

Model Choice and Model Validation

AIC, BIC, DIC, cross validation
and all that

Ram's tip for today:

Generate models on part of the data, and test them on the rest.

“... the simple idea of splitting a sample in two and then developing the hypothesis on the basis of one part and the testing it on the remained may perhaps be said to be one of the most seriously neglected ideas in statistics.” G. A. Barnard as quoted in Common Errors in Statistics by Good and Hardin.

You need a theory.

- Without a working theory, then all deviations look the same. You cannot target the questionable data.

You need several theories:

- Chamberlin ([1890] 1965) advocated the concept of “multiple working hypotheses.” Here, there is no null hypothesis; instead, there are several well-supported hypotheses (equivalently, “models”) that are being entertained.
- Chamberlin, Thomas. [1890] 1965. “The Method of Multiple Working Hypotheses.” *Science* 148:754-9.

Methods of Validation

- Independent verification (new data from same or other population).
- Splitting sample (one for calibration , and the other for verification)
- Resampling (this is difficult with auto-correlated data).

Independent verification

- Best approach, use completely independent data to test model using new data, historic data, or data from other populations.

Sample splitting

- One part to generate hypotheses, one part to test
- Hold back $1/3$ or $1/4$ of data for validation
- This is REALLY simple, and thus easy to explain, and harder to make subtle errors (as it is easy to do in bootstrapping).
- There is a small loss of efficiency, but it is not too bad.

Resampling

- Bootstrapping (sampling with replacement)
- K-fold analysis (divide data into K samples)
- Leave one out analysis
- Jackknife (a version of leave one out, where the analysis is done with all data).
- Delete-d (see aside k% for testing).

In any resampling method suggests that model is unstable, then you reformulate model.

It is impossible for semi-complex data sets to generate and test hypotheses on the same data.

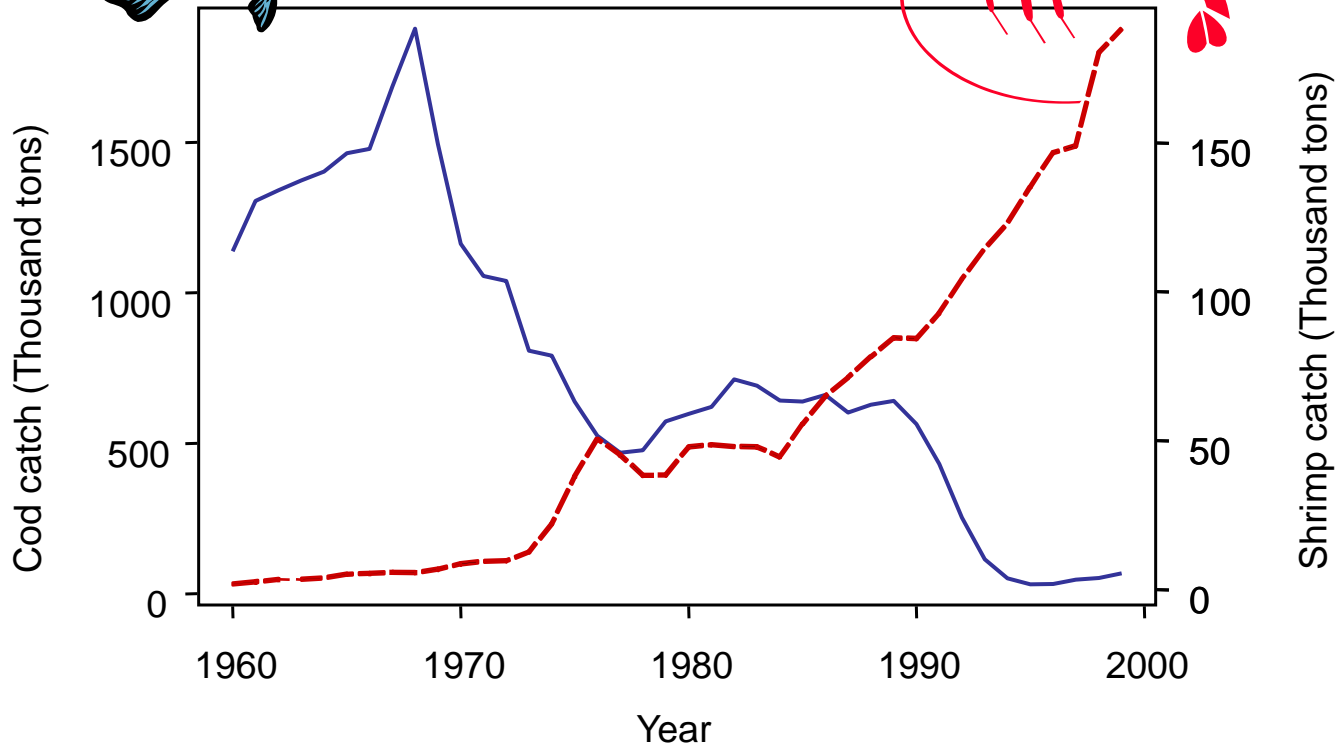
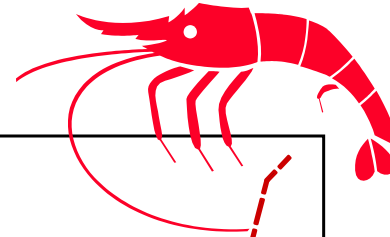
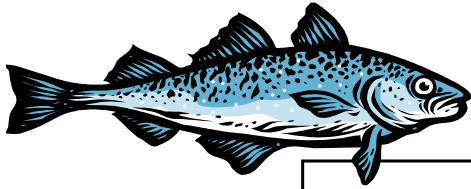
- An example, I am interested in what factors affect cod recruitment on Georges Bank.
- I have a 30 year time series of cod recruitment data.
- I have access to 10 environmental variables with data on each month.
- I consider 3 possible year lags.
- This is $10 \times 12 \times 3$ possible correlations with single variables.

- We have approximately $20 \cdot 12 \cdot 3$ squared 2 factor combinations

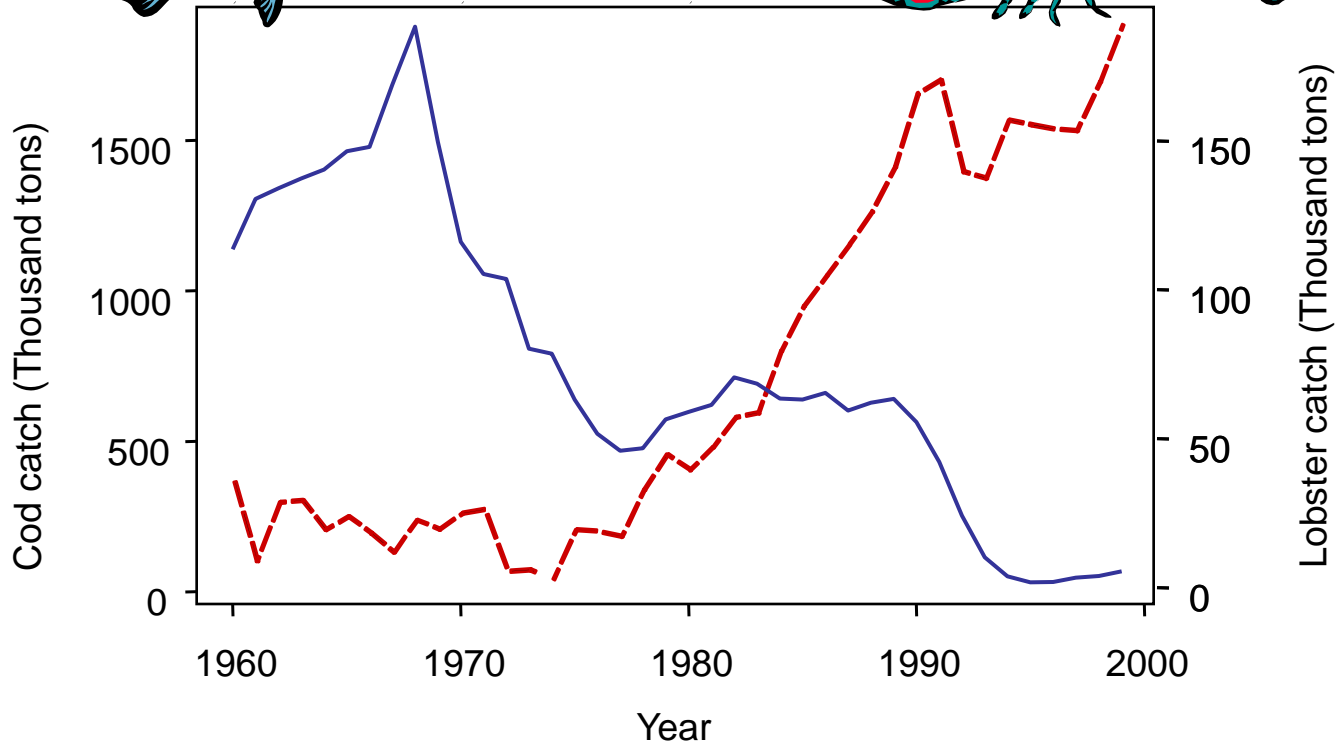
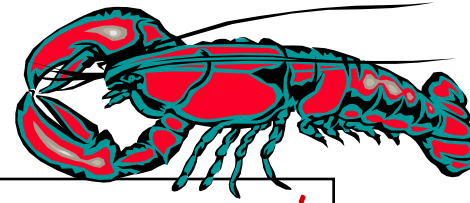
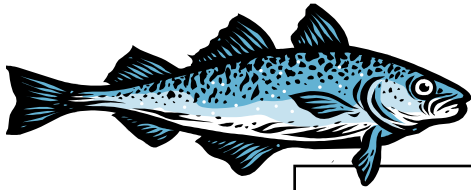
But it is much worse than that

- These time series are generally auto-correlated, i.e. there are fewer degrees of freedom than you think.

