Dance with statistics – quick, quick, slow

with simple and complex figures

Cajo J.F. ter Braak, 17 October, Trondheim symposium

WAGENINGEN UR
For quality of life

Cajo.terbraak@wur.nl

My dance with statistics – quick, quick, slow

with simple and complex models and algorithms

Cajo J.F. ter Braak, 17 October, Trondheim symposium
with Paul Eilers & Rodríquez-Álvarez (P-spline work)

WAGENINGEN UR
For quality of life

Cajo.terbraak@wur.nl

Simple models – work horses of applied stats

- ANOVA (mean) models
- simple linear regression
- multiple regression
- generalized linear models (1972, but use from late ’80s)

More structure

- Additive random terms → mixed / hierarchical models
- Bilinear (random) terms→ Principal components, Factor

Simple models – simple/complex algorithms

- ANOVA (mean) models
- multiple regression
- generalized linear models
- mixed models
- bilinear models

- means and sums of squares (balanced design/Genstat, ...)
- sweeps, QR, ...
- iteratively reweights least-squares (Glim, ...)
- nonlinear optimization for variance components*
- criss-cross/reciprocal linear regression, singular value decomposition

*not so simple.... e.g. in large data sets

Gaussian (constrained ordination) model

- model for unimodal response w.r.t. manifest variables
- latent variable model for unimodal multivariate response

Unimodal model in ecology

Theory:

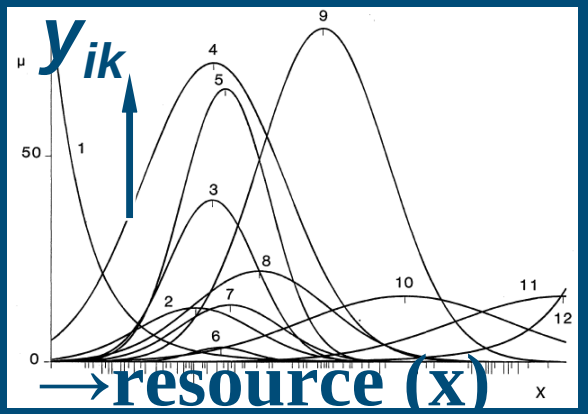
"a plant species does not grow when it is either too wet or too dry"

Liebig's law: a species requires a minimum amount of a resource (e.g. N): agriculture

Shelford's law of tolerance (1919): but also does not tolerate more than a certain maximum

Ecological niche: region where species actually grows

Niches vary among species



preference model in psychology

DATA:

CtB, Ecology 1986

Spider abundance $\mathbf{Y}_{n\times m}$ and environment $\mathbf{Z}_{n\times p}$

- n = 28 sites (pitfalls), m = 12 spider species ,
  p = 6 environmental variables

species

$\mathbf{Y}^T=$

	Site numbers																											
	15	19	20	16	17	18	2	8	21	5	6	14	4	7	13	3	1	9	12	25	11	10	28	23	22	27	24	26
Species																												
Arct lute										1	2	1	1	3	1	1				1				1				
Pard lugu	2	3	3	2	1	2	1	7	4	1		1	1	1	1	1				2								
Zora spin	1	1	1	2	1		3	1	1	4	5	5	5	4	4	1	2			2								
Pard nigr		1		1			3	1		9	5	3	5	9	7	4	3	1	1	2								
Pard pull							6	1	1	8	4	8	9	9	8	6	6	1	2		1							
Aulo albi							5	2		3	2	2	4	4	4	3	2			1	1							
Troc terr	5	4	4	5	4	5	8	5	4	9	7	9	9	9	9	8	7	1	3	4	2	1	1	1	1			1
Alop cune		1	1	1		1	1	3	1	4	2	1	2	2	6	4	3	1	3	1	1							
Pard mont							1	1	1	1	3	3	2	5	4	5	7	5	9	3	9	4	2	2	1	1	1	
Alop acce										1			1	1	1	3	5	1	4	3	3	1	3	4	2	5	3	1
Alop fabr															1	1				3	1	1	3	3	4	3	4	2
Arct peri																							1	2	1	2	2	4
Environmental variable																												
Water Content	9	7	8	8	9	8	8	6	7	8	9	8	6	8	9	6	5	5	5	3	4	4	0	0	1	0	2	0
Bare Sand	0	0	0	0	0	0	0	0	0	0	5	0	0	0	3	0	0	0	0	7	0	8	7	6	7	5	7	9
Cover Moss	1	3	1	1	1	0	2	2	1	0	5	4	5	1	1	5	7	9	8	2	9	7	8	9	9	8	9	4
Light Refl	1	0	0	0	2	2	3	1	0	5	1	2	6	5	7	8	8	7	8	5	8	8	8	9	8	8	9	9
Fallen Twigs	9	9	9	9	9	9	3	9	9	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Cover Herbs	5	2	0	0	5	5	9	6	2	9	6	9	9	9	9	9	9	6	8	8	7	5	6	6	0	6	5	2

sites

env. var.

$\mathbf{Z}^T=$

row/column order from CCA

discretized data

WAGENINGEN UR

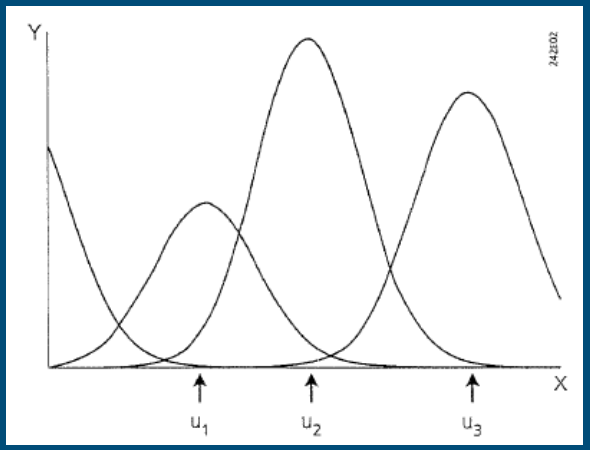
Gaussian ordination

- Gaussian response model with (latent) predictor **x**

$y_{ik} \sim \text{Poi}(\mu_{ik})$, for example

$Ey_{ik} = c_k \exp\{-½ (x_i - u_k)^2 / t_k \}$

Gaussian curves for 4 species



- u_k optimum of a species
- c_k maximum
- t_k tolerance, niche width
- x_i manifest/latent or combination of manifest environmental variables **x**=**Zb**

Unimodal model

For manifest x_i: glm y ~ pol(x,2),... t_k >0

For latent x_i much harder to solve, particularly multi-d. **x**

Ecologists devised

A simple alternative algorithm

Method of weighted averaging (WA) to obtain ‘ optima’

- preference or indicator value of a species wrt to a physical gradient, *e.g.* moisture
  - WA of moisture values at sites where species occurs
- estimate of moisture value at a site ‘ latent values’
  - WA of indicator values of species that occur there

Gause (1930), Ellenberg (1948), Whittaker (1948)

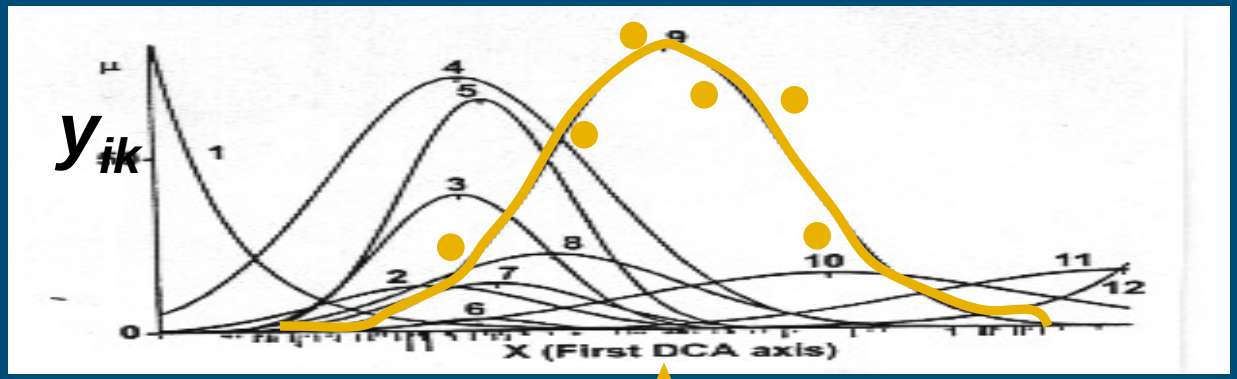
Idea: iterate to replace a manifest gradient by the best latent one

→ Reciprocal averaging (Hill, 1973,74) == Correspondence Analysis

Recall: Criss-cross/reciprocal regression (Gabriel, Biometrika, 1998) , svd

Instead of glm: y ~pol(x,2), ecologists proposed

Weigthed Averaging (WA) to estimate optimum



$i \rightarrow$ sites

$k \rightarrow$ species

$y_{ik} \geq 0$

$\rightarrow$ **X**

- WA: $u_k = \sum_i y_{ik} x_i / \sum_i y_{ik}$

For presence/absence data y_{ik} is 1 or 0.

Example: species *k* found at sites with moisture values x = 20, 30 and 40 and nowhere else, then

u_k = (1*20+1*30+1*40)/(1+1+1)= 90/3 = 30

Not the best, or even a bad method, but....

Disregard s absences (zero inflated model...)

WAGENINGEN UR

And the reverse:

Weigthed Averaging (WA) to estimate x_i

- WA: $x_i = \sum_k y_{ik} u_k / \sum_k y_{ik}$

with $\{u_k\}$ known optima/ "indicator values"

$i \rightarrow$ sites

$k \rightarrow$ species

$y_{ik} \geq 0$

For presence/absence data y_{ik} is 1 or 0.

Example:

Let site i contain the species with optima 75, 80, 85 wrt moisture,

then its moisture value is estimated as

x_i = (1*75+1*80+1*85)/(1+1+1)=80 (the average of the optima of species present)

Gause (1930), Ellenberg (1948), Whittaker (1948)

Weighted averaging → correspondence analysis

- Idea: iterate to replace a manifest gradient by the best latent one

→ Correspondence Analysis == Reciprocal averaging

Recall: Criss-cross/reciprocal regression (Gabriel, Biometrika, 1998) , svd

With constraints on the rows (sites) **x** = **Zb**:

→Canonical correspondence analysis (CCA)

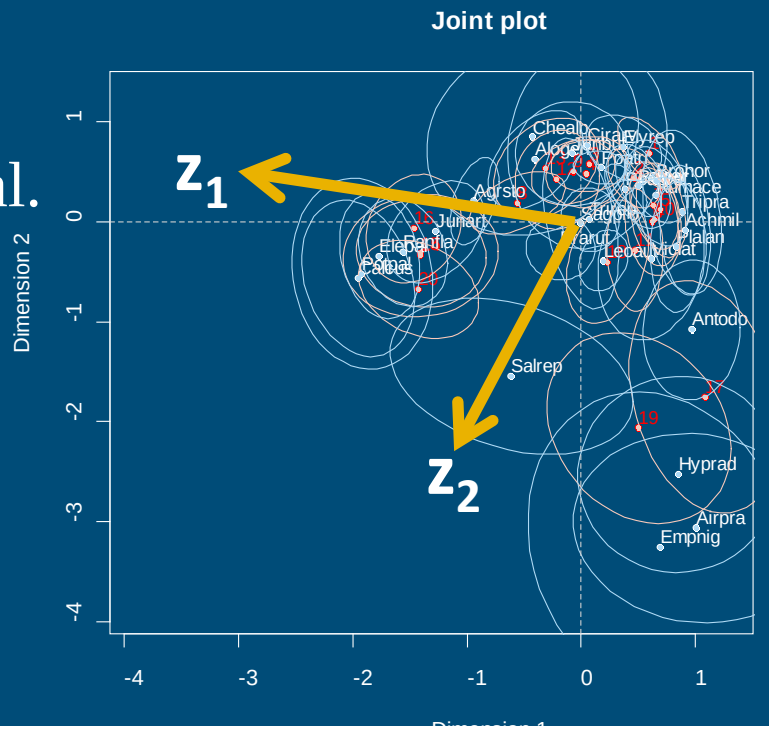
Algorithm: extra regression step in cross-cross or via svd

M.O. Hill (1973/74), ter Braak (1986)

CCA in R packages anacor, vegan, ade4

#e.g.:
library(anacor) # Jan de Leeuw & Patrick Mair
# CCA
res = anacor(dune, row.covariates = dune.env)
plot(res)

library(vegan) # Jari Oksanen et al.
cca(Y~ Z)
# also with a formula interface
# e.g factors A, B and C
cca(Y~ A*B + Condition(C))



and... in Canoco 5 (www.canoco.com)

WAGENINGEN UR

Paradox

- How can a (bi)linear method (essentially an svd) fit unimodal response surfaces???
- It can be understood from relationship of CCA :
  - with canonical variate analysis/discriminant analysis,
  - the equi-tolerance Gaussian model and its relation to the generalized bi-linear (mixed) model

$$\log(E y_{ik}) = \alpha_k + \beta_k^T x_i + \gamma_i \quad \textit{with } x_i, \alpha_k, \beta_k \textit{ and } \gamma_i \text{ random}$$

PeerJ. 2013; 1: e95. PMCID: PMC3709111
Published online Jul 9, 2013. doi: 10.7717/peerj.95

Generalized linear mixed models can detect unimodal species-environment relationships

Tahira Jamil[1,2] and Cajo J.F. ter Braak[✉1]

WAGENINGEN UR
For quality of life

From equi-tolerance Gaussian to generalized bilinear

- Gaussian ordination with $t_k = t = 1$:

$$Ey_{ik} = c_k \exp(-(x_i - u_k)^T (x_i - u_k))$$

distance model

with x=Zb (ideal point discriminant analysis)

Can be rewritten [**work out the square** in 1-d] as:

- Goodman's RC model

$$Ey_{ik} = r_i c_k \exp(u_k^T x_i)$$

$$\log(Ey_{ik}) = \alpha_k + u_k^T x_i + \gamma_i$$

generalized bilinear (mixed) model

- (Canonical) Correspondence Analysis is approx. to this model (ter Braak 1988, Takane 1987)

Goodman showed this for small λ, ter Braak for large λ

Exact ML methods (treating x=Zb without error)

- Yee (2004) – ML Canonical Gaussian ordination and
- Yee (2006) Constrained additive ordination, a spline response model / GAM w.r.t. latent x= Zb!!

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

Ecology,

A NEW TECHNIQUE FOR MAXIMUM-LIKELIHOOD CANONICAL GAUSSIAN ORDINATION

THOMAS W. YEE[1]

CONSTRAINED ADDITIVE ORDINATION

THOMAS W. YEE[1]

Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

- both in R {VGAM}, nevertheless, little usage...
- problems with noisy and big data sets; inherent instability when used unconstrained ($Z = I_n$); new P-spline attempts by Schnabel/Eilers et al 2014

Why use an approximate method while exact exists?

- **The approximate methods (CA/CCA and related) are good enough for what ecologists want (an overview)**

.....but I still wish to promote ....

Stable, quick methods needed for simple related models, e.g. for

the generalized bilinear mixed model $\log(E y_{ik}) = \alpha_k + \beta_k^T x_i + \gamma_i$

with x_i **random (and** γ_i **,** α_k **and** β_k**)** ... $\beta_k = u_k$

(see Hui ..& Warton, MEE, 2014, who use a stochastic EM approach).

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014

doi: 10.1111/2041-210X.12236

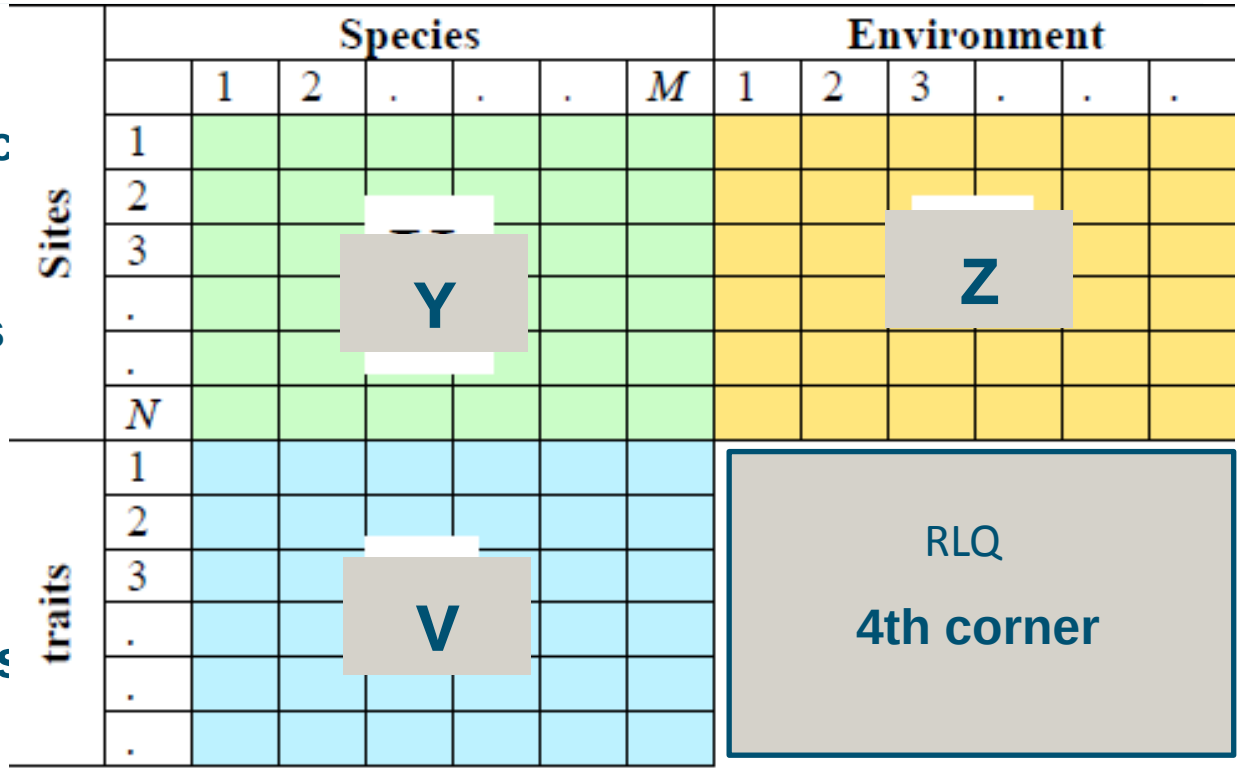
SPECIAL FEATURE PAPER: NEW OPPORTUNITIES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Model-based approaches to unconstrained ordination

