UNIVERSITY OF HOHENHEIM

Models and model selection in science

3502-440 Methods of Scientific Working for Crop Science

WS 2024/2025

Prof. Dr. Karl Schmid
Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

Platt's view of science

Platt (1964): The inductive-deductive method is important in science.

Strong inference about a phenomenon:

1. Devise alternative hypotheses
2. Devise crucial experiments with alternative possible outcomes. Each outcome excludes at least one hypothesis
3. Carry out experiment to get clean results
4. Iterate procedure to refine hypotheses that remain

Consequences of strong inference

Strong inference leads at every turn to an experimental testing of parts of the complete theory (Step 3) and further logical extension (Step 1) of the theory

Platt argues that

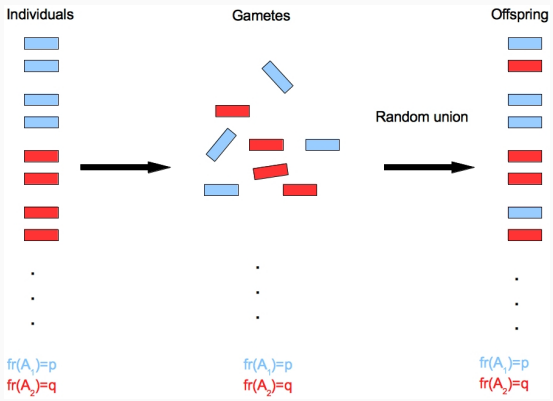
- strong inference is responsible for most of the rapid development in any field of science
- work spent on theory-building and experimental design that is not linked to the key principless is not well-invested
- strong inference uses the minimal number of steps to achieve scientific progress
- Thinking about several hypotheses stimulates search of evidence for /each/ hypothesis

Models and hypotheses

- Hypothesis: A statement. A good hypothesis allows falsification in the Popperian sense.
- Model: Formal description of knowledge about a phenomenon. May be simplified in contrast to reality, but allows predictions (e.g., hypotheses) which can be falsified. May be expressed in mathematical language.

Example for models: Hardy-Weinberg population model, Fisher-Wright model, SIR model

Population model: Random union of gametes



From model to hypothesis

Allele type:	A_1	A_2
Frequency:	p	q

$\Downarrow$

Genotype:	A_1A_1	A_1A_2	A_2A_2
HW frequency:	p^2	$2pq$	q^2

Main assumptions: Infinite population size, random mating

The model describes allele and genotype frequencies at a biallelic locus in a very large diploid population with no selection and mutation.

$\Downarrow$

Hypothesis 1: After one generation of random mating, genotypes are in Hardy-Weinberg frequencies

A second model leads to a second hypothesis

In a second model with the same assumptions, male and female individuals are differentiated with random mating only between male and female individuals

$\Downarrow$

Hypothesis 2: After one generation of random mating, the genotype frequencies are

Genotype:	A_1A_1	A_1A_2	A_2A_2
HW frequency:	p_mp_f	$p_mq_f + q_mp_f$	q_mq_f

,

where p_f, q_f are allele frequencies of A_1, A_2 in the female population and p_m, q_m the A_1, A_2 frequencies in the male population.

Falsifying a hypothesis by an experiment

Design experiment to differentiate hypotheses 1 and 2.

Take very big, equal sized sets of male/female individuals from two inbred lines of a dioecious plant (e.g., pistacia) which are fixed for different alleles A_1/A_2 ($p_m = 1$, $p_f = 0$, overall frequency $p = \frac{1}{2}$ for A_1). Mate the two lines randomly. The resulting F_1 population will consist on of heterozygous individuals at the observed locus.

- Hypothesis 1 predicts a frequency of $2pq = \frac{1}{2}$ of heterozygotes in F_1,
- Hypothesis 2 predicts a frequency of 1.
- The experiment falsifies hypothesis 1,
- Hypothesis 2 is consistent with the data.

Checking hypotheses by looking at samples

- Hypotheses make general prediction (e.g., all plants have DNA)
- Can be falsified by experiments
- For some (e.g., categorical) hypotheses, a single counterexample is sufficient
- For other hypotheses, a sample allows to gather evidence
  $\Rightarrow$ Estimate probability that a hypothesis is true with statistical framework

Application of Bayesian inference

Instead of looking at whole F_1 population, collect a sample of 10 individuals. All individuals are heterozygous

$\Rightarrow$ Consistent with hypotheses 1 and 2 Compare posterior probabilities of hypotheses 1 and 2 given the sample with Bayes' theorem in ratio form:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}.$$

Consider hypotheses 1 and 2 equally likely before the experiment, therefore prior probabilities are $P(H_1) = P(H_2) = 0.5$.

Application of Bayesian inference

Therefore,

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)0.5}{P(D|H_2)0.5} = \frac{P(D|H_1)}{P(D|H_2)}.$$

Likelihood of $P(D|H_1)$ is binomial probability of picking 10 heterozygous individuals $P(D|H_1) = (\frac{1}{2})^{10}$, the likelihood $P(D|H_2)$ is 1.
Ratio of posterior probabilities:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{1}{2^{10}} = \frac{1}{1024}$$

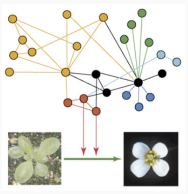
Conclusion: Given the data of the experiment, Hypothesis 2 is 1024 times more probable than Hypothesis 1.
A frequentist analysis is also possible!