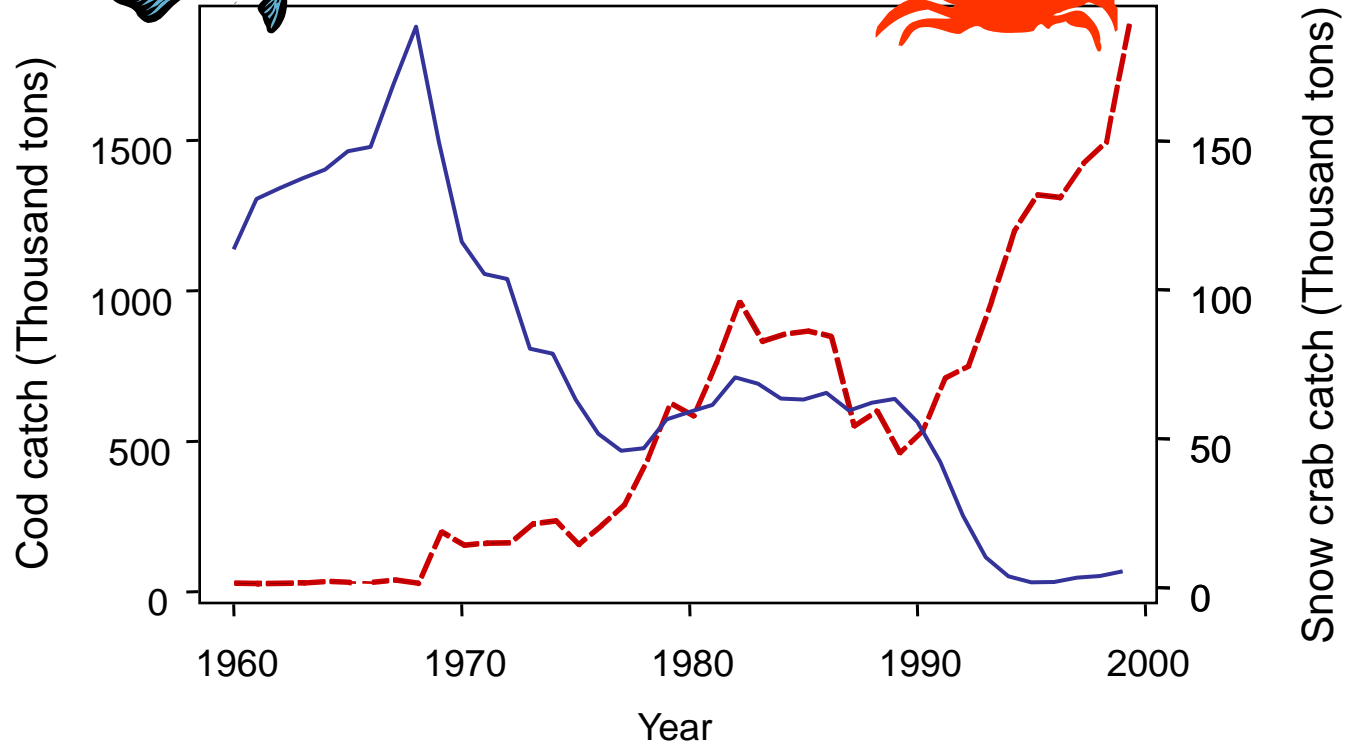
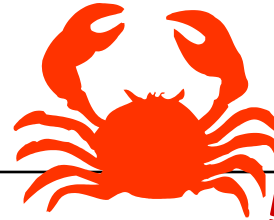
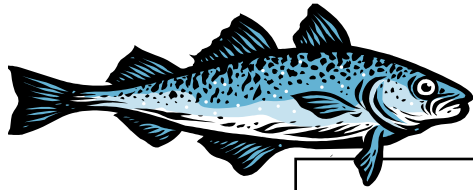
Cod versus shrimp catches in all NAFO areas combined



