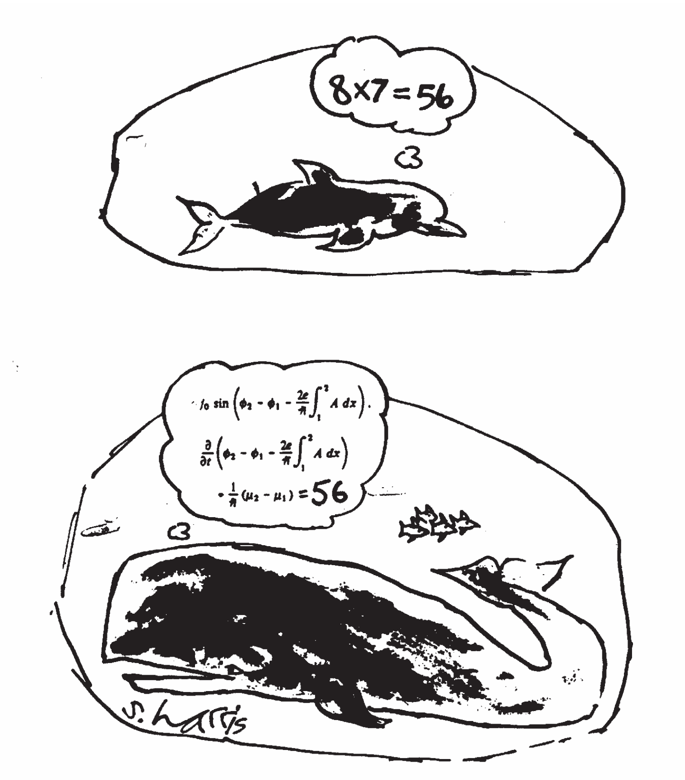
Model Selection Methods

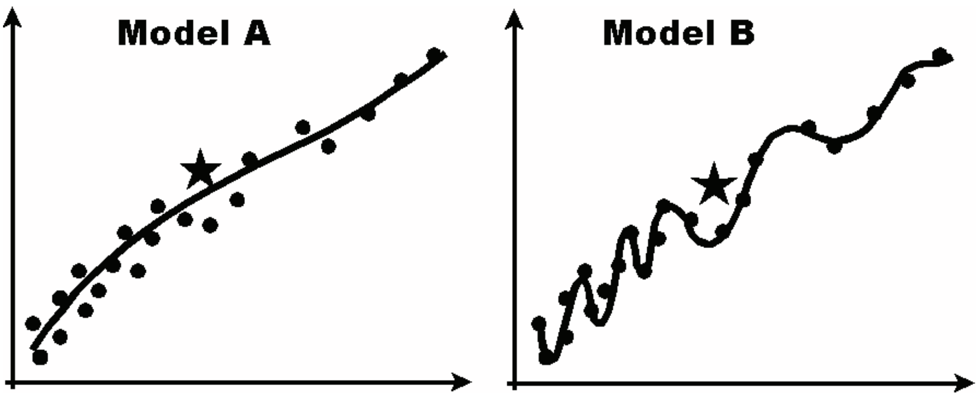
Jay Myung
Ohio State University

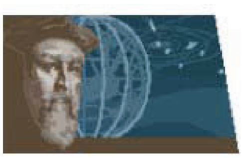
In collaboration with **Mark Pitt**

Amsterdam Workshop on Model Selection
Aug 27-29, 2004

Whale's Views of Model Selection







Preview of Conclusion:

"Thou shall not select the **best-fitting** model but shall select the **best-predicting** model."

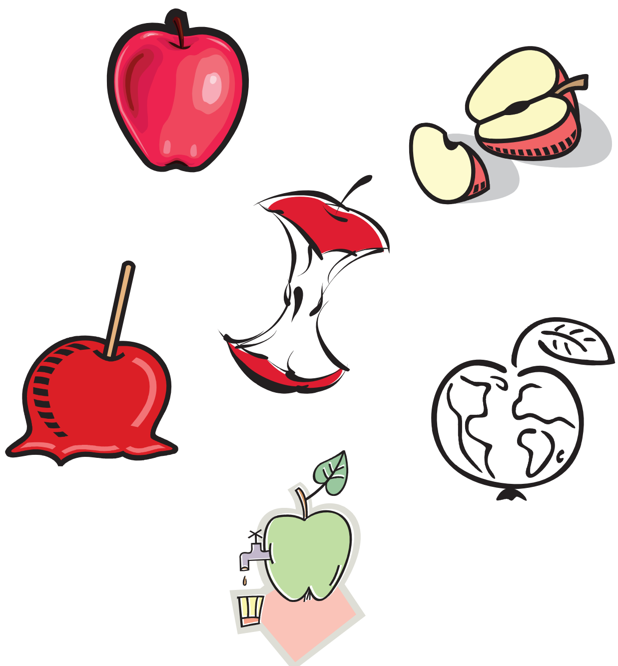
Overview

- **Part 1: Non-technical Introduction to Model Selection**
- **Part 2: "Technical" Tour of Model Selection Methods**
- **Part 3: Example Application**
- **Part 4: Conclusions**

Part 1:
Non-technical Introduction to Model Selection

Terminology

- **Model Selection**
- **Model Choice**
- **Model Comparison**



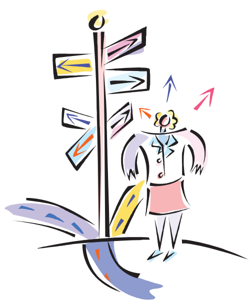
What is a Quantitative Model?

- Mathematical instantiations of key assumptions and principles embodied in the theory from which it evolved.
- A formalization of a theory that enables the exploration of its operation.

Why Modeling?

- To infer the underlying structural properties of a mental process from behavioral data that were thought to have been generated by that process.

Often entertain **multiple models** as possible explanations of observed data



Model Selection Problem

- **Q:** How should we choose between differing explanations (models) of data?

- **A:** Select the one, among candidate models, that "best" captures the underlying regularities.

How to identify such a model?

Goodness of Fit (GOF) Measures as Methods of Model Selection

Examples of GOF measures:

- Percent Variance Accounted For (**PVAF**)

$$PVA = 100 * \left(\sum_{i=1}^{m} (obs_i - pred_i)^2 / \sum_{i=1}^{m} (obs_i - obs_{mean})^2 \right)$$

- Root Mean Square Deviation (**RMSD**)

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} (obs_i - prd_i)^2}{N}}$$

Problems with GOF as a Model Selection Method

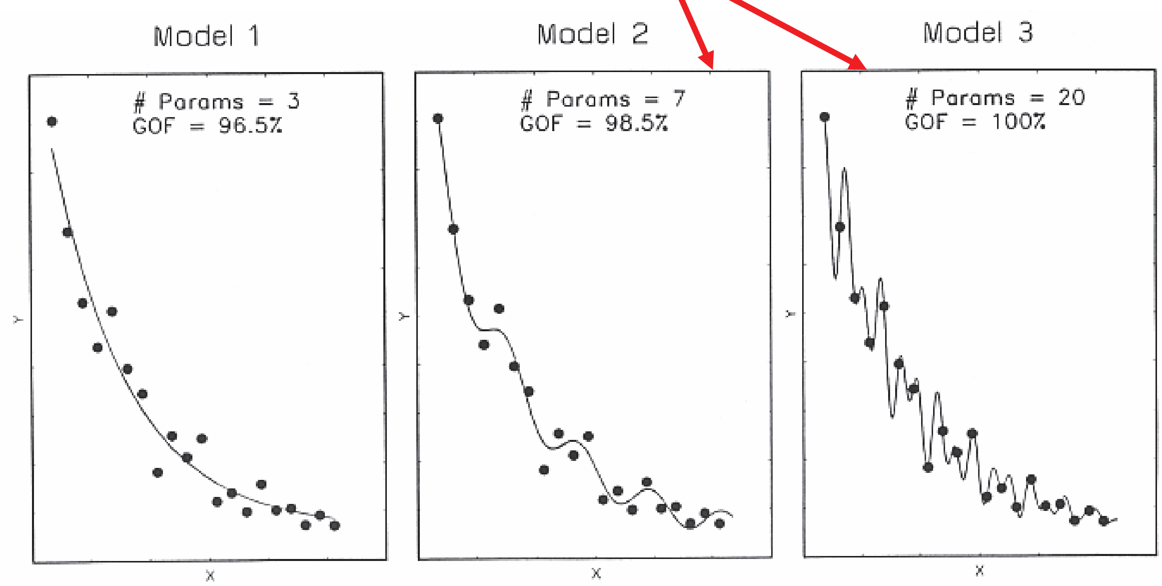
Data: **Noise** **&** **Regularity**

(sampling error) (underlying mental process)

GOF = fit to noise + fit to regularity

Properties of the model that have nothing to do with its ability to capture the underlying regularities can improve fit.

