Taking information out of an ordination

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

So far, we have covered three types of ordination.

- Unconstrained ordination
- Constrained ordination
- Concurrent ordination

Outline

So far, we have covered three types of ordination. Choosing between these is relatively straightforward:

- Unconstrained ordination: when you do not have covariates
- Constrained ordination: when there is no residual covariation
- Concurrent ordination: when you have covariates and residual covariation

Outline

What I will cover in this lecture:

- The different formula interfaces
- The impact of adding effects outside of the ordination (and what role it plays)
- Partial ordination
- How to choose where to include covariates
- How this impacts variation explained

Outline

What I will cover in this lecture:

- The different formula interfaces
- The impact of adding effects outside of the ordination (and what role it plays)
- Partial ordination
- How to choose where to include covariates
- How this impacts variation explained

The goal is to get you to think about model structure and formulation

Questions so far?

Ordination with nested design

Imagine a study with nested design (e.g., plots $k = 1 \dots K$ in sites $i = 1 \dots n$), and our usual model:

$$\eta_{ijk} = \beta_{0j} + \mathbf{u}_{ik}^\top \boldsymbol{\gamma}_j \tag{1}$$

Here, we are incorporating the plots into the ordination (which can be quite messy).

Ordination with nested design

But, perhaps we think an ordination "lives" at the site-level. We define $\mathbf{u}_{ik} = \mathbf{u}_i + \mathbf{u}_k$:

$$\eta_{ijk} = \beta_{0j} + (\mathbf{u}_k + \mathbf{u}_i)^\top \boldsymbol{\gamma}_j \qquad (2)$$

Now we have two (connected by loadings) ordinations in the model: 1) $\mathbf{u}_i^\top \boldsymbol{\gamma}_j$ and 2) $\mathbf{u}_k^\top \boldsymbol{\gamma}_j$.

- 1) is a site-specific ordination
- 2) is a plot-specific ordination
- We assume these have the same dimensions
- We assume these have the same loadings
- So the covariance matrix remains similar: $\boldsymbol{\Sigma} = 2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$

Connection to ordination with predictors

The ordination with nested design from before:

$$\boxed{\eta_{ijk}} = \boxed{\beta_{0j}} + \boxed{(\mathbf{u}_k + \mathbf{u}_i)}^\top \boxed{\boldsymbol{\gamma}_j} \tag{3}$$

Is a type of constrained ordination, because we can write the latent variable $\mathbf{u}_{ik} = \mathbf{B}^\top \mathbf{x}_{ik}^{lv}$ with

$$\mathbf{B} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_i \\ \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_k \end{bmatrix} \qquad \mathbf{X}^{lv} = \begin{array}{c|ccc|ccc} & \text{site 1} & \text{site 2} & \text{site 3} & \text{plot 1} & \text{plot 2} & \text{plot 3} \\ \hline 1.1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 2.3 & 0 & 1 & 0 & 0 & 0 & 1 \\ 3.1 & 0 & 0 & 1 & 1 & 0 & 0 \end{array}$$

Ordination with nested design

Following similar logic, we can relax the assumptions a little:

$$\eta_{ijk} = \beta_{0j} + \mathbf{u}_i \boldsymbol{\gamma}_j + \mathbf{u}_k \boldsymbol{\theta}_j \tag{4}$$

Ordination with nested design

Following similar logic, we can relax the assumptions a little:

$$\eta_{ijk} = \beta_{0j} + \mathbf{u}_i \boldsymbol{\gamma}_j + \mathbf{u}_k \boldsymbol{\theta}_j \tag{4}$$

- Now we have two separate ordinations
- These can have different dimensions
- The covariance matrix is now: $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \boldsymbol{\Theta}\boldsymbol{\Theta}^\top$

Ordination with nested design

What happens if we just omit the variation at the plot-level?

Ordination with nested design

What happens if we just omit the variation at the plot-level?