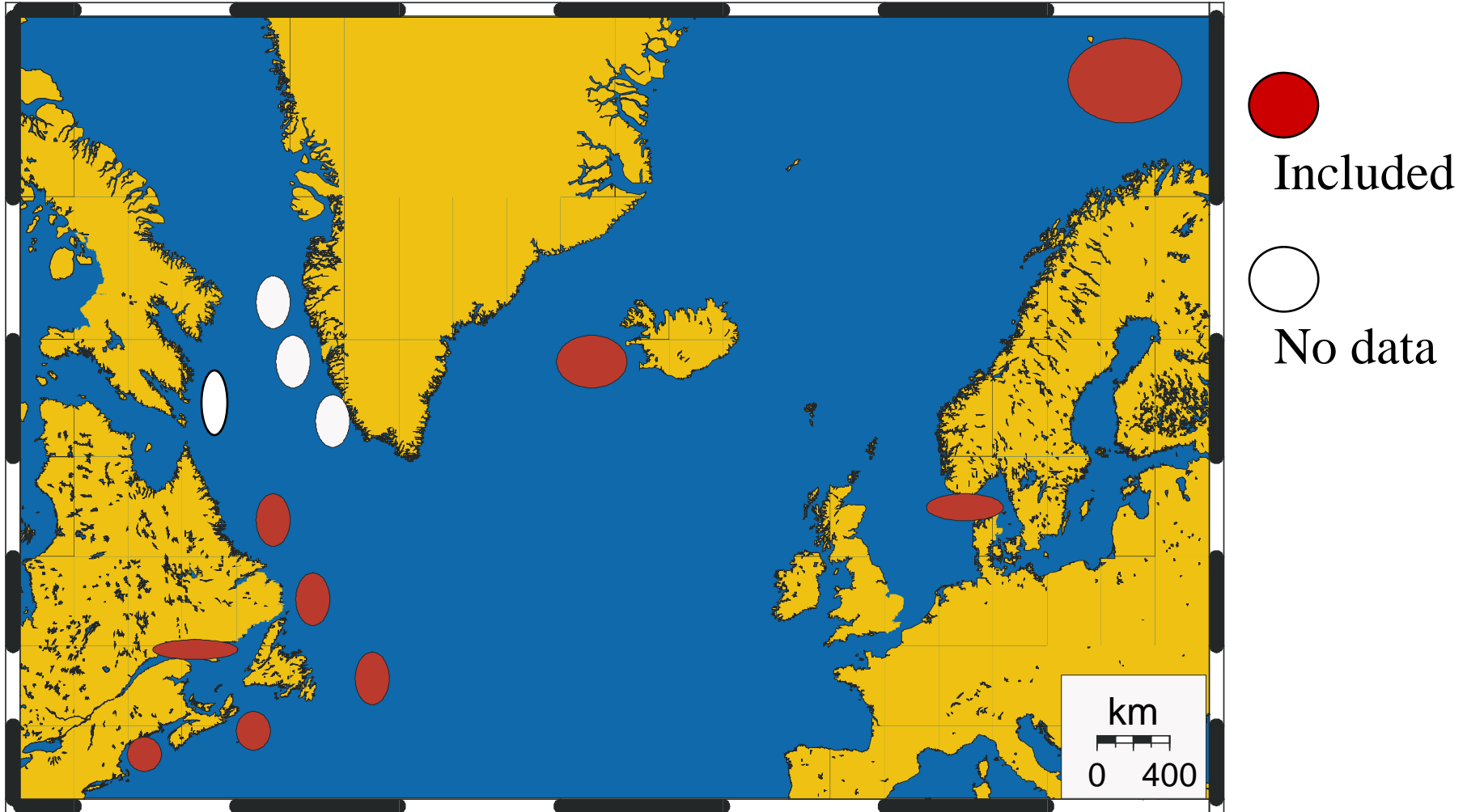
Cod versus lobster catches



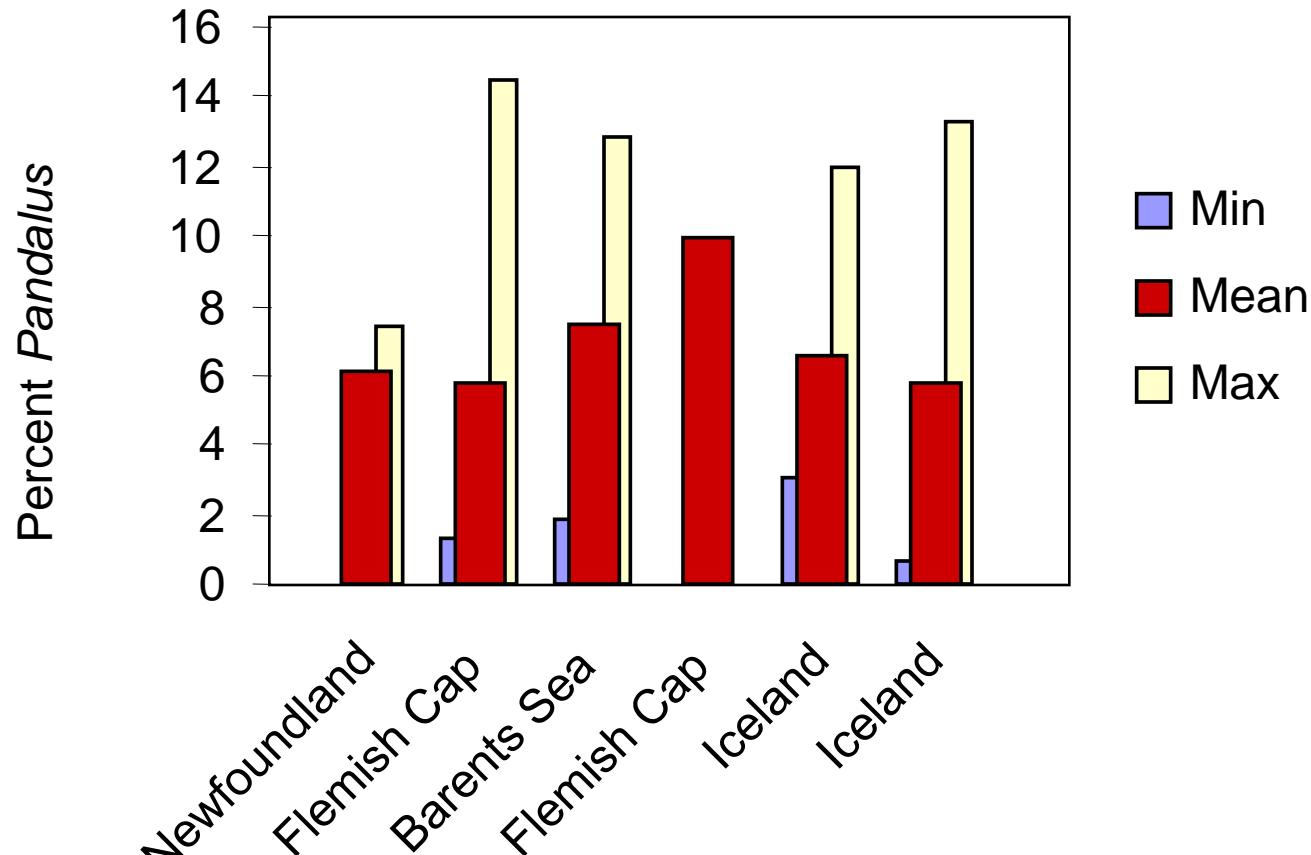
Cod versus crab catches



Major shrimp stocks in the North Atlantic

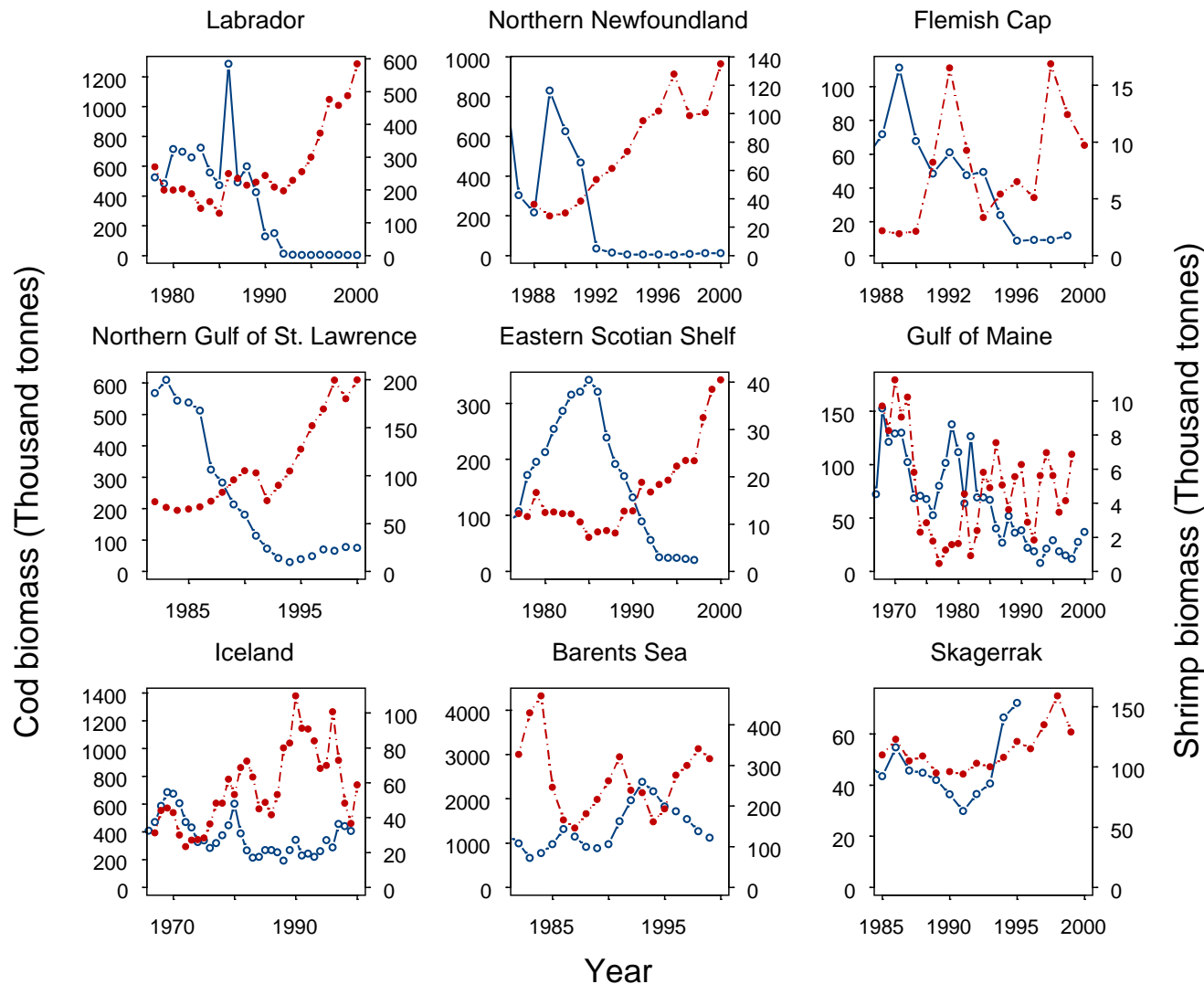


Similar cod diet across regions

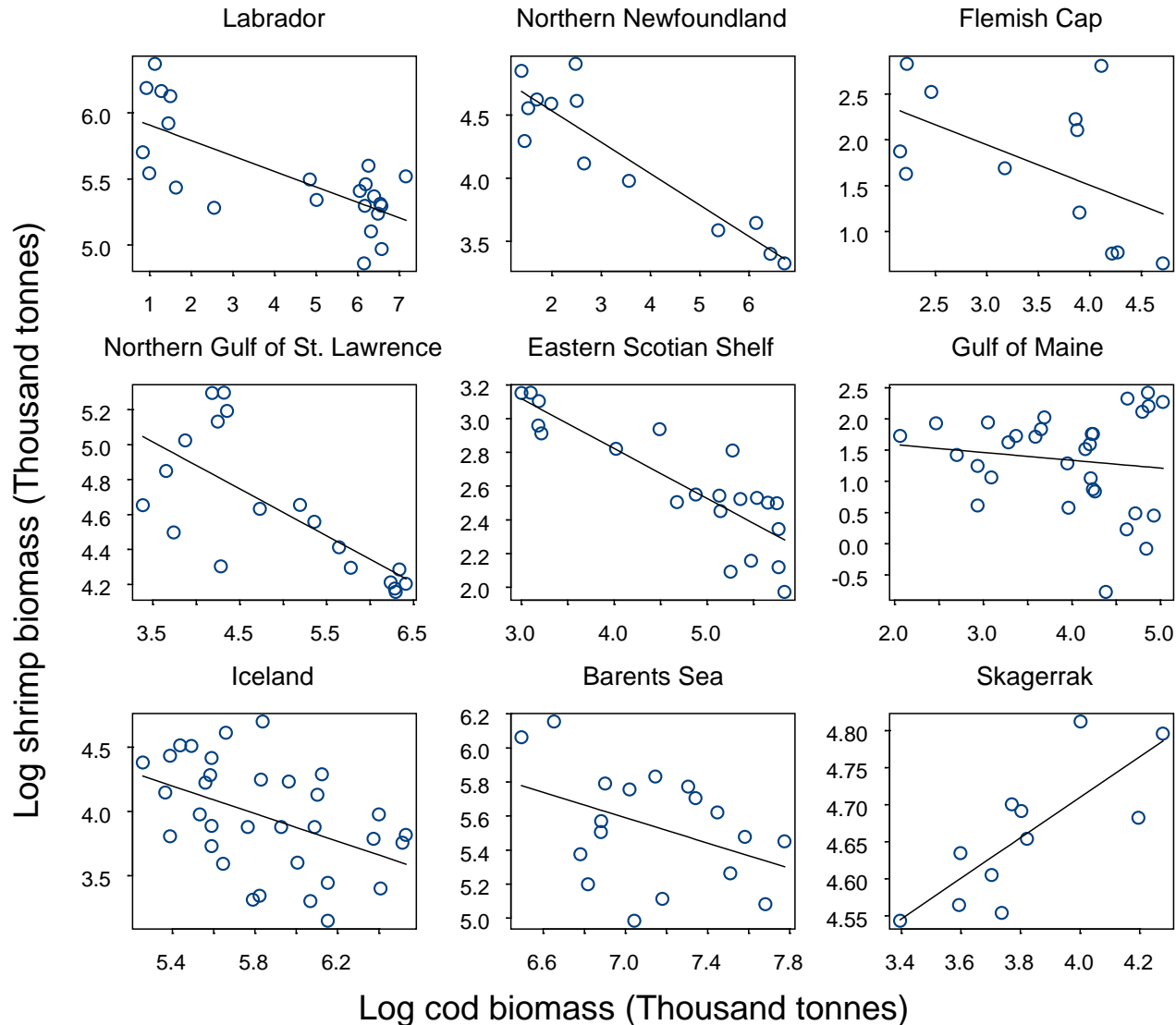


Source: Pálsson 1983, Boerje et al. 1987
Magnússon and Pálsson 1991, Rodríguez-Marín and del Río 1999
Lilly et al. 2000, Berenboim et al. 2000, Torres et al. 2000

Cod and shrimp biomass in the North Atlantic: time series



Cod and shrimp biomass in the North Atlantic: correlations



Step 1: Dealing with autocorrelation and measurement error

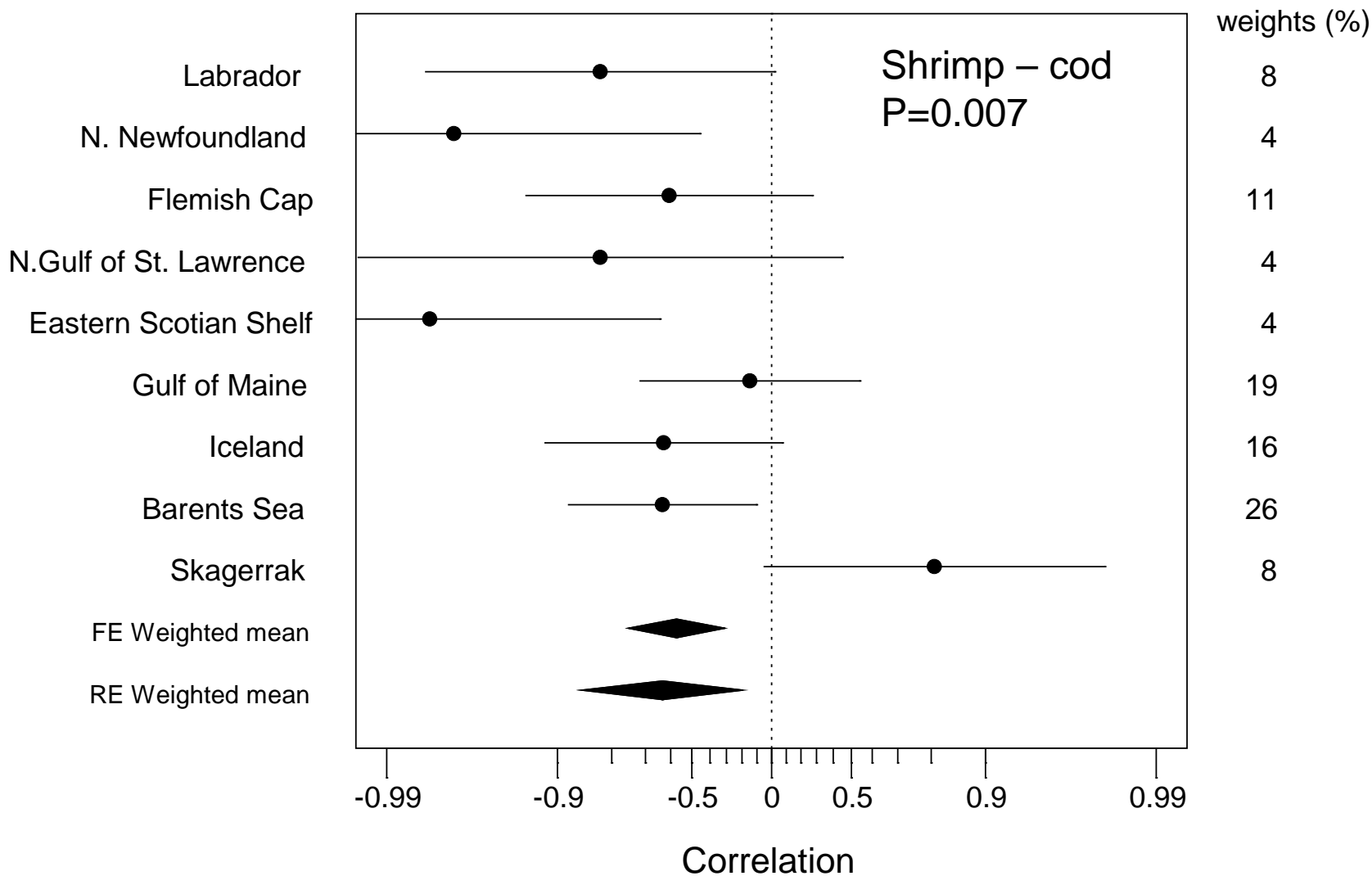
Simple analysis

Corrected analysis

Region	r	N	P	r^*	N^*	P^*
Labrador	-0.746	23	0.000	-0.827	4.8	0.054
N. Newfoundland	-0.911	13	0.000	-0.976	3.3	0.012
Flemish Cap	-0.526	12	0.073	-0.607	6.3	0.161
N.Gulf of St. Lawrence	-0.708	19	0.000	-0.827	3.4	0.165
Eastern Scotian Shelf	-0.856	21	0.000	-0.982	3.5	0.004
Gulf of Maine	-0.131	31	0.485	-0.147	9.3	0.701
Iceland	-0.459	33	0.006	-0.63	8.2	0.075
Barents Sea	-0.412	18	0.087	-0.635	11.7	0.023
Skagerrak	0.788	11	0.002	0.808	5.0	0.061