Fit improves with more parameters (i.e, over-fitting)



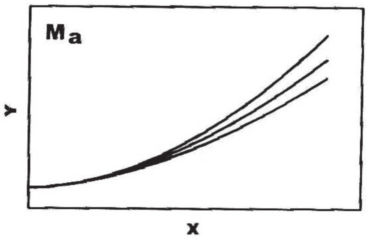
Model 1: $Y = ae^{-bX} + c$

Model 2: $Y = ae^{-bX} + c + dX^{-e} \cdot \sin(f \cdot X + g)$

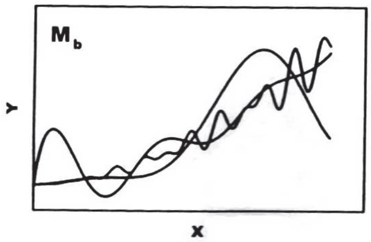
Model Complexity

Complexity: A model's inherent **flexibility** that enables it to fit a wide range of data patterns

Simple Model



Complex Model



Complexity: # of parameters + functional form

Complexity: More than number of parameters?

$$M1: y = ax + b$$

$$M2: y = ax^b$$

$$M3: y = \sin(\cos ax)^a \exp(-bx) / x^b$$

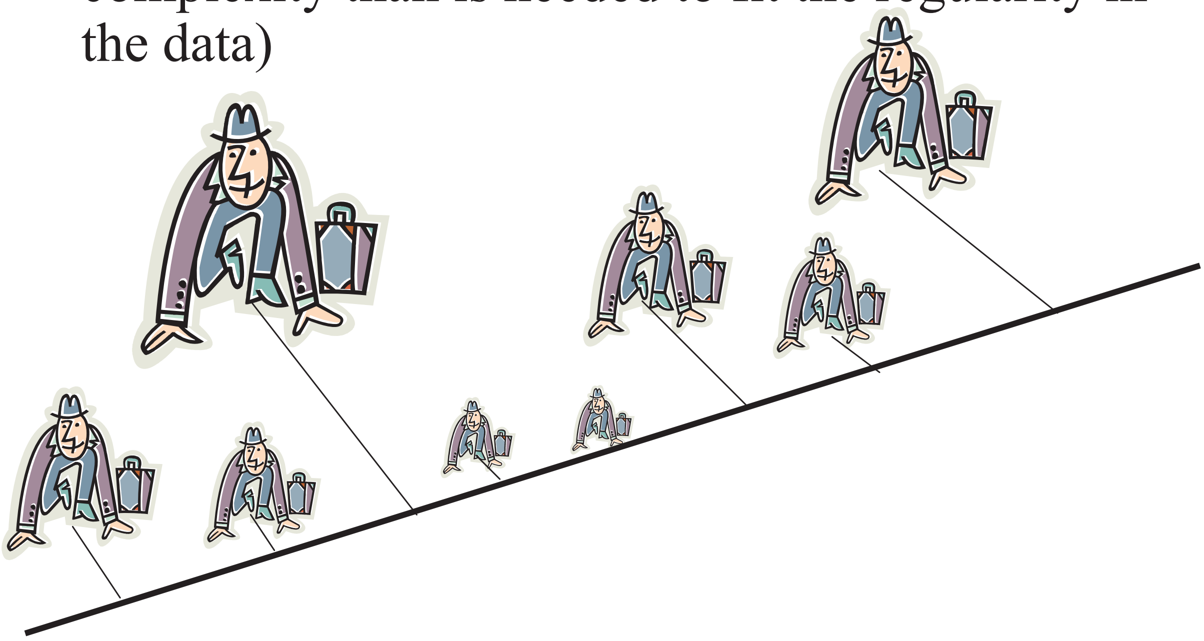
Are these all equally complex?

Wanted:

A Method of Model Selection that Takes into Account Effects of **Complexity**

Placing Models on an Equal Footing

Penalize models for excess complexity (i.e., more complexity than is needed to fit the regularity in the data)



Akaike Information Criterion (AIC) as a Method of Model Selection

\# of parameters

Akaike (1973):

$$AIC = -2 \ln f(y \mid \hat{\theta}) + 2k$$

Goodness of fit (ML) + **Model Complexity**

The model that minimizes AIC should be preferred

Bayesian Information Criterion (BIC)

Schwarz (1978):

$$BIC = -2 \ln f(y \mid \hat{\theta}) + k \ln n$$

Goodness of fit (ML) + **Model Complexity**

Selection Criteria as Formal Implementations of Principle of **Occam's Razor**

"Entities should not be multiplied beyond necessity"

(William of Occam, ca. 1290-1349)

"Select the **simplest** model that describes the data **sufficiently well**."

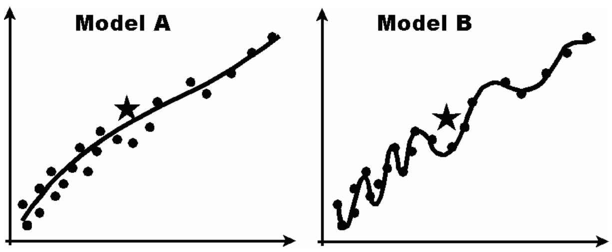
$$AIC = -2\ln f(y\,|\,\hat{\theta}) + 2k$$

$$BIC = -2\ln f(y\,|\,\hat{\theta}) + k\ln n$$

What Do AIC and BIC Measure?

They estimates a model's **generalizability** – the model's ability to fit all "future" data samples from the same underlying process, not just the current data sample.

Generalizability
= '**proximity**' to underlying process
= **Predictive accuracy**



"An important goal of scientific theorizing is to identify hypotheses that generate <u>accurate predictions</u>."

"Overfitting is a sin precisely because it undermines the goal of <u>predictive accuracy</u>."

