Statistical Inference with Complex Systems Models

Statistics and Complex Systems

What statistics is * What complex systems are * Bringing the two together

Model Fitting

Fitting models * Generalization * Over-fitting and its control * Penalization/regularization * Cross-validation * Model ensembles

Inductive Complexity

Learning theory * Inductive complexity vs. system complexity * Structural risk minimization

Information Theory and Statistics

Likelihood and entropy * Hypothesis tests * Information geometry

Evaluation of Models

Goodness of fit * Comparing alternatives * Evaluation without solution

Statistics and Complex Systems

Statistics

The science of drawing reliable inferences from partial, noisy, or otherwise imperfect data

- **Stochastic models** of the system and the data-generating process
- **Inferential procedures** going from data to models
- **Evaluation** of the procedures in terms of their errors and reliability

Early statistics, to c. 1930

Sampling: Independent, identically-distributed (IID) variables

$\Rightarrow$ CLT says averages are Gaussian

Regression: Linear relations with Gaussian noise

$\Rightarrow$ Method of least squares

Tedious calculations based on linear algebra

Very powerful in its own domain

Fossilized in various disciplines

Experimental physics, physiology, psychology, political science...

Complex Systems

Many effective degrees of freedom/coordinates, strong interactions

Not: only a few effective coordinates

Not: many *independent* variables

Does this mean we can't use statistics?

Of course we can use statistics!

Modern statistics has
- stochastic models for strong, nonlinear dependence among many variables,
- inferential procedures for fitting such models,
- and methods to evaluate the inferences

So we're good

Complex systems models can also serve as stochastic models

Ergodic deterministic systems might as well be stochastic

Some of them are special cases of standard statistical models

e.g., Agent-based models are "interacting hidden Markov models" or "dynamic Bayes nets with latent variables"

Other lecturers will handle models in detail

Let's talk about inference and evaluation

Why isn't all this standard?

On the one hand:

Biologists, psychologists, etc., learn old statistics that doesn't apply to these problems

Theoretical physicists learn no statistics at all

   OK when you have experimental physicists

   Not OK when you try to analyze data

On the other hand:

Statisticians don't know that "complex systems" exists

Model Fitting

Elements of Model Fitting

A class of models

A measure of inaccuracy or error (**loss function**)

    Error rate (classification)

    Mean squared error (regression) [MSE]

    Likelihood (general probability models)

    Or: negative log-likelihood [Λ]

Data - must have data!

Inference rule

    Picks model from class depending on the data

Generalization

In-sample or **empirical loss:** loss of the model on the data

Out-of-sample or **generalization error:** expected loss on *new* data

We really care about generalization error

We know in-sample error

This is a problem

Over-Fitting

A model is **over-fit** when the generalization risk is worse than the in-sample error
Why does this happen?

(in-sample error) = (generalization) + (sample noise)

Noise term can be negative

Low in-sample error could be

good generalization,
or large negative fluctuation
or both

Empirical Risk Minimization

Pick the model with the best in-sample performance

Problem 1: The generalization risk is usually higher

Problem 2: The best-generalizing model isn't usually the best in-sample model

But, often, ERM converges on the best risk and, less generally, even on the best model!

Control of Over-Fitting

ERM often works in the long run - "but in the
long run we are all dead"
Limit over-fitting to to cut the long-run short:
- penalization or regularization
- cross-validation
- ensembles
- capacity control

Penalization

How is over-fitting possible?

    Next slide: Polynomial fits of order 0, 1, 5, 15

    Slide after next: More points from same distribution, old fit (dashed line), new fit (blue line)