Source: Hedges & Olkin 1985, Pyper & Peterman 1998

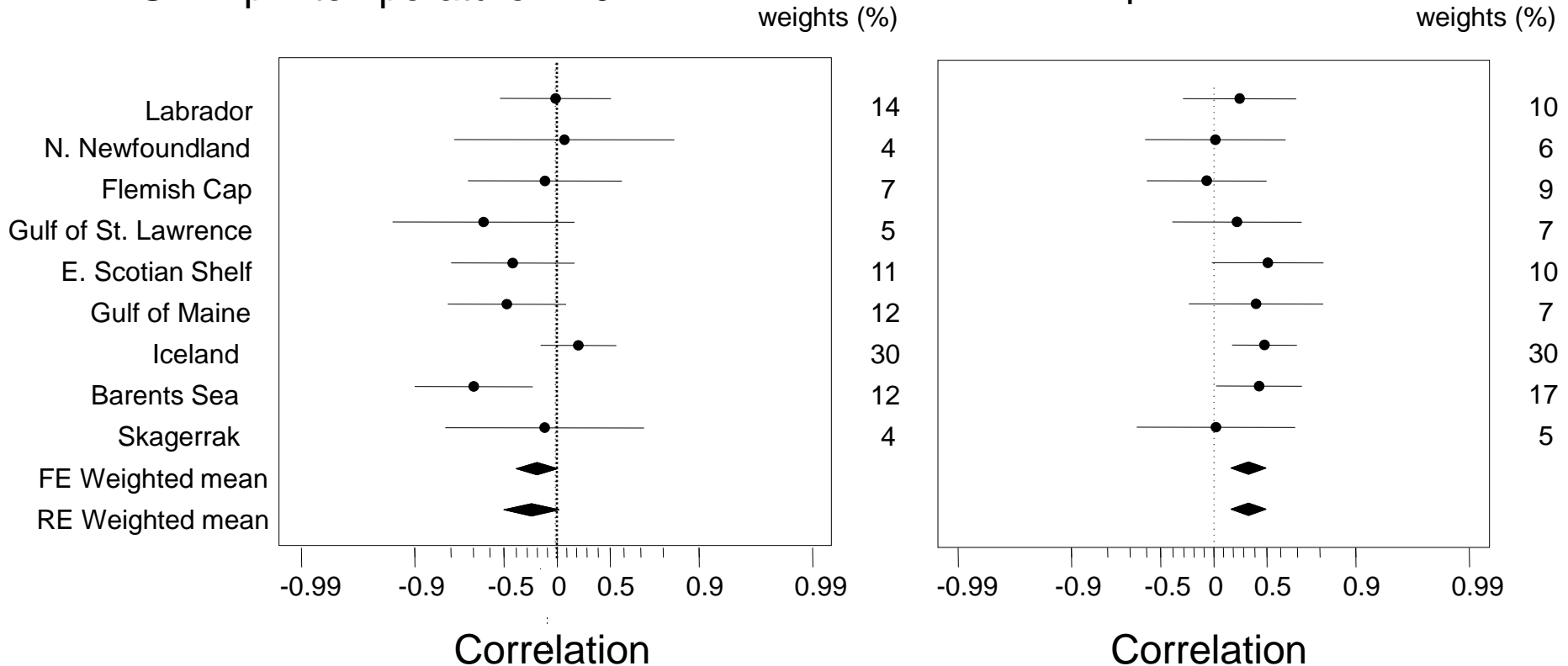
Step 2: Random-effects meta-analysis



Step 3: Testing environmental forcing

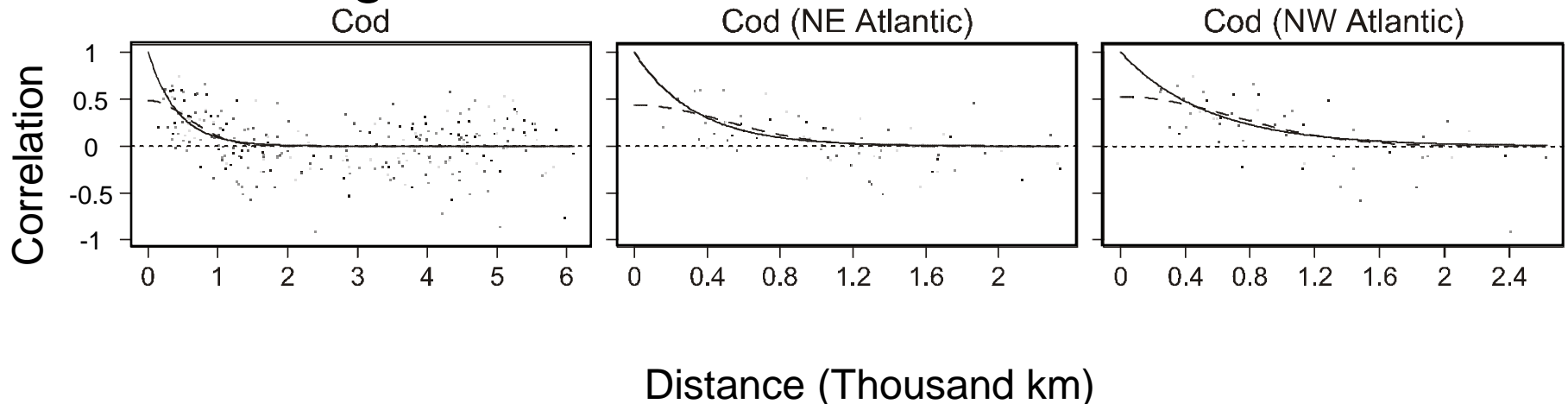
Shrimp – temperature $P=0.174$

Cod – temperature $P=0.001$



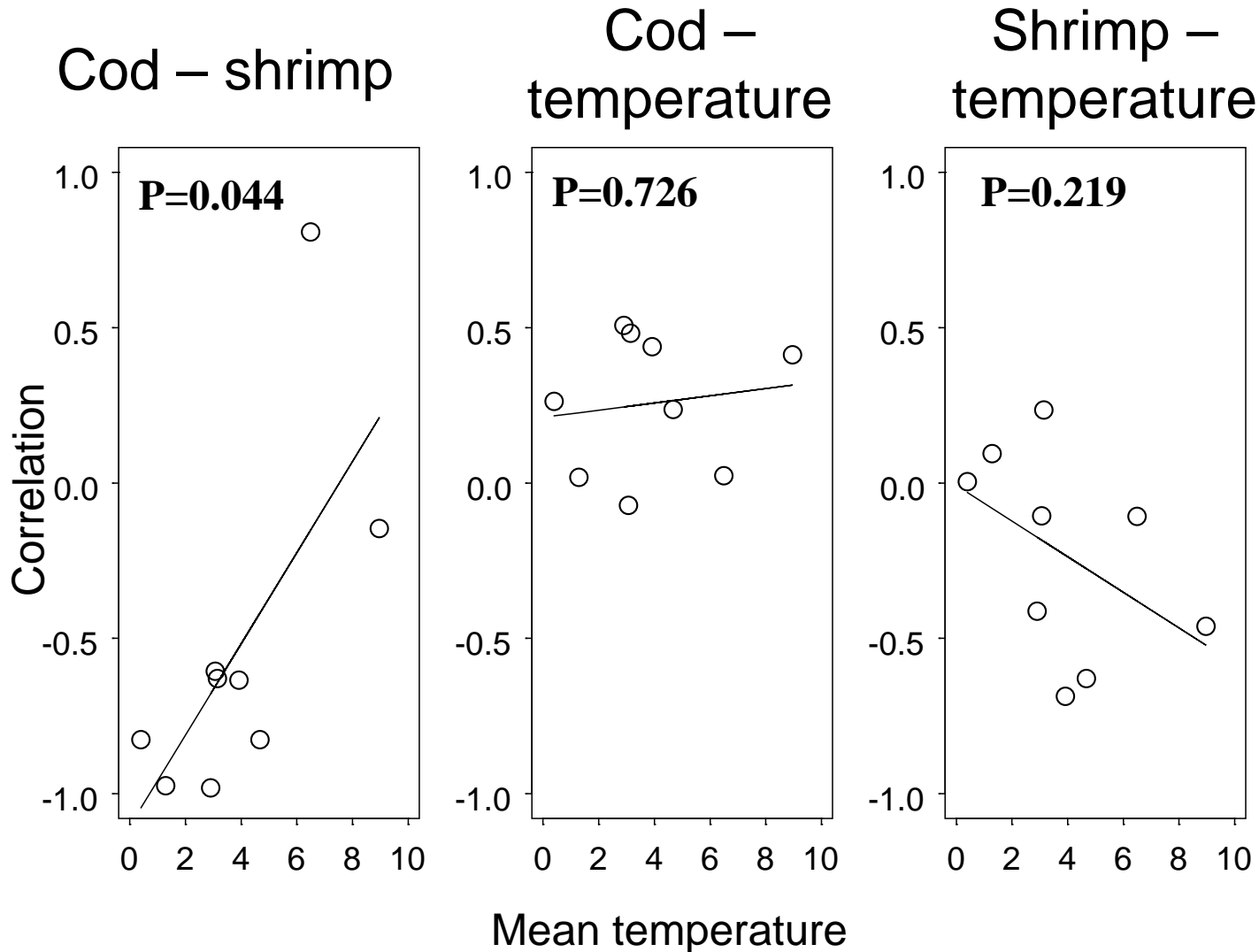
Step 4: Examining spatial correlation

- Cod recruitment is correlated on scales <500 km
- Stocks are not entirely independent
- Sensitivity analysis shows that this does not change results

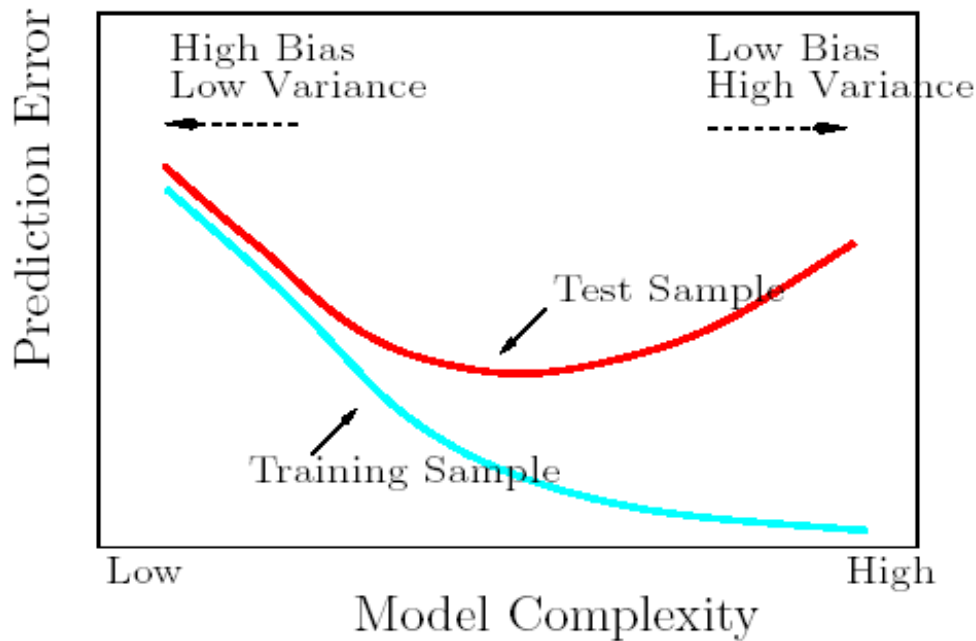


Source: Myers et al. 1997

Step 5: Testing for latitudinal gradients



Bias, Variance, and Model Complexity



- Bias-Variance trade-off again
- Generalization: test sample vs. training sample performance
 - Training data usually monotonically increasing performance with model complexity

Figure 7.1: Behavior of test sample and training sample error as the model complexity is varied.