(both from Hitchcock & Sober, 2004)

Definition of generalizabilty

Formally, a model's generalizability may be defined as

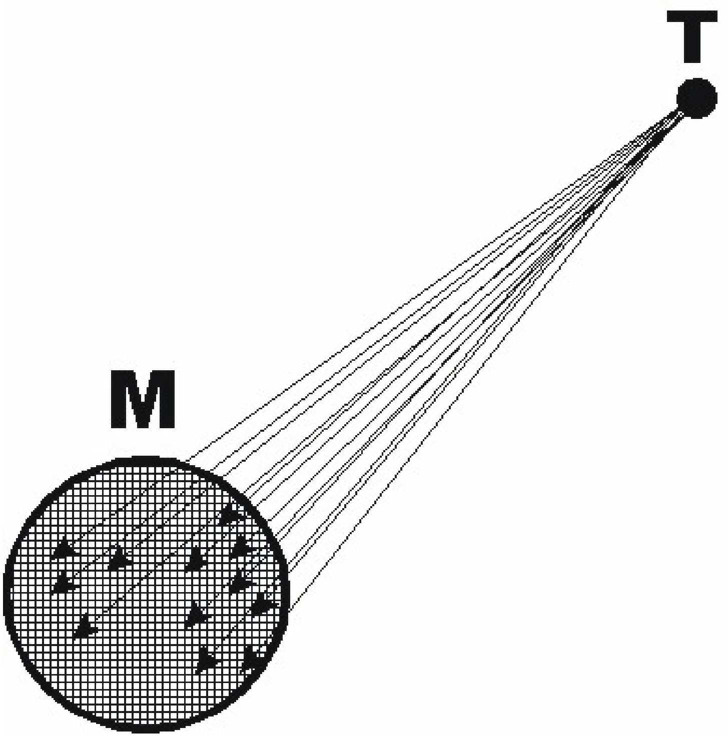
$$E\left[D(M,T)\right] = \int D(f_M(y\,|\,\hat{\theta}), f_T(y))\, f_T(y)\, dy$$

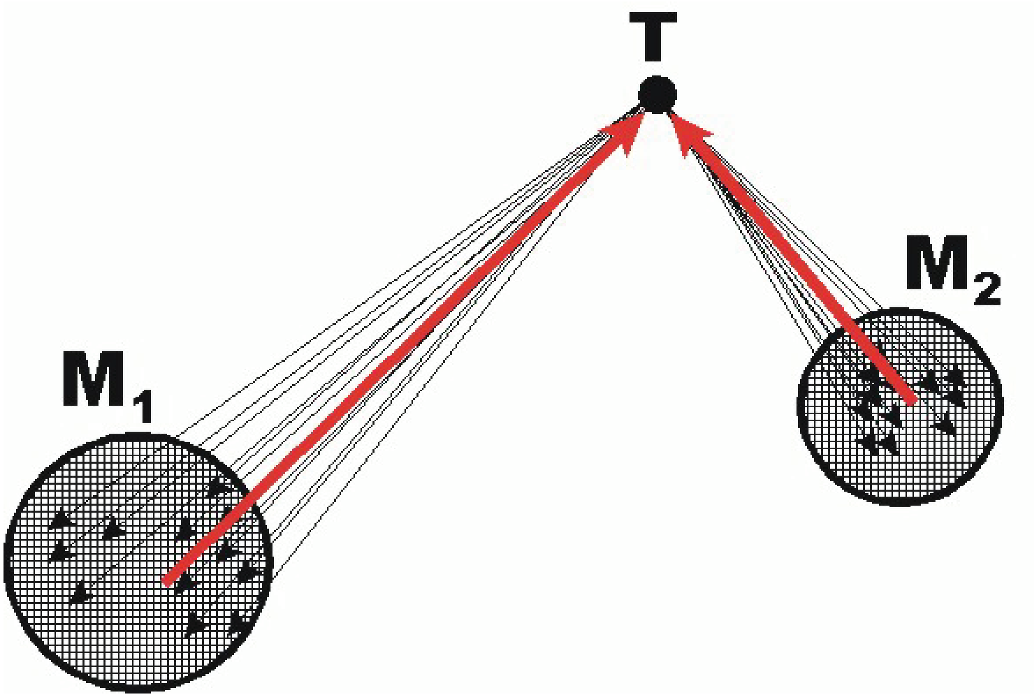
As mean discrepancy between the model of interest and the true model under some *discrepancy function* D satisfying

$$D(f,g) > D(f,f) = 0 \;\; for\; f \neq g$$

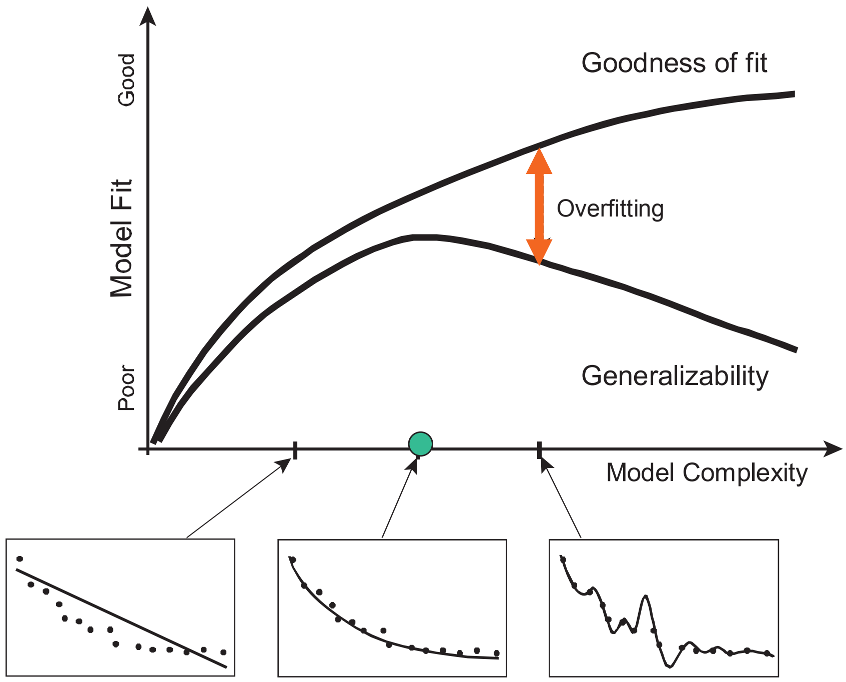
(e.g., Kullback-Liebler information distance)

“Geometric” Definition of Generalizabilty





Relationship between Goodness of Fit and Generalizability



Part 2:
"Technical" Tour of
Model Selection Methods

Selections Methods to be discussed

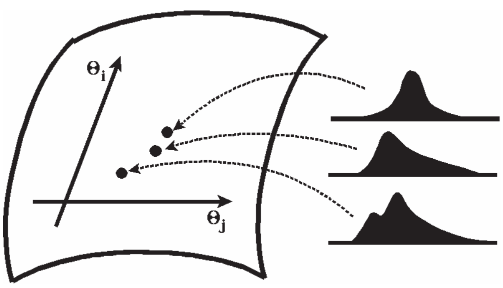
- AIC
- Cross-validation
- Bootstrap
- Bayesian Methods (Bayes Factor, BIC)
- **Minimum Description Length**

Formal Definition of A Statistical Model

A model is defined as a parametric collection of probability distributions, indexed by model parameters:

$$M = \{f(y \mid \theta) \mid \theta \in \Omega\}$$

forming a *Riemannian manifold*, embedded in the space of probability distributions (Rao, 1945; Efron, 1975; Amari, 1980)



Akaike Information Criterion (AIC)

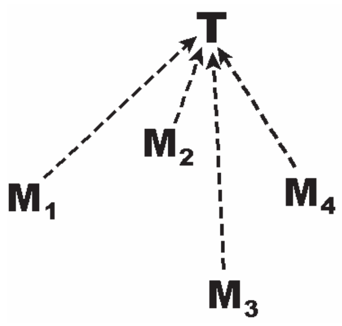
(Akaike, 1973)

AIC derived as asymptotic approximation of Kullback-Liebler information distance between the model of interest and the truth:

$$KL(M,T\mid x)=\int f_T(y)\ln\frac{f_T(y)}{f_M(y\mid\hat{\theta}(x))}dy$$

$$2\cdot E\left[KL(M,T\mid x)\right]=2\cdot\int KL(M,T\mid x)\,f_T(x)\,dx$$

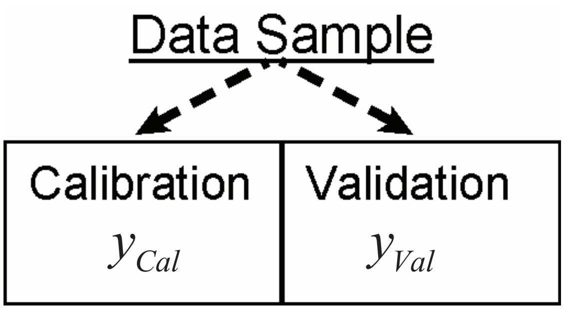
$$=\mathrm{AIC}+(\textit{higher order terms})$$