    Points are pure noise, so order-0 fit is actually best

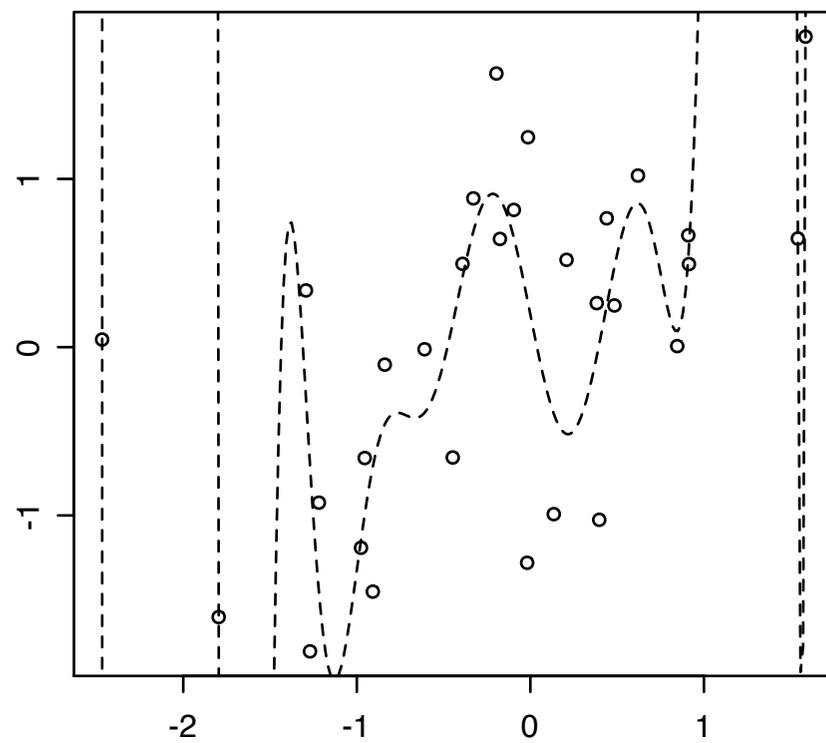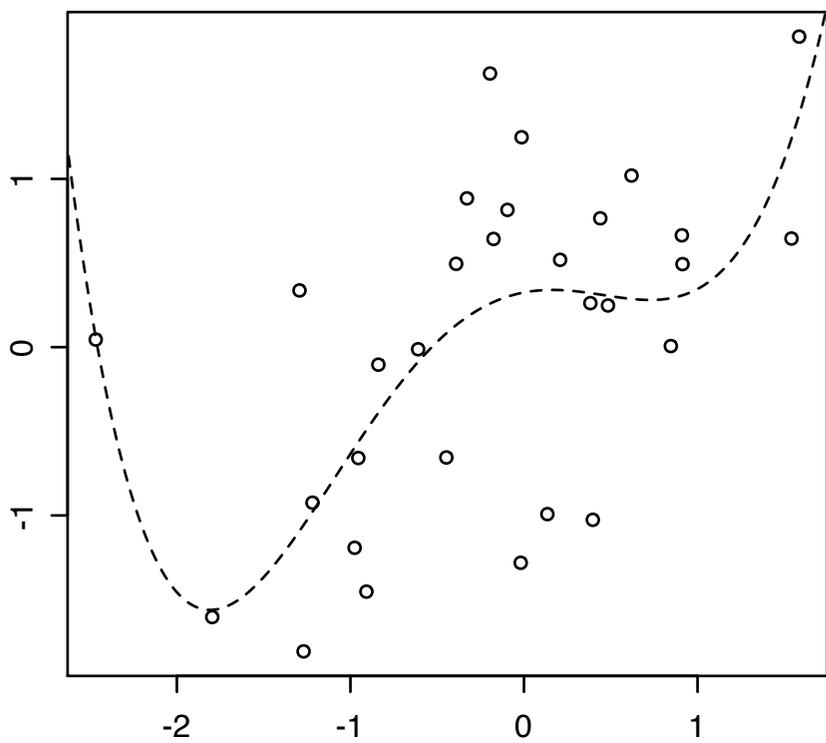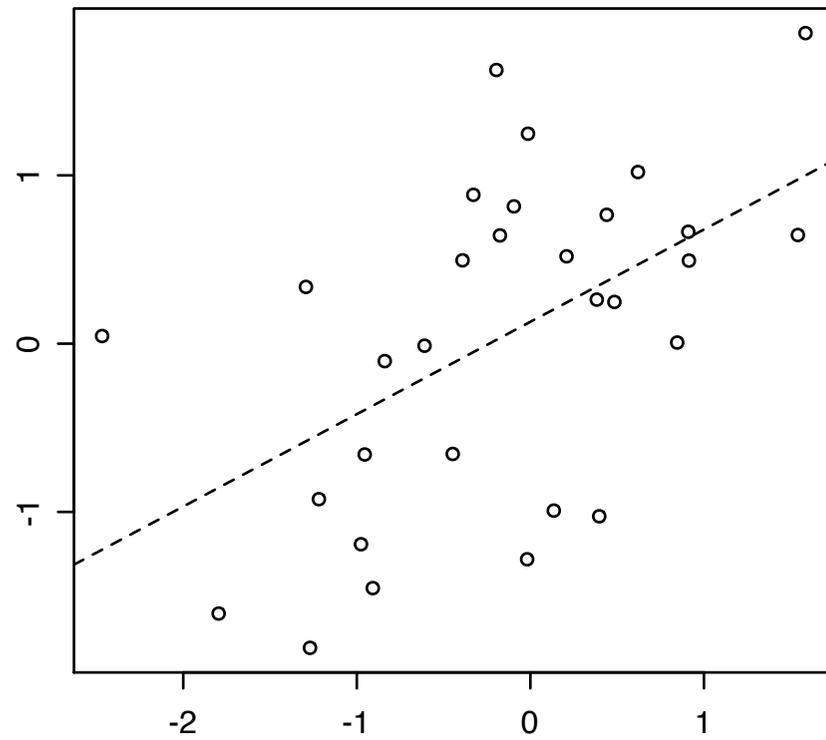