Training Error

- Training error - Overfitting
 - not a good estimate of test error
 - consistently decreases with model complexity
 - drops to zero with high enough complexity

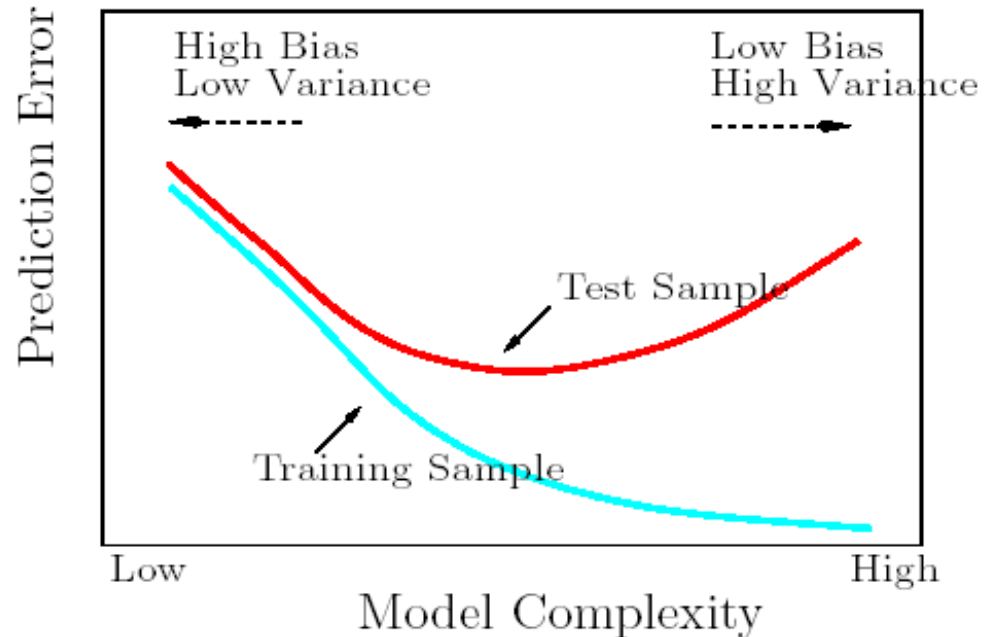
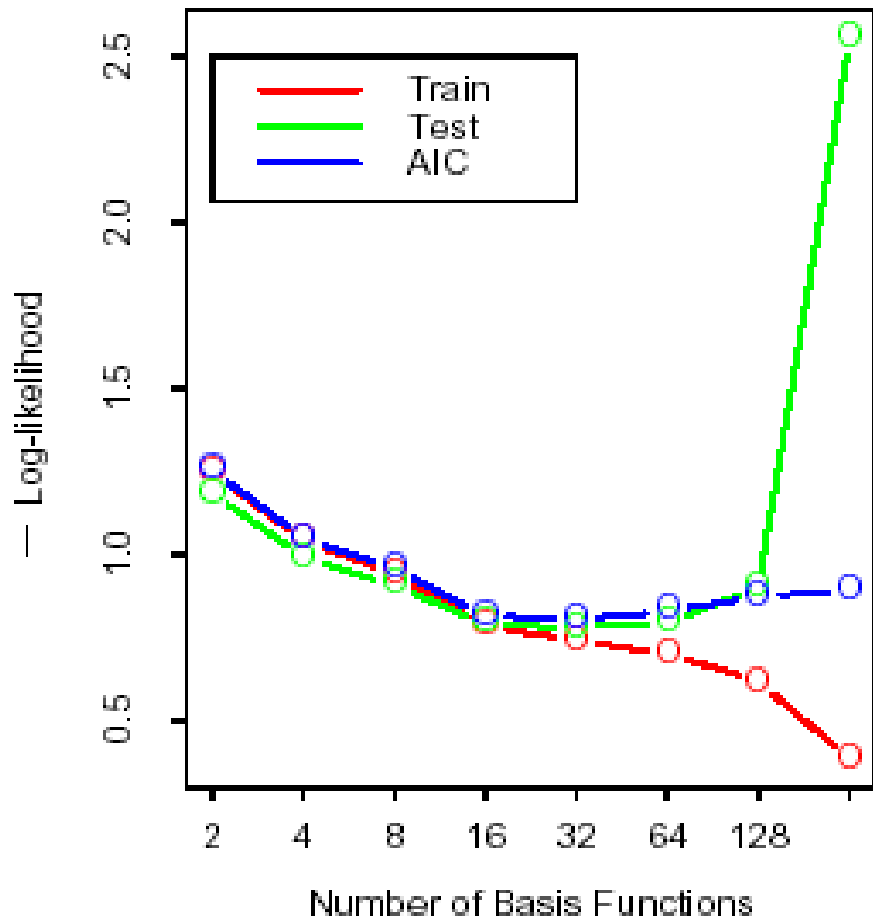


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

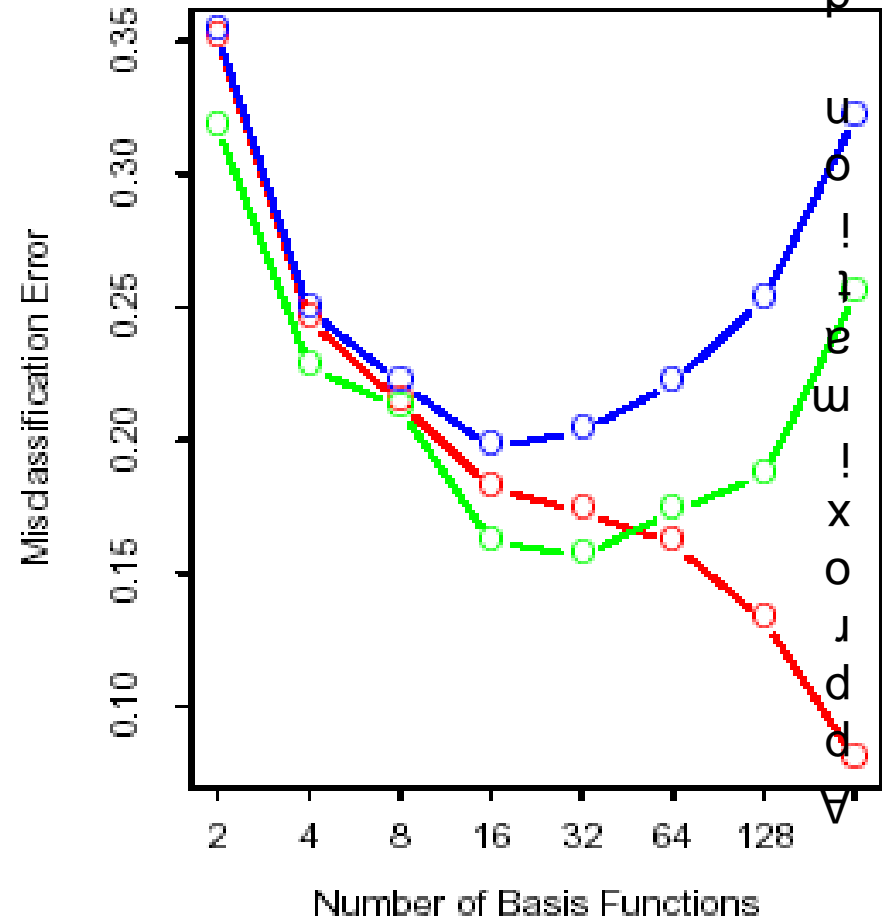
Using AIC to select the # of basis functions in a spline regression

1
0
U
S
e
o
p

Log-likelihood Loss



0-1 Loss

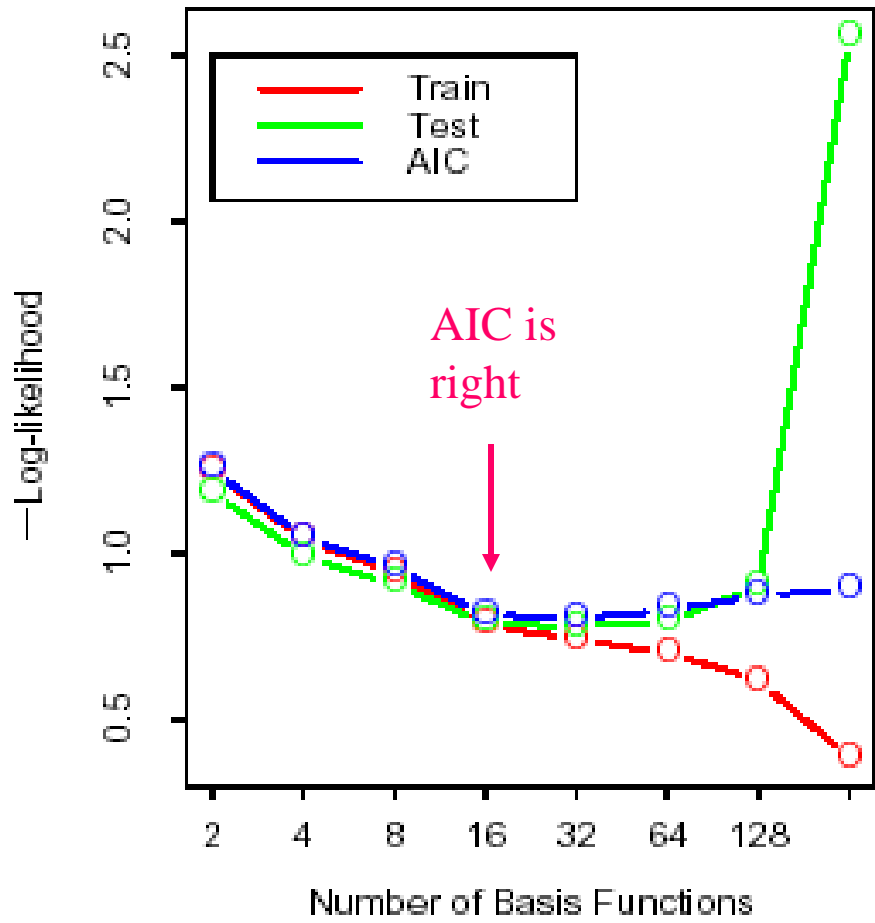


U
S
e
o
p
X
o
J
d
o

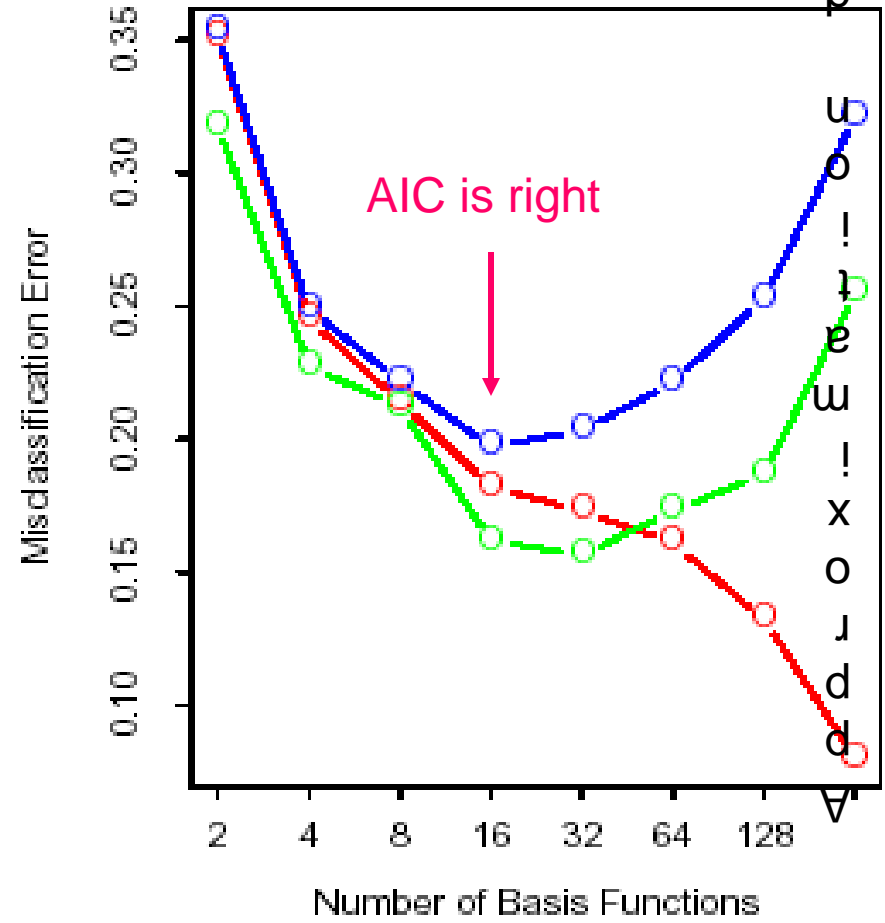
Using AIC to select the # of basis functions in a spline regression