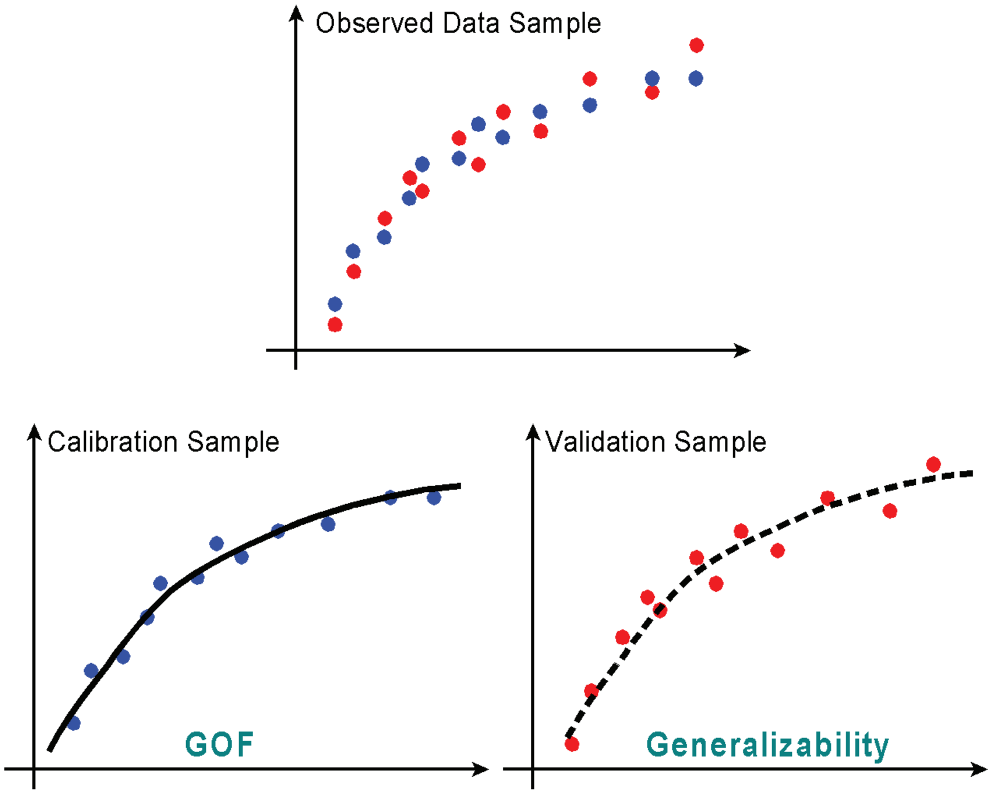
Cross-validation (CV)

(Stone, 1974; Geisser, 1975)

- Sampling-based method of estimating generalizability
- No explicit measure of model complexity, unlike AIC



$$CV=-\ln f(y_{Val}\mid\hat{\theta}(y_{Cal}))$$



Features of CV

- <u>Pros</u>
  - Easy to use
  - Sensitive to functional form as well as number of parameters
  - Asymptotically equivalent to AIC

- <u>Cons</u>
  - Sensitive to the partitioning used
    - Averaging over multiple partitions
    - *Leave-one-out CV*, instead of *split-half CV*
  - Instability of the estimate due to “loss” of data

Bootstrap Model Selection (BMS)

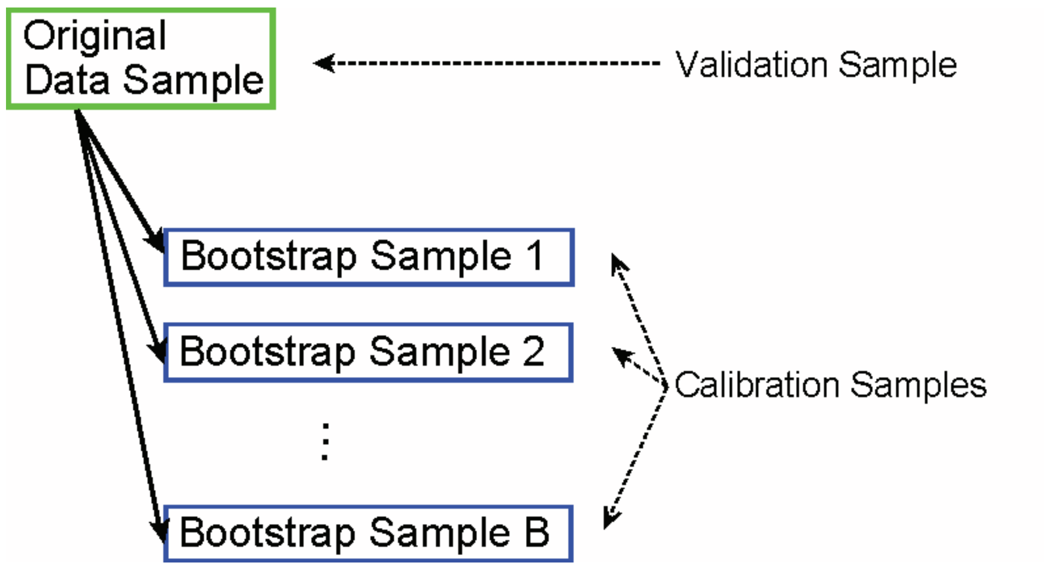
(Efron, 1983; Shao, 1996)

Similar to CV,

- **Re**sampling-based method of estimating generalizability
- No explicit measure of model complexity

Unlike CV,

- Full use of data sample in estimating generalizability



$$BMS = -\frac{1}{B}\sum_{i=1}^{N} \underline{\ln f(y_{Original} \mid \hat{\theta}(y_{Boot_i}))}$$

CV with n, not n/2

Bayesian Methods

(Kass & Raftery, 1995)

- In Bayesian model selection, each model is evaluated based on its *marginal likelihood* defined as

$$P(y_{obs} \mid M_j) = \int f(y_{obs} \mid \theta, M_j)\pi(\theta \mid M_j)d\theta, \quad j = 1, 2, \ldots, J$$

- Model selection is then based on the ratio of two marginal likelihoods or *Bayes factor (BF)*

$$BF_{ij} \equiv \frac{P(y_{obs} \mid M_i)}{P(y_{obs} \mid M_j)}$$

- Under the assumption of equal model priors, BF is reduced to the *posterior odds*:

$$\begin{aligned} BF_{ij} &= \frac{P(y_{obs} \mid M_i)}{P(y_{obs} \mid M_j)} \\ &= \frac{P(M_i \mid y_{obs})}{P(M_j \mid y_{obs})} \Big/ \frac{P(M_i)}{P(M_j)} \quad (from \quad Bayes\, rule) \\ &= \frac{P(M_i \mid y_{obs})}{P(M_j \mid y_{obs})} \end{aligned}$$

- Therefore, the model that maximizes marginal likelihood is the one with highest probability of being “true” given observed data

Features of Bayes Factor

- <u>Pros</u>
  - No optimization (i.e., no maximum likelihood)
  - No explicit measure of model complexity
  - No overfitting, by averaging likelihood function across parameters

- <u>Cons</u>
  - Issue of parameter prior (virtue or vice?)
  - Non-trivial computations requiring numerical integration

BIC as an approximation of BF

- A large sample approximation of the marginal likelihood yields the easily-computable *Bayesian information criterion (BIC)*:

$$-2\ln P(y_{obs} \mid M_j) = -2\ln \int f(y_{obs} \mid \theta, M_j)\pi(\theta \mid M_j)d\theta$$

$$= \underbrace{-2\ln f(y_{obs} \mid \hat{\theta}, M_j) + k\ln n}_{BIC} + (higher\ order\ terms)$$

Selections Methods to be discussed

- AIC
- Cross-validation
- Bootstrap
- Bayesian Methods (Bayes Factor, BIC)
- **Minimum Description Length**

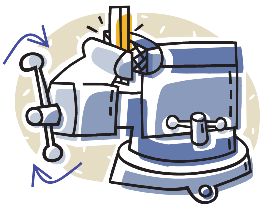
Minimum Description Length (MDL)

(Rissanen, 1978, 1989, 1996, 2001; Hansen & Yu, 2001)

- Algorithmic coding theory
- Models and data as compressible codes
- Regularities (redundancy) can be used to compressed the data

The MDL Principle:

"The best model is the one that provides the shortest description length of the data in bits by "compressing" the data as tightly as possible."



