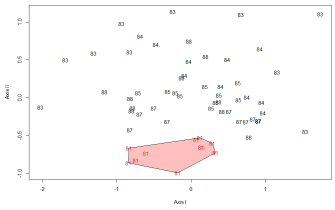
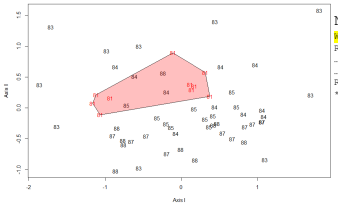


```
NMDS
autotransform=F
trymax=500

Run 40 stress 0.1847481
... Procrustes: rmse 0.0004282455  max
resid 0.00212223
... Similar to previous best
```



```
NMDS (defaults)
Wisconsin double standardization
Run 0 stress 0.1977875
...
...
Run 20 stress 0.2092379
*** No convergence -- monoMDS stopping criteria:
    20: stress ratio > sratmax
```

Model Selection

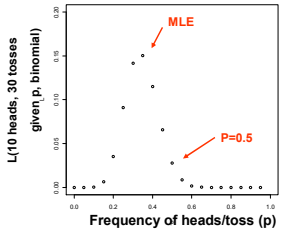
- Models are meant to be a representation of the real world, they are:
  - **Imperfect** – always some uncertainty
  - **Simplified** – goal is to represent as much of the real world as possible in a way that is understandable
- Tradeoff between model fit and model parsimony

Model Selection

- Given the observed data, we can quantify model fit in a number of ways
  - R^2
  - RMSE
  - Maximum Likelihood – the most likely set of parameters and model given the data

10 heads from 30 tosses of a coin,

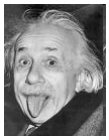
10/30 = 0.3 → MLE



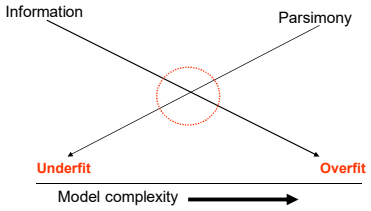
Model Selection

- Parsimony ~ model simplicity
  - **Occam's razor** -14th-century English Franciscan friar (William of Ockham) "shave away all that is unnecessary"
  - **Einstein** – "Everything should be made as simple as possible, but no simpler"

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \frac{8\pi G}{C^A}T_{\mu\nu}$$



Model Selection



"When multiple competing hypotheses are equal in other respects, the principle recommends selecting the hypothesis that introduces the fewest assumptions and postulates the fewest entities."

Overfit

- Include unnecessary variables
- Lack of parsimony
- Difficult interpretation

Underfit

- Lack important variables
- Minimal explanatory power
- Inaccurate parameter estimates

Goal is to strike a balance between these two.

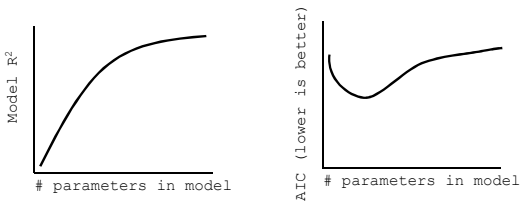
Parsimony – Fit in CCA

- We used similar logic earlier in CCA
- Stepwise regression and VIF score examination were aimed at eliminating redundancy (avoid overfit)
- Selecting the best variables and maximizing % variance explained (avoid underfit)

Model Selection and Parsimony in Ecology

- Much of ecology is observational
- Hypotheses often of the form: "which of these observed variables best explains the observed ecological process or pattern"
- **Traditional approach**
  - H_o: response variable y differs significantly with respect to variables a, b and c
  - Test the significance of H_o
- **Information-theoretic approach**
  - What set of observed variables form the most parsimonious and accurate model predicting the response variable
  - No significance test

Akaike Information Criterion – Parsimony vs. Fit



Akaike Information Criterion

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null Hypothesis Testing: Problems, Prevalence, and an Alternative. *Journal of Wildlife Management* **64**(4): 912-923.