1
0
U
S
E
O
P

Log-likelihood Loss



0-1 Loss



U
S
E
O
P
X
O
J
D
A

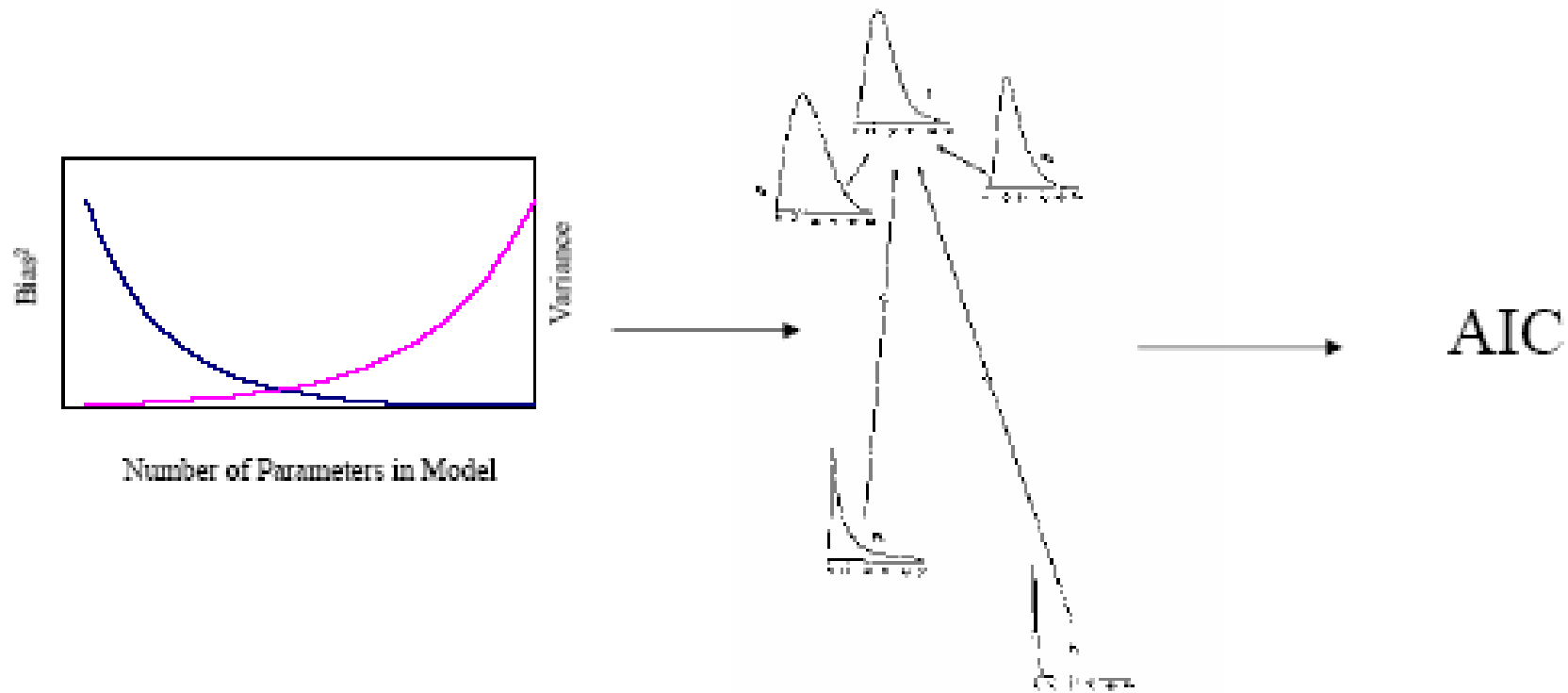
Compare fit of fit without log(lead)

- `glm(response ~ lead, family = binomial)`
- Null Deviance: 101.3 (Null the same)
- Residual Deviance: 56.23 AIC: 76.71

- `glm(response ~ log(lead), family = binomial)`
- Null Deviance: 101.3 (Null the same)
- Residual Deviance: 2.784 AIC: 23.26
- Residual deviance and AIC MUCH smaller with log(lead)

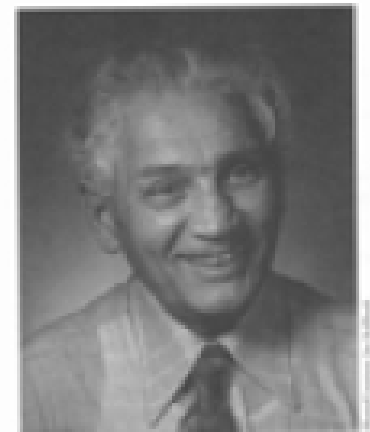
AIC estimates the expected value of the relative K-L Distance

$$AIC = -2 \ln(\mathcal{L}(\hat{\theta})) + 2K$$



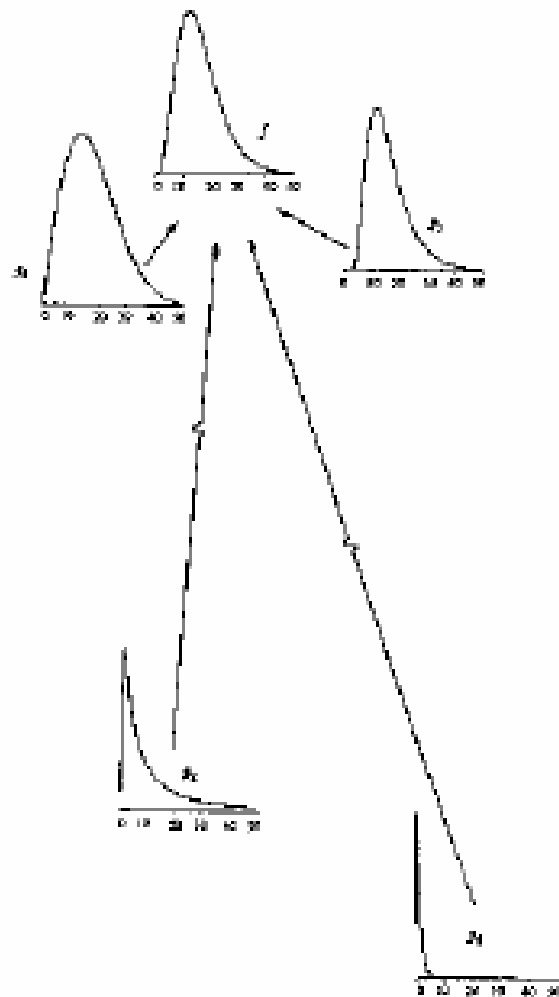
AIC provides a fundamental basis for evaluating the strength of evidence for models in data.

The Kullback-Leibler Distance



Solomon Kullback

Solomon Kullback (1907-1994)

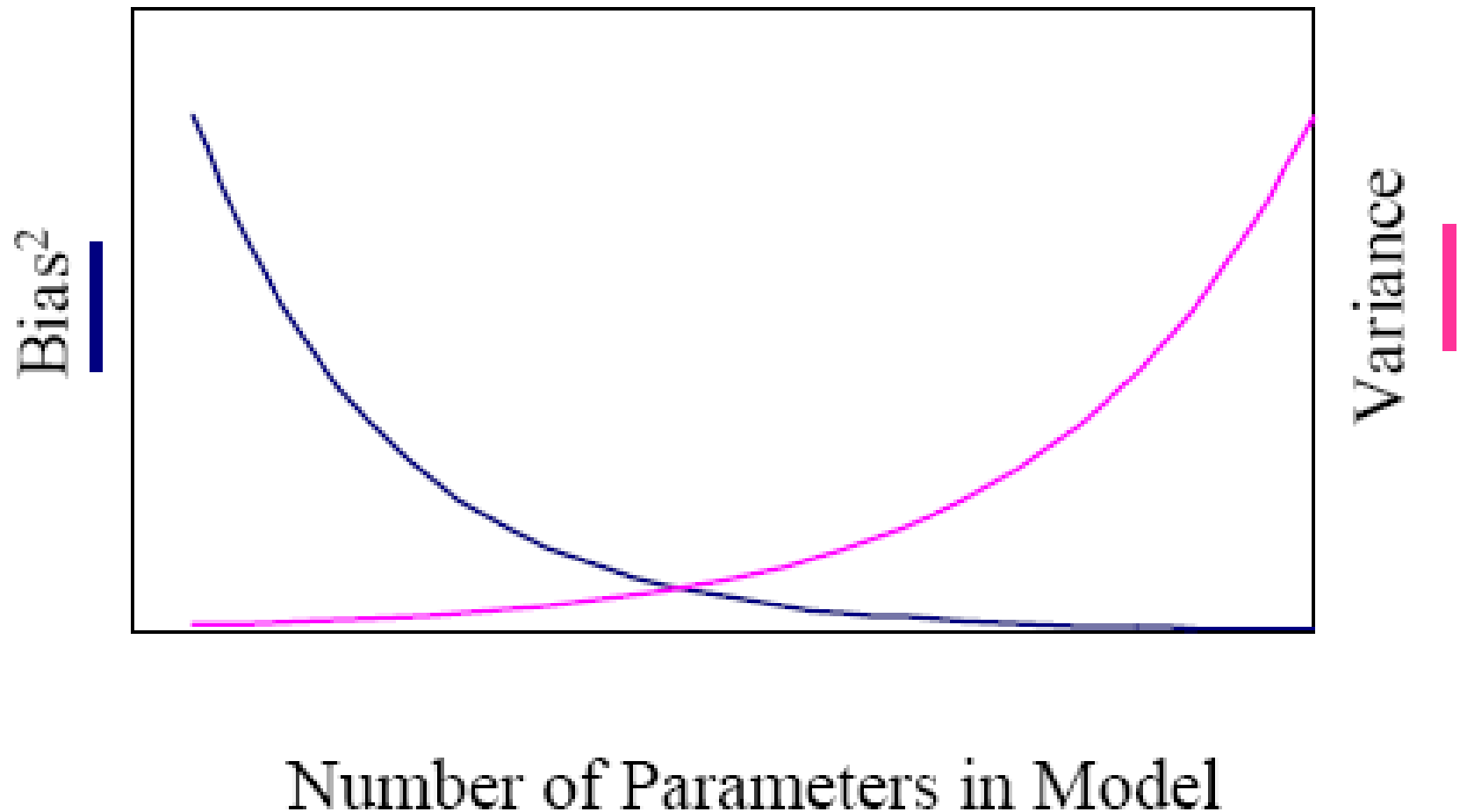


The Kullback-Leibler discrepancy is a directed distance from the “true” model (f) to candidate models (g_i).

A simple way to think about AIC.

- Think about a nested parameter model and you want to test if a parameter is statistically significant.
- The likelihood ratio test: $\log(\text{ratio of models with an extra parameter}) \sim \chi_1^2$
- Thus, we keep a parameter is $qchisq(.95, 1)/2 = 3.84/2 \sim 2$.

The Principle of Parsimony



Sakamoto et al. 1986

"True model:" $y = e^{(x-0.3)^2} - 1 + \varepsilon,$

Generated 10 data sets sampling from normal distribution with mean = 0 and variance = .01

Fit 5 approximating models to the 10 data sets

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5$$

What creates “noise” in models?