Information Theoretic Motivation

MDL can be motivated from a communication game:

Task: A sender tries to transmit data to a receiver

How many bits are needed to allow the receiver to fully reconstruct the data?

Goal: What is the most efficient (shortest) coding strategy?

MDL idea: "Find a code (i.e., model) that takes advantage of the structure in the data, thereby requiring fewer bits to describe the data."

The Basic Idea

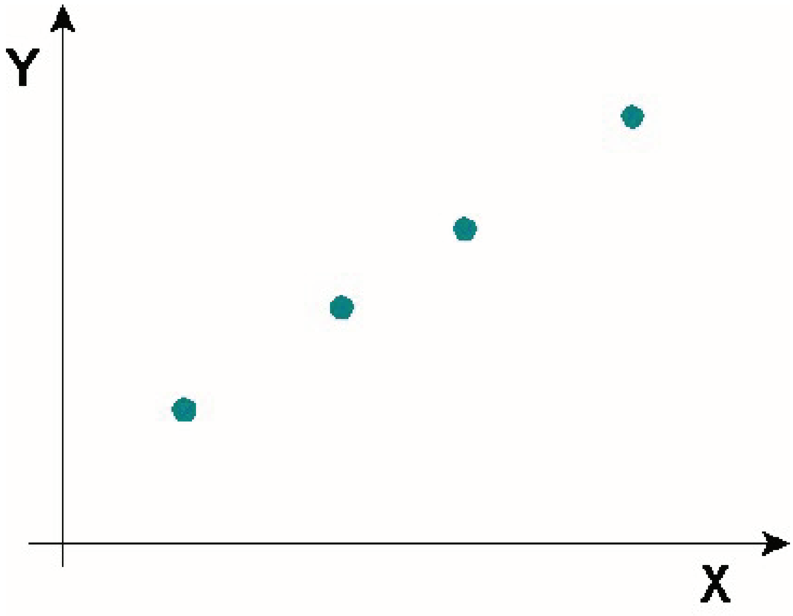
Seq 1: 000100100010001…000100010001

Seq 2: 011101001101001…100110100101

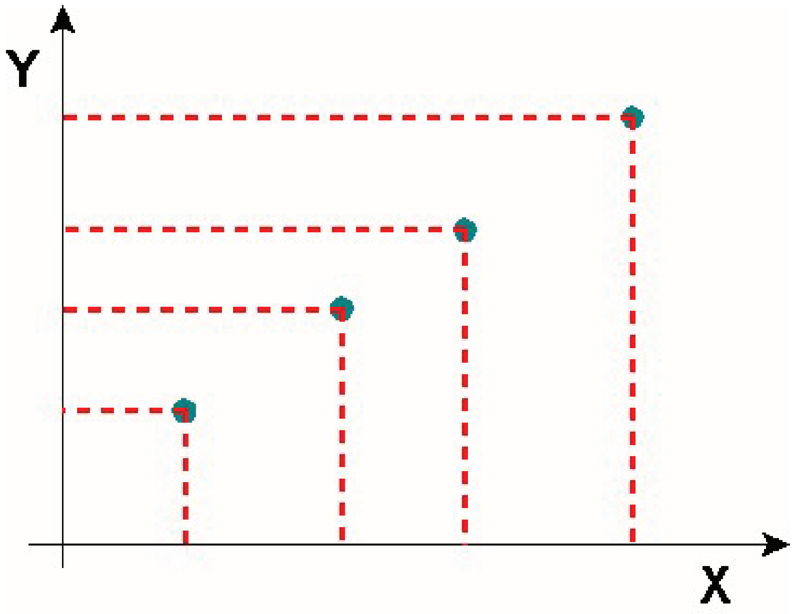
(Coded Seq 1): *"for i= 1 to 100; print '0001'; next; halt"*
(Coded Seq 2): (not compressible! -- coin tossing outcomes)

- More regularity or redundancy in *Seq 1* than *Seq 2*
- Shorter description for *Seq 1* than *Seq 2*

How to *describe* data?



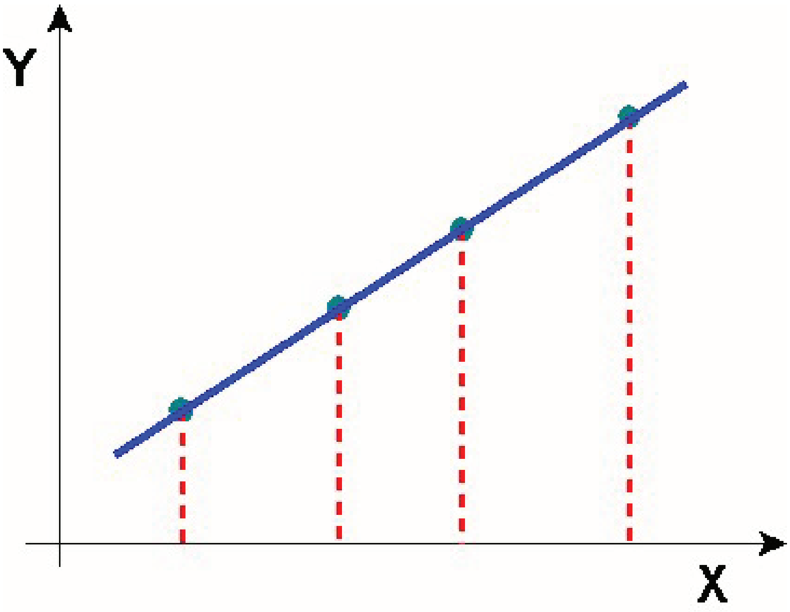
Raw method (no compression)



Overall description length (**ODL1**):

ODL1 = $DL(X_1,Y_1) + DL(X_2,Y_2) + \ldots + DL(X_n,Y_n)$

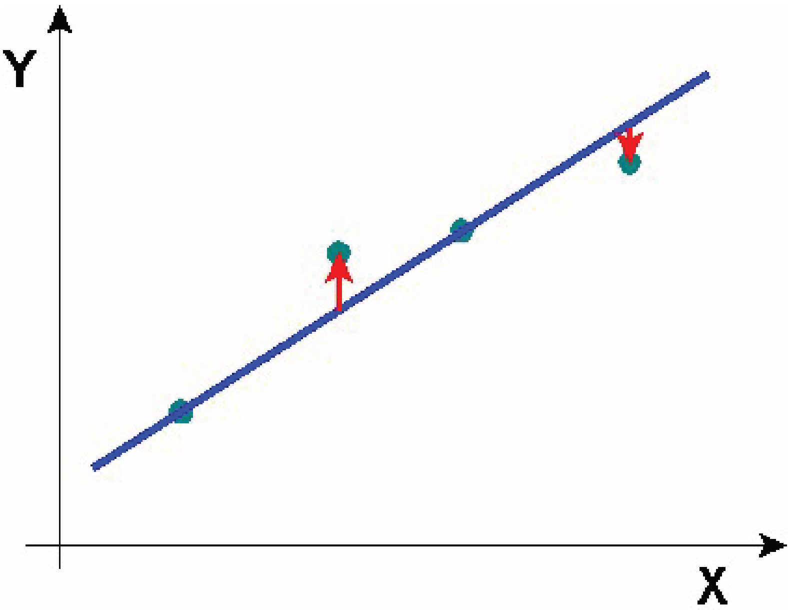
Regularity-based Method (compressed)



Overall description length (**ODL2**):

OLD2 = $DL(X_1) + DL(X_2) + \ldots + DL(X_n) + DL(Y_i = aX_i+b,\ _I = 1,,,n)$

How about ***noisy*** data?