It depends on the true model, but for example:

$$\eta_{ijk} = \beta_{0j} + \mathbf{u}_i^\top \boldsymbol{\gamma}_j + \epsilon_{kj} \tag{5}$$

We could be omitting a plot-level random effect.

Example 1: Wadden data

Wadden sea data Dewenter et al. (2023)

- Abundance (counts) or Biomass of macrozoobenthos
- Covariates
- **Transects at islands (Norderney, Spiekeroog, Wangerooge)**

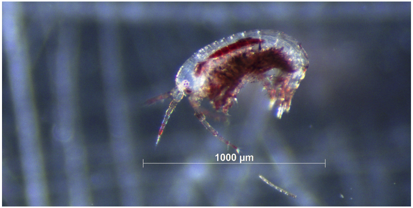


Figure 1: nioz.nl

```
Y <- read.csv("../data/waddenY2.csv")[,-c(1:2)]
Y <- Y[, colSums(ifelse(Y==0,0,1))>2]
X <- read.csv("../data/waddenX.csv")
X[,unlist(lapply(X,is.numeric))] <- scale(X[,unlist(lapply(X,is.numeric))])
```

Example 1: Study design

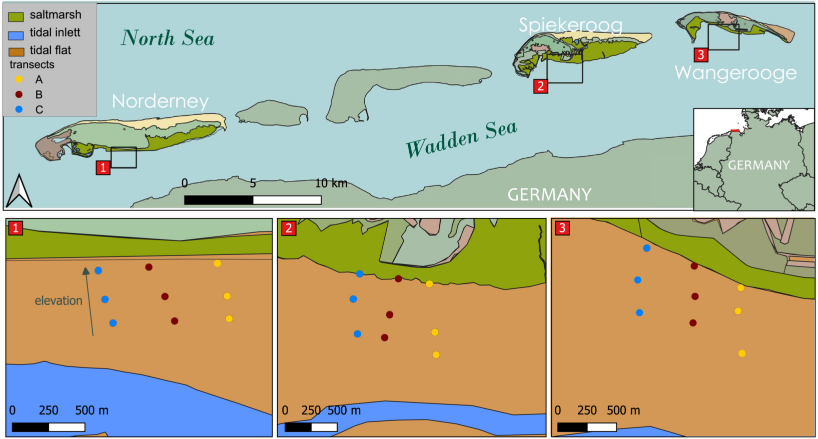
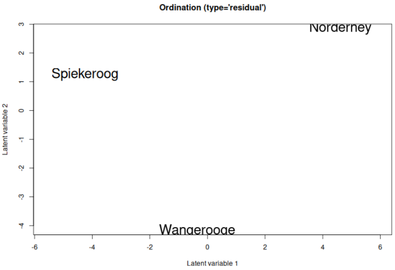


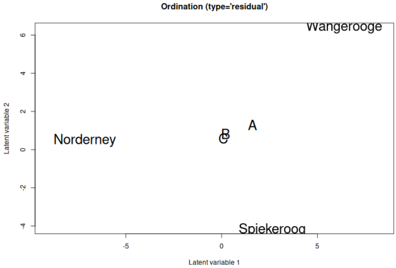
Figure 2: Dewenter et al. Fig 1.

Example 1: group-level ordination

```
model1 <- gllvm(y = Y, num.lv = 2,
                lvCor = ~(1|island), studyDesign = X,
                family = "tweedie", Power = NULL, n.init = 3, disp.formula = rep(1,ncol(Y)))
model2 <- update(model1, lvCor = ~(1|island) + (1|transect))
```



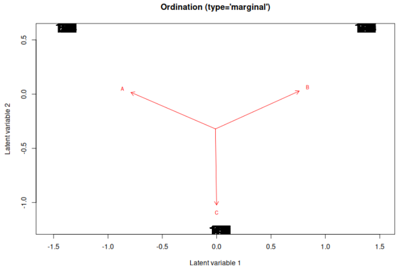
This is an ordination at the island-level.