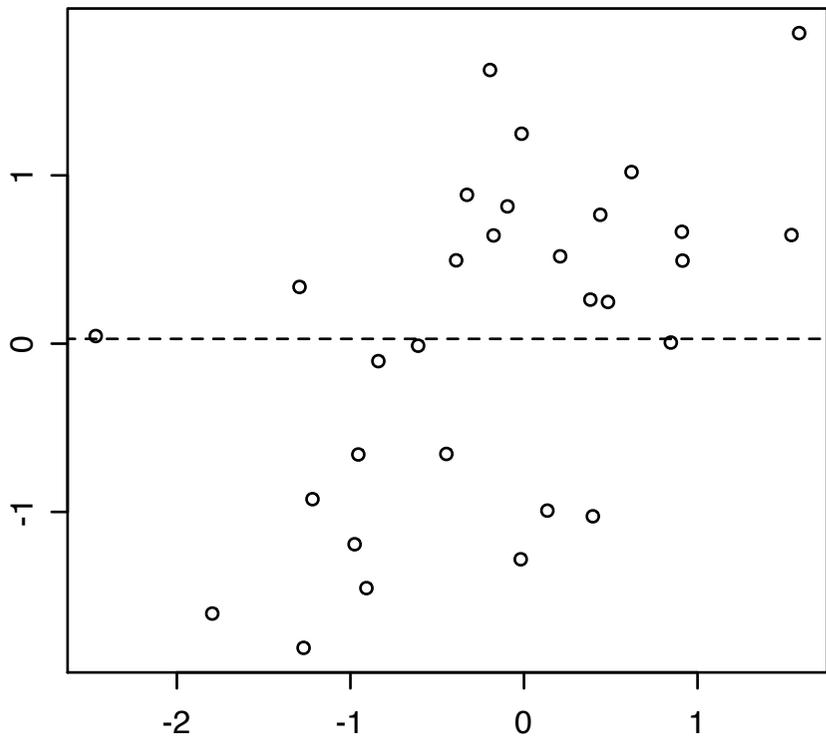The more flexible the model, the more capacity it has to match the data...
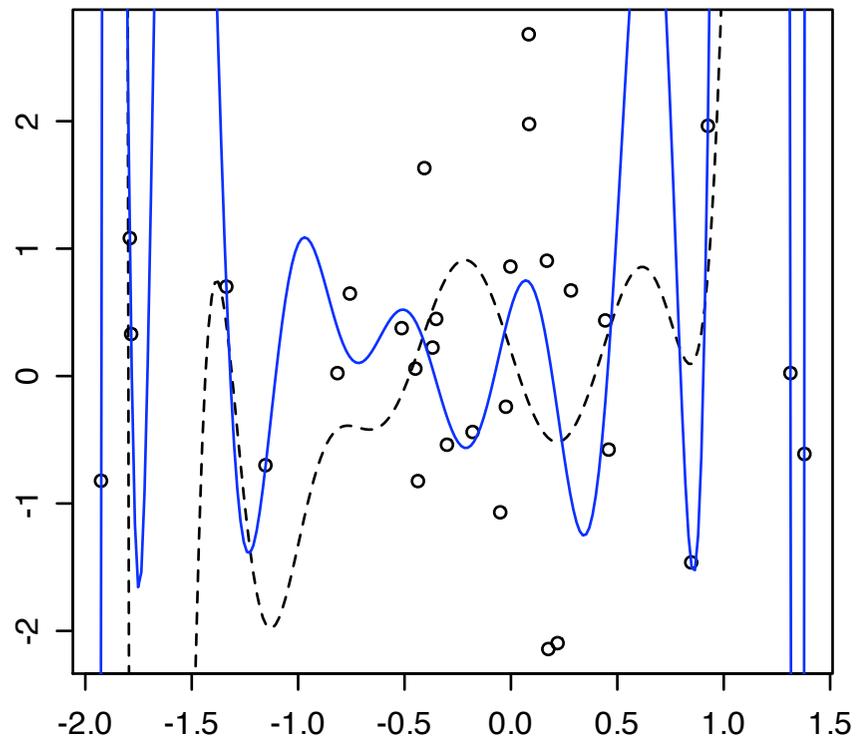
...and to match fluctuations