Francis K.C. Hui [1,2,*], Sara Taskinen[3], Shirley Pledger[4], Scott D. Foster[2] and David I. Warton[1,5]

Species, environment and trait data: CA →RLQ

- Species data Y_{ij}
  - Presence-absenc
  - Abundance
  - species biomass
- Environmental Z
  for each site
- Species traits V
  for each species



- Test for association between V and Z complicated as they are measured on different units→ three joint null hypotheses
- Solution: use a sequential test (Goeman and Solari, 2010, ter Braak et al 2012)

Correspondence Analysis →trait-environment relations

CA with constraints on the rows of Y (sites): **x** = **Zb**:

→Canonical correspondence analysis (CCA)

With also constraints on columns (species) **u** = **Vc**

→doubly constrained CCA, or after neglecting correlations **Z** and **V**

→RLQ, which maximises the weighted correlation **uV'YZb**

$\mathbf{Z}=\mathbf{z}_1=\mathbf{x}_1$ and $\mathbf{V} = \mathbf{u}_1$ →4th corner statistics (weighted correlations)

Dolédec et al 1996, Legendre et al 1997, Dray et al 2014

RLQ factorial diagram (Dune meadow data)

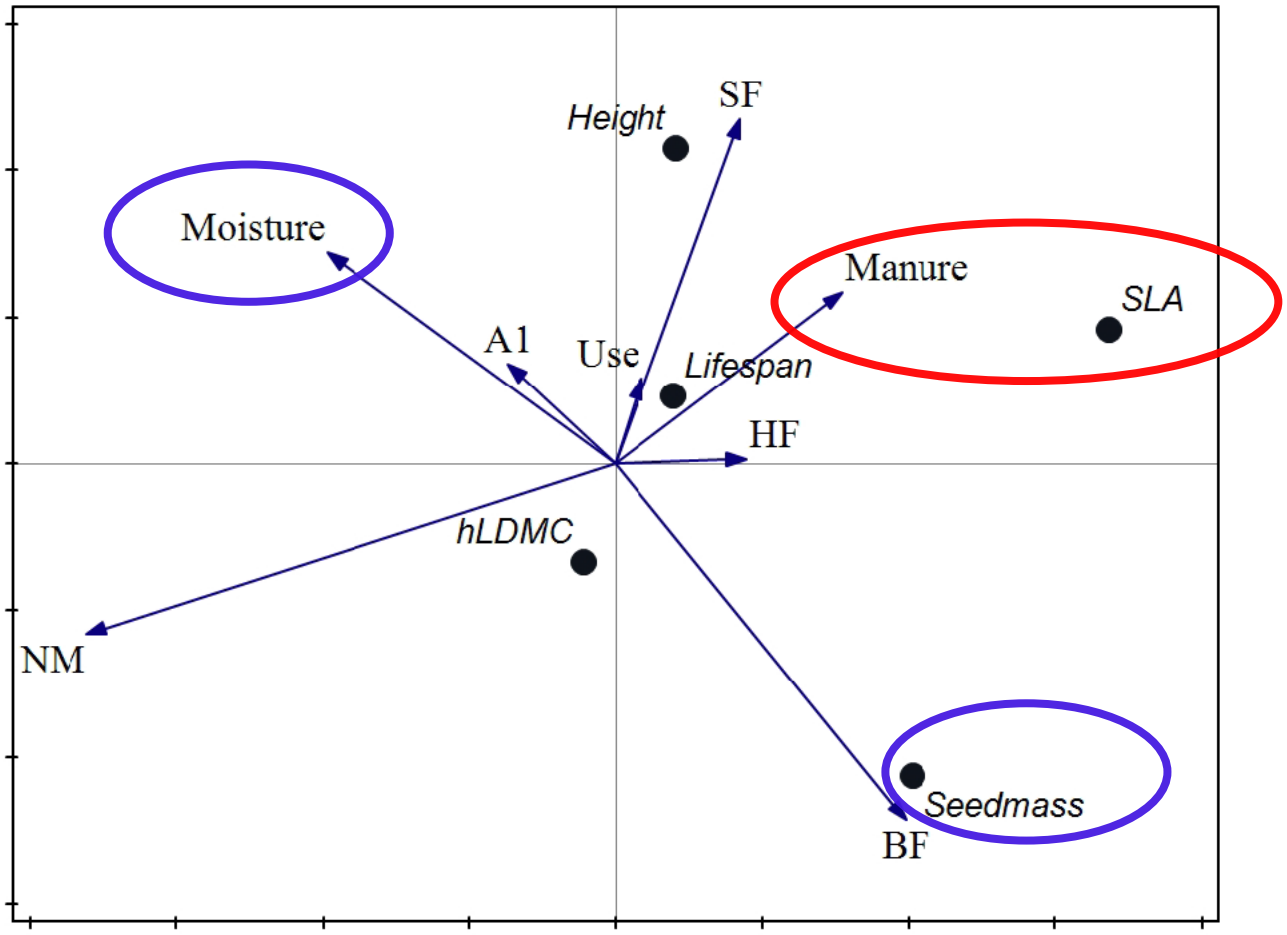


Fig. 4. RLQ biplot of the Dune Meadow data explaining 90% of the variance in the fourth corner statistics.

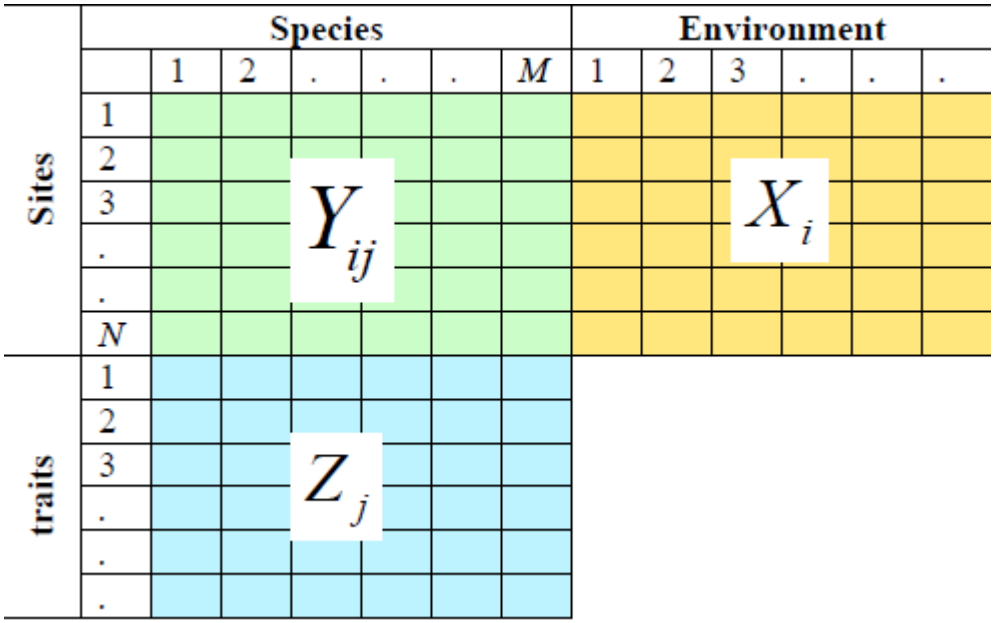
biplot

From 4th corner / RLQ → glmm's

Jamil et al. (2013) treat the problem (without dimension reduction) as a GLMM

- link(Ey_{ij})= $\alpha_k + \beta_j^T x_i + \gamma_i$

with γ_i , α_j and β_j random terms and x_i given predictor(s).



Traits included by the submodel:

$\beta_j \sim N(b_0 + b_1^T z_j, \Sigma^2_\beta)$

Motivated by link between glmm and the equi-tolerance Gaussian....

Fitted by lme4

Example result (Dune meadow data)

Table S2. The final GLMM model (after model selection) for the Dune Meadow data.