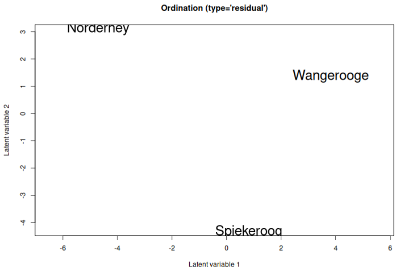


These are two ordinations with the same loadings.

Example 1: two ordinations with different loadings

```r
model3 <- gllvm(y = Y, num.lv = 2, lvCor = ~(1|island), studyDesign = X,
            num.RR  = 2, lv.formula = ~diag(1|transect), X = X, randomB = "LV",
            family = "tweedie", Power = NULL, n.init = 3, disp.formula = rep(1,ncol(Y)))
```





These are two ordinations with different loadings.

Ordination with nested design

Another model may be:

$$\eta_{ijk} = \beta_{0j} + \alpha_k + \mathbf{u}_i^\top \gamma_j \tag{6}$$

or even:

$$\eta_{ijk} = \beta_{0j} + \mathbf{z}_k^\top \lambda_j + \mathbf{u}_i^\top \gamma_j \tag{7}$$

What is the difference between these two models? What assumptions do we make in these models?

Conditioning

In classical methods, we can use `Condition` to remove effects due to a covariate from the ordination. Here, we adjust our model with terms outside of the ordination:

$$\eta_{ij} = \beta_{0j} + \boxed{\ldots} + \mathbf{u}_i^\top \boldsymbol{\gamma}_j \tag{8}$$

Which can be a fixed or random effect (fixed in classical ordination)

Conditioning with covariates

Sometimes, we want to deliberately **remove information** from the ordination:

- Observer effects
- Nested designs
- Spatial/temporal effects
- Confounders

To improve interpretability of the ordination: this is called "conditioning" or "partial" ordination.

Constrained and concurrent ordination instead **include information** in the ordination.

Conditioning with covariates

Sometimes, we want to deliberately **remove information** from the ordination:

- Observer effects
- Nested designs
- Spatial/temporal effects
- Confounders

To improve interpretability of the ordination: this is called "conditioning" or "partial" ordination.

Constrained and concurrent ordination instead **include information** in the ordination.

We can combine these two concepts by conditioning on covariates.

Conditioning: example

Imagine an example where we have treatments (say fertilizer use) under different conditions (say soil type).

If we are interested in the effect of the treatment, but we need to **control** for soil type, we include both in a model:

Conditioning: example

Imagine an example where we have treatments (say fertilizer use) under different conditions (say soil type).

If we are interested in the effect of the treatment, but we need to **control** for soil type, we include both in a model:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^{\text{soil type}\top}\beta_j^{\text{soil type}} + \mathbf{x}_i^{\text{fertilizer}\top}\beta_j^{\text{fertilizer}}$$

Conditioning: example

Imagine an example where we have treatments (say fertilizer use) under different conditions (say soil type).

If we are interested in the effect of the treatment, but we need to **control** for soil type, we include both in a model:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^{\text{soil type}\top}\beta_j^{\text{soil type}} + \mathbf{x}_i^{\text{fertilizer}\top}\beta_j^{\text{fertilizer}}$$

Now, the effect of fertilizer is **conditional** on the effect of soil type. In essence, we keep soil type constant for determining the effect of fertilizer.

What happens if we do not control for soil type?

Conditioning: example

Imagine an example where we have treatments (say fertilizer use) under different conditions (say soil type).

If we are interested in the effect of the treatment, but we need to **control** for soil type, we include both in a model:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^{\text{soil type}\top}\beta_j^{\text{soil type}} + \mathbf{x}_i^{\text{fertilizer}\top}\beta_j^{\text{fertilizer}}$$

Now, the effect of fertilizer is **conditional** on the effect of soil type. In essence, we keep soil type constant for determining the effect of fertilizer.

What happens if we do not control for soil type? If we administer fertilizer to already rich soils we probably find no effect.

