y=Y, Family="poisson", nLVs = 2
LatentINLA::biplot(model)
```

vignettes

Community-level basis function models


RESEARCH ARTICLE

Methods in Ecology and Evolution

Spatiotemporal joint species distribution modelling: A basis function approach

Francis K. C. Hui[1] | David I. Warton[2] | Scott D. Foster[3] | Christopher R. Haak[4]

- Most recent developments: GAM for multiple species
- Specifically targeted on spatio,temporal or spatio-temporal analysis
- This is something GLLVMs are not -terribly- good at yet (but very much an area of interest)
- Based on the idea of LVMs, but not with LVs
- Fitting using `TMB`
- I.e., JSDM-oriented, not ordination

CBFM

`CBFM`	`gllvm`
Frequentist	Frequentist
Penalized Quasi-likelihood	Approximate marginal likelihood
For large spatio-temporal problems	Not an option (yet?)
Wide range of response types	Wide range of response types
Post-hoc ordination	Is an ordination method
Can include extra "random effects" as smooths	Many other random effects
Parallelisation	Parallelisation
No traits of Phylogeny	Traits and Phylogeny

Sorry, no example yet. Bird data does not have coordinates, and CBFM only fits models with space it seems?

Software summary

Package	cran[1]	UO[2]	CO[3]	CN[4]	RE[5]	CI[6]	traits	Phylogeny	Space	framework[7]
gllvm	yes	yes	yes	yes	yes	yes	yes	yes	Not really	F
Boral	yes	yes	no	no	some	yes	yes	no	yes	B
HMSC	yes	yes	no	no	yes	yes	yes	yes	yes	B
ecoCopula	yes	yes	no	no	no	no	kind of	no	no	F
VGAM	yes	yes	yes	no	no	some	new?	no	no	F
glmmTMB	yes	yes	yes	no	yes	yes	yes	soon	Kind of	F
gmf	no	yes	no	no	no	no	no	no	no	F
RCM	no	yes	yes	no	no	no	no	no	no	F
latent INLA	no	no	no	yes	yes	yes	soon	no	yes	B
CBFM	no	no	no	no	yes	yes	no	no	yes	B

[1]cran: Package available on CRAN. [2]UO: Unconstrained ordination. [3]CO: Constrained. [4]CN: Concurrent. [5]RE: Random effects. [6]CI: Confidence/Credible intervals. [7]framework: The underlying framework of the model (F: Frequentist, B: Bayesian).

When to use what package?

- `HMSC` for extensive support for JSDMs
- `VGAM` is what you want is not supported by `gllvm`
- `glmmTMB` for many (structured) random effects
- `ecoCopula` if you have a huge dataset and `gllvm` is too slow
- `CBFM` for large spatial/temporal models

`gllvm` for all your ordination needs

Summary

New software implementations are continuously being developed. Dimension reduction methods for ecology have entered a new era.

- It is important that we continue to explore new and better methods
- Especially the application of ordination methods are still a bit stuck in the past
- Generally speaking, there is still a lot of work to be done on multivariate methods for ecommunity ecology
- There are more packages for model-based analysis that I have not mentioned
- E.g., `jSDM`, `sjSDM`, `BayesComm`