

Evaluating Model Fit and Selecting among Multiple Models

9 Evaluating Model Fit and Selecting among Multiple Models	217
9.1 Preselection of Candidate Models	218
9.2 The Many Ways of Measuring Model Fit	219
9.3 The Widely Applicable Information Criterion (WAIC)	225
9.4 Variable Selection by a Spike and Slab Prior	228
9.5 Reduced Rank Regression (RRR)	242

DO NOT try all possible model combinations and apply “blindly” model selection tools!

9.1 Preselection of Candidate Models

While we present several methods that can be used to select the ‘best model’ based on quantitative criteria, we do not recommend applying the blind strategy of running all possible model variants through an automated model selection pipeline. Instead, we emphasise that an important task for the ecologist applying HMSC – or any statistical modelling approach – is to use their prior knowledge of the system to restrict the model structure and the sets of candidate variables to be included. The first reason for this is to ensure that the models make sense

- You will risk being distracted from addressing your study questions in a focused way and entering a huge fishing expedition. Do you need all those candidate models to address your study questions?
- “Trying out all models and comparing them” is usually not feasible.

Why it is difficult to compare all possible models?

Example: HMSC model with 100 species, 5 candidate environmental predictors, 3 traits with phylogenetic and spatial data.

- **Environmental covariates.** Linear and quadratic effects: 243 candidate models. Also interactions among predictors: 59,049 candidate models.
- **Traits.** Linear and quadratic effects: 27 candidate models.
- **Phylogeny.** Included or not, 2 candidate models.
- **Spatial random effect.** Included or not, 2 candidate models.

In total: $59,049 \times 27 \times 2 \times 2 = 6.4$ million models.

Why it is difficult to compare all possible models?

We can do variable selection for environmental variables separately for each of the 100 species!

In total: $(59,049)^{100} \times 27 \times 2 \times 2$ models =

17618582001700359442475172675156470941571030498691410334496059522871717892794646050478568081315676174447112208841094659068078029639111380160571997476997561501569
44774357952234230745751915104122148443241661297531164413757238960313637948213981133438032921677861643870953220366622269097782519004062798470228209860198071616844
10644975153284363186771260083520294710564518869103428122084286628829643618674752307429430028138561312071954544047652787826683269041811590433387440350300672558298
86258278380123340042051320529355473239400368549718369349403440786771448670771733259453566311814910108604930427555797660364630216376483172486541474084704289352301
60733956808121797556591856733429166469922426204191393536052768919812774281488801383192325372011252176815119721397582153418505155921002880337476441685054820887861
16735596303234501435261058213413780905856531592805279062905253858004878975217245072663474507195209396534928474557199531030405101984097009674210577912258417591478
39394036961279463780451214604016488100830907972785286071632627158758326374363627333147464565653059738269606395583954549191080257378364970422919470620131113681835
23150119410373134427592668813338707687598987426017112798438056690194295608895253946953082267931627221656600590531983546995534455129068325409881399974906616673561