Theory Selection

- Theory selection is a more general term for model selection
- What should a model achieve?
- Which type of model should we use to study a scientific problem?
- How complex should the model be?

Which model should be used?

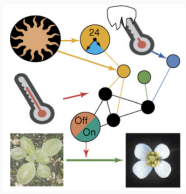
Example: Which model best describes the transition to flowering in plants?



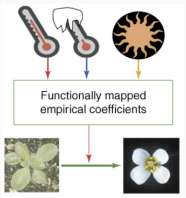
Complete network of regulatory genes

⇓

Full model



Simplified network

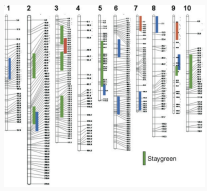


Empirical model

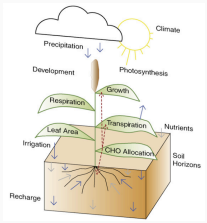
⇓

Black box model

How many parameters should be included in a model?



Genes



Physiology and Environment

Yield

Are some parameters more important than others?

William of Ockham: The parsimony principle

"Among theories that fit the data, choose the simpler one"

William of Ockham (c. 1288- c.1348)

Aspects of theory choice

Data and model are important!

Additional criteria:

- **Parsimony:** How complex does not model have to be?
- **Predictive accuracy:** How good are the predictions?
- **Explanatory power:** How well can the model explain results?
- **Testability:** Can the model be tested?
- Is it possible to generate new insights and knowledge?
- Is it coherent with other scientific and philosophical beliefs?
- Can the results be repeated?

Aspects of parsimony

- No insight is possible without simplification
- Insight by focussing on important points? (What are they?)
- Improve accuracy (How?)
- Increase efficiency: Save time and money by focussing on what is important

Prefer a simpler model versus Prefer a simple model!

Example of parsimony: Copernican revolution

- Geocentric vs. heliocentric model of planet movement
- Copernicus preferred heliocentric model solely due to parsimony
- Much later: Other reasons were discovered (stellar parallax, Bessel, 1838)

History of Parsimony: Medieval time

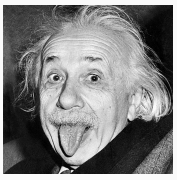
- Grosseteste (Oxford) and Thomas Aquinas (Paris) made statements about parsimony.
- William of Ockham: “Experience can serve to justify plurality”

Differences between Grosseteste and Ockham:

- Ontological definition: Nature is simple
- Epistemiological definition: Keep your theories simple

History of Parsimony: Modern times

Newton, Leibniz and Einstein proposed both ontological and epistemiological interpretations of parsimony.



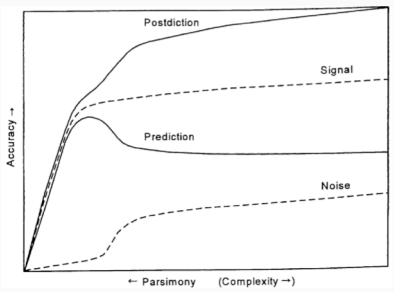
Einstein’s famous quote:

Everything should be made as simple as possible, but not simpler.

Why simplicity and parsimony?

- Scientists showed that simple theories tend to make reliable predictions
- Simpler models use data better
- Complicated models may model noise

Principles of Parsimony: Key Terms



- Signal and Noise
- Sample and Population
- Prediction and postdiction

Signal and noise

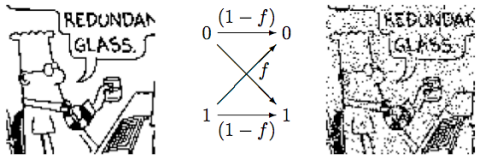
- Terms come from radio transmission theory:

  The fundamental problem of communication is that of reproducing at one point either eactly or approximately a message selected at another point. (Claude Shannon, 1948)

- Communication channels:

modem	$\rightarrow$	phone line	$\rightarrow$	modem
parent cell	$\rightarrow$	daughter cells		
computer memory	$\rightarrow$	storage device	$\rightarrow$	computer memory
nature	$\rightarrow$	scientist		

Signal and noise



A binary data sequence of length 10,000 transmitted over a binary symmetric channel with noise level $f = 0.1$.

Application in plant breeding

- Genetics $\rightarrow$ signal
- Environment $\rightarrow$ signal or noise?
- Experimental error $\rightarrow$ noise

Sample and Population

- **Sample:** a subset of the population
- **Induction:** Use sample to make statements about the population
- Data should be random and unbiased
- How large should the sample be?
- Maximize signal, minimize noise and experimental effort

Prediction versus Postdiction - A paradox?