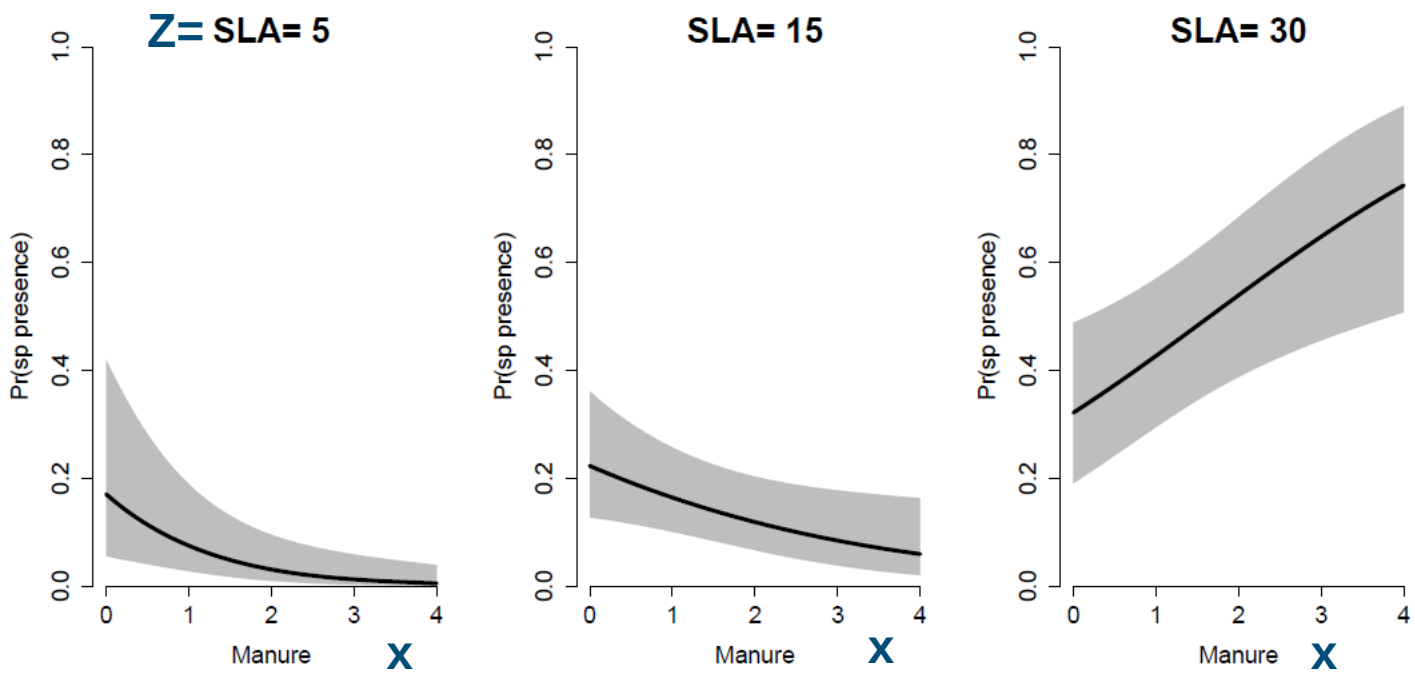
```
Generalized linear mixed model fit by the Laplace approximation
Formula: y ~ (1 | site) + (1 + Moist + Manure | sp) + Moist + Manure
+      SLA + Seedmass + Manure:SLA + Moist:Seedmass
   Data: Data
   AIC   BIC logLik deviance
 540.9 601.4 -256.4    512.9
Random effects:
 Groups Name        Variance Std.Dev. Corr
 sp     (Intercept) 1.43789  1.19912
        Moist       2.82409  1.68050  -0.231
        Manure      1.33610  1.15590   0.299  0.511
 site   (Intercept) 0.32543  0.57046
Number of obs: 560, groups: sp, 28; site, 20

Fixed effects:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.618291   0.313967  -5.154 2.55e-07 ***
Moist          -0.031705   0.380945  -0.083   0.9337
Manure          0.009563   0.299437   0.032   0.9745
SLA             1.276773   0.298861   4.272 1.94e-05 ***
Seedmass       -0.274162   0.270879  -1.012   0.3115
Manure:SLA      0.904098   0.259725   3.481   0.0005 ***
Moist:Seedmass -0.824813   0.327925  -2.515   0.0119 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model after selecting best environmental variables and then best traits

WAGENINGEN UR
For quality of life

Example result (Dune meadow data)



RLQ factorial diagram (Dune meadow data)

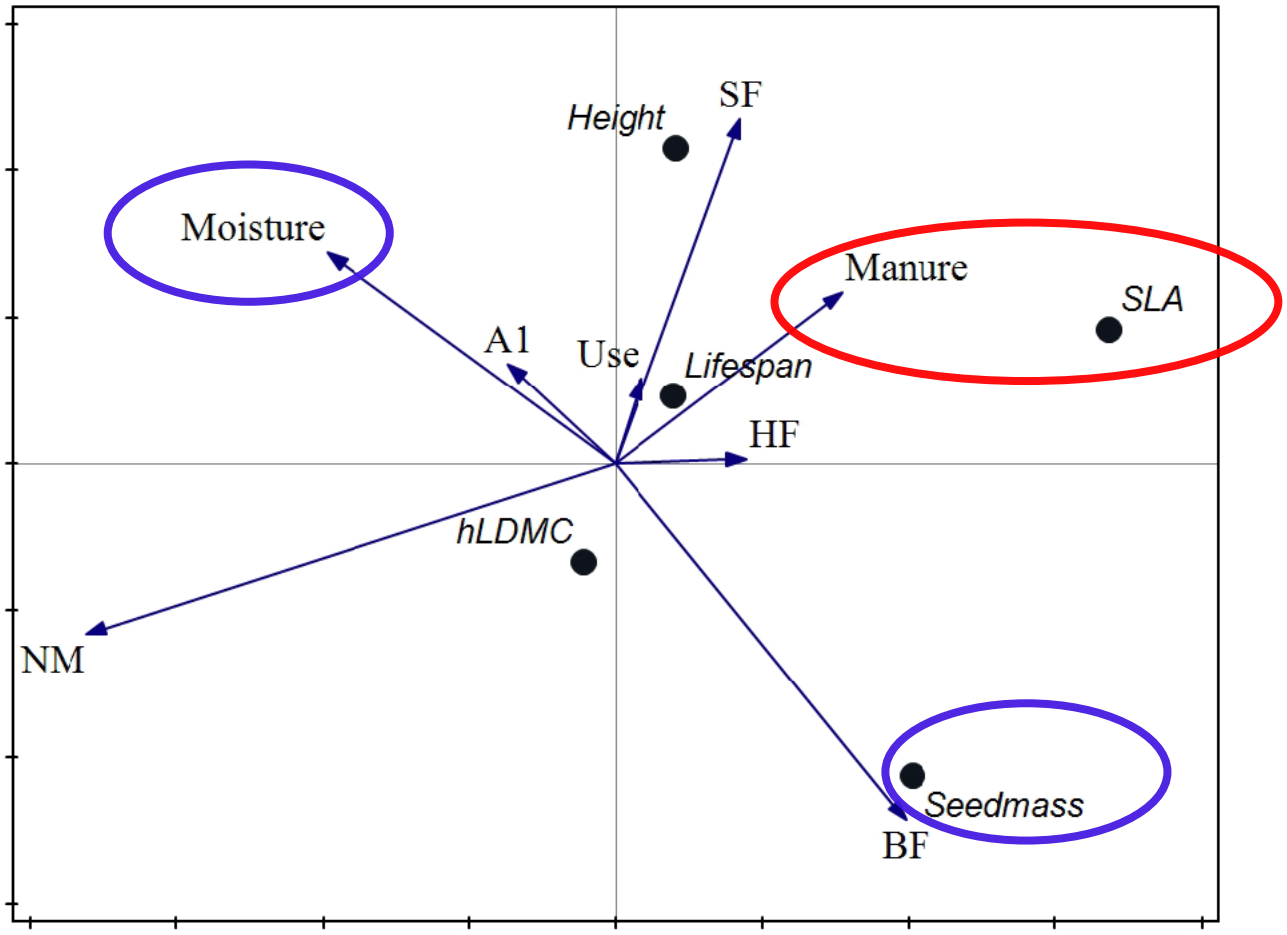


Fig. 4. RLQ biplot of the Dune Meadow data explaining 90% of the variance in the fourth corner statistics.

- Correlation based
- No model selection
- But many ecologist may prefer this diagram over the glmm graphs...

Trait-environment relationships

- Non-linear models: 2-d splines, but with proper error terms for the species component

From:

$\text{link}(E y_{ij}) = \alpha_k + \beta_j x_i + \gamma_i$ with $\beta_j = b_0 + b_1 z_j + \varepsilon_\beta$

To something like...

$\text{link}(E y_{ij}) = \alpha_k + f_j(x_i) + \gamma_i$ with $f_j(x_i) = g(x_i, z_j) + \varepsilon_j(x_i)$

Should nowadays by doable... INLA?

Bayesian P-splines/GAM (1)

Motivation / Wanted

- Automatic choice of penalty / smoothing parameter
- Average prediction uncertainty over uncertainty in the penalty
- No over-smoothing (as ML in GLMM !?) and stability
- Choice of reference curves
  - constant, straight line, (cf. rw1, rw2)
  - parabola (rw3?) for fitting densities/ niche models in ecology
- Large datasets... (preferable)
- Quick ... try avoid MCMC (BayesX), Gibbs? Try INLA..

INLA is like GLIM in the 1970-1990's
It unifies and is practical

Bayesian P-splines/GAM (2)

- INLA has already high rank approach: rw1, rw2, ...,
- But what about rw3... and large datasets? [inla.group]

- We wanted to explore the low rank approach P-splines (Eilers & Marx 1996)....

Remarks:

- Alternatively, O-splines (O'Sullivan 1986) can be used, closer to integrated squared derivatives and better extrapolation according to Wand & Ormerod (2008).
- Multidimensional P-splines:
  - Currie et al 2006 GLAMM, Rodríguez-Álvarez et al 2014 SAP (Separation of Anisotropic Penalties)

P-splines/GAM à la Eilers &Marx 1996

- B-splines of chosen degree (default 3, cubic, B_3)
- abundant number of **equi-spaced** knots
- y= Bα + ε
- link(Ey)= Bα (GLM)
- *r*th order difference penalty on the coefficients

$$Q = ||y - B\alpha||^2 + \lambda||D\alpha||^2 \qquad Q = \frac{||y - B\alpha||^2}{\sigma^2} + \frac{||D\alpha||^2}{\tau^2}$$

$$\lambda = \sigma^2/\tau^2$$

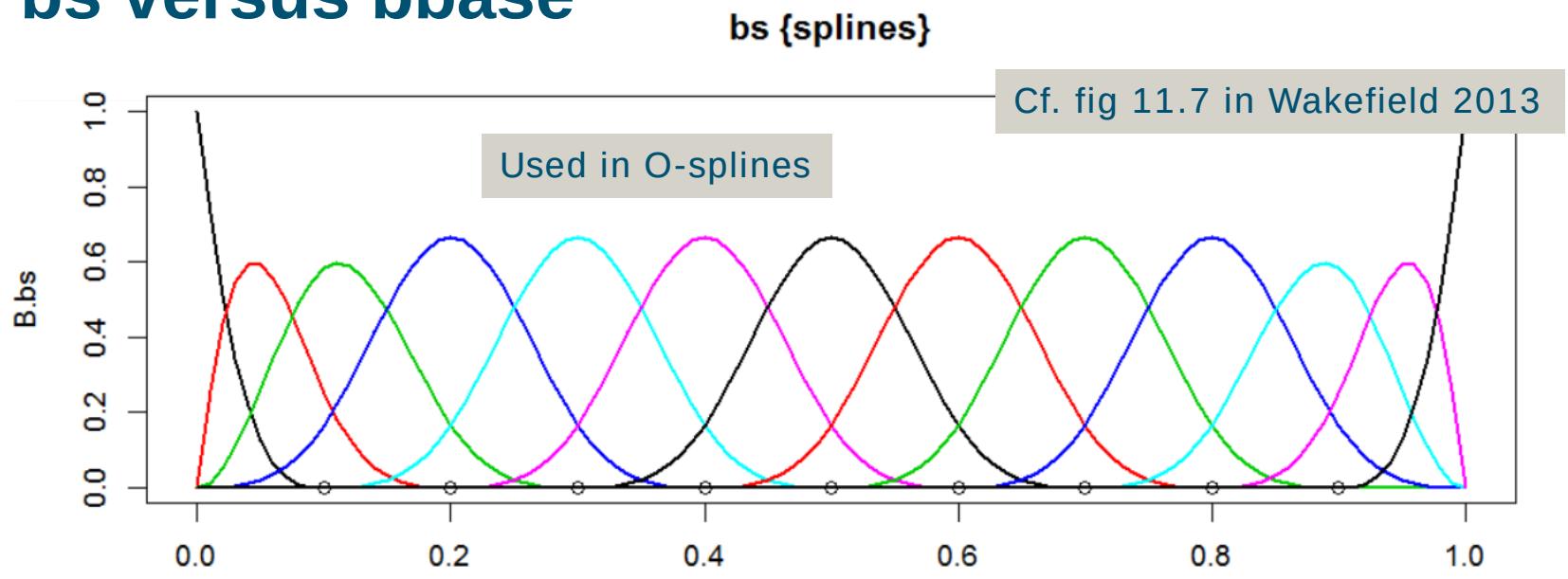
- not quite bs {splines} though ...

Penalty matrix = $D'D$

WAGENINGEN UR
For quality of life

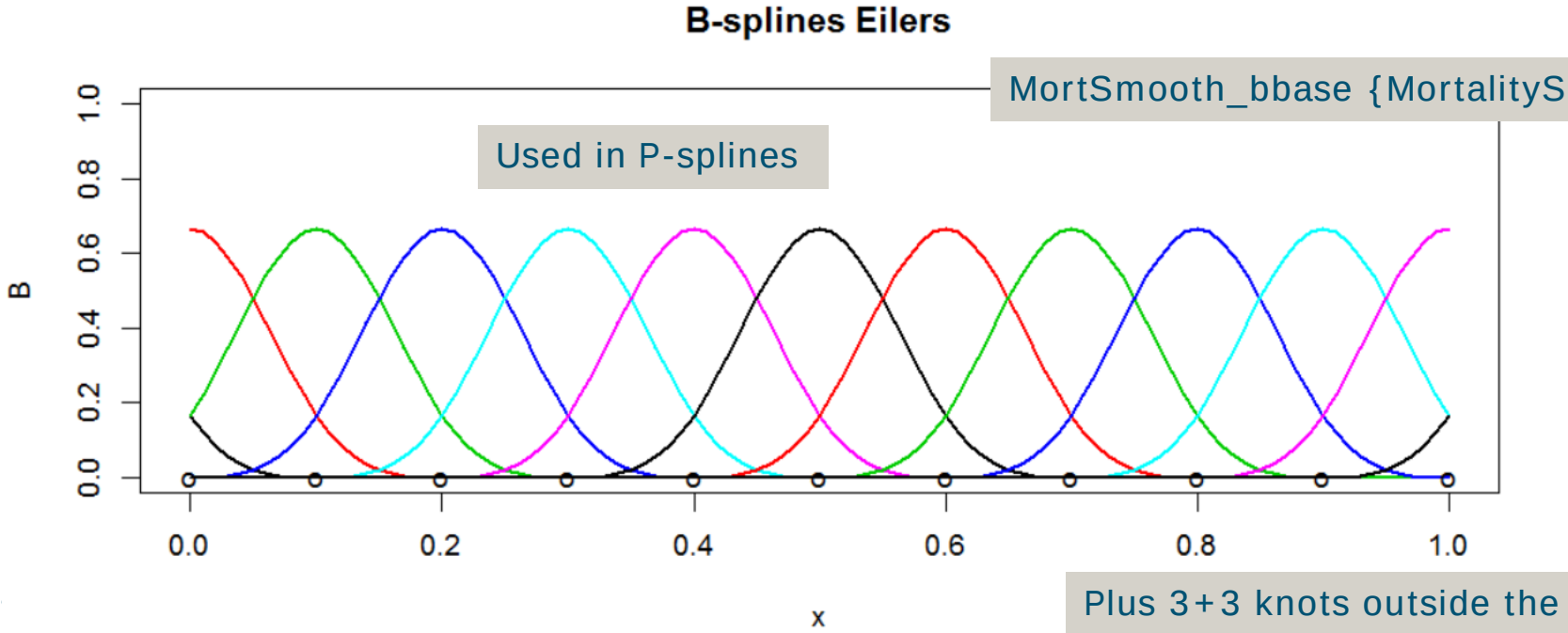
rw2 on regular grid: B_0 or B1 with 2th order diff. pen.

bs versus bbase



Cf. fig 11.7 in Wakefield 2013

Used in O-splines



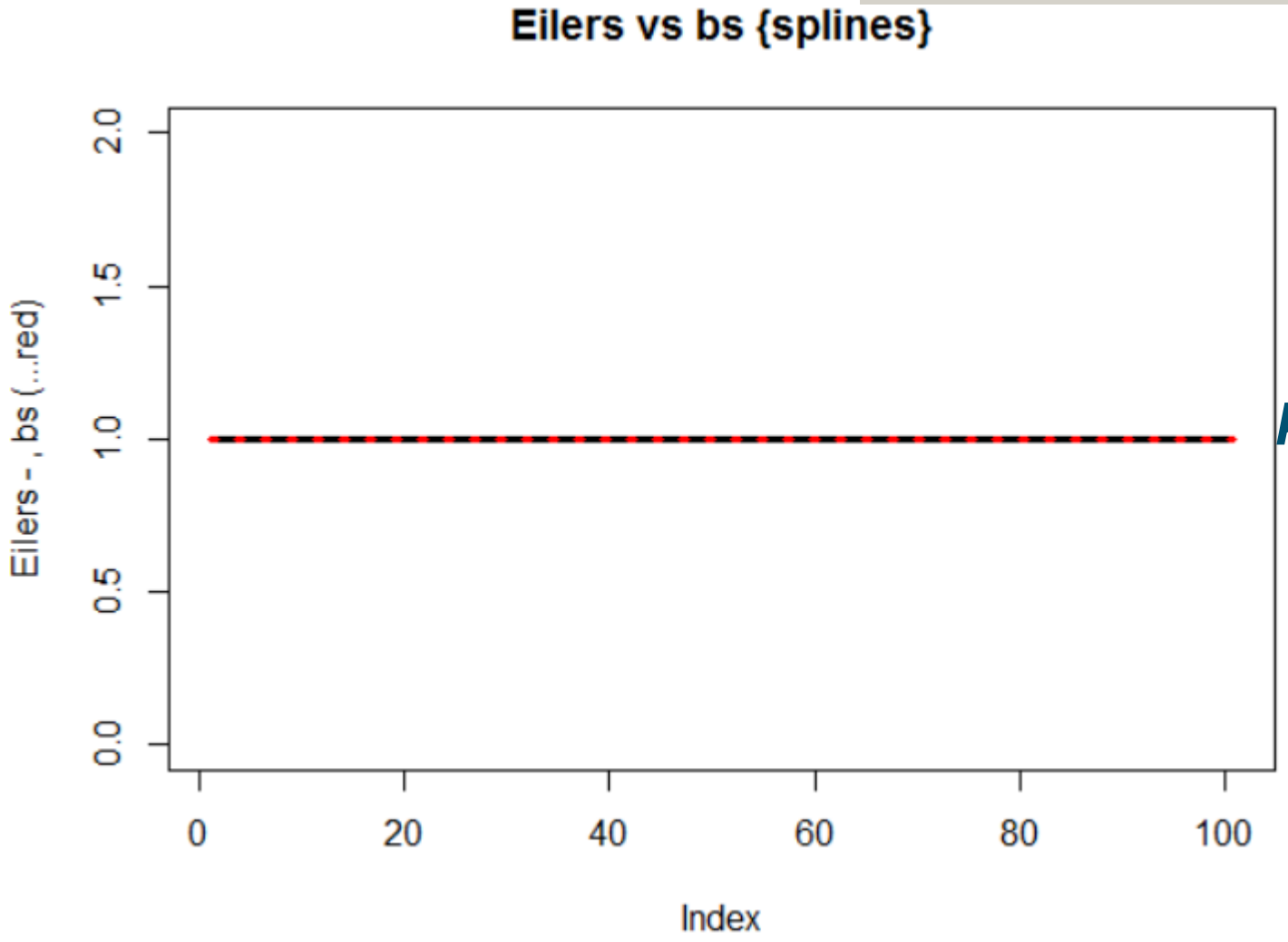
MortSmooth_bbase {MortalitySmooth}

Used in P-splines

Plus 3+3 knots outside the range

Effect on linear combination Bα with $\Delta^k \alpha = 0$

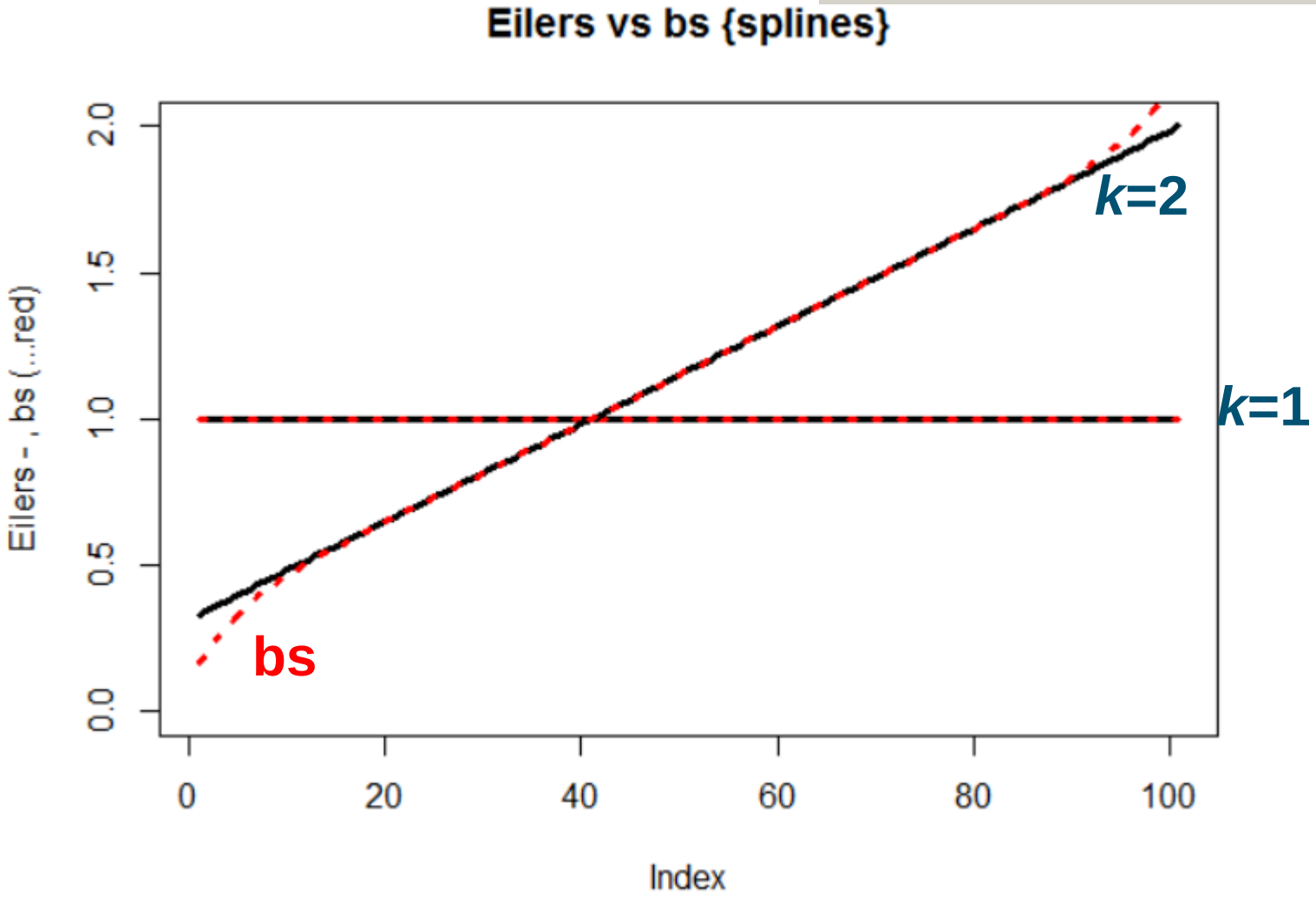
diff(alpha, diff = k) =0



k=1

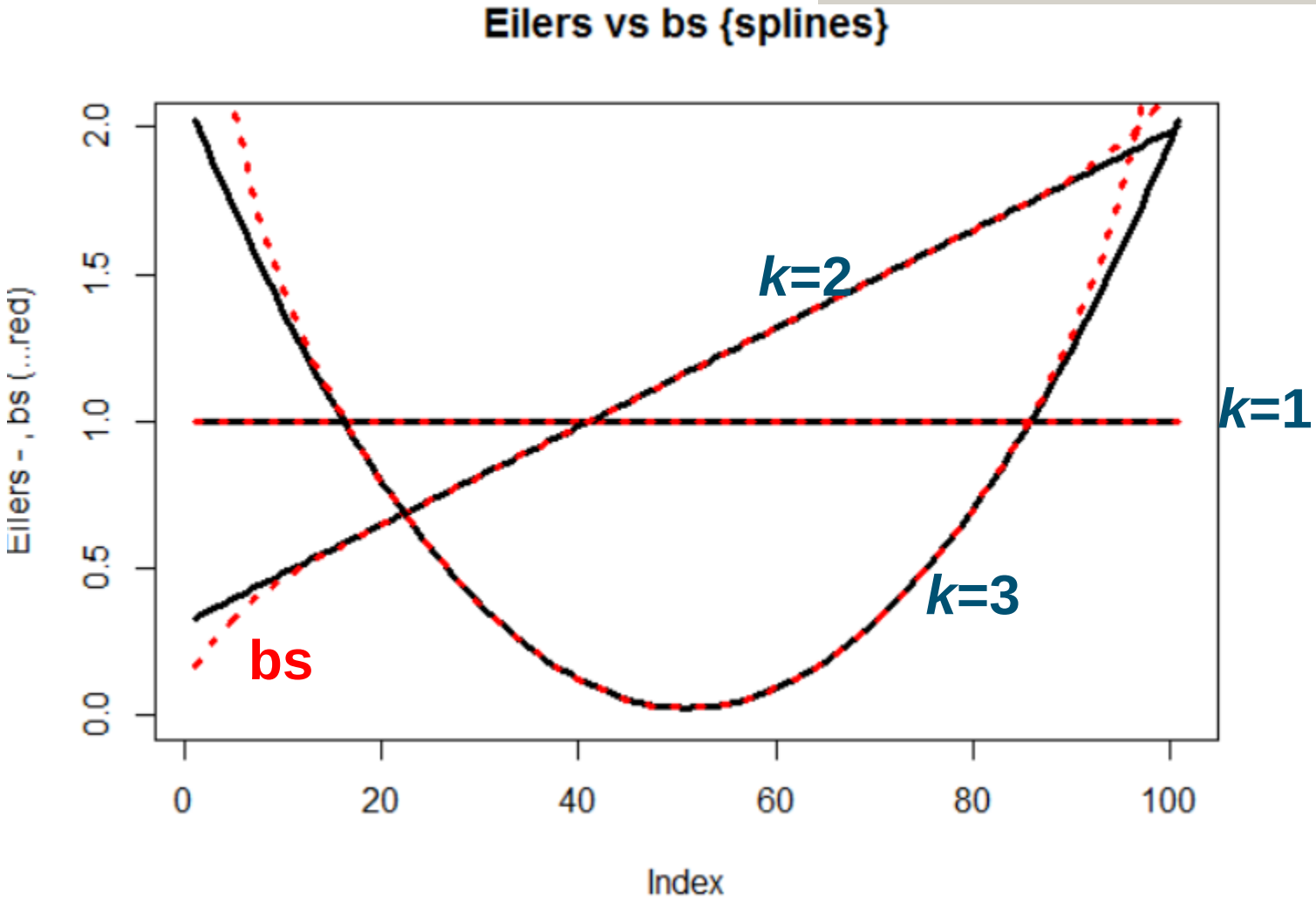
Effect on linear combination Bα with Δ^k α = 0

diff(alpha, diff = k) =0



Effect on linear combination Bα with $\Delta^k \alpha = 0$

diff(alpha, diff = k) =0



bs basis needs a slightly more complicated penalty matrix to give these reference lines; Wand & Ormerod (2008)

P-splines/GAM à la Eilers &Marx 1996

$$Q = \|y - B\alpha\|^2 + \lambda\|D\alpha\|^2 \qquad Q = \frac{\|y - B\alpha\|^2}{\sigma^2} + \frac{\|D\alpha\|^2}{\tau^2}$$

$$\lambda = \sigma^2/\tau^2$$

- **Penalty parameter λ chosen by**
  - **Crossvalidation, gcv, aic ....**
  - **Mixed models (marginal likelihood or REML)**
    - **Allows for fixed effects, other smooth terms and additional random effects**
    - **Inference conditional on the estimated variance components (i.e. on the hyper-parameters)**

Jones, (1996):
P-splines “on a par with ... various other approaches” ,
Speed less important than “developing better selectors”

Our Bayesian P-splines/GAM: how we learned..

- Rue Bilthoven/Utrecht course, January 2013 J. Illian July

(P-)splines are standard implemented via mixed model formulation software (SemiPar, mgcv) ...

- Qu: can INLA do Bayesian mixed models?

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon}$

$\boldsymbol{\beta}$ fixed, $\mathbf{a} \sim \mathbf{N}(\mathbf{0}, \sigma_a^2 \mathbf{G})$, $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{R})$, ☺ with R=I

standard **mixed model** with matrix input.
Could you point me how I can do the matrix version of the standard **mixed model** in INLA: Y = X beta + Zb+ e. with X and Z matrices with real numbers, ...
09/07/2013 by Cajo J.F. ter Braak - 5 posts by 3 authors - 132 views

Håvard Rue: yes, via A-matrix approach...

now also via z-model

WAGENINGEN UR
For quality of life

Our Bayesian P-splines/GAM

- First attempt then via mixed models
  - Done also by Wakefield (2013, Ch 11) Bayesian and Frequentist Regression Methods, Springer.
- This looses sparseness
- But INLA exploits precision matrices
  - allowed to be singular
- So, a quite natural A-matrix approach is:

$\mathbf{y} = \mathbf{Z}\mathbf{a} + \boldsymbol{\varepsilon}$ with $\mathbf{a} \sim \mathbf{N}(\mathbf{0}, \sigma_a^2 \mathbf{P}^-)$, $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ ☺