83178561623357698181150240900897751871034427997126526186984654059383771423353646415505507857850189898072741794194277268627391174768228263134679681882167883166912
45952160722689757891593365112254328376784013036393403084941596656775905256337284769472141717492130658131604828453544225828531510753517049735048534762565083843821
97852616421561313605563179580839115301301543727993666882523443206527186567283631638641701265617959750160325608420443653587492687052787393060270158572417437095561
41074151750120277808085920464806934510772875872673679461950424445979434582440133470324827573791936450437396559997179266880948261095793222978086955426974661721691
86110220884952035747080128587850348439210889775568230683441626377620231665943527235312843508833393421570659706286843697909616472702073339437427252873272384454389
94734289725772926576745858408806934348813964097519531818638782541848721783719356766970418177251763736415734844080639460544868427749332751479074732928121534683949
92813258073229866003742458064508053752406952142279058225999371938109635873233000316297638326574827289445992376726171271034216223319631613011430135075520096881106
29506927872865989175546550144173003901400436256718056192772143341150385554271471028295587230420276079374404062201044230562036642118389011115586254155172669932754
03662776969967128393711510205471060860182370641725333811740273141716499471848053146310135697007881276981443494351018477288080261834155680561429940751303615567463
94608488645023915026024574036918674376157178199128269125494179050478378617616328326703480876862261682749856957080044305560669184764843449835371495624254466734685
84415882191316035953880925632663132295334722738129252835197223055003109592715984014000187560054351814019448026766624845821319095271327861467912170675590714957181
51157281952610245776788855327297526791620715913224314956742234087009211440551211711842950715742859100671446455403508078106866791688202059388317892250588988777656
39975017716131699794500776337066257847875805521634677805225348828607876237893702300080837700431920800800575369962041927801497530945224374338014394674655536085217
04774461060367264691625962737412616947630013869195150635459859922211277713177144777908022048319633010912800480698273043576732895824474390412729438641183614842853
71795088629797334312403548644391804429665480783882188341829713098035724119968917886538964955681970725906940896193146535386552289405244008942613905124207851986819
97998526946697564743495626364016997405217567274094429812305551965932621615535687229745451492926189784480297727702031473160459173333815448641633640808319422857431
60035379750414435647452750614291298633273119015339134371177892275687821379715832677459060843129209939161780419077434026556348926849462319593987388722641087929151
77923217503127424739845993717998774423697977686011492574283191373461778802465982439904936571260668303811205549003649459184011268085492752762680354601959856409286
58848613836574638386269700914807242321675105850970263883083980114265121348754085555879762577365869431426749190367069975892233454152758813016136172881051306725444
37900417308534362240424974490984245862445971816447179869274165567363032577105112346141906824701594918517289895336874254789003682192531735831548113676788989948453
41479891096195940327196050202547016132618328621222235329027555083180465738368128511057728397374402864929666157084447740568737232648376963582673185491021973350015
000461314948547572861766427281602376241995444071677831704805588443519630072147555462081795362307637600108 models

