

Generalization (or Overfitting)

Nico Formánek

High Performance Computing Center (HLRS)

May 27th, 2025

Hume's problem

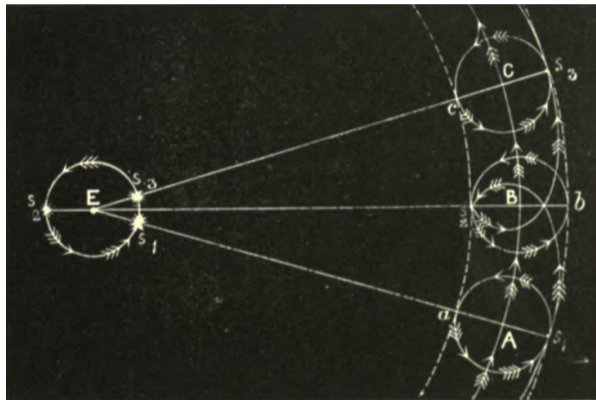


David Hume (1711 - 1776)

Thus, not only our reason fails us in the discovery of the ultimate connexion of causes and effects, but even after experience has inform'd us of their constant conjunction, 'tis impossible for us to satisfy ourselves by our reason, why we shou'd extend that experience beyond those particular instances, which have fallen under our observation.

A Treatise of Human Nature (1739)

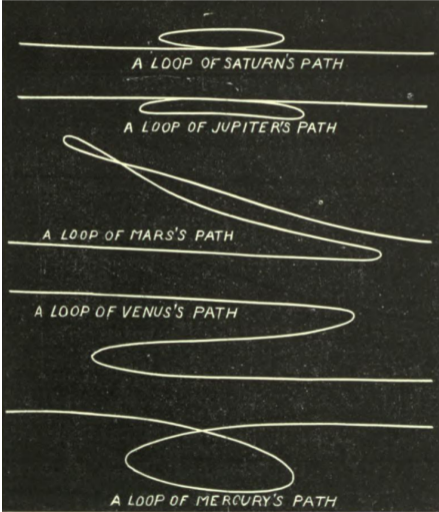
Hume's problem in modern terms



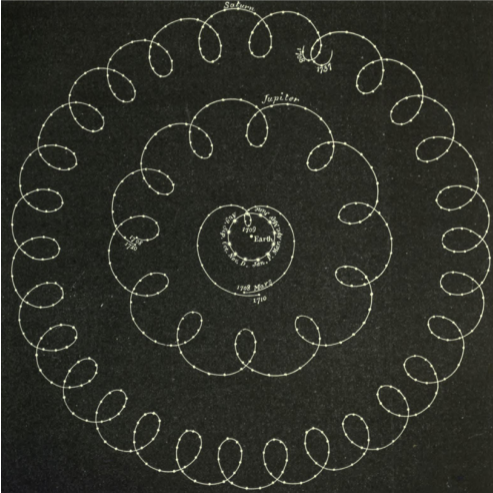
Some orbital modelling.

- Model should generalize from *seen* to *unseen* data.
- How do we know the model really generalizes?
- Either: **Assume some things** and mathematically prove a guarantee (SLT, inductive logic, stats).
- Or: **Test** the model with unseen data.

Seen Data



This is what you actually see in a telescope.

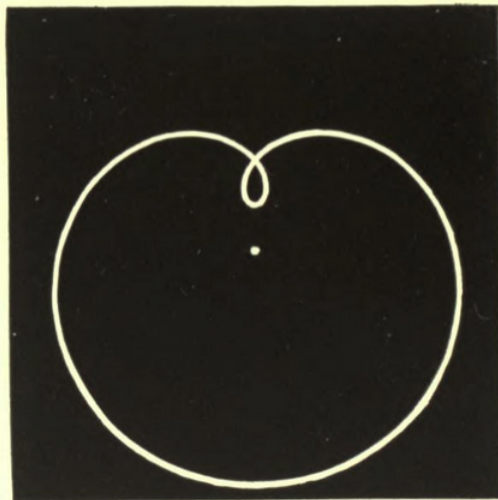


Creating the model from seen data

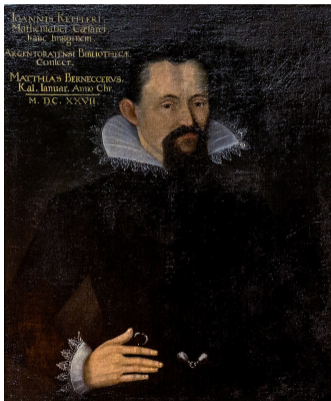
FIG. 2.—VENUS.



FIG. 3.—MARS.



Textbook Kepler



Johannes Kepler (1571 - 1630)

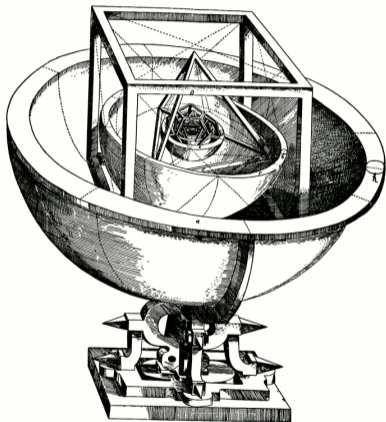
- Kepler “obtains” Brahe’s data.
- Notices discrepancy with epicyclical models.
- Produces a better fitting model (elliptical orbits).

The worry about overfitting

I sometimes have a nightmare about Kepler. Suppose a few of us were transported back in time to the year 1600, and were invited by the Emperor Rudolph II to set up an Imperial Department of Statistics in the court at Prague. Despairing of those circular orbits, Kepler enrolls in our department. We teach him the general linear model, least squares, dummy variables, everything. He goes back to work, fits the best circular orbit for Mars by least squares, puts in a dummy variable for the exceptional observation-and publishes. And that's the end, right there in Prague at the beginning of the 17th century.