```
# Prepare basis and penalty matrix
B = bbase(x, xrange[1], xrange[2], nseg, degree)
D = diff(diag(ncol(B)), diff = diff.order)
P = t(D) %*% D
# specify the linear combination for predictions on a grid
x_grid = seq(xrange[1], xrange[2], length = ngrid)
B_grid = bbase(x_grid, xrange[1], xrange[2], nseg, degree)
# In A-matrix approach, the output is c(eta*, eta), so shift B_grid...
B_grid_plus =cBind(Matrix(0,nrow = nrow(B_grid), ncol = nrow(B)),B_grid)
# set up INLA call using the A-matrix approach
mod.P = inla(
    y ~ -1 + f(id.b, model="generic", Cmatrix = P, constr = F, hyper = hyperB),
    data = list(y = y, id.b = 1:ncol(B)),
    control.predictor = list(A = B, compute = TRUE),
    lincomb = inla.make.lincombs(Predictor = B_grid_plus))
Pred = mod.P$summary.lincomb.derived
```


For O-splines, change bbase and P

Our Bayesian P-splines/GAM

- **A-matrix approach**

$\mathbf{y} = \mathbf{Za} + \boldsymbol{\varepsilon}$ with $\mathbf{a} \sim \mathbf{N}(\mathbf{0}, \sigma_a^2 \mathbf{P}^-)$, $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I})$ ☺

In practice include fixed intercept, for numerical stability...

Extend to GLM:

$\text{link}(\mathbf{Ey}) = \mathbf{1} + \sum_j \mathbf{Z}_j \boldsymbol{a}_j$ with $\boldsymbol{a}_j \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\tau}_j \boldsymbol{P}_j)$ [precision matrix notation]

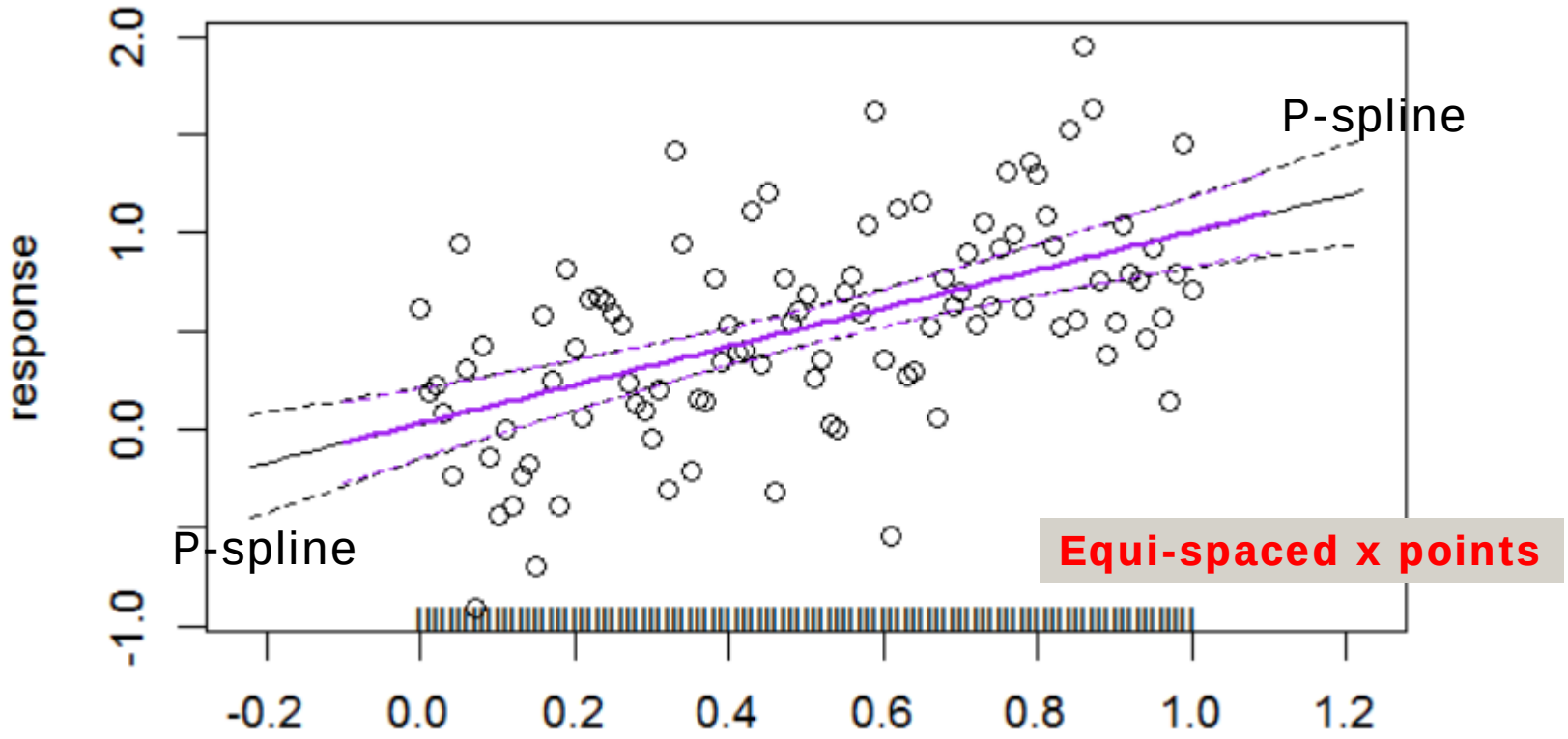
- **add any fixed and random effects**

BayesianPsplines library

- A-matrix approach gets quite complicated, so ...
- Wrapper for this approach with:
  - Component -plus- residuals plots
    - Based on component predictions on a grid
- the trick we use
  - B_grid_all is block matrix of component B_grid's
    - with a intercept added for easier interpretation

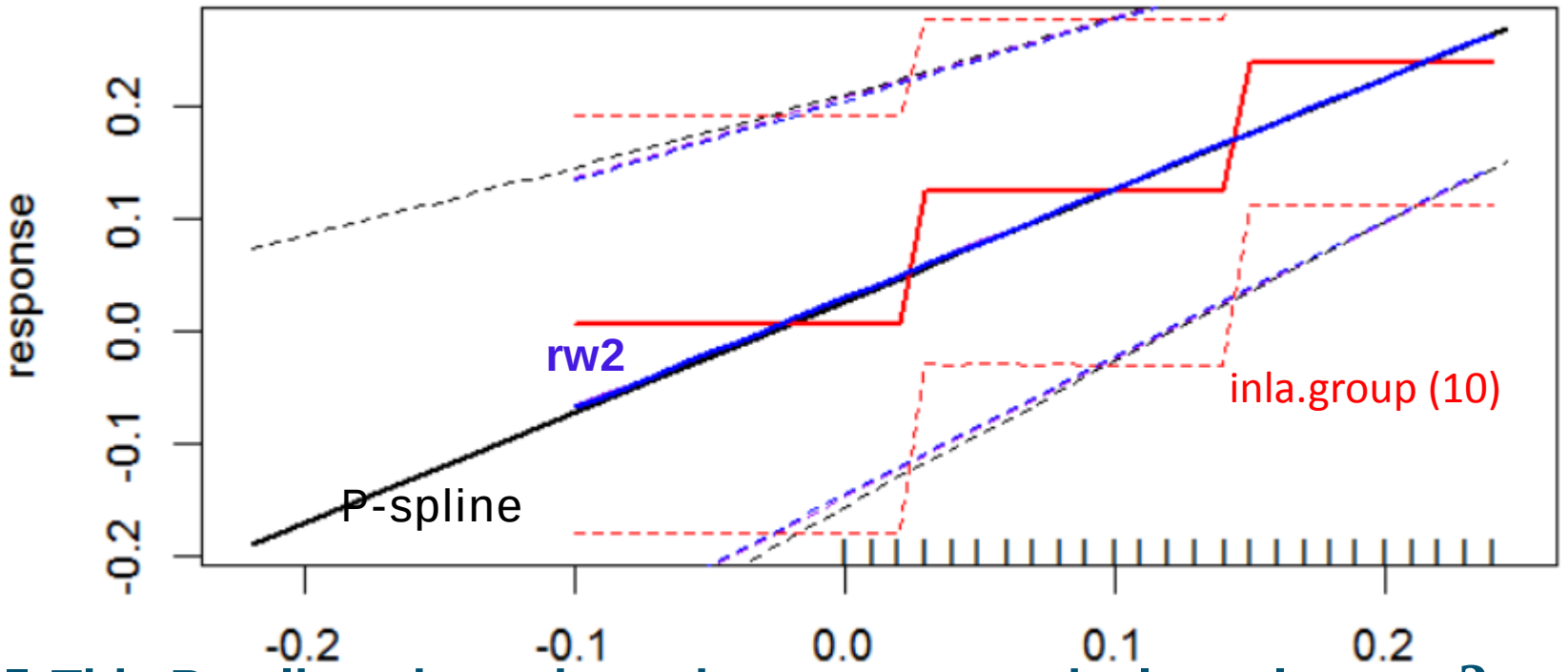
How do the P-splines extrapolate?

- Data x =seq(0,1,0.01), y = x + rnorm(n, 0.5)
- Compare with linear regression (purple)



Zoom in on [-0.2, 0.25]

- and compare with rw2 and inla.group(x, n= 10)

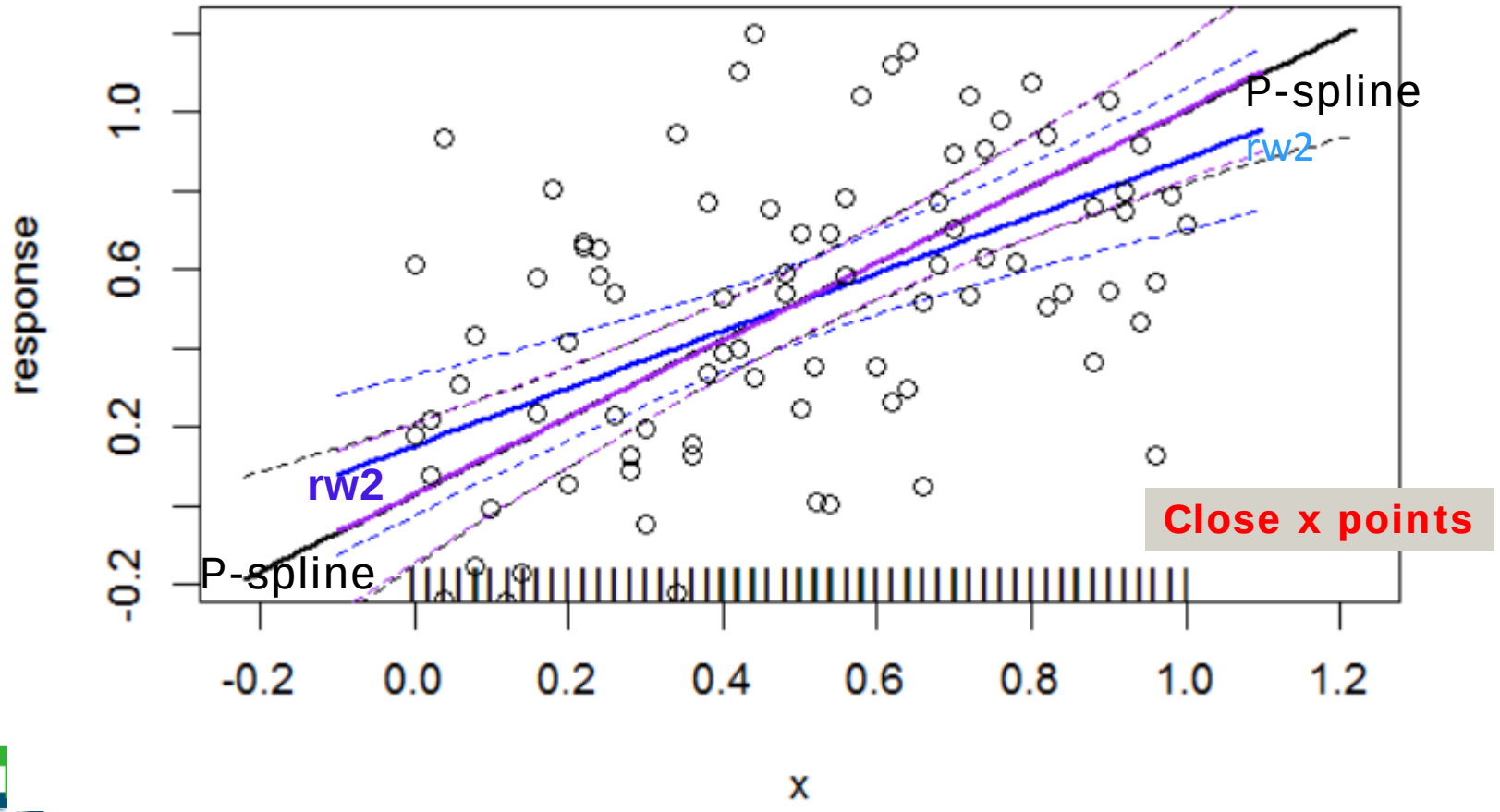


- This P-spline gives about the same numerical result as rw2
- inla.group gives a step-function...

Inla.group with same number of coeffs as P-spline

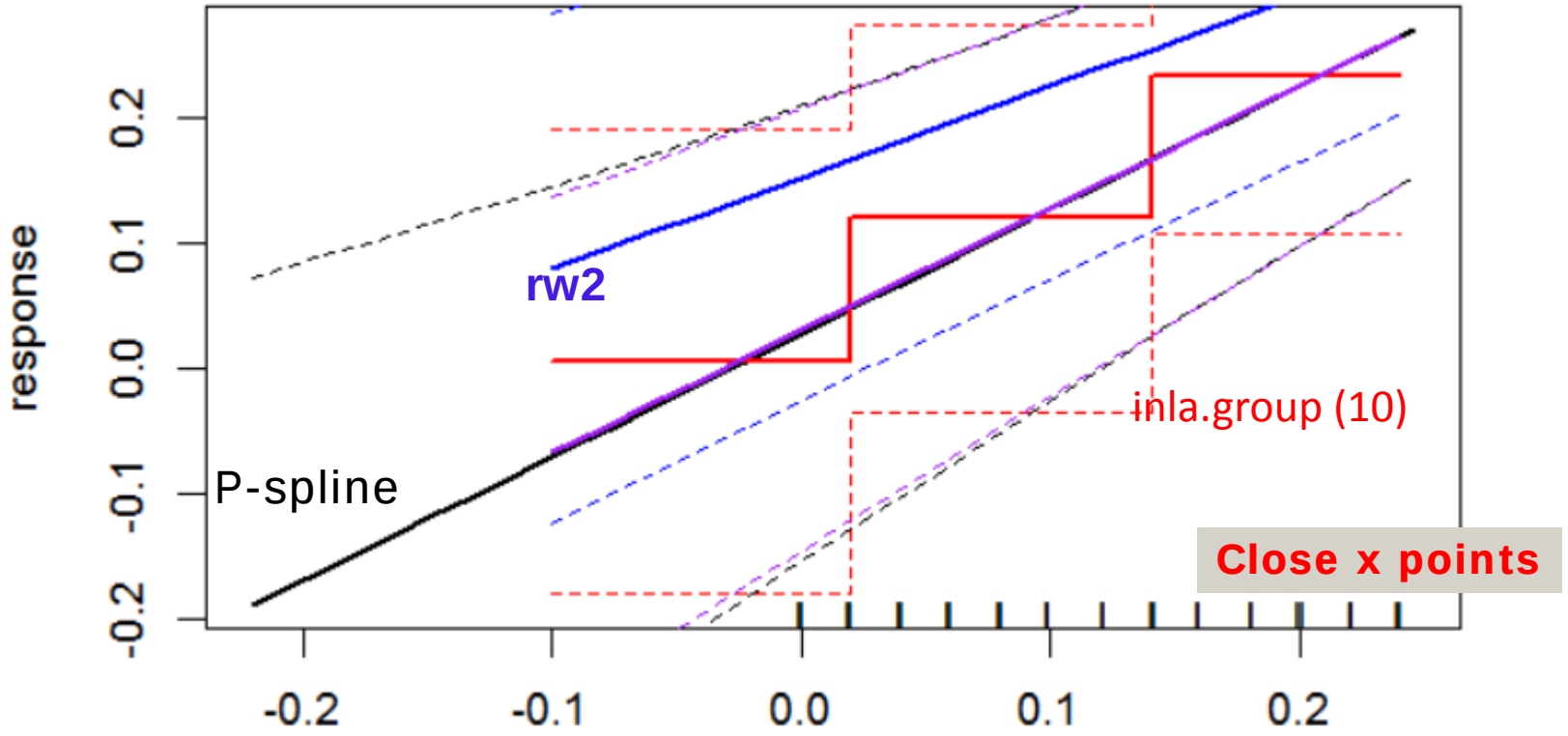