- Mathematical estimate of the balance between parsimony and fit of a model.

$$AIC = \underbrace{-2\log\big(L(model|data)\big)}_{\text{Log likelihood of model given the data}} + \underbrace{2k}_{\text{Model complexity}}$$

Akaike Information Criterion

- There are modifications to this (weights for variables, sample size adjustments etc.), but the general idea is the same.
  - AIC_c
  - BIC
- Minimizing AIC = finding the set of explanatory variables that explain the most about the response variable(s).
- Addition of parameters is penalized
- **Uses**
  - Compare models (each model represents a competing hypotheses)
  - Compare levels of treatments

AIC in R

- Function AIC() will calculate AIC for any model
- Simple application
  - Response variable A
  - Factors B, C and D
  - Compare AIC values for models
    - A ~ B*C*D (global)
    - A ~ B*C
    - A ~ B*D
    - A ~ B
    - A ~ 1 (null)
- The absolute size of AIC values relative to **other competing models** is what matters.
- AIC values **not** comparable across studies

AIC in Ecology

- Ask good questions
- Conduct well designed studies
- Develop a set of ***plausible*** models to explain data
- Calculate AIC values for this set of models
- Rank models by AIC value (lowest to highest)
- Compute the difference in AIC values between the first and each subsequent model
  - AIC difference $\Delta_i = AIC_i - AIC_{min}$
  - Δ_i values: 0-2 = good support, 4-7 = less support, >10 none
- **Remember, AIC values within a set of competing models are only comparable among themselves**

AIC in Ecology

- AIC weights

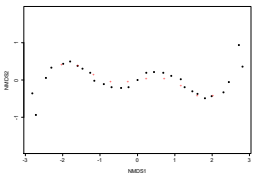
$$w_i = \frac{e^{-0.5*\Delta_i}}{\sum e^{-0.5*\Delta_i}}$$

- w_i = the weight of evidence for a particular model relative to other models
- Typical AIC table:

Model g_i[a]	K_i	Δ_i	w_i
$\{\phi_l\ p.\}$	3	0.00	0.673
$\{\phi_{s+l}\ p.\}$	4	2.07	0.239
$\{\phi_{s*l}\ p.\}$	5	4.11	0.086
$\{\phi.\ p.\}$	2	12.71	0.001
$\{\phi_s\ p.\}$	3	14.25	0.001
model averaged			

Model Comparison, example dataset

- NMDS of community data
- What environmental data best predicts axis 1 scores?
- Global model:




```
Call:
lm(formula = metanmds$points[, 1] ~ ., data = envdata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.37338 -0.11365 -0.01083  0.14677  0.25822

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.135265   0.427212  -7.339 3.68e-06 ***
spatialb     0.178619   0.243453   0.734   0.4752
spatialc     0.148344   0.444862   0.333   0.7437
env1        -0.300768   0.261051  -1.152   0.2686
env2        -0.544517   2.419253  -0.225   0.8252
env3         0.292868   0.165836   1.766   0.0992 .
env4         0.003002   0.227521   0.013   0.9897
env5        -0.046674   0.222875  -0.209   0.8371
env6        -0.192800   0.222696  -0.866   0.4012
env7         6.516912   0.722254   9.023 3.29e-07 ***
env8         0.018205   0.233328   0.078   0.9389
env9         0.169447   0.181752   0.932   0.3670
env10        0.488443   2.456476   0.199   0.8452
env11       -0.246350   0.198231  -1.243   0.2344
env12       -0.211397   0.200459  -1.055   0.3095
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.24 on 14 degrees of freedom
Multiple R-squared: 0.9906,   Adjusted R-squared: 0.9813
F-statistic: 105.7 on 14 and 14 DF,  p-value: 1.037e-11
```

Model Comparisons

- **Competing models (each should represent a hypothesis):**
  - Null – no variables
  - Global – all variables
  - Spatial variable
  - Environmental variable 7
  - Spatial and environmental variable 7
- Function AICctab in package bbmle

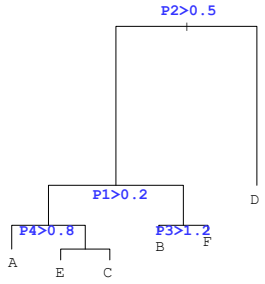
```
models<-list(Null,Spatial,Seven,Spatial*Seven,Global)

AICctab(models,nobs=29,mnames=c("Null","Spatial","Seven","Spatial*Seven","Global"),base=T,weights=T,delta=T,logLik=T)
```

	logLik	AICc	dLogLik	dAICc	df	weight
Seven	2.2	2.5	59.3	0.0	3	0.69
Spatial*Seven	7.6	4.1	64.6	1.6	7	0.31
Global	10.7	55.9	67.7	53.4	16	<0.001
Spatial	-25.7	61.1	31.3	58.6	4	<0.001
Null	-57.0	118.5	0.0	116.0	2	<0.001

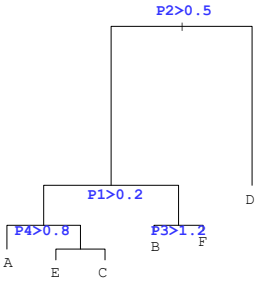
Classification Trees and Regression Trees

