Some Remarks on the Model Selection Problem

Branden Fitelson & Justin Sharber

Department of Philosophy

&

Center for Cognitive Science (RuCCS)

Rutgers University

branden@fitelson.org

- We'll adopt a simple framework today. Our assumptions:
  - A **model** ($\mathcal{M}$) is a family of *hypotheses*.
  - A **hypothesis** (H) is a *curve* plus an associated *error term* ϵ. For simplicity, we'll assume a common $\mathcal{N}(0,1)$ *Gaussian* ϵ.
  - To fix ideas, we will focus today on this family $\mathcal{F}$ of four (parametric) models with *univariate, polynomial hypotheses.*

  (LIN) $y = \mathsf{a}\,x + \mathsf{b} + \epsilon$.

  (PAR) $y = \mathsf{c}\,x^2 + \mathsf{d}\,x + \mathsf{e} + \epsilon$.

  (CUB) $y = \mathsf{f}\,x^3 + \mathsf{g}\,x^2 + \mathsf{h}\,x + \mathsf{i} + \epsilon$.

  (QRT) $y = \mathsf{j}\,x^4 + \mathsf{k}\,x^3 + \mathsf{l}\,x^2 + \mathsf{m}\,x + \mathsf{n} + \epsilon$.

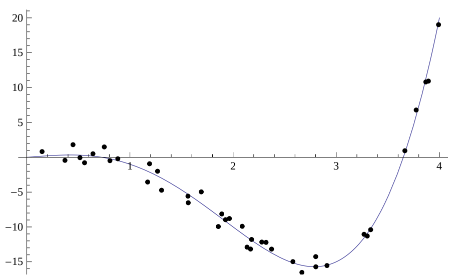
  - Note: these are *nested* models: LIN $\subset$ PAR $\subset$ CUB $\subset$ QRT. *E.g.*, LIN = PAR with $\mathsf{c} = 0$; PAR = CUB with $\mathsf{f} = 0$, *etc.*
- We remain neutral on the origin/status of ϵ. Perhaps ϵ is due to observational error, perhaps it's more metaphysical.
- We can visualize hypotheses, as polynomials with super-imposed $\mathcal{N}(0,1)$ ϵ-distributions. Examples, below.

- We will assume that there is a (single) **true hypothesis** ($\mathfrak{t}$), which generates all of the data sets that we will discuss.
- The smallest model containing $\mathfrak{t}$ is the **true model.** [This makes the alternative models *mutually exclusive* ([6], [4]).]
- A **data set** (E) is a set of $\langle x, y\rangle$ points, generated by $\mathfrak{t}$.
  - **Example #1**. The true hypothesis is the following:

    ($\mathfrak{t}_1$) $y = x^4 - 4x^3 + x^2 + x + \epsilon$.

    Here, the *only* true model is (QRT). Suppose we observe this 40-point data set (E_1) with $x \in [0,4]$, generated by $\mathfrak{t}_1$:

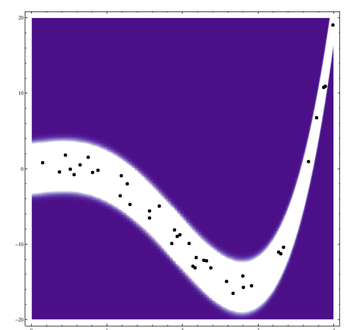
(E_1)



- The **likelihood** of a hypothesis H — relative to data set E — is the probability of E, conditional on the truth of H.

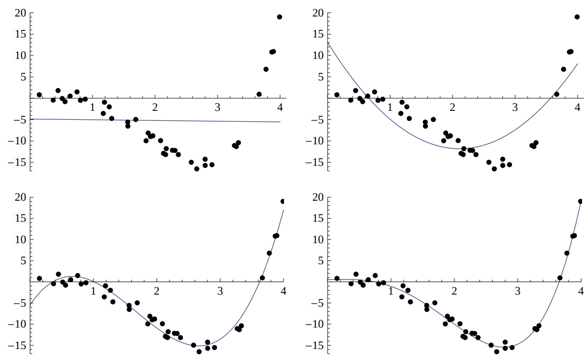
  Likelihood of H, relative to E $\stackrel{\text{def}}{=} \Pr(E \mid H)$.

- Here's what the dataset E_1 and the hypothesis $\mathfrak{t}_1$ look like (white region is such that $\mathfrak{t}_1$ *has a non-negligible likelihood*).