Pre-selection of candidate variables (both environmental and species traits)

- Include only those that make ecological sense, not all just because they were measured.
- Remove highly correlated variables by choosing only one of them, or using summary variables such as PCAs or means (but see for RRR later).
- You may run a preliminary analysis e.g. by RDA or species richness analyses to see which variables seem to explained a largest part of the variation.

Model/variable selection strategies

- You may consider also alternative model structures, not only alternative sets of predictors (e.g. models with or without spatial random effects).
- You may decide to keep some of the candidate variables in any case (whether supported or not), e.g. those that control for variation in sampling effort.
- You may select variables hierarchically, e.g. first asking whether “some of the five climatic variables” should be included in the model (models with all and without any climatic variables), and if yes, then examining which of those five explain most variation to select only a subset.

Specific techniques implemented in HMSC for model comparison / variable selection

- Cross-validation (which model has the highest predictive power)?
- WAIC (which model has the lowest WAIC?)
- Variable selection by spike and slab prior
- Reduced rank regression

Variable selection by spike and slab prior

“Usual” prior:

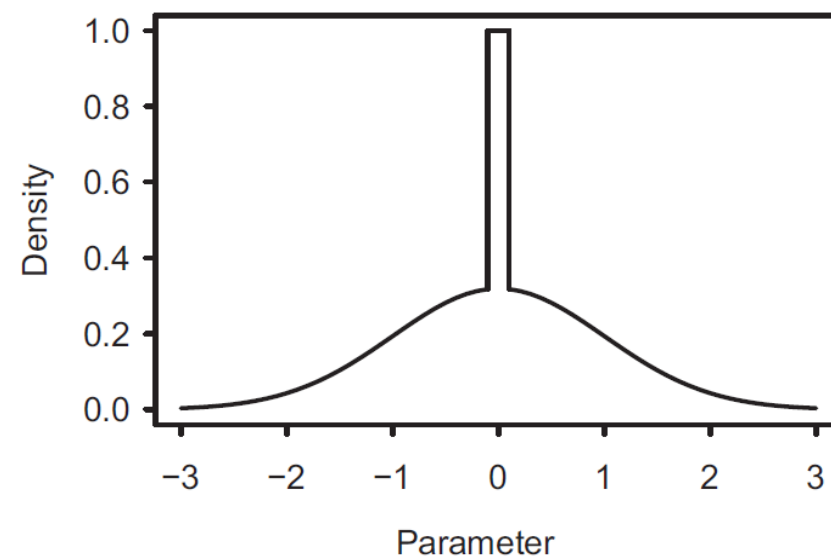
$$\beta \sim N(0, \sigma^2)$$

Spike and slab prior:

$$\beta = p\hat{\beta}$$

$$p \sim \text{Bernoulli}(q)$$

$$\hat{\beta} \sim N(0, \sigma^2)$$



HMSC implements spike and slab prior for the β -parameters

9.4.2 Simulated Case Study with HMSC

Hmsc(Y=...,
XData=...,
XFormula=...,
XSelect = ...,
...)

```
qq = 0.1 #prior probability for a covariate to be included

#1: No variable selection
XSelect.FULL = NULL

#2: Variable selection jointly for all species
XSelect.JOINT = list()
for (k in 2:nc){
  covGroup = k
  spGroup = rep(1, ns)
  q = rep(qq, max(spGroup))
  XSelect.JOINT[[k-1]] = list(covGroup = covGroup,
    spGroup = spGroup, q = q)
}

#3: Variable selection separately for each species
XSelect.SEPARATE = list()
for (k in 2:nc){
  covGroup = k
  spGroup = 1:ns
  q = rep(qq, max(spGroup))
  XSelect.SEPARATE[[k-1]] = list(covGroup = covGroup,
    spGroup = spGroup, q = q)
}
XSelect = list(XSelect.FULL, XSelect.JOINT, XSelect.SEPARATE)
```

HMSC implements spike and slab prior for the β - parameters

9.4.2 Simulated Case Study with HMSC

Explanatory power:

##			Model FULL	Model JOINT	Model SEPARATE
##	Data	FULL	0.929	0.822	0.729
##	Data	JOINT	0.712	0.470	0.603
##	Data	SEPARATE	0.713	0.286	0.587

Predictive power:

##			Model FULL	Model JOINT	Model SEPARATE
##	Data	FULL	0.375	0.228	0.263
##	Data	JOINT	0.154	0.284	0.074
##	Data	SEPARATE	0.099	0.067	0.106

Reduced Rank Regression (RRR)

- PCA (and other such techniques) can in a first step be applied to the original covariates to reduce their dimensionality & collinearity, and then in a second step use a small set of dominating PCAs as predictors in \mathbf{X} .
- RRR combines these two steps so that instead explaining most variation in \mathbf{X} , the dimension-reduced covariates explain as much variation as possible of \mathbf{Y}

Reduced Rank Regression (RRR)