Comparison of P-spline with rw2 for close x-points

- Data x = c(x1,x1+ε) x1=seq(0,1,0.02), y = x + rnorm(n, 0.5) ε= 1e-5
- Compare with linear regression (purple)



Zoom in on [-0.2, 0.25]

- and compare with linear regression, rw2 and inla.group(x, n= 10)

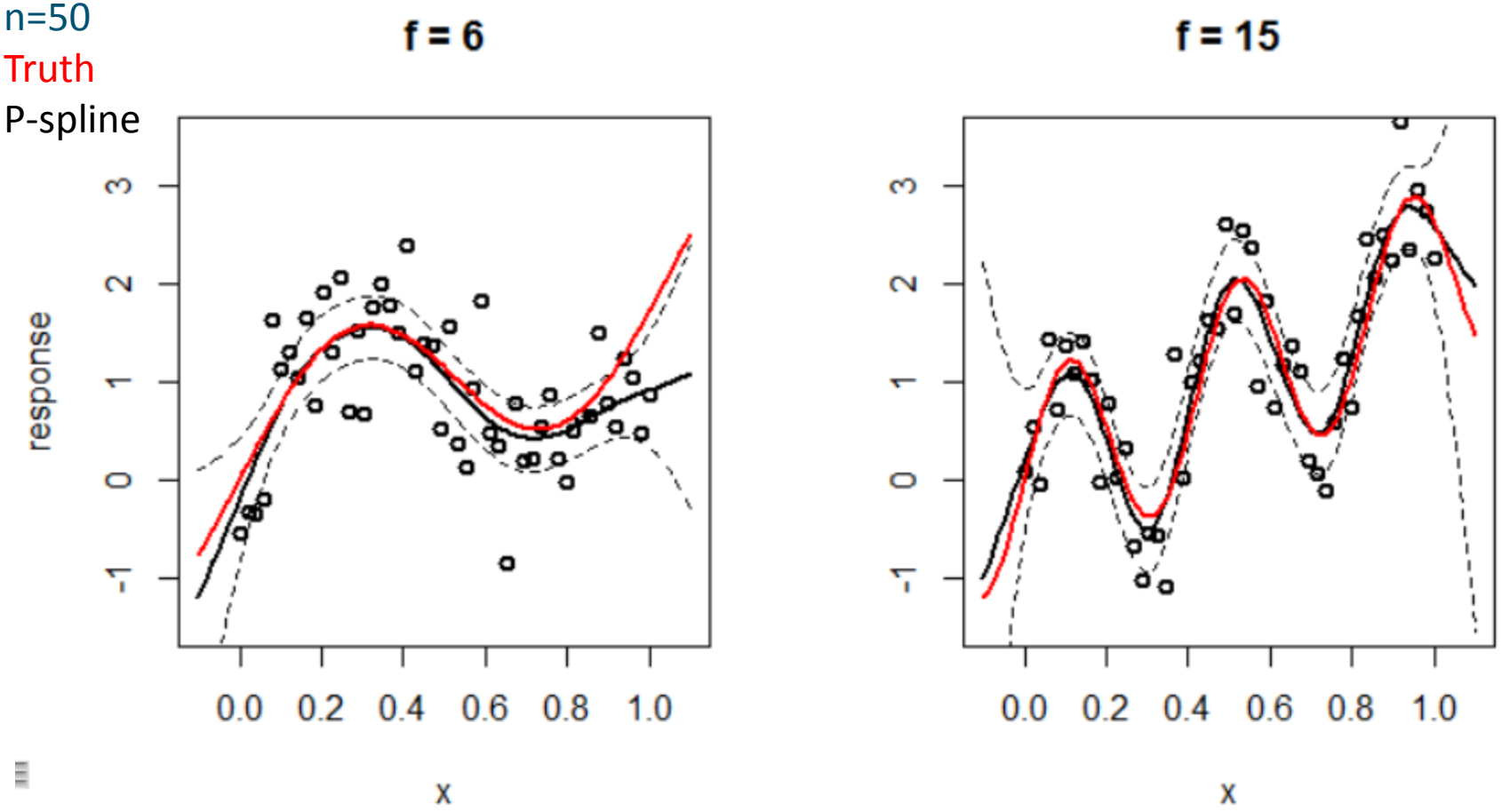


- (c)rw2 in error because of near singularity? OK for ε=0!

using control.fixed=list(prec.intercept=0.1),....

Simulated example y = sin(fx)+2x+e, sd(e)=0.5

n=50
Truth
P-spline

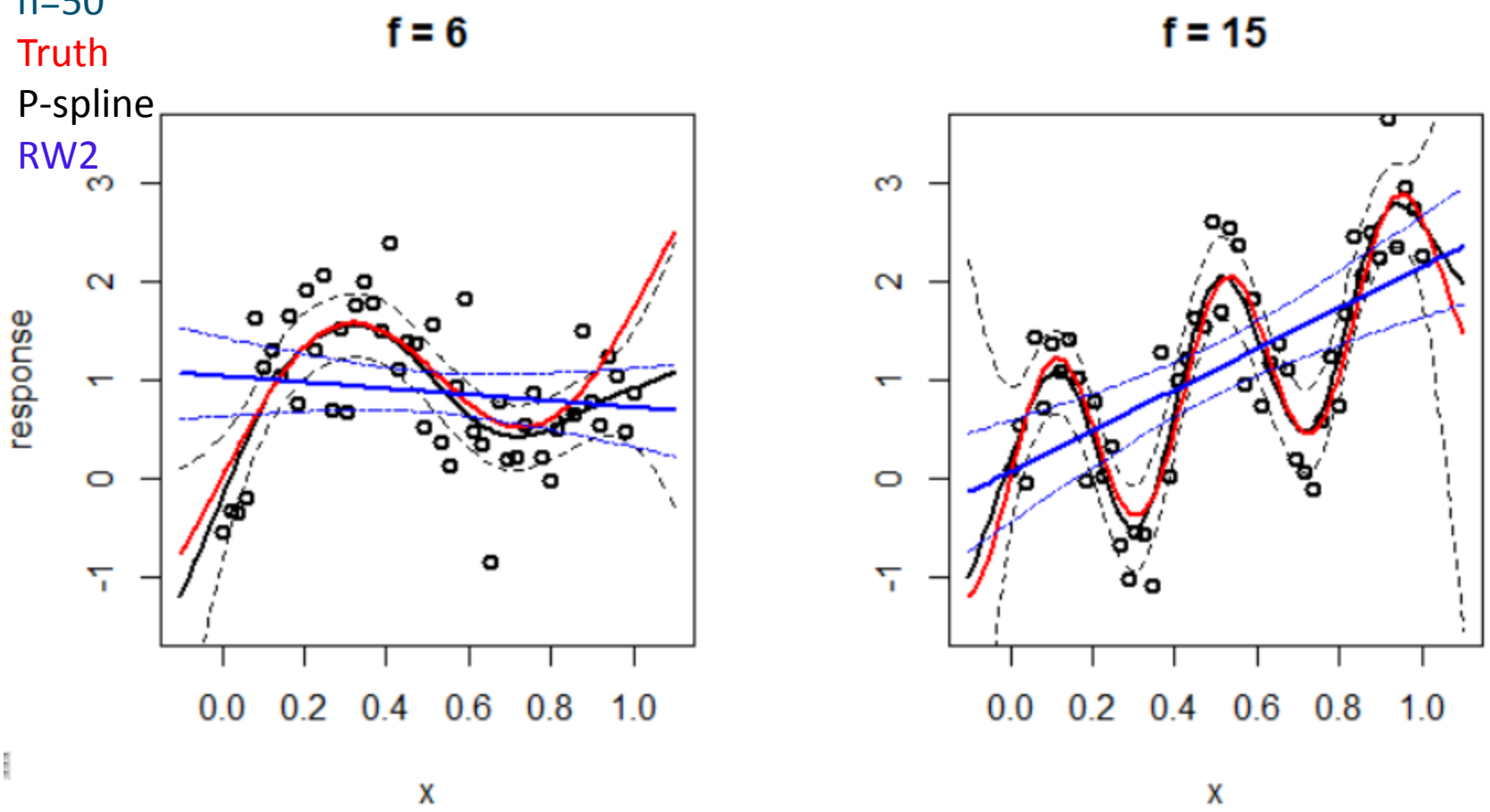


priors: flat for precision of errors

$$\tau_j \sim G(1,0.01)$$

Simulated example y = sin(fx)+2x+e, sd(e)=0.5

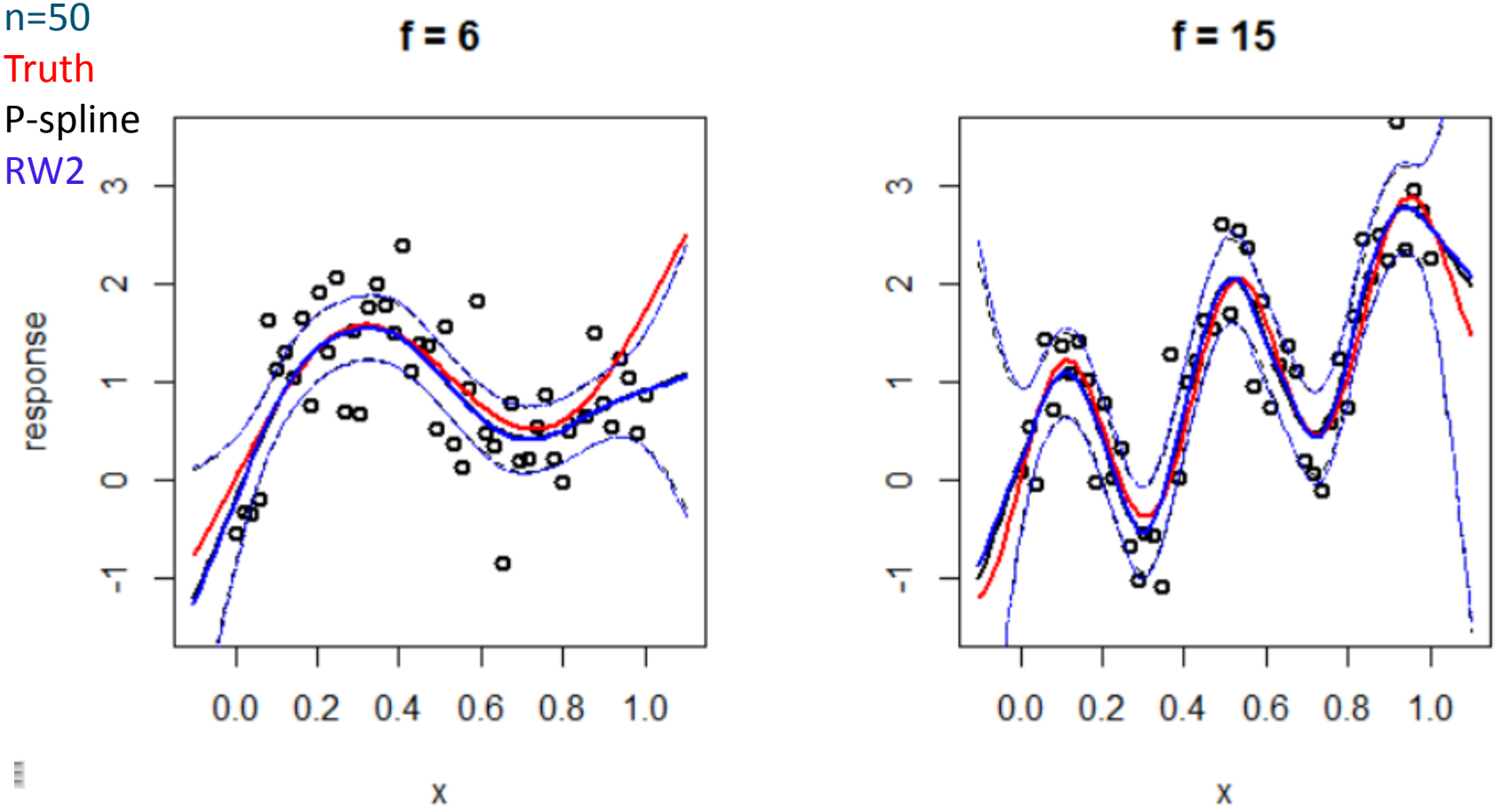
n=50
Truth
P-spline
RW2



priors: flat for precision of errors
RW2: default, $\tau_j \sim G(1,5e{-}05)$....

Simulated example y = sin(fx)+2x+e, sd(e)=0.5

n=50
Truth
P-spline
RW2



priors: flat for precision of errors

Also RW2: $\tau_j \sim G(0,0.1)$

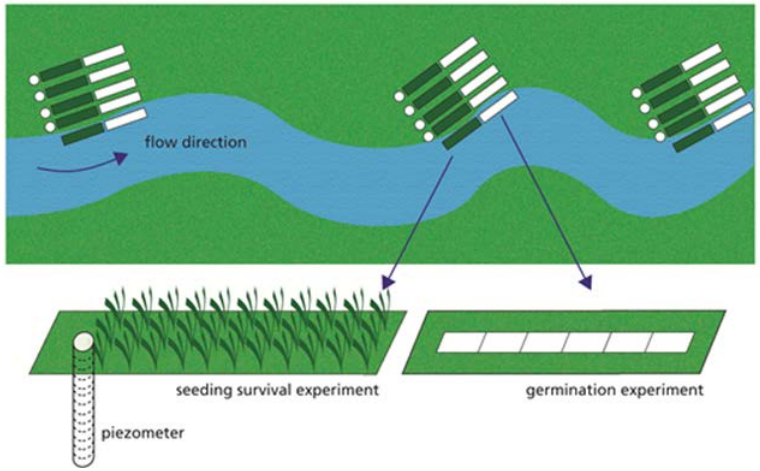
Germination experiment along riparian transects

- 3 streams, 3 transects per stream →9 Transects
- 5 plots per transect → different water level (wl)
- 20 seeds per plot
- 17 species.... We need to do it not just once....
- y~ Beta-Bin(20,p), logit(p) = Transects + f(wl) + ...
- Expected: unimodal effect of wl → 3th order differences
- Gumbel (λ) type II prior

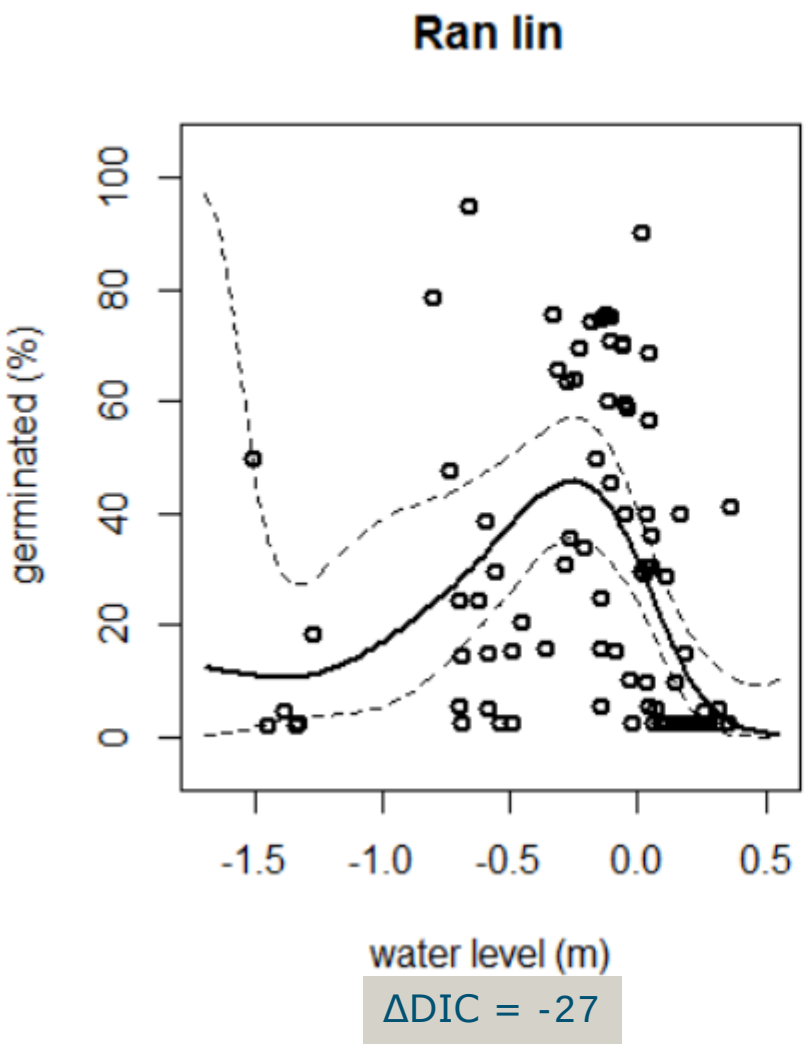
λ = 1.427 from α=0.01, sd(u)= 1

(Martins et al 2014)

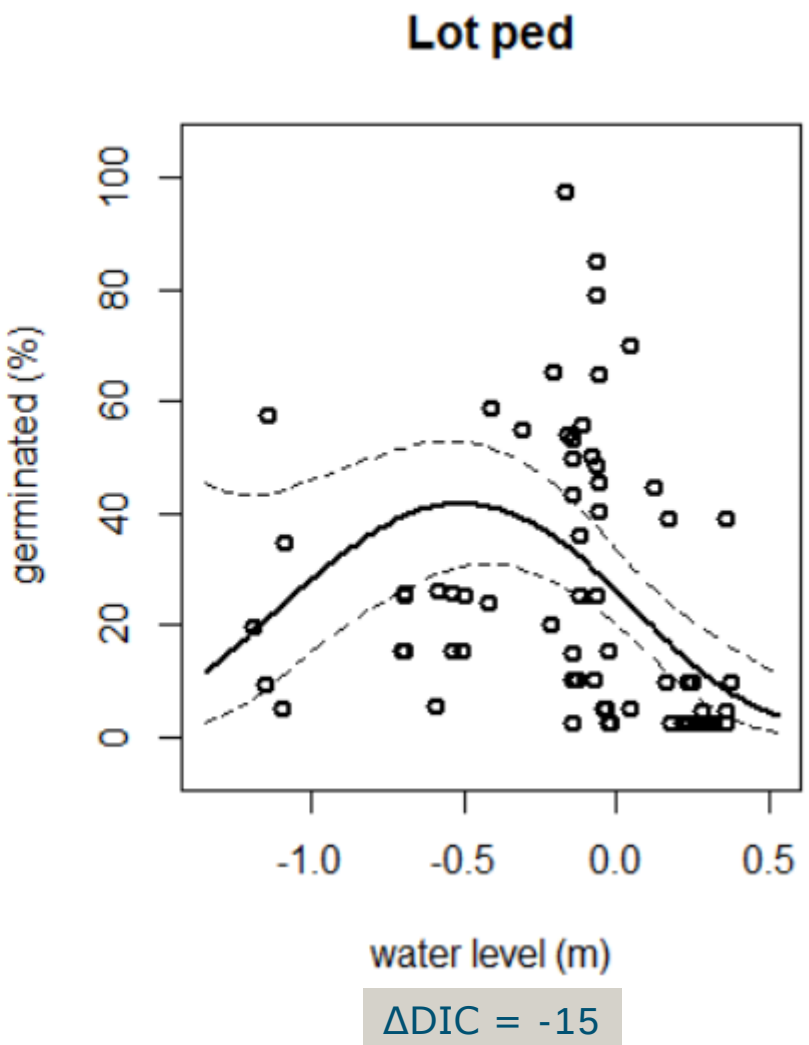
... λ = 1 or 3, similar results



Two species against water level

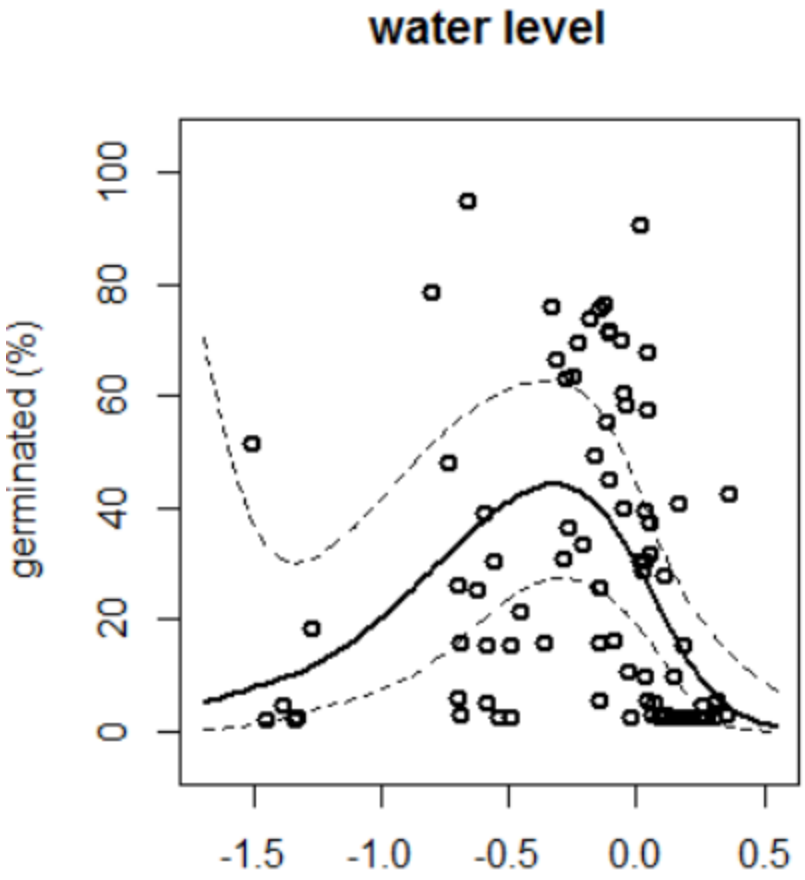


ΔDIC = -27

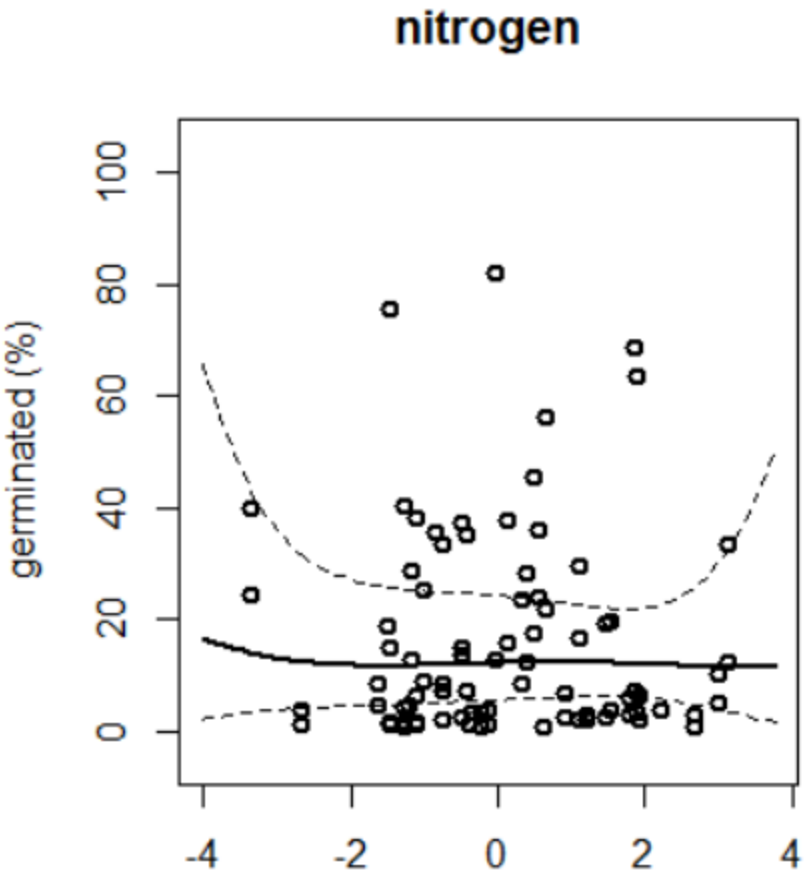


ΔDIC = -15

Component-plus-residual plots for Ran lin

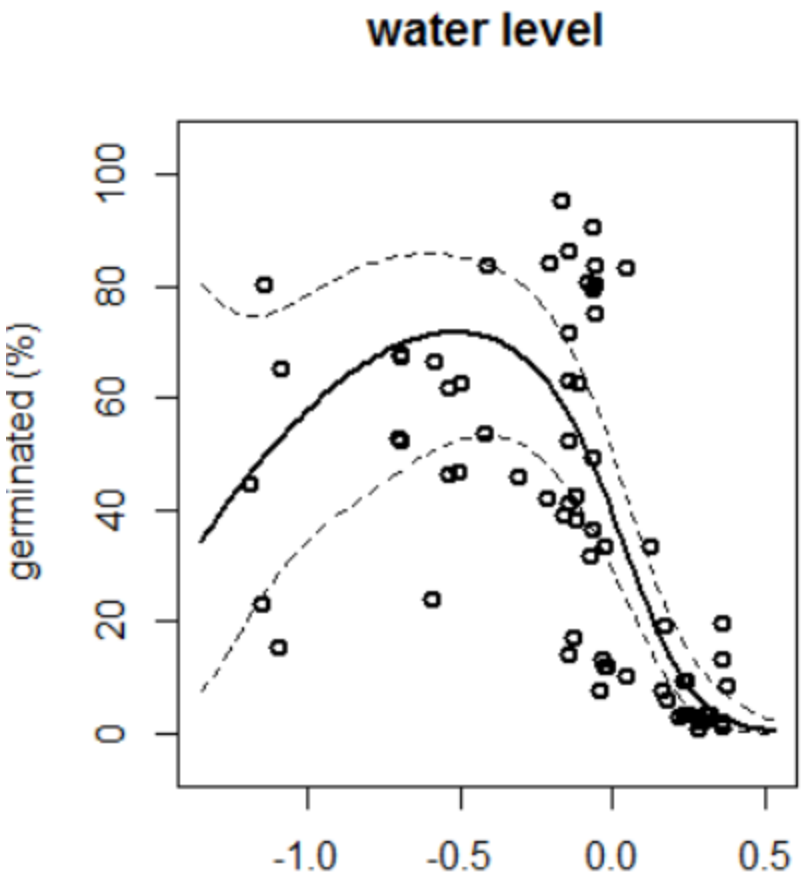


ΔDIC = -27

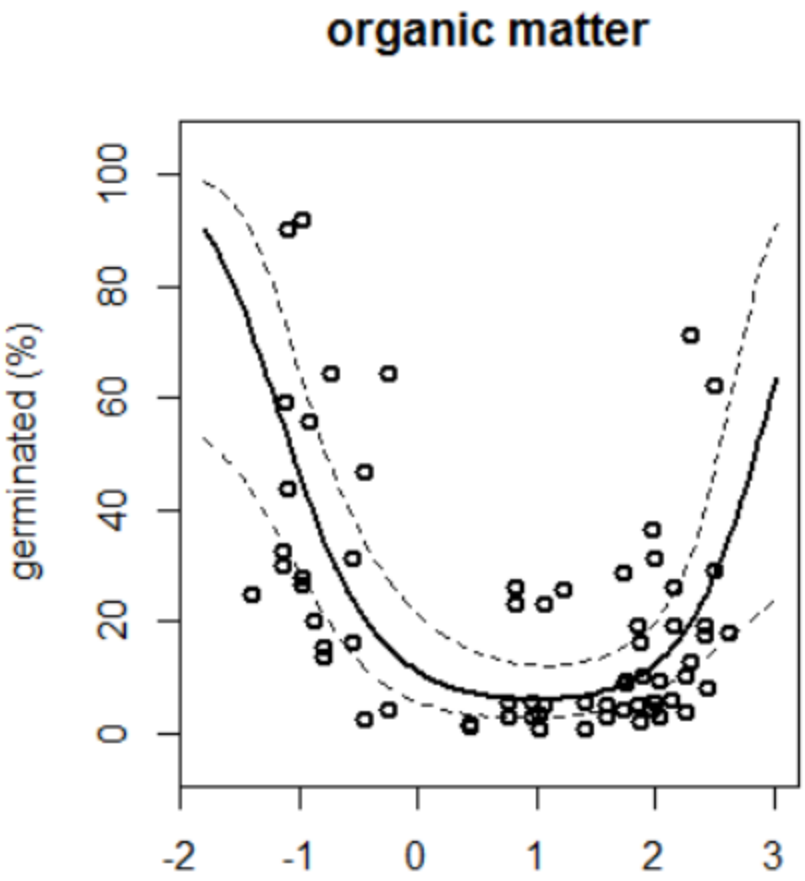


ΔDIC = -6

Component-plus-residual plots for Lot ped

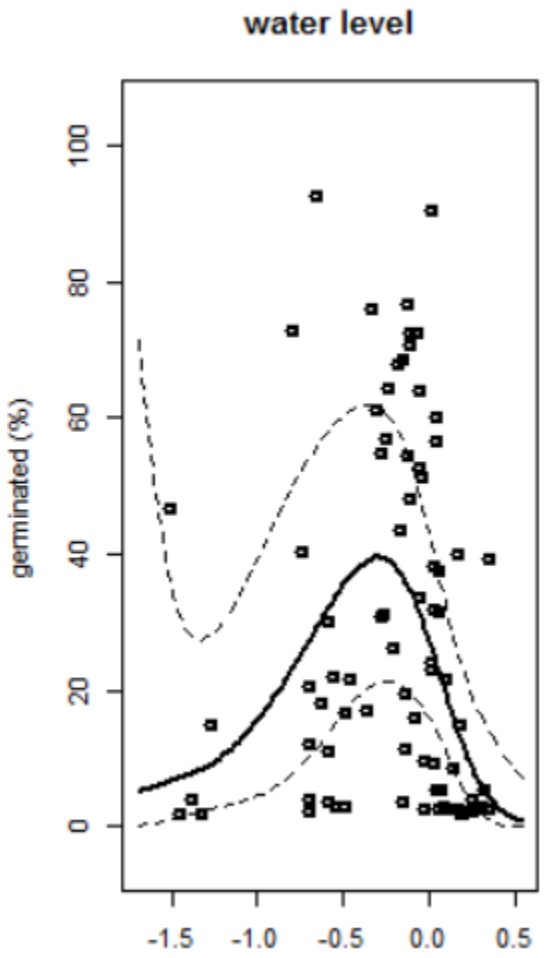


ΔDIC = -15

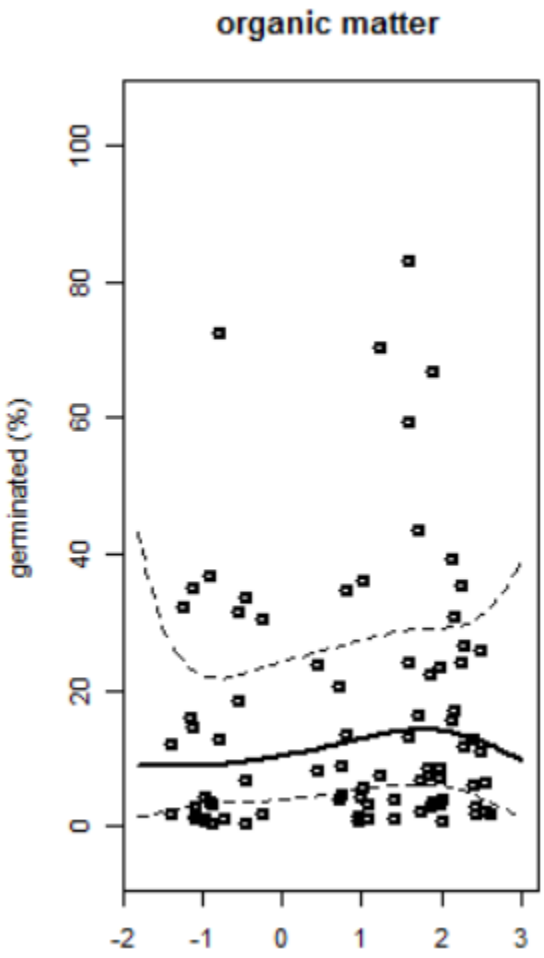


ΔDIC = -21

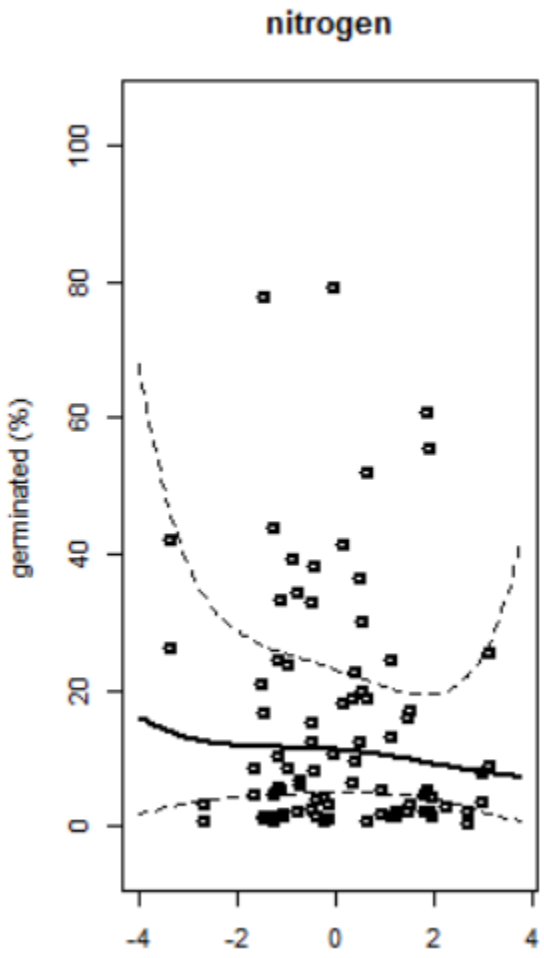
Component-plus-residual plots for Lot ped



ΔDIC = -15



ΔDIC = -21



ΔDIC = +2

WAGENINGEN UR
For quality of life

How to compare?

- We wanted: automatic choice of the penalty parameters, but .... need

- Choice of prior
  - Scale dependent (Sørbye, Scaling IGMRF)