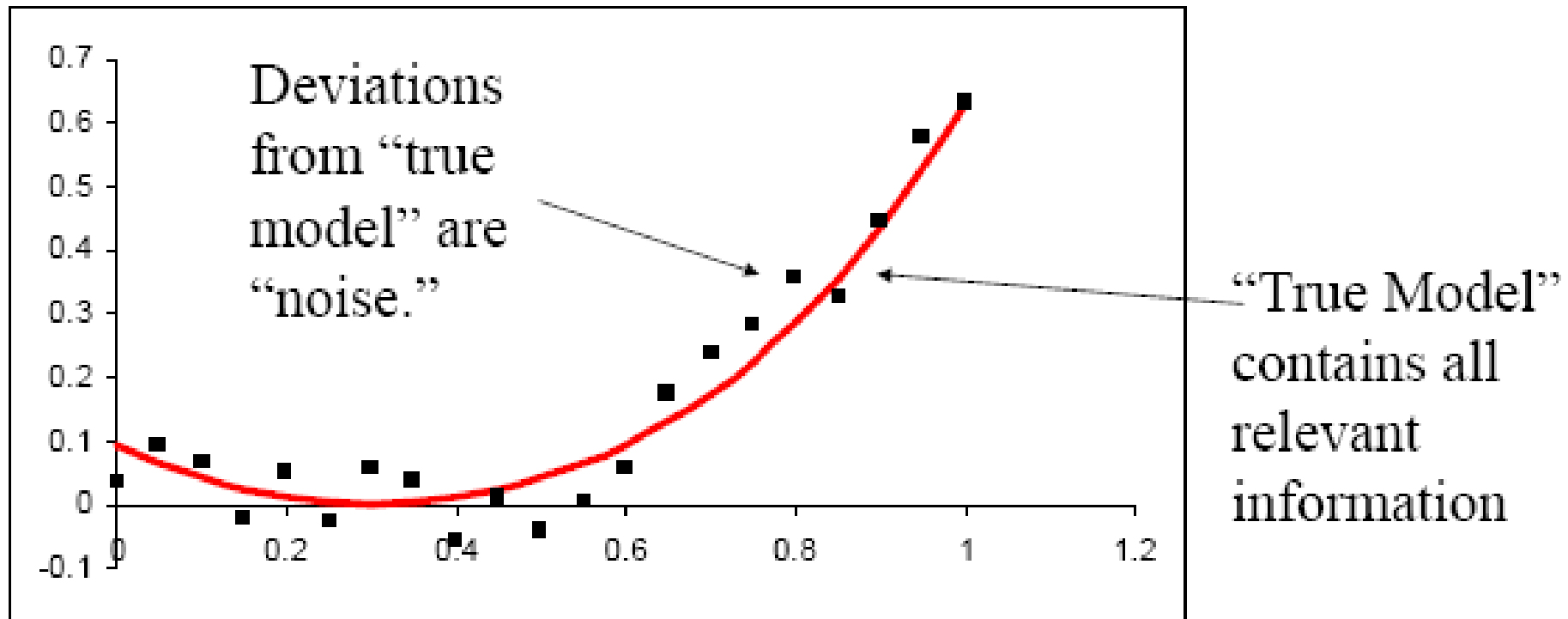
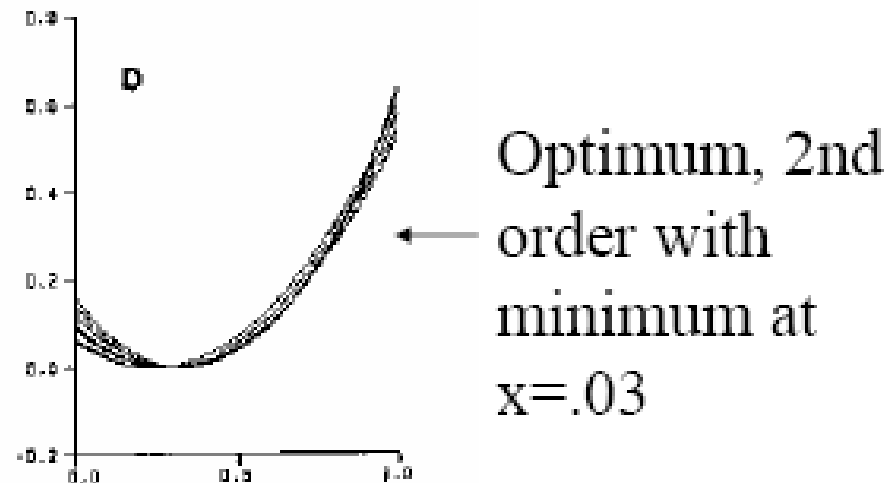
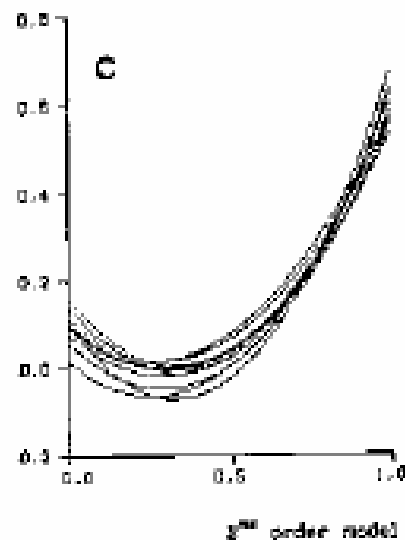
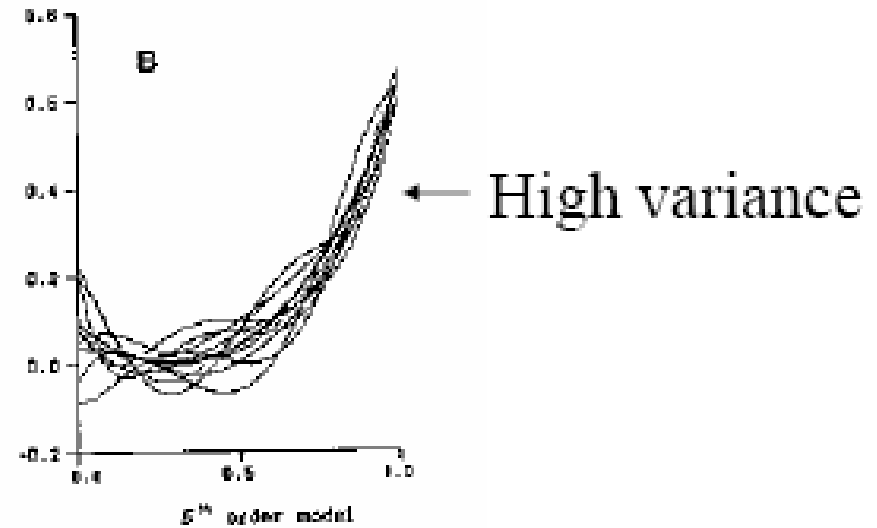
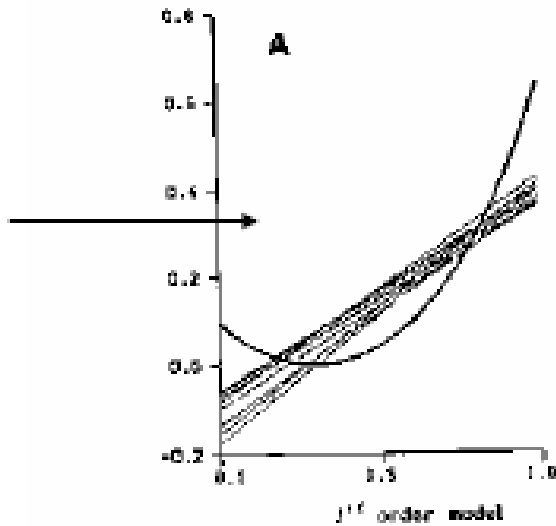
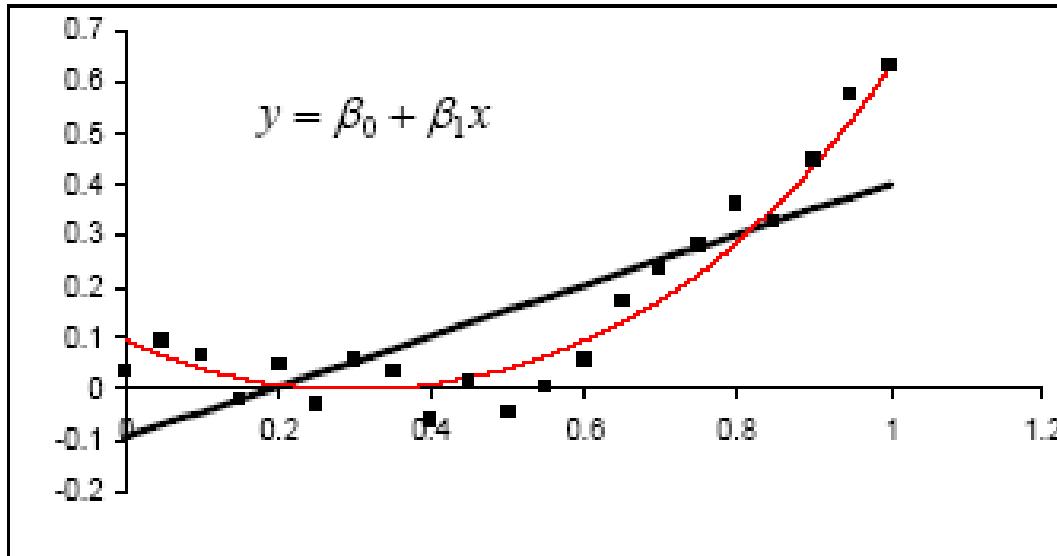