Overall description length (**ODL3**):

OLD3 = ODL2 + **DL(deviations)**

Formally, the description length of data consists of two components:

- **DL(M)**: Description length of the model itself
- **DL(D|M)**: Description length of the data when encoded with the help of the model

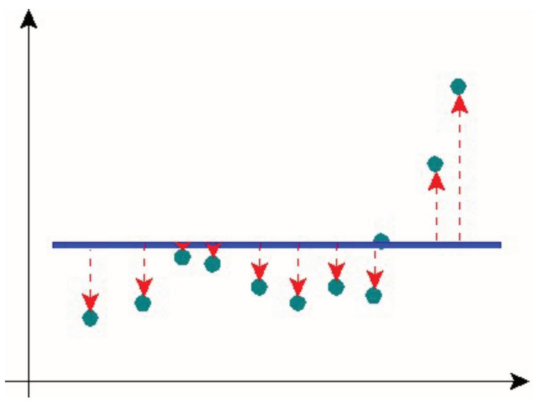
Overall description length (OVD):

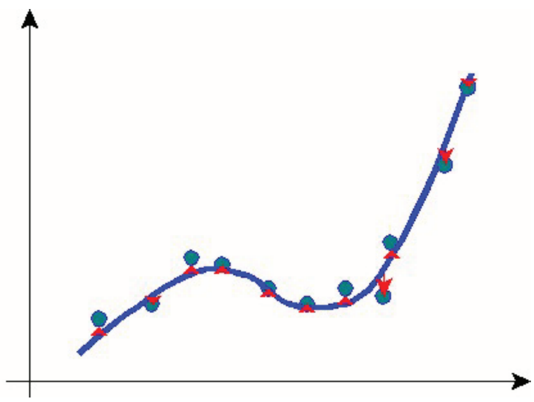
$$\text{OVD} = \textbf{DL(M)} \quad + \quad \textbf{DL(D|M)}$$

(expected patterns)	(deviations)
(model complexity)	(fit)

$$\text{M1: } y = \theta_0 + e$$

$$\text{M2: } y = \sum_{i=0}^{k} \theta_i x^i + e$$





	DL(M)	DL(D|M)	OVD
M1	**2.0**	**14.5**	**16.5** bits
M2	**7.8**	**3.1**	**10.9** bits

Two Implementations of MDL for Parametric Models

- FIA (Fisher Information Approximation; Rissanen, 1996)
- NML (Normalized Maximum Likelihood; Rissanen, 2001)

$$FIA = -\ln f(y \mid \hat{\theta}) + \frac{k}{2}\ln\frac{n}{2\pi} + \ln\int\sqrt{|I(\theta)|}\,d\theta$$

$$NML = -\ln\frac{f(y \mid \hat{\theta})}{\int f(z \mid \hat{\theta}(z))\,dz}$$

Fisher Information Approximation (FIA)

Rissanen (1996):

$$FIA = \underbrace{-\ln f(y \mid \hat{\theta})}_{\text{Goodness of fit (ML)}} + \underbrace{\frac{k}{2}\ln\frac{n}{2\pi} + \ln\int\sqrt{|I(\theta)|}\,d\theta}_{\text{Model Complexity}}$$

Goodness of fit (ML) + **Model Complexity**

The model that minimizes MDL should be preferred

$$FIA = \underbrace{-\ln f(y \mid \hat{\theta})}_{\{\text{ - Goodness of fit}\}} + \underbrace{\frac{k}{2}\ln\frac{n}{2\pi} + \ln\int\sqrt{|I(\theta)|}\,d\theta}_{\{\text{ Model Complexity}\}}$$

$$\boxed{\underbrace{\frac{k}{2}\ln\frac{n}{2\pi}}_{\text{Complexity due to number of parameters (k) (e.g., AIC, BIC)}} + \underbrace{\ln\int\sqrt{\det I(\theta)}\,d\theta}_{\text{Complexity due to functional form of the model equation}}}$$

Complexity due to number of parameters (k)
(e.g., AIC, BIC)

Complexity due to *functional form* of the model equation

Complexity: More than the number of parameters?

$M1: y = ax + b$

$M2: y = ax^b$

$M3: y = \sin(\cos ax)^a \exp(-bx)/x^b$

Are these all equally complex?

Information Geometric Interpretations of FIA

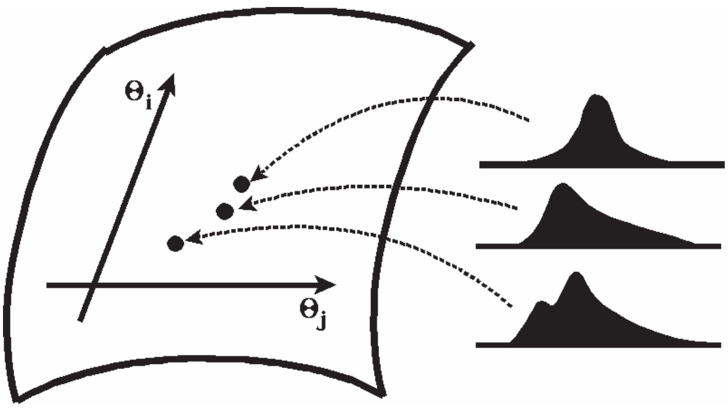
The geometry of the space of probability distributions provides a well-justified and intuitive framework of model complexity, the central concept in model selection.

In this approach, we construct "geometric" complexity of a model by counting the number of different distributions it indexes.

(e.g.) Data space = {a, b, c, d, e, f, g, h}
Model A = {a, c, d} vs Model B = {b, d, e, g, h}

Space of Probability Distributions

The family of probability distributions forms a Riemannian manifold in which "similar" distributions are mapped to "nearby" points (information geometry; Rao, 1945; Efron, 1975; Amari, 1980).



Distance Metric and information Matrix

A distance metric that measures 'dissimilarity' between two neighboring distributions is defined as

$$ds^2 = \sum_{i,j} g_{ij}(\theta)\, d\theta_i\, d\theta_j$$

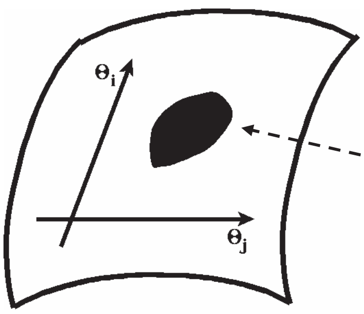
where g_{ij} is the *Riemannian metric tensor* of the form:

$$g_{ij}(\theta) = -E\left[\frac{\partial^2 \ln f(y|\theta)}{\partial\theta_i \partial\theta_j}\right]$$

which is the *Fisher information matrix*, $I(\theta)$.

Complexity and *Riemannian volume*

In a geometric context, model complexity should be related to the volume the associated manifold occupies in the space of distributions:

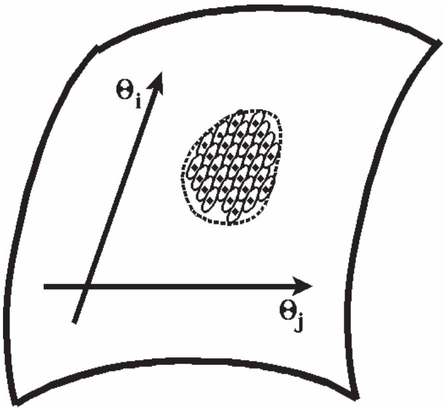


$$\sqrt{\det I(\theta)}\,\Delta\theta$$

which is known as the *Riemanninan volume* in differential geometry.