$$y_{i,t} = c_i + \sum_{j=1}^m \alpha_{i,j} y_{j,t-1} + e_{i,t},$$

PROCEEDINGS B

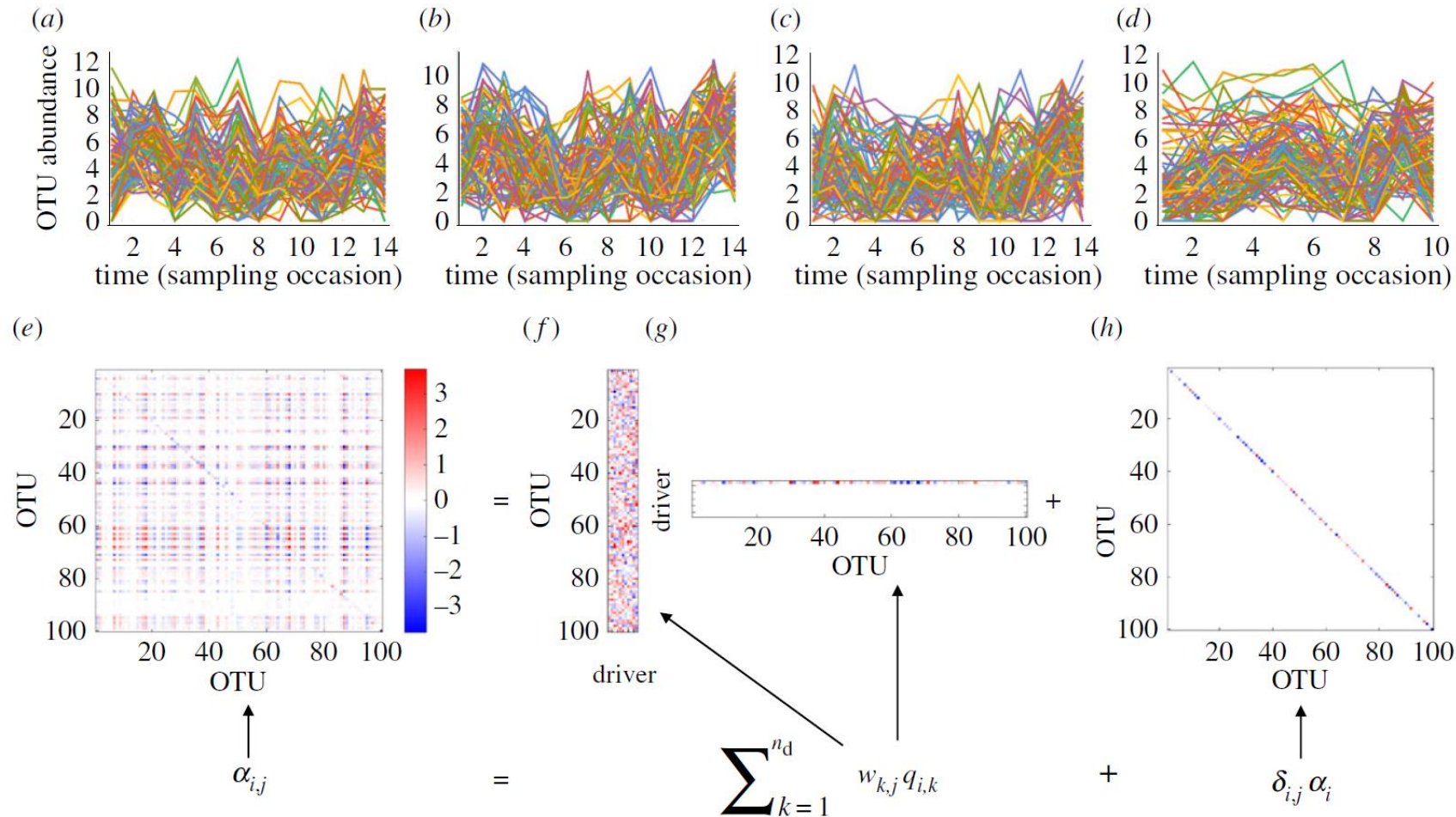
rsph.royalsocietypublishing.org

Research



How are species interactions structured in species-rich communities? A new method for analysing time-series data

Otso Ovaskainen^{1,2}, Gleb Tikhonov¹, David Dunson³, Vidar Grøtan², Steinar Engen⁴, Bernt-Erik Sæther² and Nerea Abrego^{2,5}