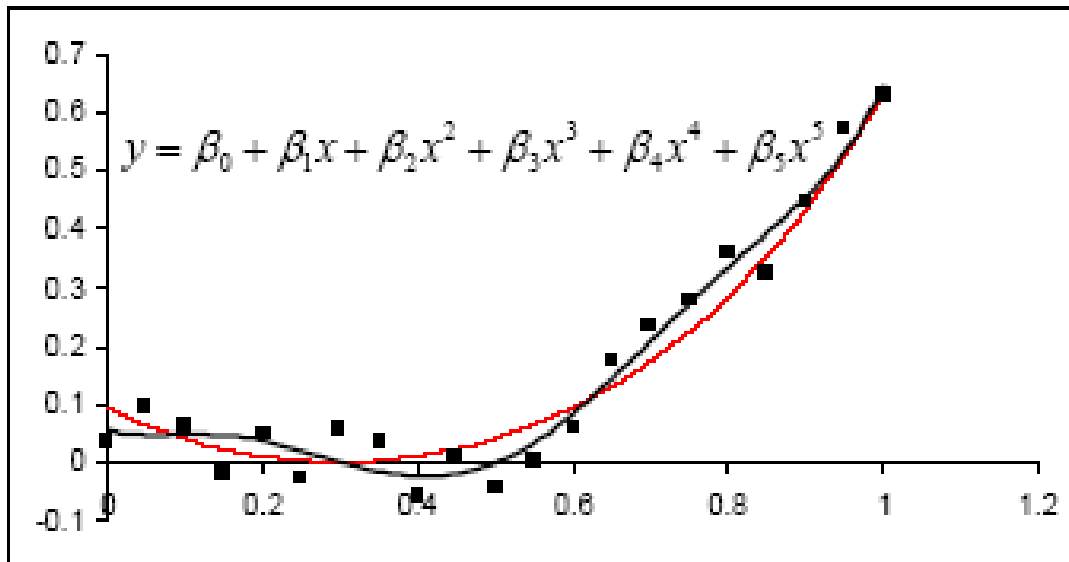
Illustration of trade off

High bias



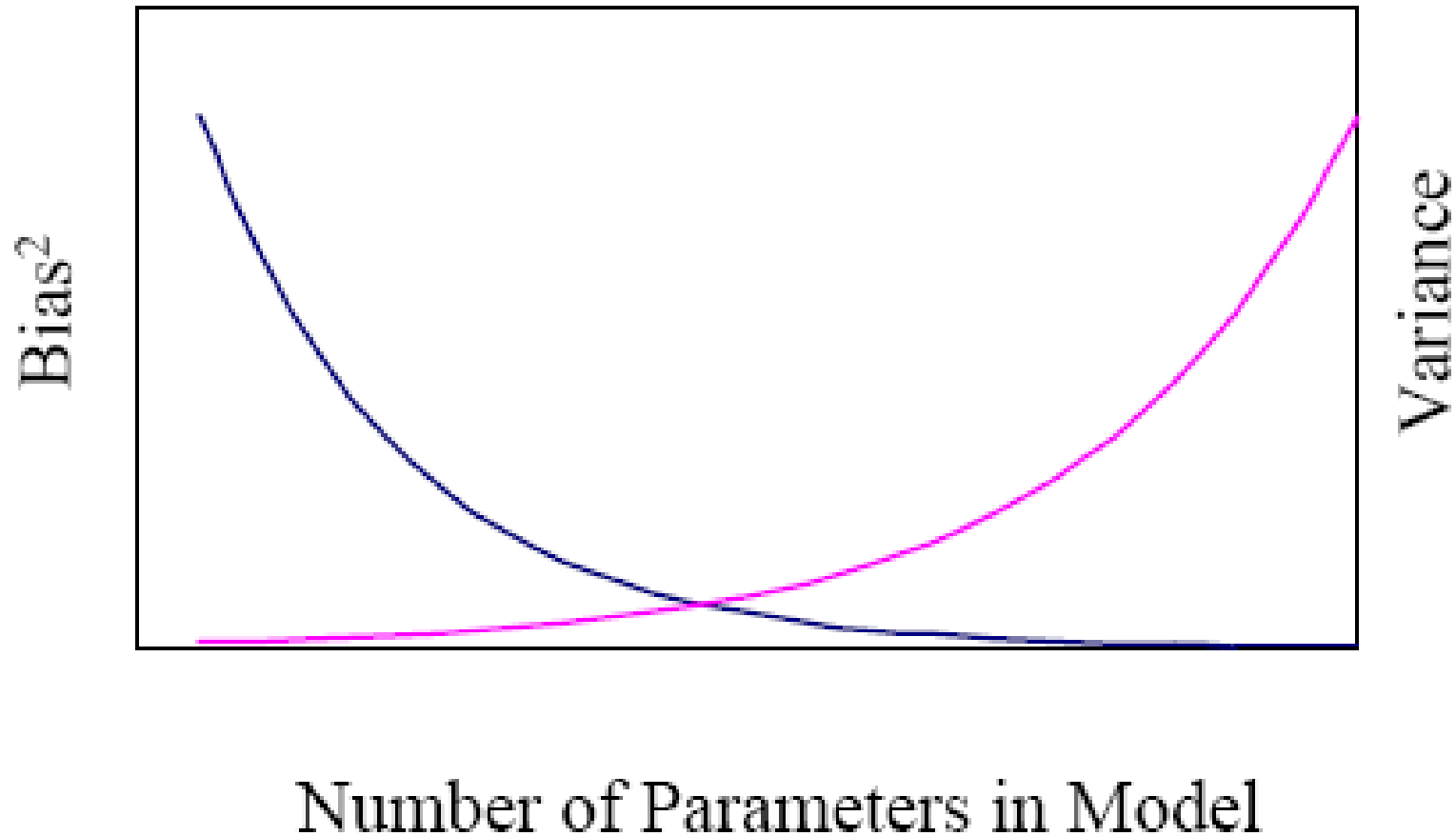


Two few parameters--
fails to respond to
information. Bias is
high.



Too many parameters--
responds to "noise."
Variance is high.

The Principle of Parsimony



Steps in Model Selection

- Develop candidate models based on biological knowledge. Lots of thinking here!
- Take observations relevant to predictions of models.
- Use data to obtain *maximum likelihood estimates* of model parameters.
- Evaluate evidence supporting alternative models using AIC.
- Evaluate estimates of parameters relative to direct measures. Are they what you think they are?

How well does AIC work?

- It tends to keep too many parameters for complex models. Alternative (Bayesian Information Criterion) keeps in fewer parameters.
- There are various corrections that can be used (these are details I will not discuss).
- For variance components and hierarchical models you need DIC = Deviance Information Criteria

The fundamental problem of science: How strong is the evidence for one view of nature (read model, hypothesis) compared with an alternative view?

- Hypothesis testing
- Model selection

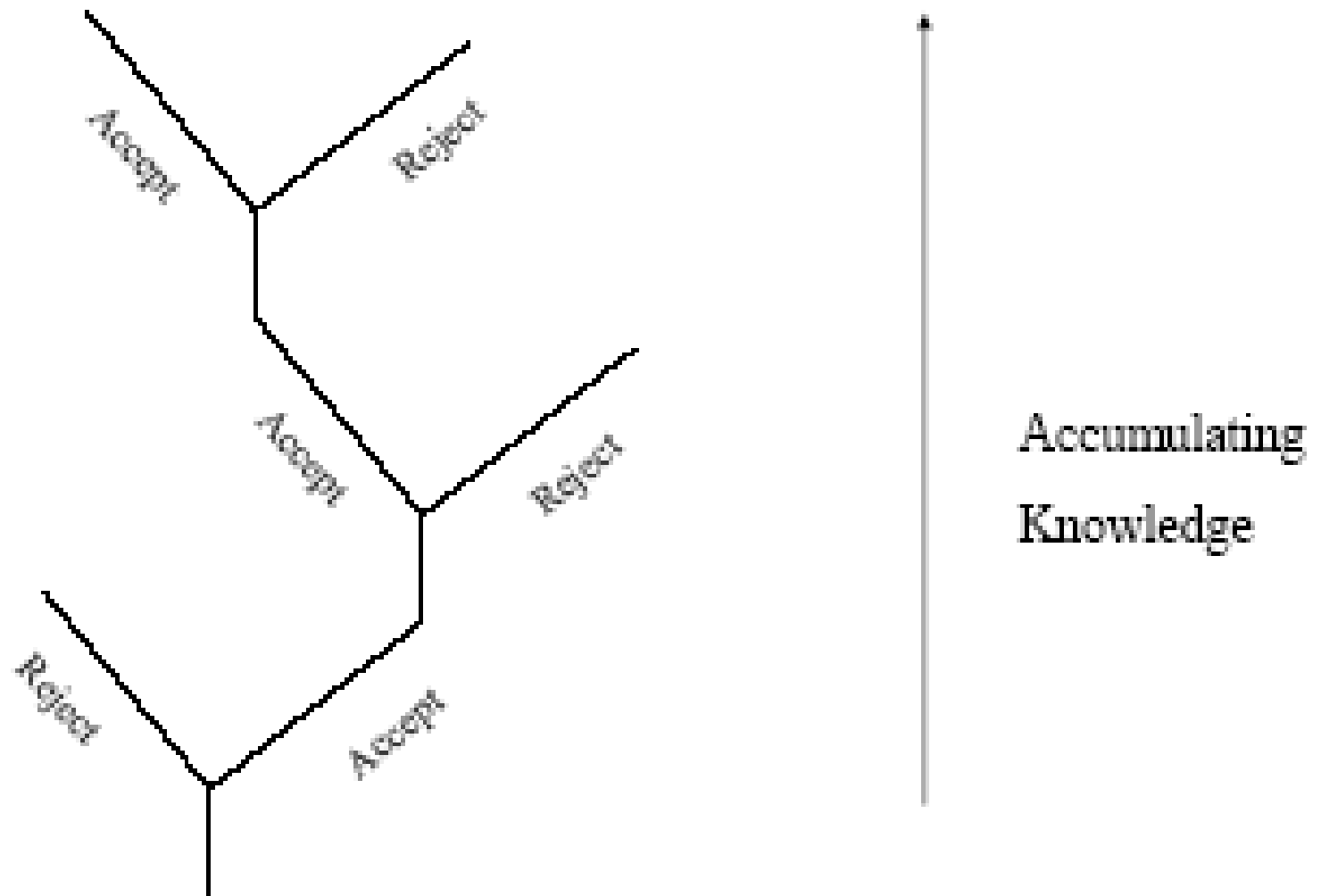
“...the reasons that students have problems understanding hypothesis testing is that they may be trying to think.”

W. E. Demming 1975

Statistical Hypothesis Testing

- Relies on falsification.
- Ultimately, we learn what is true by establishing what is false through a series of hypothesis tests.
- Most hypothesis tests depend on interpreting a P value or significance level.
- Series of “rejections” leaves a view of nature that has greatest support in observations.

Platt JR (1964) Strong inference - certain systematic methods of scientific thinking may produce much more rapid progress than others.
Science 146:347-353



Model Selection as an Alternative to Hypothesis Testing

- Hypotheses = models
- Models = approximations of complex truth?
- Purpose of science: how well do models approximate truth?
- Appreciate how different this is from “truth by rejection.”

Relativity of Evidence in Model Selection

- We are not asking if a model is right or wrong. We ask, “*Does a model have more support in the data than a competing model?*”
- The strength of evidence (support) for a model is relative.
 - Relative to other models---as models improve, support may change.
 - Relative to data at hand---as the data improve, support may change.

Model section: AIC and all that

measure of fit + complexity penalty

AIC is defined as

$$AIC_i = -2 \log L_i + 2V_i$$

L_i = Maximized log likelihood of model i

V_i = Number of free parameters

Choose the model with the smallest AIC (and perhaps retain all models within 2 of the minimum).

For “small data sets” use corrected AIC
(for number of observation/ $V < 40$)

$$AIC_c = -2 \log L + 2V + \frac{2V(V+1)}{(n-V-1)}$$

L_i = Maximized log likelihood of model i

V_i = Number of free parameters

Choose the model with the smallest AIC (and perhaps retain all models within 2 of the minimum).

BIC

- Schwarz (1978) derived the Bayesian information criterion as

$$\text{BIC} = -2 \ln(L) + V \log(n).$$

Parameter

Observations

- As usually used, one computes the BIC for each model and selects the model with the smallest criterion value.

DIC – Deviance Information Criterion

In GLMs (and elsewhere) the *deviance* is the difference in twice maximized log likelihood between the *saturated* model and the fitted model, or

$$D(\theta) = \text{deviance}(\theta) = \text{const}(\mathcal{T}) - 2L(\theta; \mathcal{T})$$

and in GLMs we use $D(\hat{\theta})$ as the (unscaled) (residual) deviance.

In practice

- Use stepAIC to pick best models, or group of models that are consistent with the data (you have to load the MASS library).

Explanation vs Prediction

This causes a lot of confusion. For explanation, Occam's razor applies and we want

an explanation that is as simple as possible, but no simpler

attrib Einstein

and we do have a concept of a 'true' model, or at least a model that is a good working approximation to the truth, for

all models are false, but some are useful

G.E.P. Box, 1976


Explanation is like doing scientific research.


Prediction

- Prediction is like doing engineering, you only care that it works.
- If the aim is prediction, then model choice should be based upon quality of predictions.

Three types of models

•Understanding



- Theory based models (sometimes called mechanistic models).
 - Empirical Models: simple restrictions on behaviour, e.g. linear models, AR(p) processes.
 - Modern flexible models: neural networks, generalized additive models, “data mining”.
- 

Conclusions

- Have multiple theories
- There are lots of formal ‘figures of adequacy’ for a model. Some have proved quite useful, but
 - Their variability as estimators can be worrying large.
 - Computation, e.g. of ‘effective number of degrees of freedom’, can be difficult.
 - Their implicit measure of performance can be overly sensitive to certain aspects of the model which are not relevant to our problem.
- Formal training/validation/test sets, or the cross-validatory equivalents, are a very general and safe approach.

- ‘Regression diagnostics’ are often based on approximations to over-fitting or case deletion.
- Now we can (and some of us do) fit extended models with smooth terms or use fitting algorithms that downweight groups of points. (I rarely use least squares these days,)
- Use AIC, especially in simpler problems. Hence the stepAIC function for S-PLUS/R.