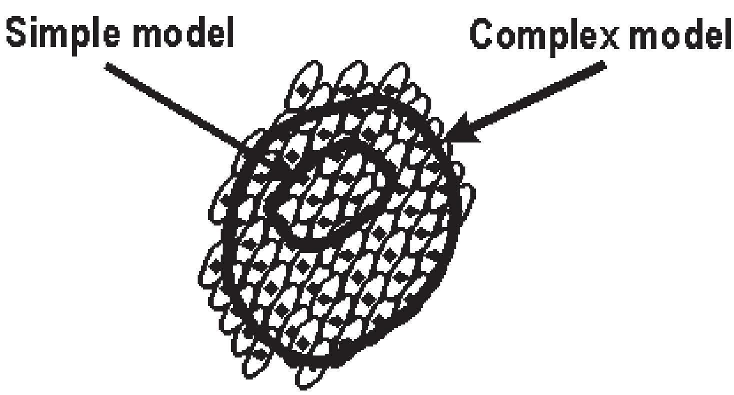
Count only “distinguishable” distributions

The Riemannian volume measure is related to the *local density* of ‘*distinguishable*’ probability distributions indexed by the model.



Overall volume: $V(f) = \int d\theta \sqrt{\det I(\theta)}$

Simple vs complex models: An information geometric view



Distributions close to the truth

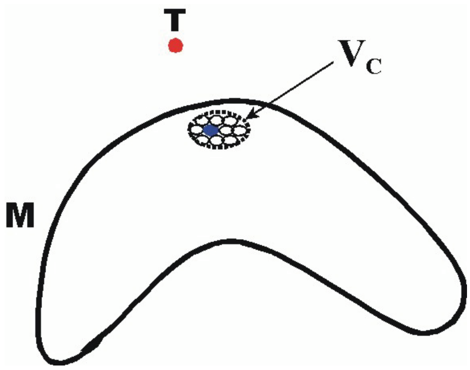
A good model should contain many distinguishable distributions that come close to the truth, in the sense.

$$C\!: a\ subset\ of\ distributions\ s.t.$$

$$f(y|\theta) \approx f(y|\hat{\theta})$$

The Riemannian volume of such region is obtained as:

$$V_C = \int dV_C \approx \left(\frac{2\pi}{n}\right)^{k/2}$$



Model complexity as volume ratio

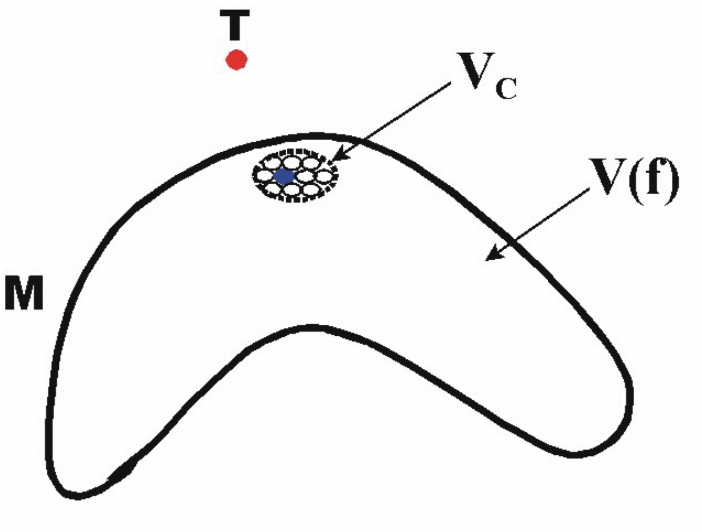
The **log volume ratio,** V(f)/Vc, gives

$$\ln\left(\frac{V(f)}{Vc}\right) = \underline{\frac{k}{2}\ln\frac{n}{2\pi} + \ln\int\sqrt{\det I(\theta)}\,d\theta}$$

"geometric complexity"

➔ *Geometric complexity* turns out to be equal to the complexity term of Fisher Information Approximation (FIA: Rissanen, 1996):

$$FIA = -\ln f(y\,|\,\hat{\theta}) + \underline{\frac{k}{2}\ln\frac{n}{2\pi} + \ln\int\sqrt{|\,I(\theta)\,|}\,d\theta}$$



$$\text{ModelComplexity}_{FIA} = \ln\left(\frac{V(f)}{V_c}\right)$$

From this view, a complex model is one containing many different distributions overall (V(f)) but relatively few ones close to the truth (Vc)

Normalized Maximum Likelihood (NML)

$$NML = -\ln \frac{f(y\,|\,\hat{\theta})}{\int f(z\,|\,\hat{\theta}(z))dz}$$

$$= -\ln \frac{\text{ML value of current data}}{\text{Sum of all ML values of all possible data}}$$

From the NML viewpoint, a good model is the one that gives relatively high ML only for current observations but low ML values for other data patterns.

NML as Minimax Coding Strategy

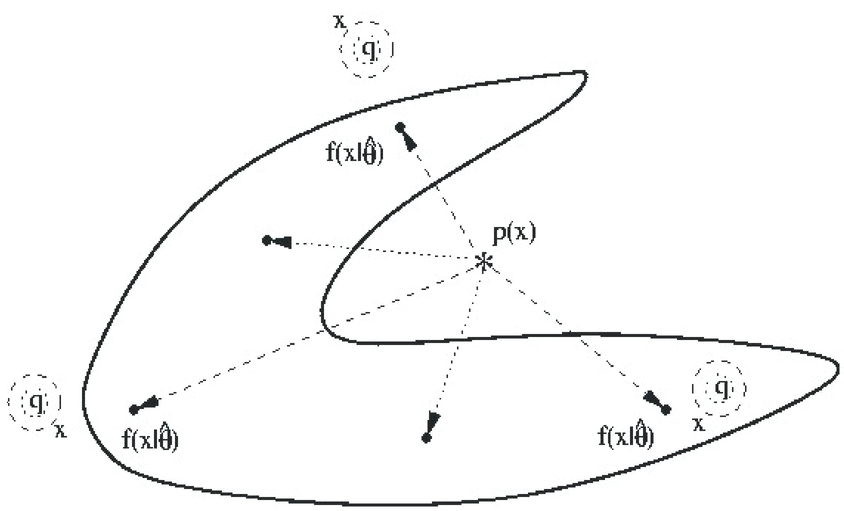
$$NML = -\ln \frac{f(y \mid \hat{\theta})}{\int f(z \mid \hat{\theta}(z))dz}$$

- NML derived as minus logarithm of a probability distribution that minimizes the maximum distance between the desired distribution and the best-fit member of the model family.