- Classification Trees
  - Descriptive technique
  - Multiple samples are categorized (by treatment, location, presences of species etc) into A-F.
  - Tree is built (recursive partitioning)
  - Eg. goal is to describe differences in samples A-F in relation to predictor variables **(P1-P5)**



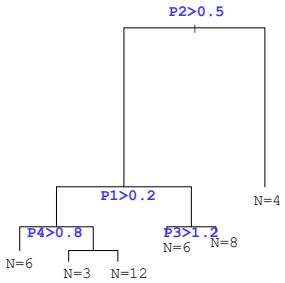
Classification Trees and Regression Trees

- Classification Tree
  - samples are divided into **discrete groups**
  - At each node, analysis finds a cutoff point for a variable
  - Accuracy measured as proportion classified correctly
- Example
  - Community data across sites. Sites divided into groups (A-F) based on geographic location. Want to describe what species (P1-P5) best differentiate groups.



Classification Trees and Regression Trees

- Regression Tree
  - Continuous response variable. Samples **not in discrete groups**.
  - At each node, analysis seeks cutoff values that group samples based on similarity in response variable.
  - Accuracy measured as homogeneity within tips.
  - Reports number of samples classified at the end of each branch.
- Examples
  - Community productivity measured at sites. Want to describe what species (P1-P5) best predict productivity levels.



Classification Tree and Regression Trees

- Basic algorithm, at each iteration
  - Continuous variables transformed to ranks (non-parametric)
  - First node is formed by:
    - Partitioning the data by levels of each variable
    - Choose the one that partitions to produce the most homogeneous groups.
  - Within each of the two existing clusters, repeat the above process
  - Iterative procedure, repeat until additional partitioning does not reduce within group heterogeneity
  - One variable can appear multiple times

Classification Tree and Regression Trees

- Function rpart (rpart package)

```
regtree<-
rpart(metanmds$points[,1]~spatial+env1+env2+en
v3+env4+env5+env6+env7,data=envdata)
```

- Formula specifies the response variable (continuous for regression trees, a factor for classification trees) and the independent variables.
- Options
  - Size= specify number of tree nodes (if not specified, tries to guess at stopping point to avoid overfitting)
  - Numerous options for controlling when the iterations stop and how clusters are formed. See ?rpart for details.

Classification Tree and Regression Trees

```
n= 29

node), split, n, deviance, yval
      * denotes terminal node

1) root 29 86.755100  1.531342e-17
  2) env7< 0.4530161 14  9.054051 -1.565784e+00 *
  3) env7>=0.4530161 15 11.342280  1.461398e+00 *
```

Boosted Regression Trees

- **Boosting** (aka machine learning) – a series of simple but poor models (recall parsimony-fit tradeoff) combined are better than a single good model.
- Boosted algorithms common in informatics (search engines, speech recognition etc.)
- Boosted regression trees (package gbm)
  - Series of simpler (more parsimony, less fit) trees combined

Boosted Regression Trees

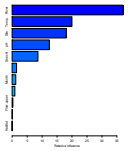
- Iterative process as before…
  - Build a regression tree as before
  - Calculate residual variation, this represents samples in the analysis that are not classified in the first tree
  - Build a regression tree classifying residuals (thus directly focusing on aspects of the dataset not addressed in first tree)
  - Combine trees
  - Repeat last three steps until a stopping point is reached.
- Often run with cross-validation (randomly select a subset of data for each iteration)

Boosted Regression Trees

- Control over the iterative process:
  - Tree complexity – how many nodes at each iterative step
  - Learning rate – contribution of each tree to the overall model.
  - Number of trees – how many trees can be combined in the overall model.
- Function gbm in package gbm

```
gbm(pascagoula$Cond. ~ Flow.upper + Flow.lower + Temp. + Seechi + DO + Depth + pH + Month +
River + Site + Habitat, data=pascagoula, distribution="gaussian", n.trees=5000)

                var     rel.inf
River           River 37.04167385
Temp.           Temp. 19.97221362
Site             Site 18.09427402
pH                 pH 12.41140162
Seechi         Seechi  8.63823356
Flow.lower Flow.lower  1.47180382
Month           Month  1.20404794
DO                 DO  0.89286338
Flow.upper Flow.upper  0.25674082
Depth           Depth  0.01674737
Habitat       Habitat  0.00000000
```