- PC-priors (Martins et al 2014)
  - Gumbel type 2 looks fine for P-splines..
    - formal derivation needed

P-spline density estimation à la Eilers & Marx 1996

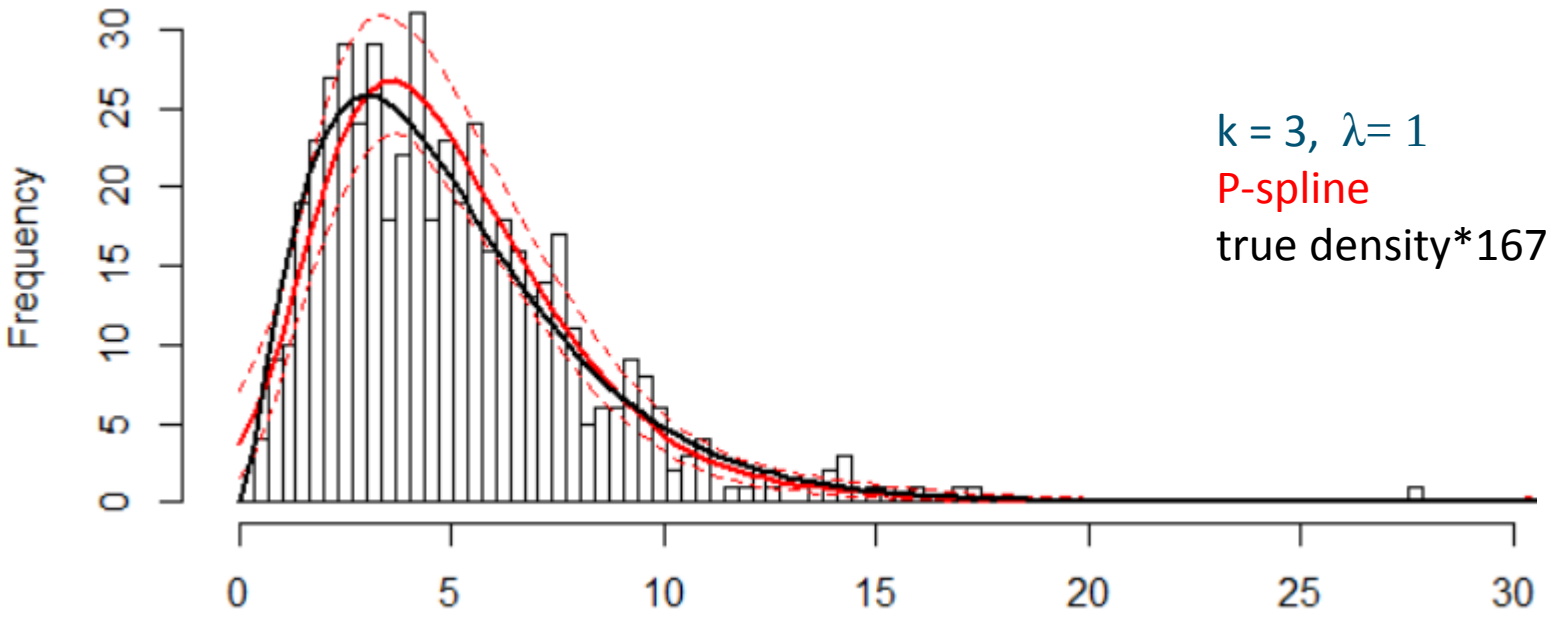
- Difference order 3 with log-link

→ reference distribution is the normal distribution

→ conserves the mean and variance of the data

- Poisson regression on binned data using B-splines
  - fine bins if you want → y= 0 or 1 per bin

u ~ rchisq(n= 500, df = 4) in 100 bins



- Point-wise credibility ...
- to draw a density from the posterior, one needs the joint posterior of the B-spline coefficients....?!
- Gaussian copula via **lincomb.derived.correlation.matrix = TRUE**
  - Alternatives?

An alternative used in INLA

rule. To represent the density $\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$, we use

$$\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \mathcal{N}\{x_i; \mu_i(\boldsymbol{\theta}), \sigma_i^2(\boldsymbol{\theta})\} \exp\{\text{cubic spline}(x_i)\}. \quad (17)$$

The cubic spline is fitted to the difference of the log-density of $\tilde{\pi}_{\mathrm{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ and $\tilde{\pi}_{\mathrm{G}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ at the selected abscissa points, and then the density is normalized by using quadrature integration.

NB: not from data but calculated densities

- **To mimic (17), log-link with distribution ?**
- **Could use x and x^2 as fixed effects (this then gives the reference normal**
- **O- or P-splines with penalty of order 2**

Going higher dimensional with P-splines

- Tensor product of B-splines $$f(x_1, x_2) = \sum_{j=1}^{c_1} \sum_{k=1}^{c_2} \theta_{jk} B_{1j}(x_1) B_{2k}(x_2)$$
  - with sum of penalty matrices, one for each dim.

$$\check{P} = \lambda_1 I_{c_2} \otimes \check{P}_1 + \lambda_2 \check{P}_2 \otimes I_{c_1}$$

- as GLM inconvenient...
- but one can exploit the grid structure: array models
  - Eilers et al (2006), Currie et al. (2006)
- and use a Schall-type algorithm of penalty estimation
  - SAP algorithm Rodríguez-Álvarez et al (2014)
- For INLA, just two hyper parameters
  - If a linear combination of precision matrices were implemented... Is it? ... Adaptive smoothing...

A wild attempt