David Freedman (1985)

Did Kepler create his model from data alone?

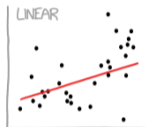


A data independent assumption of Kepler.

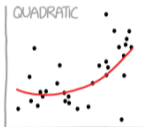
- Kepler did not create his ellipses purely from data.
- He had metaphysical and religious convictions which influenced his model selection.
- He had physical convictions which influenced his model selection.
- In short: He supported his model with data independent assumptions.

Digression: What is overfitting?

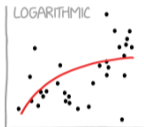
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



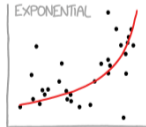
"HEY, I DID A REGRESSION."



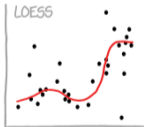
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH!"



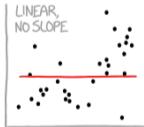
"LOOK, IT'S TAPERING OFF!"



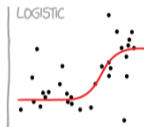
"LOOK, IT'S GROWING UNCONTROLLABLY!"



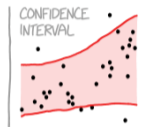
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."



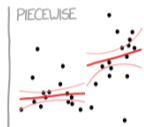
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."



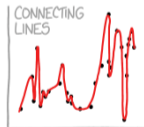
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."



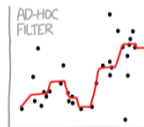
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."



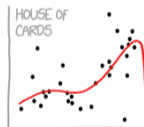
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND!"



"I CLICKED 'SMOOTH LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"



"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!"

Overfitting and simplicity (1935)

NEW BIOLOGICAL BOOKS

371

popular explanation of the Ogino-Knaus theory of the female "safe-period." He justly cautions that too much reliance cannot be placed on the "safe-period" and remarks in characteristic manner: "Trust in the Ogino-Knaus theory and have a pre-ventive jelly handy."



BIOMETRY

STATISTICAL CONFLUENCE ANALYSIS BY MEANS OF COMPLETE REGRESSION SYSTEMS.

By DORIS FRISCH, Universität Göttingen

manner of construction of the variates. In another example it is applied to measuring money flexibility from a six-variate analysis of annual consumption statistics from 1919 to 1931. Perhaps we are old fashioned but to us a six-variate analysis based on thirteen observations seems rather like overfitting.



COMPARABILITY OF MATERNAL MORTALITY RATES IN THE UNITED STATES AND CERTAIN FOREIGN COUNTRIES. *A Study of the Effects of Variations in Assigment Procedures*

Ideas about overfitting from the literature (stats, ML)

Overfitting is ...

- ... model not generalizing.
- ... model fitting noise.
- ... model interpolating (training error = 0 or below optimal).
- ... when complex model generalizes worse than simpler model (aka bias-variance trade-off).
- ... model misspecified (weak prior, wrong prior).
- ... when method of inference yields model that doesn't generalize.

Ideas about overfitting from the literature (stats, ML)

Overfitting is ...

- ... **model not generalizing**.
- ... model fitting noise.
- ... **model interpolating** (training error = 0 or below optimal).
- ... when **complex model generalizes worse than simpler model** (aka bias-variance trade-off).
- ... model misspecified (weak prior, wrong prior).
- ... when method of inference yields model that doesn't generalize.

Condensation attempt

A model is *overfit to the training data* iff...

1. ...it does not generalize.
2. ...it is too complex and does not generalize.
3. ...the training error is less than expected.

Condensation attempt

A model is *overfit to the training data* iff...

1. ...it does not generalize.
2. ...it is too complex and does not generalize.
3. ...the training error is less than expected.

A *method is overfitting* iff it yields models that are 1. or 2. or 3.

Benign overfitting

- Some authors call interpolating models *benignly overfitted*.
- Contradicts the older concepts of overfitting (1. & 2.) which were decidedly **negative**.
- Be aware!

Generalization apriori and aposteriori

- We would like to know the generalization before creating the model (apriori).
- Generalization is a relation between data, model and model selection procedure (as is overfitting!)
- Establishing generalization after having seen some data seems therefore easier.
- Hume's problem tells us that even aposteriori measures (e.g. CV) come with assumptions.

Coming to terms with Hume's problem

- Science creates new (unseen) data to test models.
- Science also creates new preferences among inductive assumptions.
- As long as inductive assumptions are an explicit part of the model they can be discussed.
- Automated methods for model selection often hide inductive assumptions in the models they select.
- We cannot discuss our preferences for inductive assumptions once they are hidden.

The End

Further reading I



Alicia Curth.

Classical Statistical (In-Sample) Intuitions Don't Generalize Well: A Note on Bias-Variance Tradeoffs, Overfitting and Moving from Fixed to Random Designs, September 2024.



Colin Howson.

Fitting Your Theory to the Facts: Probably Not Such a Bad Thing After All. *Minnesota Studies in the Philosophy of Science*, 14:224–44, 1990.



Colin Howson.

Hume's Problem: Induction and the Justification of Belief.
Oxford University PressOxford, 1 edition, November 2000.

Further reading II



Galit Shmueli.

To Explain or to Predict?

Statistical Science, 25(3):289–310, August 2010.



Tom F. Sterkenburg.

Statistical Learning Theory and Occam's Razor: The Core Argument.

Minds and Machines, 35(1):3, November 2024.



Ulrike von Luxburg and Bernhard Schoelkopf.

Statistical Learning Theory: Models, Concepts, and Results, October 2008.