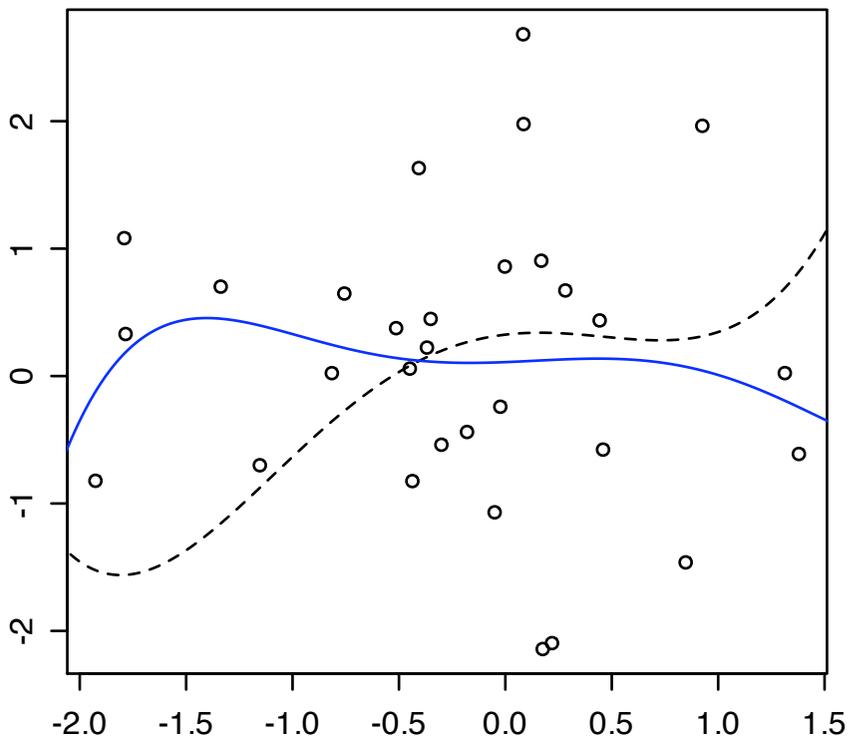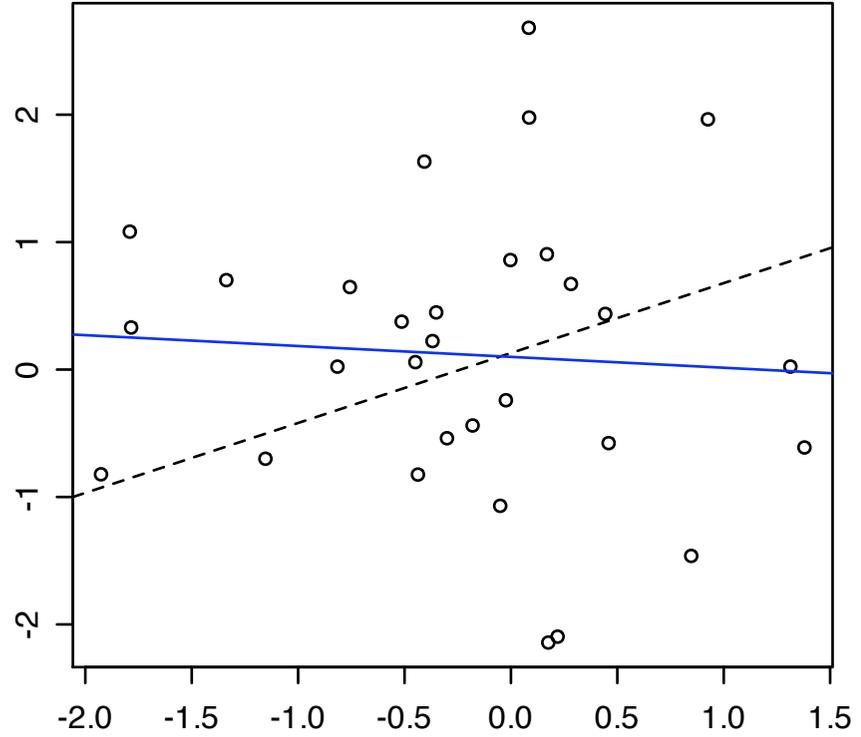   "With five parameters, I can fit an elephant to your data, and with six I can move its tail" (von Neumann)