```
# Kronecker product of B-spline bases per predictor
B = B1%x%B2
# Precision matrices
P1 = Diagonal(c2)%x%crossprod(D1)
P2 = crossprod(D2)%x%Diagonal(c1)
idx1<- idx2 <-  1:ncol(B)
data <- list(y, idx1, idx2)
formula = y ~ -1 + f(idx1, model = "generic", Cmatrix = C1) +
                   f(idx2, model = "generic", Cmatrix = C2)
inla(formula, data = data, control.predictor = list(A = B))
```

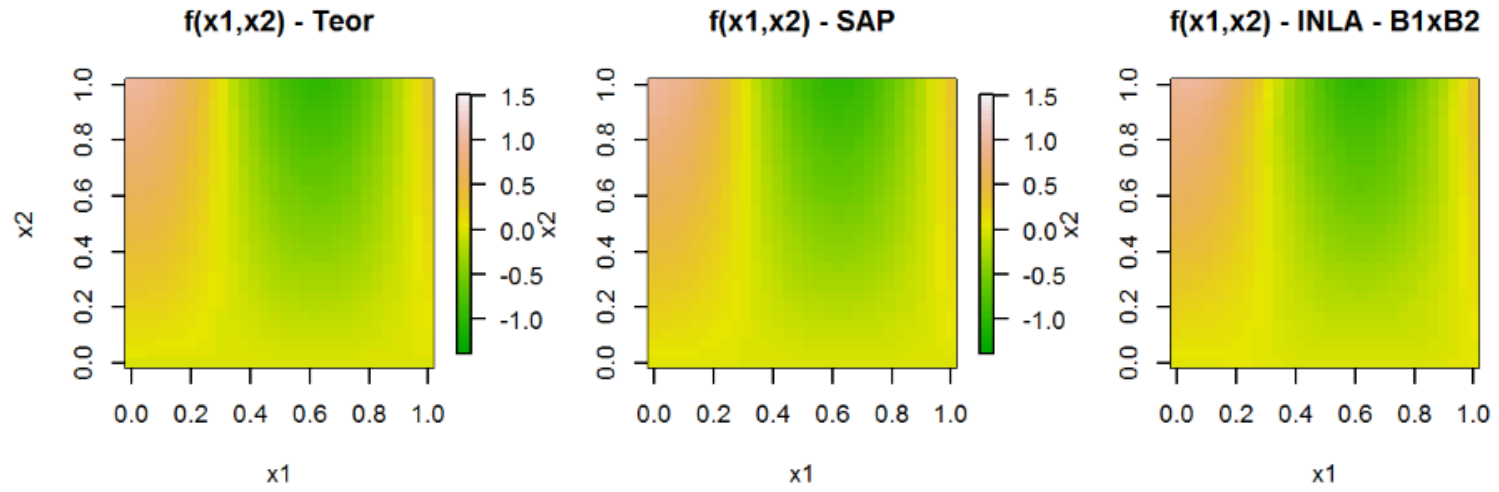
A two parameter model, but for independent u

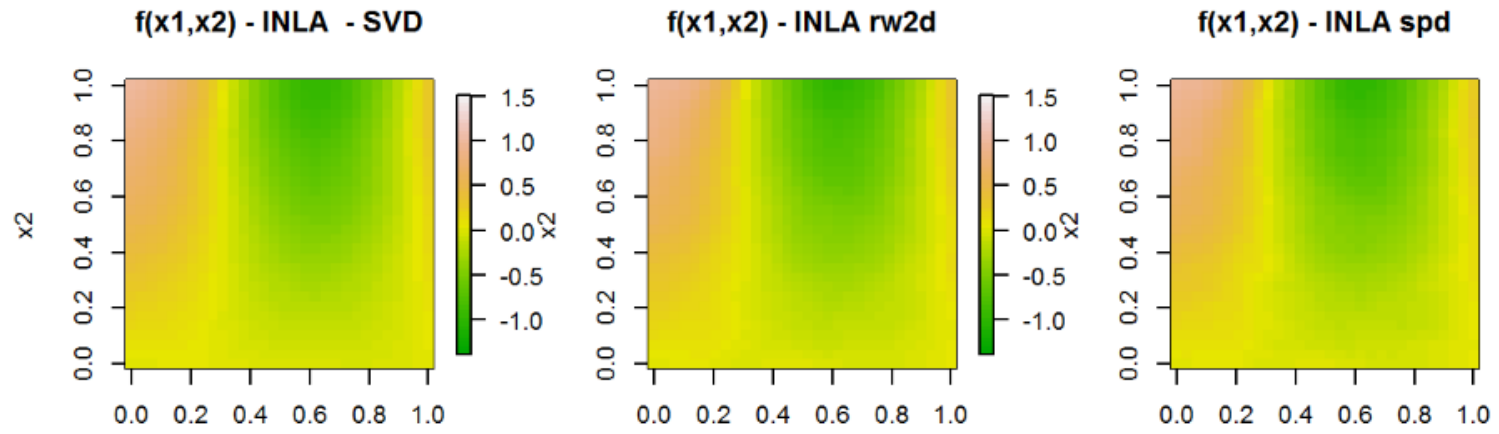
- $var(u_1 + u_2) = \sigma_1^2 + \sigma_2^2$ also with vector/matrices
- $prec(u_1 + u_2) = \left(\sigma_1^2 + \sigma_2^2\right)^{-1} \neq \tau_1^2 + \tau_2^2$

Y0 = outer(cos(5*x1), x2, '*') on 30 x 30 grid

$\sigma = 0.1$

t.SAP	t.B1xB2	t.SVD	t.rw2d	t.spde
0.05	4.17	7.90	1.51	5.96





mse

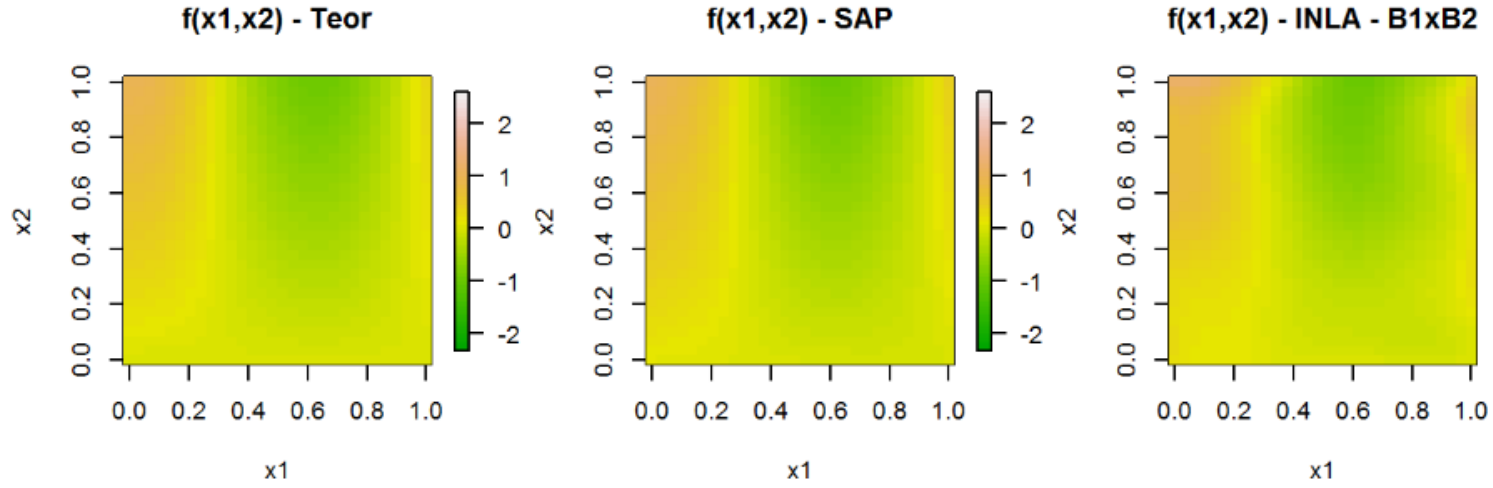
SAP	INLA - B1xB2	INLA - SVD	rw2d	spd
6.447e-05	2.201e-04	2.202e-04	3.492e-04	4.248e-04

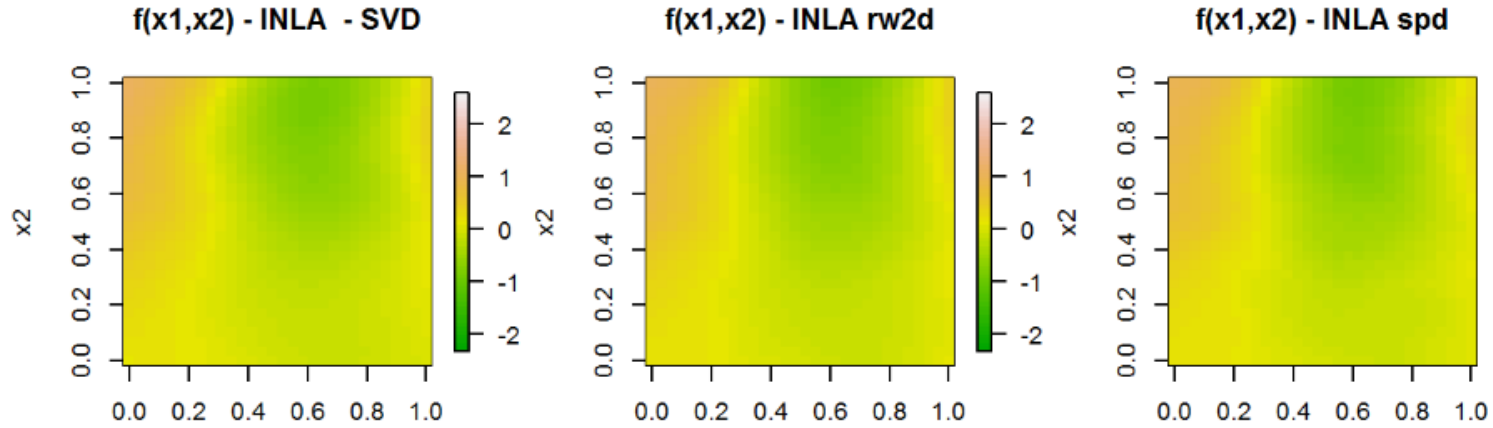
WAGENINGEN UR
For quality of life

Y0 = outer(cos(5*x1), x2, '*') on 30 x 30 grid

$\sigma = 0.5$

t.SAP	t.B1xB2	t.SVD	t.rw2d	t.spde
0.03	5.83	12.90	2.29	4.43





mse

SAP	INLA - B1xB2	INLA - SVD	rw2d	spd
0.001023	0.004221	0.003219	0.002574	0.002983

WAGENINGEN UR
For quality of life

Y0 = outer(cos(5*x1), x2, '*') on 100 x 100 grid

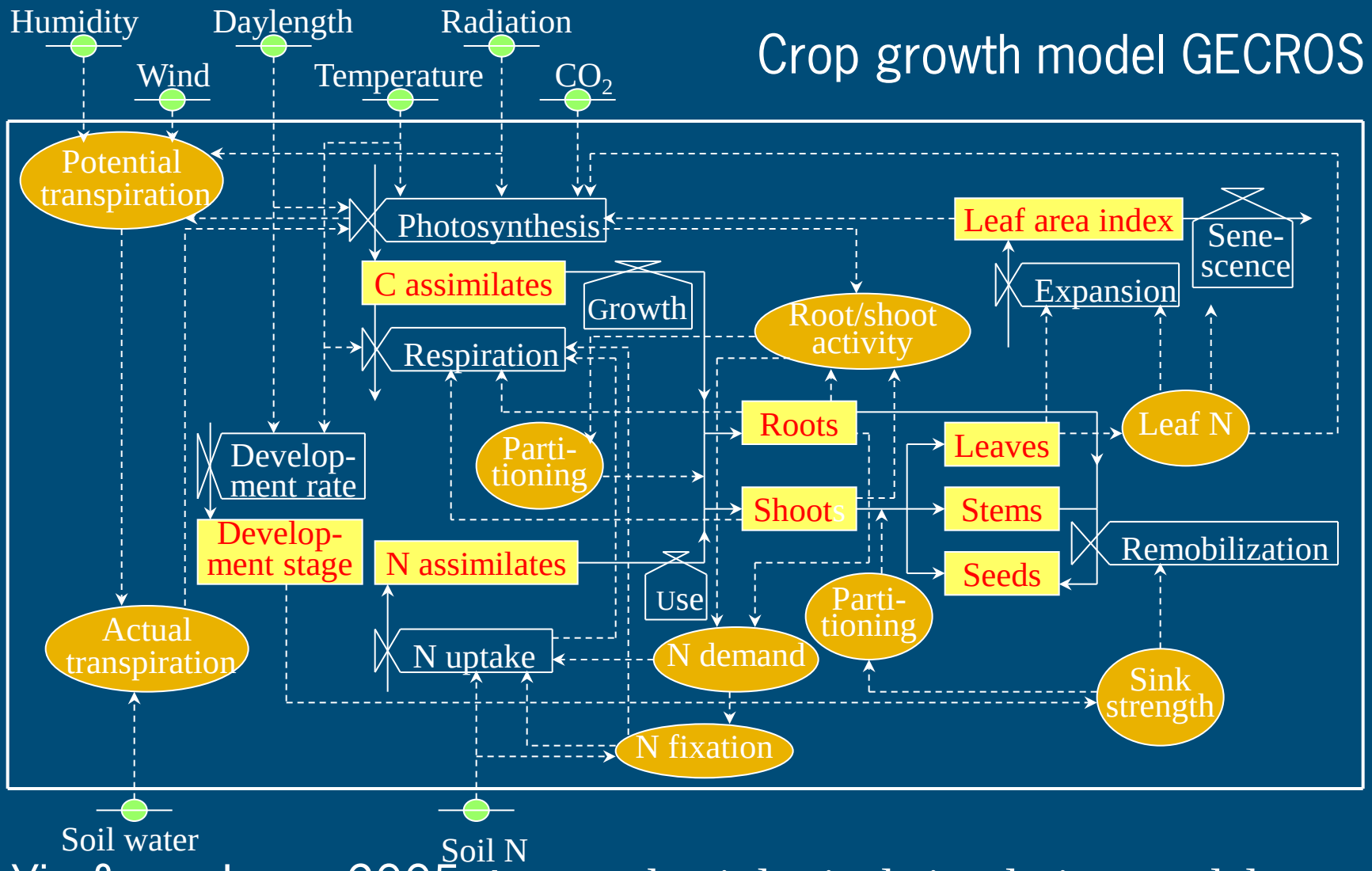
```
> mse
         SAP INLA - B1xB2   INLA - SUD         rw2d          spd
0.0002838900 0.0013100943 0.0006757228 0.0006984377 0.0008692138
> times
  t.SAP t.B1xB2   t.SUD  t.rw2d  t.spde
   0.08   39.05  539.82   36.40   11.43
>
```

Complex models

- Mechanistic models / computer models
- Just to give an impression of such a model, next slide

- Dynamics, many relationships, many rate parameters

A complex mechanistic model...



Yin & van Laar, 2005 An ecophysiological simulation model for genotype-by-environment interactions

Complex models, simple algorithms?

Model / parameter calibration

- Tune parameters of model to best fit (e.g. min residual sum of squares)
- Update parameters on arrival of new data

DANGER: turning the physics model into a planetarium (de Wit) so that the meaning of the parameters is lost

- Bayesian methods can help via the prior

Algorithms for model calibration

- Either by craftsmanship or by some optimization method

  Methods that require you to specify first and second order derivatives (algebra!) ‘Hill climbing’, local maxima??

  Other, e.g. global optimization methods,

  - Genetic Algorithms (GA: Get away with Algebra)
    - Price algorithm, Nelder-Mead simplex
  - Simulated annealing
- ☹ What about parameter uncertainty and resulting model uncertainty?

What about parameter uncertainty?

- Optimization methods typically end with a single best solution
  - populations of solutions are just generated to reach the globally best solution.
- parameter uncertainty by
  - Fisher information = inverse of Hessian of loglikelihood (see Uncertainty evaluation, this morning)
  - Or a numerical approximation thereof
    - E.g has been proposed for Nelder-Mead, seldomly used, I believe.
  - Bootstrapping
- Bayesian methods are ideal here! Just draw a sample from the posterior.
- STORE and publish the sample to be able to calculate 95% (credible) intervals for any later prediction from the model.

van Mourik et al 2014

Optimization

objective function $f(\theta)$

Aim: find **best** θ and max $f(\theta)$

- Gradient-based methods
- Alternate direction opt
- Nelder-Mead Simplex
- Differential evolution
- …

Bayesian statistical methods

posterior density $\pi(\theta) \propto$ prior density × data likelihood

Aim: find the 90% credible region of θ by sampling from π

- Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis
- Particle Filters, Sequential Monte Carlo

Bayes adds the uncertainty about θ

Multi-site vs single-site calibration

Example: Reinds et al 2008 Geoderma calibrated the VSD soil acidification model to European forest monitoring data

>122 sites (forests)

4-6 year data on $p = 5$ outcome variables

$q = 6$ model parameters to calibrate (tune).

- On data from all sites (lots of informative data) the posterior uncertainty vanishes!
- On data from each single site, huge uncertainty

Multi-site calibration with a geostatistical prior

- Single sites that are close in space are likely to have close model parameters → use a geostatistical prior for the model parameters
- This prior also has parameters (hyper-parameters),
  nb: need also a prior for these.

Non-linear regression with geo-prior

- *n* sites (index *i*) and data per site (y_{ij} ,x_{ij}) $j = 1...n_i$
- Regression model per site : $y_{ij} = h(\beta_i , \{x_i\})+\varepsilon_{ij}$

  $\varepsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. (or....)
- Prior for $\beta_i \sim N(\theta,\Sigma)$
- Geostatistical parameters: θ and parameters defining Σ
- Posterior:

$$p(\boldsymbol{\theta},\Sigma,\boldsymbol{\beta}\,|\,\mathbf{y}) \propto \underbrace{p(\boldsymbol{\theta},\Sigma)}_{\text{Prior } \boldsymbol{\theta} \text{ and } \Sigma}\ \underbrace{N(\boldsymbol{\beta}\,|\,\boldsymbol{\theta},\Sigma)}_{\text{Prior } \boldsymbol{\beta}}\ \underbrace{\prod_{i,j}^{n} N(y_{ij}\,|\,h(\boldsymbol{\beta}_i,\boldsymbol{\theta}),\sigma^2)}_{\text{Likelihood}}$$

Multi-site calibration with a geostatistical prior

- Single sites that are close in space are likely to have close model parameters → use a geostatistical prior for the model parameters
- This prior also has parameters (hyper-parameters),
  nb: need also a prior for these.
- Fit the model by a Gibbs sampling approach alternating between:
  - Sample the geo-statistical hyper-parameters given the current values of the model parameters
  - Sample the model parameters for each of the sites (with the current geo-statistical model)

Bayesian computation (in R)

- Sample the geo-statistical hyper-parameters given the current values of the model parameters
  - This is a ‘standard’ multivariate geostatistical problem with data = current values of the q (6) model parameters
  - We use one iteration of spMvLM {spBayes}, which gives a Bayesian analysis of a linear co-regionalization model
- Sample the model parameters for each of the sites (with the current geo-statistical model)