Conditioning: ordination

The same applies too the ordination. If we condition, the ordination captures what is not captured by the conditioning term, for example:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^{\text{soil type}\top} \beta_j^{\text{soil type}} + \mathbf{u}_i^\top \gamma_j \tag{9}$$

Sparse data

Ultimately, the goal of ordination is to facilitate analysis of sparse community data.

We can only have so many effects outside of the ordination.

- Each effect outside of the ordination "costs" one parameter per species
- Each effect inside the ordination "costs" one parameter per latent variable
- Conditioning effects can be of interest
- Effects that we want to represent with higher accuracy
- For example, an elevation gradient
- The ordination accounts for patterns thereafter

The model: partial ordination components

$$\eta_{ij} = \beta_{0j} + \alpha_i + \nu_{ij} + \mathbf{u}_i^\top \boldsymbol{\gamma}_j \tag{10}$$

As before, the index of $\mathbf{u}_i$ is flexible .

- We always condition on β_{0j}
- $\alpha_i = \mathbf{x}_i^{r,\top} \boldsymbol{\beta}^r + \mathbf{z}_i^{r,\top} \boldsymbol{\lambda}^r$ are species common or row effects outside of the ordination
- $\nu_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{z}_i^\top \boldsymbol{\lambda}_j$ are species-specific effects outside of the ordination

The model: partial ordination interface

$$\eta_{ij} = \beta_{0j} + \alpha_i + \nu_{ij} + \mathbf{u}_i^\top \gamma_j \tag{10}$$

As before, the index of $\mathbf{u}_i$ is flexible and controlled with ‘lvCorr‘.

- We always condition on β_{0j}
- $\alpha_i = \mathbf{x}_i^{r,\top} \boldsymbol{\beta}^r + \mathbf{z}_i^{r,\top} \boldsymbol{\lambda}^r$ is controlled with ‘row.eff‘
- $\nu_{ij} = \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{z}_i^\top \boldsymbol{\lambda}_j$ is controlled with ‘formula‘

Model specification

We can only completely eliminate an effect from the ordination if the model is specified correctly:

- Consider non-linear effects of variables (e.g., quadratic)
- Species-specific random effects are more flexible than row-specific effects
- Random if we consider a confounder “nuisance”
- Fixed if we are interested in full-rank estimation of an effect

Model specification

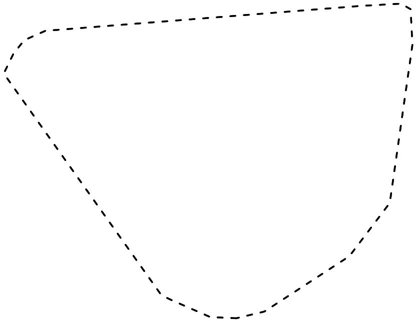
We can only completely eliminate an effect from the ordination if the model is specified correctly:

- Consider non-linear effects of variables (e.g., quadratic)
- Species-specific random effects are more flexible than row-specific effects
- Random if we consider a confounder "nuisance"
- Fixed if we are interested in full-rank estimation of an effect

Note: random effects in `formula` assume species independence, in `randomB` they do not.

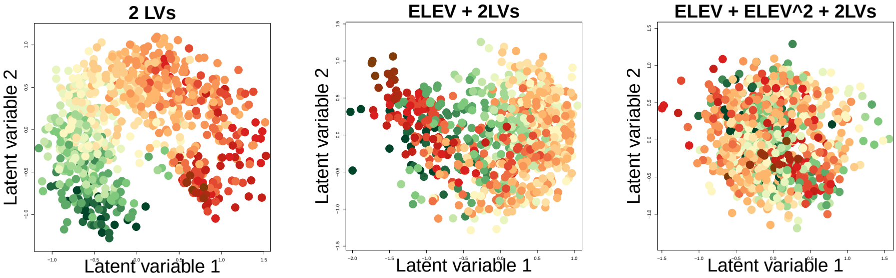
Example: alpine plants in Switzerland

- Data by D'amen et al. (2017)
- Occurrence of 175 species at 840 $4m^2$ plots
- Sampled on an elevation gradient