Add a penalty for flexibility:

    (loss) = (inaccuracy) + (flexibility)

flexibility has to pay for itself
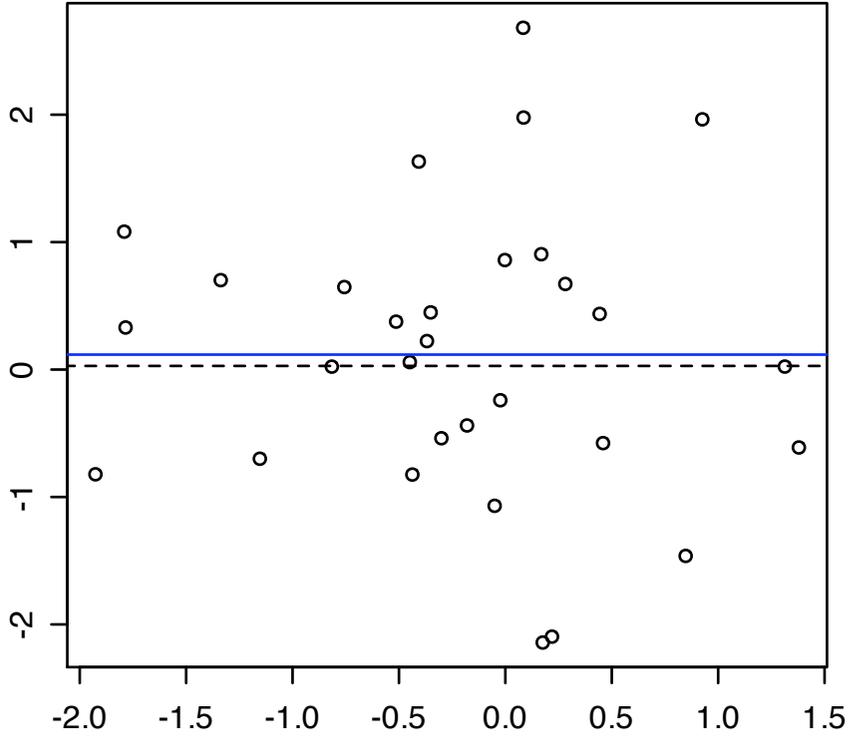
Common Penalizations

Penalization by parameters (a few among many):

Akaike information criterion: $\Lambda + p$

Bayes information criterion: $\Lambda + \dfrac{p}{2}\log n$

Soft thresholding or **Lasso:** $\text{MSE} + \displaystyle\sum_{i=1}^{p} |\theta_i|$

Many similar ones

Penalization by wiggliness $\text{MSE} + \displaystyle\int (\nabla f)^2 dx$

Penalization by coding length

description length $= \Lambda + \#$ of bits to encode model

Penalization by prejudice (Bayesian prior)

$\Lambda - \log(\text{prior probability})$

Cross-Validation

Approximates generalizing

Divide data set into **training** and **testing data**

Fit on training data

Evaluate loss on testing data

Often best to repeat with many divisions into train/test (**k-fold CV**)

One-sample test set, done n times (**leave-one-out CV**)

Sometimes can be done analytically, at least to first order (**generalized cross-validation**)

CV Model Selection

Best CV performance still gives biased estimate of generalization performance
But the biased is generally smaller
    Informally: you have to get lucky twice!
Used extensively and successfully for machine learning and statistical modeling

CV and Error Estimates

Can use CV-style sub-sampling to get range of parameter estimates

  e.g., look at the curve-fitting figures a few slides back
  Leave-one-out especially common here

Parametric bootstrap is a similar idea

For complex models, CV is often better, because closer to data-generating process

Model Ensembles

Fit many and combine them, not just one

Model averaging is the simplest way

e.g., weight by in-sample or CV performance

Or: **boosting**

Fit a model to the data

Re-sample data, emphasizing the worst-fit points

Re-fit

Iterate

Average all these models

Ensemble methods often improve accuracy

    Best model should be *nearly* optimal in-sample

    So will near-best models

    $\therefore$ Large weights (though not necessarily the largest)

Close relationship between model averaging and evolutionary search

Also caricatures of collective cognition

Inductive Complexity

Learning Theory

Probably approximately correct (PAC): can approximate best model arbitrarily closely, with arbitrarily high confidence, given enough data

Depends on:

- model class
- representation of models
- fitting procedure
- loss function
- data-generating process

Inductive Complexity

Inductive or **sample complexity**: how does N grow with accuracy level and confidence level?

$\neq$ of parameters

   One-parameter model classes which *cannot* be learned
   Million-parameter models can be easily learned

Rather depends on model class, loss function and learning rule

vs. System Complexity

Many ways to measure system complexity
None of them really match inductive complexity
Even high-complexity systems can be easily
learned if the model class is constrained

Capacity and Over-Fitting

Example: binary prediction
  1 million data points
  rule R correctly classifies all of them
  probability(err(R) > 1e-5) ≈ 4.5e-5
So if R is the first rule we look at, good odds
err(R) < 1e-5
But if we look at 1 million independent rules,
reasonable chance of getting "lucky"

How many effectively independent models are there in the model class?

This **covering number** grows with # samples

    finer discrimination among models with more data

    growth rate = **capacity** of model class

Quantified by **Vapnik-Chervonenkis dimension** (VC)

    Combinatorial definition

    Depends only on loss function and model class

    Results for lots and lots of different kinds of models: linear models, neural networks, trees, Bayes nets, ...