  - This is a ‘standard’ non-linear regression problem
  - We use one iteration of single chain DE-MC with blocks of parameters defined by sets of sites close in space

Problems with standard Metropolis or Gibbs

- Bad convergence if....
  - Step-sizes (scales) of the parameters is ill-chosen
  - parameters are (highly) correlated
- Need for adaptive algorithm to learn scale and orientation of the posterior
  - Haario, Roberts, Rosenthal ....; Gilks, Roberts, Liang...
  - I describe an algorithm related to Genetic algoritms and Adaptive Direction sampling

Aim: cross GA with MCMC for increased efficiency

Genetic algorithm

Differential Evolution (1995)

Markov Chain Monte Carlo

Metropolis et al (1953)

Differential Evolution Markov Chain (2006)

aka DREAM

adapts itself automatically

to the shape of posterior

DE: Storn & Price 1995

DE-MC: ter Braak 2006

Differential Evolution with local move

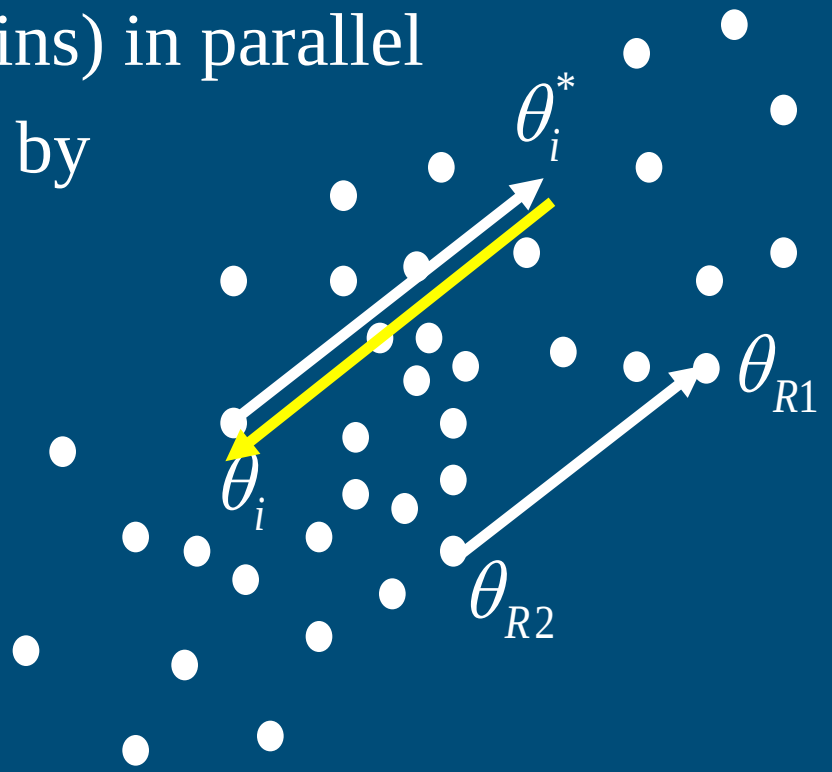
Reverse jump

equally probable

- Run N solution vectors (chains) in parallel

Generate proposal for i^{th} chain by

$$\theta_i^* = \theta_i + \gamma(\theta_{R1} - \theta_{R2})$$



In 2-dim with N = 40 chains

γ = scaling factor

Accept proposal if fitness is higher

Differential Evolution Markov Chain (DE-MC)

- Apply local DE with accept ratio

$$r = \frac{\pi(\theta_i^*)}{\pi(\theta_i)}$$

accept ratio makes π()

the invariant distribution

- DE-MC is a Markov chain on S^N . Upon convergence, joint pdf factorizes to $\pi(\theta_1)\times\ldots\times\pi(\theta_N)$ ➔ $\theta_1\ldots\theta_N$ of individual chains are independent.

➔ Can use Gelman's $\hat{R}$ for convergence checks

DE-MC: MCMC version of DE that solves two important problems with MCMC

- choosing the proposal distribution
- poor mixing with multimodal posteriors

- For DE-MC one only requires to program
  log-prior + log-likelihood
  set an initial population of θ vectors
  The rest is automatic

DE-MC does not require finite mean and covariance of θ as in Haario et al

Choice of *N*? *d*=number of parameters

- Update in min(*N*-1, *d*) space ➔ *N*> *d*, but…
- Each chain needs to travel from its initial, possibly very low-density region to a moderate- to high-density region
- This costs *N* times as much as for a single chain ➔ the startup may be costly…
- Tradeoff between exploration and exploitation
- What is the chance of having an outlier member?

Modify DE-MC so that it can work with small N (1)

as in Gibbs sampling

- Crossover- update small blocks of parameters instead of jointly updating of all d parameters
  - Example of nonlinear mixed effects models: each subject's parameters form a block
  - Efficient computation via a DAG approach (WinBugs/OpenBugs)
  - How to choose the blocks automatic in an efficient way? Random cross-over in DE = DREAM

Results in an adaptive Metropolis within Gibbs algorithm

DREAM: Vrugt, ter Braak et al. 2009

Modify DE-MC so that it can work with small N (2)

- Sampling from the past: choose θ_{R1} and θ_{R2} not just from the current time, but also from the past.

No longer a Markov chain, but by diminishing adaptation theorem (Haario et al 2001, Roberts & Rosenthal, 2007), chain is still ergodic

➔ Extreme case: single chain DE-MC (N=1) with sampling from the past only: most efficient for far off start in nicely behaved posterior distributions

Can be combined with Crossover-idea, e.g. *d = 600 in 100 blocks*

Looking for more types of updates (1)

- Can the Nelder-Mead downhill simplex algorithm be turned Bayesian?

Yes, it can, in a sense (pick random point to reflect) but one has to be careful because

reflection-expansion-contraction

is not space-conserving. Jacobians come-in to preserve reversibility of the Markov Chain.

Unfortunately, p(accept) close to 0 for $d \geq 5$.

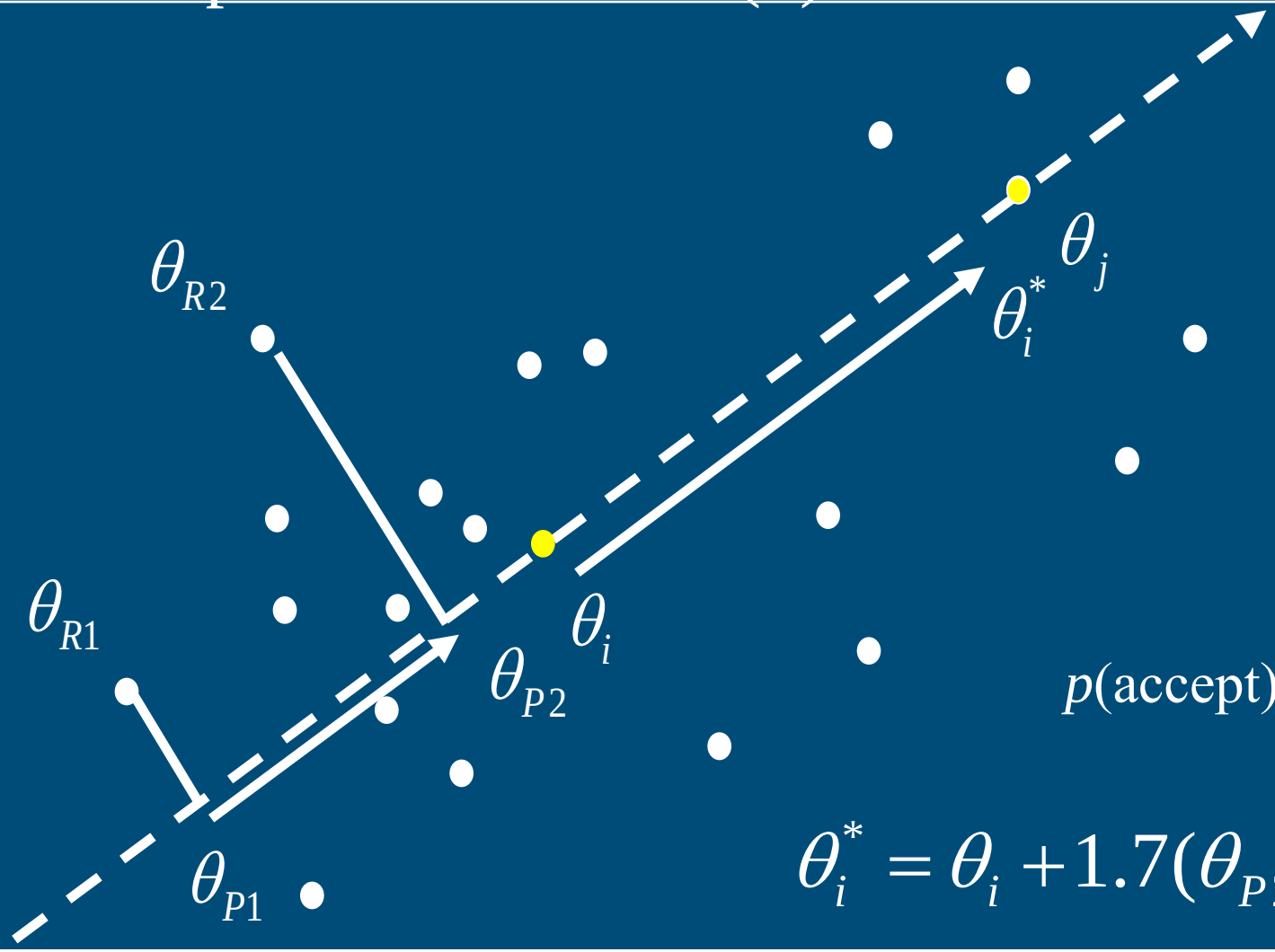
Looking for more types of updates (2)

- Snooker update

Gilks et al 1994, Liang & Wong 2001

Snooker update in DE-MC (2)

$1.7 = 2.38/\sqrt{2}$



$p(\text{accept}) \approx 0.20$

$$\theta_i^* = \theta_i + 1.7(\theta_{P2} - \theta_{P1})$$

Members j, $R1$, $R2$ can be past samples

Snooker update in DE-MC (1)

To update θ_i of the i^{th} member,

- Select another member, j say
- Sample along the line $\theta_i - \theta_j$ from the density on that line as follows.

1. Select two other random members, $R1$ and $R2$ say.
2. Project θ_{R1} and θ_{R2} on to the line $\theta_i - \theta_j$, yielding θ_{P1} and θ_{P2}.
3. Propose $\theta_i^* = \theta_i + \gamma(\theta_{P2} - \theta_{P1})$ with $\gamma = 2.38/\sqrt{2} = 1.7$
4. Calculate accept ratio
5. Accept/reject proposal

$$r = \frac{\pi(\boldsymbol{\theta}^*)\left\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_j\right\|^{d-1}}{\pi(\boldsymbol{\theta})\left\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\right\|^{d-1}}$$

Members j, $R1$, $R2$ can be past samples

Efficiency of DE-MC wrt to optimal Metropolis

for 50-*dim* Student t_3 posteriors - heavy tailed posterior

- *Eff* = 100*MSE(Metro)/MSE(DE-MC)

	start	"easy"	"far off"	"far off"	"far off"
	Sampling from	*present*	*present*	*present & past*	*present & past & snooker*
Population size *N*		150	150	10	10
median	P50	102	30	28	99
quantile	P2.5	191	1	97	1155

- Far off start and *N* = 10 now give excellent results

Start = initial population

WAGENINGEN UR

Differential Evolution Markov Chain (DE-MC)

- Adapts automatically to the optimal scale and orientation of the posterior
- Runs parallel chains that learn from each other (population MCMC, genetic algorithms)
- Start from overdispersed values, but not too far off..
- Not too many parallel chains
  - N = 1-10 is often sufficient with sampling from the past
- Apply DE-MC within Gibbs with block updating
  - Ideal in nonlinear mixed models: block = subject
- Good performance, also for multimodal posteriors

Differential Evolution Markov Chain (DE-MC)

- A one-page algorithm for complex problems

As all MCMC it can be slow...

Disadvantages:

- Start up/Burn-in:

  often better with Delayed Rejection, but how to do that properly in DE-MC?

- In a Gibbs-setting, the proposals use effectively the marginal variance instead of the conditional one.

Thank you!

The end,

but not of our

dance with statistics.

Cajo.terbraak@wur.nl

WAGENINGEN UR
For quality of life