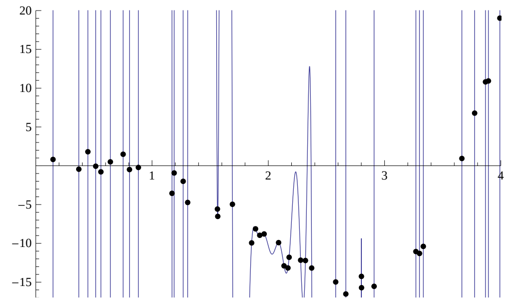


- Next, we'll look at the (standard) method for the selection of a *hypothesis* — *within* a given model — *Maximum Likelihood.*

- The **maximum likelihood** (ML) approach to *hypothesis* selection yields the maximum likelihood hypothesis H — relative to the *actual* data set E — *within* a given model $\mathcal{M}$.
- As Gauss knew [10], it is easy to find the ML $\hat{H}_M \in \mathcal{M}$ (assuming Gaussian ϵ), since $\Pr(E \mid H)$ is proportional to the *sum of squared deviations* (SOS) of H from the actual data E.
- In our Example #1, the ML-trends for the 4 models look like:



- In general, ML-hypotheses from higher-degree (more complex) models have greater likelihoods, relative to any E.
- If a data set E has n data points, then there exists a (unique) $n-1$–degree polynomial that will *interpolate* E. *E.g.*, for E_1:



- But, this is a *crazy* hypothesis! [*E.g.*, we wouldn't want to use an *interpolating* hypothesis for purposes of *prediction*.]
- This is because such complex trends tend to *over*-fit — they *perfectly* fit the *actual* data, but they tend to do very poorly (*i.e.*, have *low* likelihood) on *novel* data sets. [Show animation.]

- The **curve-fitting problem** (CFP) has two sub-problems
  - The **model selection** problem (MSP).
    - This involves selecting a model M, from a family of models.
    - ☞ Here, ML will *not* suit. If we choose the "maximum likelihood model", this implies selecting the *most complex* model. And, this will (inevitably) lead to *over*-fitting.
  - The **hypothesis selection** (within M) problem (HSP).
    - After selecting a model M, we also need to select a hypothesis $H \in M$, since H's are what make contact with E's.
    - The (HSP) is typically "solved" *via* ML. [We won't discuss (HSP) today, but one could question ML *even here*.]
- We will focus the (MSP) today. Here's the plan:
  - First, we will try to clarify what *aims* one might have in trying to "solve" the (MSP). This will lead to *two* "(MSP)"s.
  - We will discuss two kinds of approaches to (MSP) — *Naïve Bayesian* approaches (NB) and *Information-Theoretic* (IT) approaches — with respect to two popular *aims* for (MSP).

- Because the (MSP) is *part* of the (CFP), we must first get clearer on what we're going to count as a "solution" to (CFP).
- This depends on our *aims* when we're "curve-fitting".
- Here are two possible *aims* one could have w.r.t. (CFP):

  Truth. The aim is to select the *true* model/hypothesis [8].

  Predictive Accuracy. The aim is to select the *most predictively accurate* model/hypothesis [5].

  - There are other aims one could have here. *E.g.*, one could aim to select the most *explanatory* model/hypothesis, or the most *beautiful*, *etc., etc.* We'll bracket these other aims here.
- Note: there is an important logical relationship between these two aims. Truth is the *strictly more ambitious* aim.
- There are various reasons why one might *not* want to take truth to be the aim here. [*E.g.*, if the truth is not included among the alternative hypotheses, then it's *impossible*.]

- We seek **model selection procedures** that take as *input*:
  (1) An *actual (small-ish) data set* E.
    - "Small-ish", because ML is fine, *asymptotically*, as $n \to \infty$.
  (2) A *family of models* $\mathcal{F}$.

  and *outputs*:
  (3) a *selected model* $M \in \mathcal{F}$,

  which will then be input to an ML procedure that selects:
  (4) the $\hat{H}_M \in M$ that has maximum likelihood relative to E.
  - Note: it may be either M or $\hat{H}_M$ (or both) that is (ultimately) *assessed — relative to our aim regarding the (CFP).*
- This gives us a framework for *evaluating model selection procedures*. But, unless we *know the true hypothesis*, we will not be in a position to *know* which procedures are better than which — with respect to whichever aim we have.
- We'll use our toy Example #1 to illustrate "ideal" evaluation. [We'll discuss non-ideal evaluation – briefly – later on.]