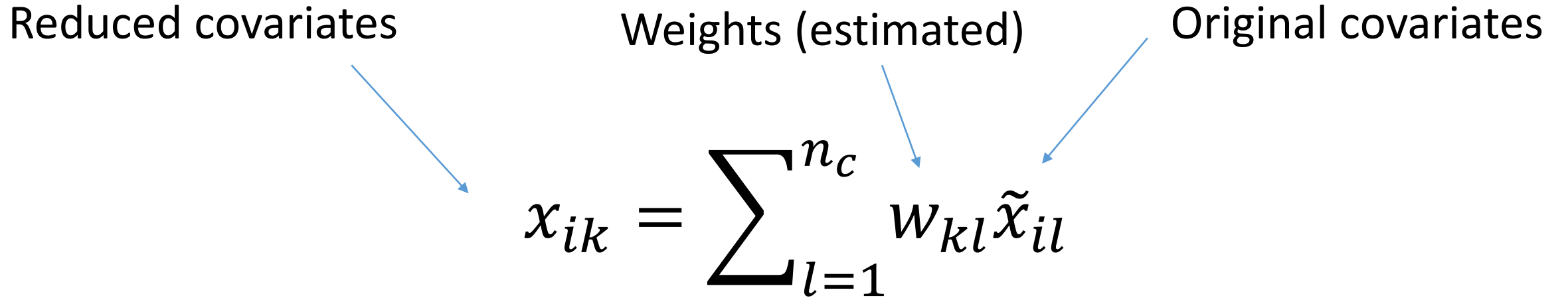


Reduced Rank Regression (RRR)

Reduced covariates

Weights (estimated)

Original covariates



The diagram illustrates the Reduced Rank Regression (RRR) equation. Three labels are positioned above the equation, each with a blue arrow pointing to a specific part of the formula: 'Reduced covariates' points to x_{ik} , 'Weights (estimated)' points to w_{kl} , and 'Original covariates' points to \tilde{x}_{il} .

$$x_{ik} = \sum_{l=1}^{n_c} w_{kl} \tilde{x}_{il}$$

$$k = 1, \dots, n_c^{RRR}$$

Prior parameters for RRR

$$w_{kl} \mid \phi_{kl}^{RRR}, \delta \sim N\left(0, \left(\phi_{kl}^{RRR}\right)^{-1} \left(\tau_k^{RRR}\right)^{-1}\right), \tau_k^{RRR} = \prod_{h=1}^k \delta_h^{RRR} \quad (9.9)$$

$$\phi_{kl} \mid v \sim \text{Ga}(v^{RRR}/2, v^{RRR}/2) \quad (9.10)$$

$$\delta_1 \sim \text{Ga}(a_1^{RRR}, b_1^{RRR}), \delta_h \sim \text{Ga}(a_2^{RRR}, b_2^{RRR}) \text{ for } h \geq 2 \quad (9.11)$$

As default values of the prior parameters, Hmsc assumes $v^{RRR} = 3$, $a^{RRR} = (1, 50)$ and $b^{RRR} = (1, 1)$.

HMSC implements RRR for the β - parameters

9.5.1 Simulated Case Study with HMSC

Hmsc(Y=...,
XData=...,
XFormula=...,
XRRRData=...,
XRRRFormula=...,
ncRRR=...,
...)

```
models = list()
for(dataset in 1:3){
  tmp = list()
  for (model in 1:3){
    switch(model,{
      m = Hmsc(Y = Y[[dataset]], XData = XData,
              XFormula = ~., distr = "normal")
    },
    {
      pc = princomp(XData)
      XData.PC = data.frame(pc$scores[,1])
      m = Hmsc(Y = Y[[dataset]], XData = XData.PC,
              XFormula = ~., distr = "normal")
    },
    {
      m = Hmsc(Y = Y[[dataset]], XData = XData, XFormula = ~1,
              XRRRData = XData, XRRRFormula = ~.-1, ncRRR=1,
              distr = "normal")
    }
  )
}
```

HMSC implements RRR for the β -parameters

9.5.1 Simulated Case Study with HMSC

Explanatory power:

##		Model FULL	Model PC	Model RRR
##	Data FULL	0.929	0.108	0.431
##	Data PC	0.667	0.374	0.392
##	Data RRR	0.797	0.070	0.607

Predictive power:

##		Model FULL	Model PC	Model RRR
##	Data FULL	0.491	-0.113	0.093
##	Data PC	0.155	0.259	0.228
##	Data RRR	0.300	-0.011	0.416