Reading

- Text: section 4.11
- Papers
  - Anderson, D.R., Burnham, K.P., and Thompson, W.L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. The Journal of Wildlife Management 64, 912-923.
  - **FYI:** Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal Of Animal Ecology 77:802-813.
  - Carassou, L., F. J. Hernandez, S. P. Powers, and W. M. Graham. 2012. Cross-shore, seasonal, and depth-related structure of ichthyoplankton assemblages in coastal Alabama. Transactions of the American Fisheries Society 141:1137-1150.
  - Sample script
    - AIC model comparison (example dataset)
    - Regression and classification trees (example dataset and Pascagoula River habitat data)
    - Boosted regression tree (Pascagoula river habitat)

Homework

- Homework
  - Ecological traits as predictors of stream/reservoir abundance

Species	Species	Ecological Traits
Reservoirs	Streams	Species
30 years of community data from 28 reservoirs	30 years of community data from 200+ streams and rivers	Goldstein and Meador (2004) – 32 traits for 429 N. American species

  - Final matrix to analyze
    - Species by ecological traits matrix
    - Response variable – log(resratio) = the ratio of mean reservoir abundance to mean stream abundance
  - What traits are the best predictors of reservoir abundance?
    - Use AIC to compare at least 5 models (including one null and one global)
    - Regression tree (boosted or non-boosted)

species	log_resrat	Herb_det	plank	omni	invert	carniv	bedrock	boulder
NOTSHU	0.000221	0	0	0	0	0	0	0
SEMATR	0.000291	0	0	0	1	1	0	0
FUNZEB	0.003205	0	0	0	0	0	0	0
NOTGRE	0.00401	0	0	0	0	0	0	0
NOTRUB	0.004355	0	0	1	0	0	1	1
FUNOLI	0.004491	0	0	1	0	0	0	0
ETHZON	0.007036	0	0	0	1	0	0	0
NOTSTR	0.007879	0	0	1	0	0	0	0
HYBPLA	0.008623	1	0	0	0	0	0	0
HYPNIG	0.009482	0	0	1	0	0	0	0
LYTUMB	0.010176	0	0	0	0	0	0	0
NOTPOT	0.010202	0	0	0	1	1	0	0
PIMPRO	0.011851	0	0	1	0	0	0	0
LEPMAR	0.01303	0	0	0	1	0	0	0
ETHRAD	0.016688	0	0	0	1	0	0	0

104 species

Log ratio of mean reservoir to stream abundance. Higher=species more abundant in reservoirs.

code	full name
NOTSHU	NOTROPIS SHUMARDI
SEMATR	SEMOTILUS ATROMACULATUS
FUNZEB	FUNDULUS ZEBRINUS
NOTGRE	NOTROPIS GREENEI
NOTRUB	NOTROPIS RUBELLUS
FUNOLI	FUNDULUS OLIVACEUS
ETHZON	ETHEOSTOMA ZONALE
NOTSTR	NOTROPIS STRAMINEUS
HYBPLA	HYBOGNATHUS PLACITUS
HYPNIG	HYPENTELIUM NIGRICANS
LYTUMB	LYTHRURUS UMBRATILIS
NOTPOT	NOTROPIS POTTERI
PIMPRO	PIMEPHALES PROMELAS
LEPMAR	LEPOMIS MARGINATUS
ETHRAD	ETHEOSTOMA RADIOSUM
PERSCI	PERCINA SCIERA
CARAUR	CARASSIUS AURATUS
ETHNIG	ETHEOSTOMA NIGRUM
CYPLUT	CYPRINELLA LUTRENSIS
MOXERY	MOXOSTOMA ERYTHRURUM
LUXCAR	LUXILUS CARDINALS
NOTBOO	NOTROPIS BOOPS
PHEMIR	PHENACOBIUS MIRABILIS

I added a file that has full species names. You can use this however you want…

Ecological/Life History traits for each species by category

Diet
- Herb_det
- Plank
- Omni
- Invert
- carniv

Substrate
- Bedrock
- Boulder
- Cobble
- Gravel
- Sand
- Mud
- Veg
- var_sub

Habitat
- Riffle
- Pool
- Run
- Backwat
- var_hab

Locomotion
- Cruis
- Accel
- Hugger
- Creeper
- Maneu

Reproduction
- Broad
- Nest
- Bearer
- migra

Original Habitat
- sm_ck
- sm_riv
- med_riv
- lg_riv

Frimpong, E.A., and Angermeier, P.L. (2009). FishTraits: A database of ecological and life-history traits of freshwater fishes of the United States. Fisheries 34, 487–495.

Goldstein, R.M., and Meador, M.R. (2004). Comparisons of fish species traits from small streams to large rivers. Trans Am Fish Soc 133, 971–982.