- **Truth-Bayesian** (TB) model selection procedures are usually described as furnishing us with a *means* to the "*truth*–aim".
- The basic idea behind (TB) is to *maximize the posterior probability that the selected model* M *is the true model* [11].
- That is, (TB) tries to achieve the truth aim *via* selecting the model M with maximum posterior probability: $\Pr(M \mid E)$.
- **Bayes's Theorem.** The posterior probability of M depends on the *likelihood* of M and the *"prior" probability* of M.

$$\Pr(M \mid E) = \frac{\Pr(E \mid M) \cdot \Pr(M)}{\Pr(E \mid M) \cdot \Pr(M) + \sum\limits_{M' \neq M \in \mathcal{F}} \Pr(E \mid M') \cdot \Pr(M')}$$

- There are three (main) problems with (NB) approaches:
  (a) Where does the "prior" of a model $\Pr(M)$ come from?
  (b) How do we calculate the "likelihood of a model" $\Pr(E \mid M)$?
  (c) Must (NB) be *truth-conducive*? Must "maximum posterior probability of truth" be (*objectively*) correlated with *truth*?

- At this point, it helps to distinguish two cases.

  Good Case (GC). The true model *is* in $\mathcal{F}$. If $\mathfrak{M} \in \mathcal{F}$, then achieving the truth aim is *possible*. And, regarding (a)–(c):

  (a) Of course, *we don't know* whether we're in (GC). So, we needn't assign probability 1 to the disjunction of the models in $\mathcal{F}$. [Also, we could assign higher priors to simpler models ([6], [4]). But, I won't do that, since I don't want the priors to do the work. So, I will go *uniform* over $\mathcal{F}$, for simplicity.]

  (b) Since models contain (uncountably) *many* hypotheses, "likelihoods of models" are going to have to be some sort of "averages" of the likelihoods of the underlying hypotheses.

  $$\Pr(E \mid M) = \text{average}(\{\Pr(E \mid H) \mid H \in M\})$$

  It's not clear *which* "average" to use. Historically, one sees attempts to "go *uniform*" over M's parameters ([6],[7],[11]). This raises difficulties, since parameters are on $[-\infty, \infty]$. I will discuss an alternative, "quasi-empirical" averaging.

  (c) Why should maximizing "$\Pr(M \mid E)$" be (objectively) *truth-conducive*? Whether it is will depend on the "priors" in (a), the "averaging" in (b), and their connection to $\mathfrak{M}$, *etc.*

Bad Case (BC). The true model $\mathfrak{M}$ is *not* contained in $\mathcal{F}$. If $\mathfrak{M} \notin \mathcal{F}$, then achieving the truth aim is *impossible*. In this case, it is odd to describe the aim as being "the selection of the true model".

- A Bayesian could add a "catch-all model" $\sim\mathcal{F}$, which asserts that the true model is not contained in $\mathcal{F}$. This won't really help.
- Now, they will need to (a) assign a "prior" [$\Pr(\sim\mathcal{F})$] to $\sim\mathcal{F}$ and also (b) calculate a "likelihood" [$\Pr(E \mid \sim\mathcal{F})$] for $\sim\mathcal{F}$.
  - It's not at all clear where these probabilities are going to come from (or what grounds them as "guides to truth").
- Moreover, step (4) of the (CFP) will no longer be achievable, since there seems to be no principled way to calculate "the maximum likelihood hypothesis", given $\sim\mathcal{F}$, relative to the actual data E.
- One can proceed *as if* $\mathfrak{M} \in \mathcal{F}$, and apply "Good Case" methods *always*. But, then problem (c) becomes *even more pressing*.
- For the sake of simplifying the discussion, I'll just *hope* we are in (GC). And, I will use Example #1 to illustrate one flavor of (TB).