    Alternatives exist, like **Pollard pseudo-dimension**

Statistical Learning

Distribution-free bounds: Bounds risk of over-fitting in terms of # of samples, confidence, VC dimension, *but no system properties*

- Worst-case: distribution that's hardest to learn
- More for IID data, but now some for ergodic data
- Implies PAC
- Need to know VC dim. of model class
- Some non-VC distribution-free results

VC bound example

Binary classification

Rule R has error rate p on n samples

VC dimension of rule class $= d$

With probability at least $1-\delta$,

$$\text{error}(R) \le 2p + \frac{4}{n}\left(d\log_2\frac{2en}{d} + \log_2\frac{4}{\delta}\right)$$

Distribution-dependent bounds: Can often prove faster rates of convergence if you assume certain distributional form.

> Statistcal mechanics methods can sometimes give **average case** learning rates (instead of distribution-free worst-case results)
>
> **Large deviations theory** gives rates of convergence for the limit theorems of probability; very useful here

Data-dependent bounds: Intermediate case, depends on sample data but not distribution

> **e.g. margin bounds** on inaccuracy of classifiers by distance of sample points from class boundaries

Structural Risk Minimization

Use many model classes, $M1 \subset M2 \subset M3 \subset ...$

Use model class that minimizes

  (in-sample error) + (VC bound on over-fitting)

Penalize by model class, not by model

Gives *much* better performance on real-world data than other penalization schemes

  Domingos 1999

Likelihood and Entropy

$$\Lambda = -\sum_{i=1}^{n} \log p(x_i) \equiv -\sum_{x} c(x) \log p(x)$$

$$= -n \sum_{x} \frac{c(x)}{n} \log p(x) \equiv -n \sum_{x} \hat{p}(x) \log p(x)$$

$$= -n \sum_{x} \hat{p}(x) \log p(x) \frac{\hat{p}(x)}{\hat{p}(x)}$$

$$= -n \sum_{x} \hat{p}(x) \log \hat{p}(x) + n \sum_{x} \hat{p}(x) \log \frac{\hat{p}(x)}{p(x)}$$

$$= nH[\hat{p}] + nD(\hat{p}\|p)$$

First term: source entropy, the same for all models
Second term: relative entropy, model-dependent
$\therefore$ Minimizing Λ is the same as minimizing relative entropy

Hypothesis Tests

Optimal test of distribution p vs. q has error rate which goes like $2^{-nD(p\|q)}$

Likelihood-based tests often attain this rate

Relative entropy = how easy is it to tell these distributions apart?

Large deviations theory shows similar results for more complicated models

Information Geometry

Space of statistical distributions

Model classes = manifolds

MLE = projection of data onto manifold

Fisher information

    Curvature = how easy is it to distinguish nearby models?

Cramer-Rao bound

    Curvature $\propto$ optimal parameter estimation rate

Applications to learning

Others to measuring complexity

Evaluating Models

A Bad Example
(a paper I rejected in March)

Bursts of activity in EEG in sleep

Experimental claim: these intermittently synchronized across the brain

Model: found intermittent synchrony in coupled oscillators (Kuramoto model) at critical point

Theoretical claim: therefore, the sleeping brain is a critical network of coupled oscillators

What's wrong with this?

Random constants

    many arbitrary thresholds; what if they were different?

Bad power laws

    intermittency = power-law distribution of stable periods, and they drew a straight line on a log-log plot

No assumption-checking

    Neurons aren't phase oscillators - does that matter?

    Neuronal networks aren't all-to-all - does that matter?

    etc.

    No reason to believe in universality here

No checking of alternatives

    There other ways to produce intermittent synchrony

What should you do?

Check assumptions
    Fit small parts of the model to data *separately*
    Model sensitivity/robustness analysis

Rule out alternative explanations
    Try things already in the literature
    Try boring things
    Remove mechanisms, interactions, etc., see if it breaks

Can you predict things you didn't fit initially?

Can you predict things without fitting?

Goodness-of-fit Tests

Need to measure divergence between data and model's expectations

    Kolmogorov-Smirnov test is a g.o.f. test

p-value = Probability of diverging as far from the model as the actual data

    low p-value $\Rightarrow$ bad fit

Bootstrap:

    Simulate the model many times

    Find distribution of divergence between model and simulated data

A Caution

Low p-value $\Rightarrow$ bad fit

  either the model is wrong, *or* you are very unlucky

Does high p-value $\Rightarrow$ good fit?

Only if the test has high *power*

  A test which *always* passes your model has false alarm rate 0, but no power at all

  Using R^2 to tell if something is a power law is like this

This is why you should really compare against serious alternative models

Sensitivity Analysis

How much does the model's output change under small changes in assumptions?