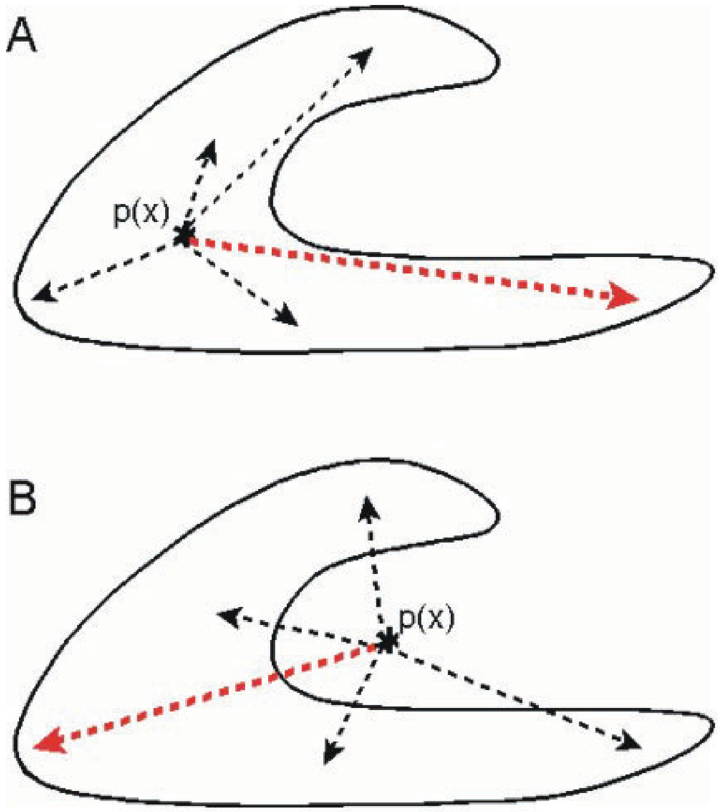
$$NML = -\ln p^*(y)$$

$$where \quad p^*(y) \triangleq \arg\inf_p \sup_q E^q\left[\ln \frac{f(x \mid \hat{\theta})}{p(x)}\right]$$

Minimax Problem in a Model Manifold

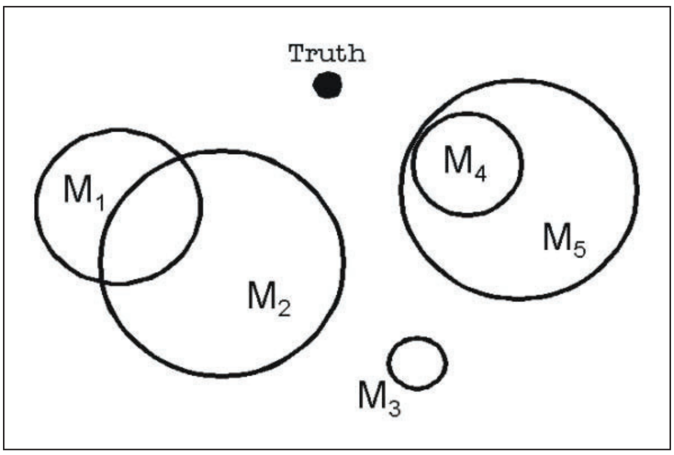


p*(x): "Universal" distribution in the sense that it can mimic the behavior of the entire model class of probability distributions.



Derivation NML as a solution to the minimax strategy does not require that:

- Models be nested within one another;
- None of the models be “true”;
- NML solution be a member of the model family



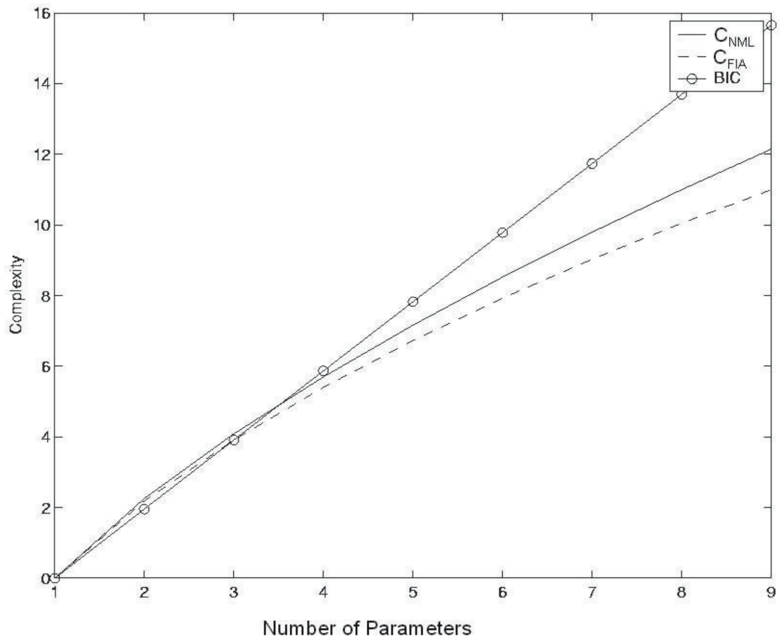
Model complexity: A NML view

$$NML = -\ln f(y \mid \hat{\theta}) + \underline{\ln \int f(z \mid \hat{\theta}(z))dz}$$

C_{NML}:

- Normalizing constant of NML distribution
- Minimax distance achieved
- Sum of all “best” (ML) fits
- Sensitive to number of parameters, sample size, functional form, experimental design, etc.

Complexity Comparison



Other model selection methods

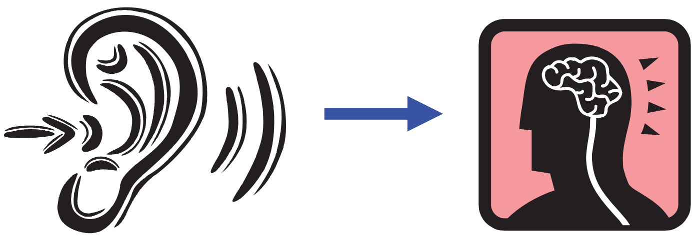
- ICOMP (Bozdogan, 1990)
- RMSEA (Steiger, 1990)
- AGFI (Jöreskog & Sörbom, 1986)
- NIC (Murata et al, 1994)
- DIC (Spiegelhalter et al, 2002)
- FIC (Claeskens & Hjort, 2003)
-
-
-

Overview

- Part 1: Non-technical Introduction to Model Selection
- Part 2: “Technical” Tour of Model Selection Methods
- **Part 3: Example Application**
- Part 4: Conclusions

Example Application in Psychophysics

Models of psychophysics describe the relationship between physical dimensions (e.g., tone intensity) and their psychological counterparts (e.g., loudness).



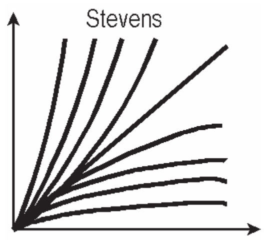
Physics (X) Psycho (Y)

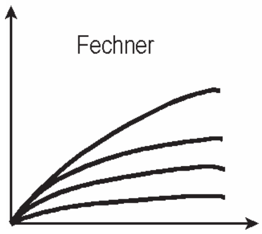
$$Y = f(X, \theta)$$

Psychophysics models

$$Stevens\ law: \quad y = ax^b$$

$$Fechner's\ law: \quad y = a \log(x + b)$$





Complexity difference:

$$C_{MDL,Stevens} - C_{MDL,Fechner} = 3.8$$

The difference in complexity must be due to the effects of functional form

$$FIA = -\ln f(y \mid \hat{\theta}) + \frac{k}{2} \ln \frac{n}{2\pi} + \ln \int \sqrt{|I(\theta)|} d\theta$$

Model Recovery Simulation (% recovery)

Selection Method	Data From: / Model Fitted:	Stevens	Fechner
AIC (BIC)	Stevens	**100**	63
	Fechner	0	**37**
FIA	Stevens	**99**	2
	Fechner	1	**98**

Model Recovery Simulation (% recovery)

Selection Method	Data From: / Model Fitted:	Stevens	Fechner
AIC (BIC)	Stevens	**100**	63
	Fechner	0	**37**
FIA	Stevens	**99**	2
	Fechner	1	**98**

Conclusions

- Models should be evaluated based on **generalizability**, not on **goodness of fit**

"Thou shall not select the **best-fitting** model but shall select the **best-predicting** model."

- Other non-statistical but *very important* selection criteria:
  - Plausibility
  - Interpretability
  - Explanatory adequacy
  - Falsifiability

"All models are wrong, but some are useful."

(G. P. E. Box, 1978)

"Model selection methods can help identify *useful* models, in the sense of *predictive accuracy* or *generalizability*." (J.I.M.)

Thank You for Your Attention!