- Here's a "quasi-empirical" Bayesian approach to our Example #1.
  - First, go *uniform on* $\mathcal{F}$: $\Pr(M_i) = c$, for each model $M_i \in \mathcal{F}$.
  - Then, we need to calculate "likelihoods" for the M_i, relative to the actual data set E_1. Here's a "quasi-empirical averaging" method:
    - For each model M, the error distribution ϵ induces a multivariate probability distribution $\mathcal{P}$ over the parameter values of the ML-hypotheses $\hat{H}_M \in M$. This distribution $\mathcal{P}$ is *itself Gaussian*.
    - We can use E_1 to calculate an *estimate* ($\hat{\mathcal{P}}$) of $\mathcal{P}$. This involves averaging the sample mean (and variance) of the $\hat{H}_M$ parameters over the 40 "leave one out" data sets that can be generated from the full data set E_1 (this is a *bootstrapping* approach [9], [3]).
    - ☞ Think of our $\hat{\mathcal{P}}$ as estimating the "average" (ML) hypothesis $\bar{\hat{H}}_M$ (from M), over all the hypothetical data sets E generated by $\mathfrak{t}_1$.
    - Once we have $\hat{\mathcal{P}}$, we can use it to provide the "weights" in our calculation of the "average likelihood" for each model M.
    - Then, plug-in these "average likelihoods" as the $\Pr(E_1 \mid M)$-terms in our Bayes's Theorem calculation of the posteriors $\Pr(M \mid E_1)$. Finally, *select the model with the maximal value of* $\Pr(M \mid E_1)$.

- Information-Theoretic (IT) approaches to model selection take *predictive accuracy* (as opposed to truth) as their aim [5].
- Predictive accuracy can be thought of as some sort of "distance/divergence from the true hypothesis".
- So, on (IT) approaches, the aim of (CFP) is to select a hypothesis that *minimizes divergence from the true hypothesis*.
- There are many different information-theoretic measures of "divergence" or "distance" from the true hypothesis. And, each of these could be used to ground an (IT) approach to (MSP).
- One commonly used measure is called the *Kullback-Leibler* (KL) *divergence*. The KL-divergence is intimately connected with *likelihood*, and so it is a natural choice in the present setting [1].
- Various information-theoretic criteria for model selection have been proposed [2]. They all involve minimizing some estimate of (some sort of) divergence from the true hypothesis.

- Even if the true hypothesis $\mathfrak{t}$ is not contained in the family of models $\mathcal{F}$, each model $M \in \mathcal{F}$ will nonetheless contain a hypothesis that is "closest to the truth" among the $H \in M$.
- Specifically, each model M will contain a hypothesis $H_M^\star$ that is *closest in KL-divergence* to to the true hypothesis. (IT)-based approaches aim to select the *overall closest* hypothesis in $\mathcal{F}$.
- In closing, I'll discuss an intimate connection between the "quasi-empirical" (NB)-approach above and (IT)-approaches.
- The parameter values of $H_M^\star$ are just the mean parameter values, under the ϵ-induced distribution $\mathcal{P}$ that we discussed above.
- In other words, if one averages the parameter values of the ML-hypotheses $\hat{H}_M$ over many data sets E generated by $\mathfrak{t}$, these averages will converge to the values of the parameters of $H_M^\star$.
- Thus, our "quasi-empirical" approach above borrows elements from both the classical Bayesian and Information-Theoretic approaches.

[1] H. Akaike. 1973. "Information theory and an extension of the maximum likelihood principle", in B.N. Petrov and F. Csaki (*eds.*), *Second International Symposium on Information Theory*.

[2] K. Burnham and D. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer.

[3] B. Efron and R. Tibshirani. 1994. *An Introduction to the Bootstrap*, CRC.

[4] M. Forster. 1996. "Bayes and Bust: Simplicity as a Problem for a Probabilist's Approach to Confirmation", *British J. for the Philosophy of Science*, 46: 99–424.

[5] ———. 2001. "The New Science of Simplicity", in A. Zellner, H. A. Keuzenkamp, and M. McAleer (*eds.*), *Simplicity, Inference and Modeling*, Cambridge.

[6] H. Jeffreys. 1961. *Theory of Probability*, Oxford University Press.

[7] R. Kass and L. Wasserman. 1996. "The Selection of Prior Distributions by Formal Rules", *J. of the American Statistical Association*, 91: 1343–1370.

[8] R. Rosenkrantz. 1977. *Inference, Method, and Decision: Toward a Bayesian Philosophy of Science*, Synthese Library, Vol. 115, Springer.

[9] J. Shao. 1996. "Bootstrap Model Selection", *Journal of the American Statistical Association*, 91: 655–665.

[10] S. Stigler. 1990. *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard University Press.

[11] L. Wasserman. 2000. "Bayesian Model Selection and Model Averaging", *Journal of Mathematical Psychology* 44: 92–107.