If the changes are large (high sensitivity)

   You'd better be pretty sure about model inputs

   Comparing model to data is informative about system

If the changes are small (low sensitivity)

   Errors in model inputs matter less; relax

   Harder to learn about system details by comparing model to data

   (a bit like Cramer-Rao bound)

A Bad Word for Econophysics

Agents' decision-rules are easy to analyze, if you
know a lot of statistical mechanics
How *people* make decisions has been extensively
studied experimentally
It is *nothing at all* like econophysical rules
Nobody knows how robust these models are
(Rational-choice economics isn't much better)

Checking Alternatives

Go beyond "stylized facts"

  There are usually *many* ways to get "qualitative
  agreement", so that's not helpful

Use *serious* alternatives

  Go to the literature and see what's been proposed
  Go to the literature in other fields

Use *boring* alternatives

  Try to remove your favorite, most exciting part of the
  model; can you tell the difference?
  Example: **neutral models** in biology
  Example: **surrogate data** in nonlinear dynamics

About Toy Models

Logistic map, Ising model, minority game, ...

Toy models can be useful

- Easier to understand
- Easier to prove results about
- Easier to simulate
- Give an idea of what is logically *possible*
- What kind of process *could*, qualitatively, give us X?

Toy models are of limited use

- Toys give quantitative accuracy in physics because of universality *at critical points*
- This doesn't even apply in most of physics!

Boring Alternatives

Neutral models: What would this model do without any of the *adaptive* mechanisms?

   Extensively used in evolutionary biology, ecology
   Some examples in social science but not many

Surrogate data: Tests for chaos, intermittency, etc.

   Fit a non-chaotic model (say, linear stochastic process)
   Make sure it's a reasonably good match
   Simulate many times
   Calculate measure of chaos
   Find p-value

Common idea: get rid of the exciting bit

Getting at Mechanisms

Does your model predict things *other than* its own inputs?

Good example: The model of foraging in Oaxaca

    Inputs = plant properties, terrain, how agents make choices

    Outputs = distribution of plant waste in caves

Can *other approaches* do this too?

    e.g., Many mechanisms predict power laws (or Gaussians, etc.)

If so, look for something else which does separate the models

    e.g., Different processes for power-laws have different dynamics, so try to look at growth histories

Evaluation without Fitting

Find results about model outputs valid for *all* inputs

Then check those in your data

These results are often inequalities

   Not as precise a prediction, but better than nothing

Example: thermodynamics

   Could come up with a detailed microscopic model of a steam engine, with combustion, condensation, friction, ...

   Or you could use Carnot's Law and measure two temperatures

Sutton applies this idea in economics

Summary

1. Modern statistics has tools to fit and evaluate large models with strong interactions

2. These tools can and should be used with complex systems models

3. Over-fitting is a serious but controllable problem

4. Be careful about model evaluation

References

General Sources on Modern Statistics

David Hand, Heikki Mannila and Padhraic Smyth (2001), *Principles of Data Mining* (MIT Press)
> Emphasizes very large data sets, computational efficiency, and pattern discovery.

Trevor Hastie, Robert Tibshirani and Jerome Friedman (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer-Verlag)
> Emphasizes statistical aspects, especially model evaluation and control of over-fitting.

Larry Wasserman (2003), *All of Statistics: A Concise Course in Statistical Inference* (Springer-Verlag)
> Based on an extremely successful interdisciplinary course at Carnegie Mellon. Teaches necessary probability theory as it goes.

Modern Models and Inference

Nello Cristianini and John Shawe-Taylor (2000), *An Introduction to Support Vector Machines, and Other Kernel-Based Learning Methods* (Cambridge University Press)
> Great tutorial on an important class of models especially designed to exploit learning-theoretic principles, which are simply explained here

Peter J. Green, Nils Lid Hjort, and Sylvia Richardson, eds. (2003), *Highly Structured Stochastic Systems* (Oxford University Press)
> Conference proceedings, with contributions on models for many traditional topics of complex systems (epidemics, time series, etc.), and techniques for inference.

Michael I. Jordan, ed. (1998), *Learning in Graphical Models* (MIT Press)
> Conference proceedings, with many valuable tutorial and review papers.

Brian D. Ripley (1996), *Pattern Recognition and Neural Networks* (Cambridge University Press)
> Actually covers many other model classes as well.

Achilleas Zapranis and Apostolos-Paul Refenes (1999), *Principles of Neural Model Identification, Selection and Adequacy: With Applications to Financial Econometrics* (Springer-Verlag)
> Specialized to neural networks in finance, but *very* clear on general principles, and why they matter in practice.