Example: elevation effect

```r
model4 <- gllvm(Y, num.lv = 2, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
model5 <- update(model4, X = X, formula=~ELEV)
model6 <- update(model4, X = X, formula=~ELEV + I(ELEV^2))
```



Example: elevation effect (2)

```
##                2 LVs   ELEV + 2LVs  ELEV + ELEV^2 + 2LVs
## ELEV    0.8266222  0.0008194027          0.0003363343
## ELEV^2 0.6612104 -0.7311265317          0.0001051139
```

- Without conditioning the ordination reflects the elevation gradient
- While conditioning on the linear term, the ordination still approximately reflects a quadratic effect of elevation
- Conditioning on both, the elevation effect is filtered from the ordination

Example: variation explained

- The variation explained by the unconstrained ordination is 390.26
- The variation explained by the first residual ordination is 400.16
- The variation explained by the second residual ordination is 270.91

```
VP(model4, group = c(1,1))
```

```
## Effect Mean.explained.varia
## LV1/LV2 100.0%
```

```
VP(model5, group = c(1,2,2))
```

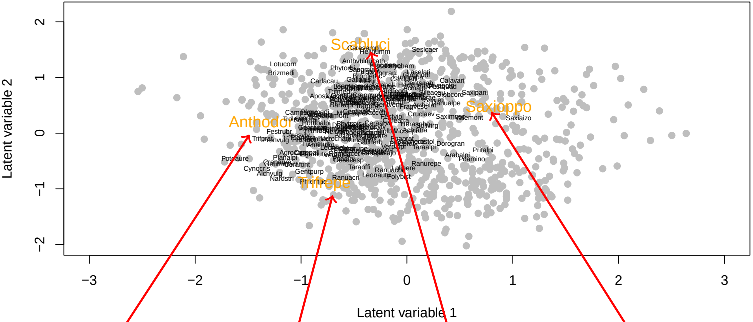
```
## Effect Mean.explained.varia
## ELEV 45.3%
## LV1/LV2 54.7%
```

```
VP(model6, group = c(1,1,2,2))
```

```
## Effect Mean.explained.varian
## ELEV/I(ELEV^2) 81.2%
## LV1/LV2 18.8%
```

There is more going on than just elevation! But no strong correlation with any observed covariates; Moisture, degree days above zero, slope, solar radiation, topography index.

Example: inferring the new gradients



Anthoxanthum odoratum

Trifolium repens

Scabiosa lucida

Saxifraga oppositifolia

The gradient of interest

Do you want to condition on ordination here?

- Not if you want to see the effect of elevation in the ordination
- But if you want to explore or find “alternative” gradients

Neither model here is right or wrong, it depends on your goal.

What we have to keep in mind: omitted variable bias.

Partial ordination

So far, we filtered unconstrained ordinations with covariates (discrete, or continuous).
When we combine these concepts with constrained or concurrent methods, we get a **partial** ordination.

Combining multiple concepts: taking information out of the ordination, **and** specifying what information we want inside the ordination.

I.e., separating drivers of community composition.

Partial constrained ordination

We again take our model from before:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \beta_j \tag{11}$$

Note, that the grey term can be represented as a constrained ordination. We take some of those covariates, and put them into an ordination:

Partial constrained ordination

We again take our model from before:

$$\eta_{ij} = \beta_{0j} + \boxed{\mathbf{x}_i^\top \boldsymbol{\beta}_j} \tag{11}$$

Note, that the grey term can be represented as a constrained ordination. We take some of those covariates, and put them into an ordination:

$$\eta_{ij} = \beta_{0j} + \boxed{\mathbf{x}_i^\top \boldsymbol{\beta}_j} + \boxed{\mathbf{x}_i^{lv,\top} \mathbf{B}} \boxed{\boldsymbol{\gamma}_j} \tag{12}$$

Partial constrained ordination

We again take our model from before:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \beta_j \tag{11}$$