Computational and Statistical Learning Theory

Pedro Domingos (1999), "The Role of Occam's Razor in Knowledge Discovery", *Data Mining and Knowledge Discovery* **3**: 409--425, http://www.cs.washington.edu/homes/pedrod/papers/dmkd99.pdf
> Great paper on how and why different penalization schemes work, or fail, in practice.

A. Engel and C. Van den Broeck (2001), *Statistical Mechanics of Learning* (Cambridge: Cambridge University Press)
> Emphasizes methods developed for disordered systems, like spin glasses.

Michael J. Kearns and Umesh V. Vazirani (1994), *An Introduction to Computational Learning Theory* (MIT Press)
> Well-written textbook. Some parts presume a first course in analysis of algorithms.

Ron Meir (2000), "Nonparametric Time Series Prediction Through Adaptive Model Selection", *Machine Learning* **39**: 5--34
> Excellent paper on applying learning theory to stochastic processes

David Pollard (1984), *Convergence of Stochastic Processes* (Springer-Verlag), http://www.stat.yale.edu/~pollard/1984book/
> Presumes advanced knowledge of probability theory.

Vladimir N. Vapnik (2000; 2nd ed.), *The Nature of Statistical Learning Theory* (Springer-Verlag)
> Opinionated but insightful introduction by one of the founders.

Sara van de Geer (2000), *Empirical Processes in* M-*Estimation* (Cambridge University Press)
> Analyzes convergence rates for different penalization schemes. Presumes advanced knowledge of theoretical statistics.

T. L. H. Watkin, A. Rau and M. Biehl (1993), "The Statistical Mechanics of Learning a Rule", *Reviews of Modern Physics*,**65**: 499--556

Information Theory and Statistics

Shun-ichi Amari and Hiroshi Nagaoka, *Methods of Information Geometry*
> Amari's summary of his pioneering work in information geometry. Self-contained, but mathematically demanding.

Nihat Ay (2001), "Information geometry on complexity and stochastic interaction", http://www.mis.mpg.de/preprints/2001/prepr9501-abstr.html

---- (2002), "An information-geometric approach to a theory of pragmatic structuring", *Annals of Probability* **30**: 416--436, http://www.mis.mpg.de/preprints/2000/prepr5200-abstr.html

Harald Cramer (1946), *Mathematical Methods of Statistics* (Princeton University Press)
> Classic textbook on the principles of theoretical statistics, including Cramer's original derivation of the Cramer-Rao bound

Solomon Kullback (1968; 2nd ed.), *Information Theory and Statistics* (New York: Dover).
> Pioneering work by the co-inventor of relative entropy.

Rudolf Kulhavy, *Recursive Nonlinear Estimation: A Geometric Approach* (Springer-Verlag)
> The easiest and most practical introduction to the geometric approach to information theory and statistics.

Jorma Rissanen (1998; 2nd ed.), *Stochastic Complexity in Statistical Inquiry* (World Scientific)
> Enthusiastic account of the "minimum description length" approach to inference by its inventor.

Model Evaluation, etc.

Robert P. Abelson (1995), *Statistics as Principled Argument* (Hillsdale, New Jersey: Lawrence Erlbaum Associates)
> Written by a psychologist for psychologists, but really applicable to any scientific field.

Colin F. Camerer (1995), "Individual Decision Making", pp. 587--703 in John H. Kagel and Alvin E. Roth (eds.), *The Handbook of Experimental Economics* (Princeton University Press)
> How people actually make economic decisions. (In short: nothing like what econophysics assumes.) Camerer has published a follow-up book on experimental game theory, but I've not read it yet.

John H. Gillespie (1998), *Population Genetics: A Concise Guide* (Baltimore, Maryland: Johns Hopkins University Press)
> Nice introduction, with a good treatment of neutral models.

Stephen P. Hubbell (2001), *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press)
> Neutral models in aspects of ecology. Some of Hubbell's ideas are controversial, but he does a good job of laying out why neutral models are useful.

Holger Kantz and Thomas Schreiber (1997), *Nonlinear Time Series Analysis* (Cambridge University Press)
> Very good on surrogate data methods (along with everything else).

Stanley Lieberson (2000), *A Matter of Taste: How Names, Fashions, and Culture Change* (Yale University Press)
> Neutral, i.e. non-adaptive, models in cultural evolution, though he doesn't use that phrase.

Deborah G. Mayo (1996), *Error and the Growth of Experimental Knowledge* (University of Chicago Press)
> First-rate methodological examination of the model evaluation process, null models, and learning from error.

John Sutton (1998), *Technology and Market Structure: Theory and History* (MIT Press)
> Detailed example of Sutton's inequalities-valid-for-many-models approach, confronted with extensive empirical data about the evolution of particular industries. Terrific but long and involved.

---- (2000), *Marshall's Tendencies: What Can Economists Know?* (MIT Press)
> Short, informal presentation of Sutton's methodological ideas.