Note, that the grey term can be represented as a constrained ordination. We take some of those covariates, and put them into an ordination:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \beta_j + \mathbf{x}_i^{lv,\top} \mathbf{B} \gamma_j \tag{12}$$

Now, we are conditioning our constrained ordination on $\mathbf{x}_i^\top \beta_j$

Hybrid ordination

Our next step could be to add residual latent variables:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \beta_j + \mathbf{x}_i^\top \mathbf{B} \gamma_j + \mathbf{u}_i^\top \boldsymbol{\gamma}_j \tag{13}$$

- Combined constrained and unconstrained ordination is **hybrid** ordination
- It is related to conditioning; we condition the unconstrained ordination on the constrained ordination
- In essence, we reduce the parameters of the conditioning part for when we have sparse data

Hybrid ordination: spillover

The "residual" latent variables may absorb covariate effects (from the constrained ordination)

There are K covariates and we have d constrained dimensions.

The remaining $K - d$ are incorporated into the residual ordination.

So, inference is more challenging, but it does improve the model

(*and can change the constrained ordination*).

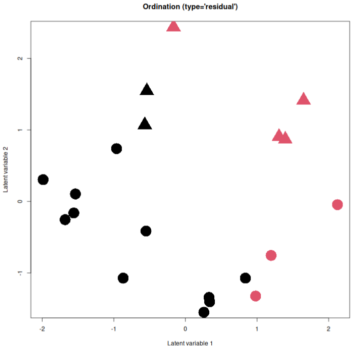
Example: Dutch Dune data

Achimill	Agrostol	Airaprae	Alopgeni	Anthodor	Bellpere	Bromhord	Chena
1	0	0	0	0	0	0	
3	0	0	2	0	3	4	
0	4	0	7	0	2	0	
0	8	0	2	0	2	3	
2	0	0	0	4	2	2	

- A classic dataset, originally by Jongman et al. (1995)
- Ordinal classes for 30 plant species at 20 sites
- 5 covariates; A1, Moisture (5 groups), Management (4 groups), Use (3 groups), Manure (3 groups

Example: unconstrained ordination (from this morning)

```
model7 <- gllvm(Y, num.lv = 2, family = "ordinal")
gllvm::ordiplot(model7, symbols = TRUE,
                s.colors = model.matrix(~0+., dune.env)[,5]+1,
                pch = model.matrix(~0+., dune.env)[,7]+16, s.cex = 4)
```

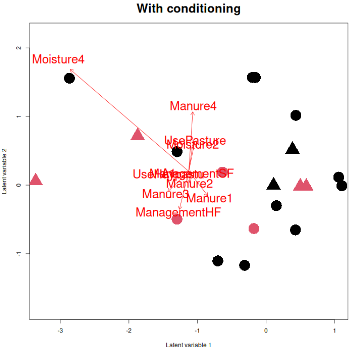


- Color by Moisture 5
- Shape by Management NM

What happens when we condition on Moisture 5 and Management NM?

Example: partial constrained ordination

```
X = model.matrix(~., dune.env)[, -1]
model8 <- gllvm(Y, X, formula = ~Moisture5 + ManagementNM,
               lv.formula = ~A1 + Moisture2 + Moisture4 + ManagementHF + ManagementSF + UseHaypastu + UsePast
Manure1 + Manure2 + Manure3 + Manure4,
               num.RR = 2, family = "ordinal", randomB = "LV", n.init = 3)
```

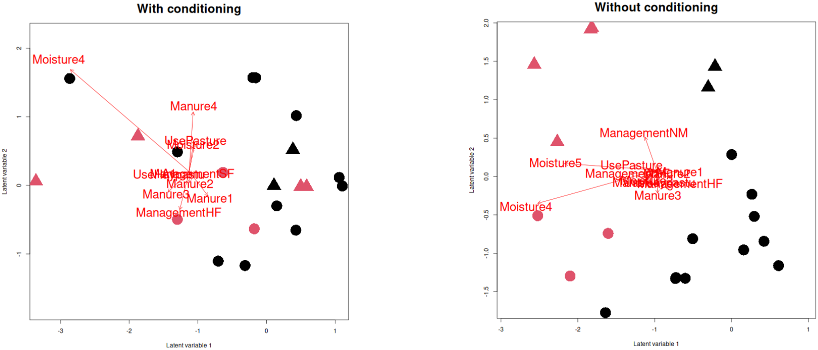


- Color by Moisture 5
- Shape by Management NM

Now, the arrow for Moisture 5 and Management NM have disappeared, and the groups have been pulled together.

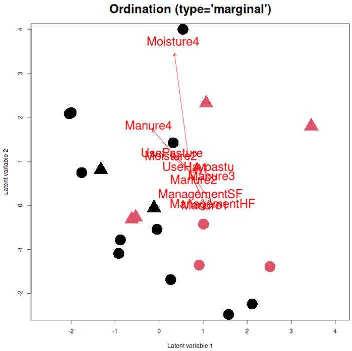
Example: partial constrained ordination

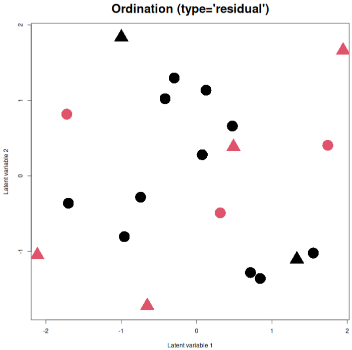
```r
X = model.matrix(~., dune.env)[, -1]
model8 <- gllvm(Y, X, formula = ~Moisture5 + ManagementNM,
            lv.formula = ~A1 + Moisture2 + Moisture4 + ManagementHF + ManagementSF + UseHaypastu + UsePast
Manure1 + Manure2 + Manure3 + Manure4,
            num.RR = 2, family = "ordinal", randomB = "LV", n.init = 3)
```



Example: hybrid ordination

```
model10 <- update(model8, num.lv = 2, jitter.var = 0.1, seed = 936, n.init = 1)
```





The residual ordination contains spillover from the covariate effects. How do we interpret it?

Example: hybrid ordination VP

```
VP(model10, group=c(1,2,3,3,3,3,3,3,3,3,3,3,3,4,4), groupnames = c("Moisture5", "ManagementNM", "CLV", "LV"))
```

```
## Effect       Mean.explained.variance
## Moisture5    38.0%
## ManagementNM 37.6%
## CLV          14.7%
## LV            9.8%
```

It does give us an impression how much residual information there still is. A frame of reference for another constrained LV:

```
model11 <- update(model10, num.RR = 3, n.init = 100, n.init.max = 10, seed = 1)
VP(model11, group=c(1,2,3,3,3,3,3,3,3,3,3,3,3,4,4), groupnames = c("Moisture5", "ManagementNM", "CLV", "LV"))
```

```
## Effect       Mean.explained.variance
## Moisture5    40.8%
## ManagementNM 33.2%
## CLV          16.7%
## LV            9.3%
```

Mis(specification)

There is a lot to choose in constructing your ordination:

- Effects inside or outside
- Fixed effects or random effects
- Constrained, concurrent, unconstrained or a combination

We have some tools to assist our decision:

- Information criteria
- Variance partitioning
- Predictive performance

Mis(specification)

We need to go through a careful procedure to ensure we have the "right" or "best" ordination.

- Usually it will capture the dominant gradients
- But what happens after that?
- If we specify the model incorrectly, our conclusions may be wrong or incomplete
- Inflated standard errors, biased parameter estimates, and so on.

Tomorrow, Audun will talk about a workflow for finding your model.

The "true" model

Be aware:

- Yes, if your mdoel is specified incorrectly there may be issues
- Be careful not to get caught in the "best model trap"
- Your expectation for "the true model" should lead the model formulation, not blind use of some tool
- Is it (ecologically) realistic to have covariates outside and inside the ordination?
- Conditioning is a powerful tool when the ordination is our primary inference vehicle
- Also when we have (too) sparse data (and many effect to estimate)

With great power..