- Scientific statements are usually made to generalize and to make predictions about whole populations.
- But: Analysis of models (i.e., parameter estimation) is based on the postdiction of existing data
  $\rightarrow$ Models are fitted to their data
- Pre- and postdiction are statistical terms, related to population and sample, and use different models
- Relationship of signal and noise:

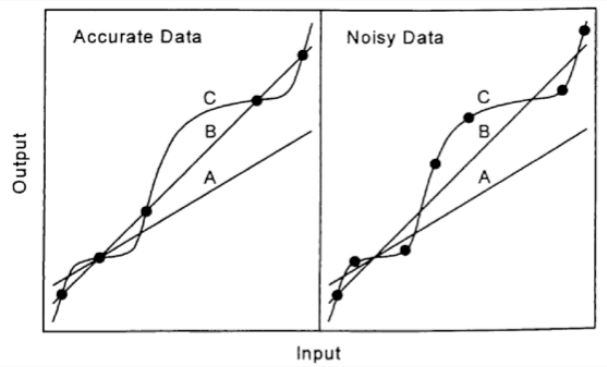
$$\begin{aligned} Data_1 &= Signal + Noise_1 \\ Data_2 &= Signal + Noise_2 \end{aligned}$$

- By definition, noise has no predictive quality
- In quantitative genetics, noise (or error) is often modeled as being normally distributed with a mean of 0: $e = N(0, \sigma)$

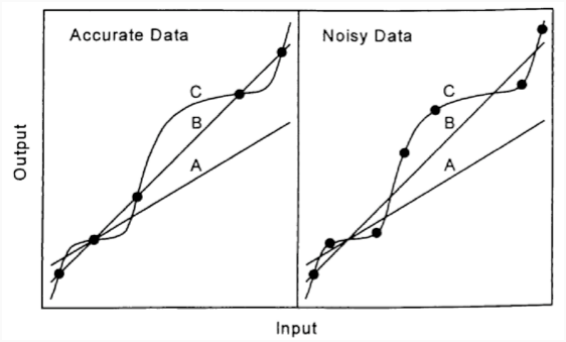
Prediction versus Postdiction - A paradox?

- If postdiction is the goal: Use a full, complex model that contains the data
- But: A full model is too specific (i.e., it models noise)
- A good and accurate model is more parsimonious than a full model.
  $\rightarrow$ Different model choice

Curve fitting problem



Curve fitting problem - Which model to choose?



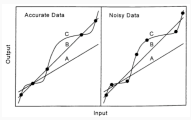
- Choose the simpler model
- Residuals from a simpler model contain information

Advantages and disadvantages of simpler models

1. Simple models are more vulnerable
2. Simple, and more comprehensive description
3. Complex models: more choice, can be advantageous or disadvantageous

- **Vulnerability:** Fewer parameters, less flexibility, Risk of falsification is good according to Popper.
- **Comprehensive description:** But it is not necessarily the case - parsimonious models can be too descriptive.

Summary: Curve-fitting

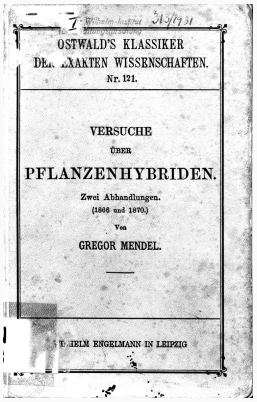


- Goodness of fit to data → Good postdiction
- Parsimony → Improved prediction
- Model selection becomes more difficult with noisy data
- Is additional knowledge necessary for model selection?
- One needs to estimate the accuracy of the data (by replication or simulation)

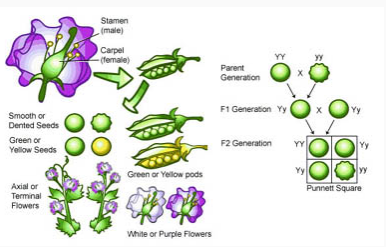
Another issue:

- Number of model parameters versus size of data set (p vs n problem)
- Trade-off between parameter numbers and statistical power

Parsimony at work: Mendel's Peas



Parsimony at work: Mendels Peas

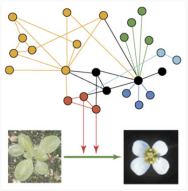


Parsimony at work: Mendels Peas

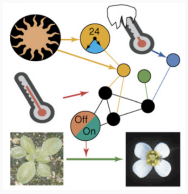
Trait	Ratio
Round : wrinkled shape	2.96 : 1
Yellow : green endosperm	3.01 : 1
Dark : white seed coats	3.15 : 1
Inflated : constricted seed pods	2.95 : 1
Green : yellow unripe pods	2.82 : 1
Axial : terminal flower positions	3.14 : 1
Tall : short plants	2.84 : 1
Average	2.98 : 1
Dominance : Recessivity	3 : 1

Mendel's conclusion: " *...an average ratio of 2.98:1 or 3:1*"

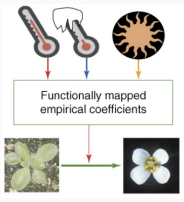
Modeling crop growth



Network control



Simplified network



Empirical model

Grand challenge: Which crops do we need for future plant production in a rapidly changing world?

→ Prediction of future climate change and its effect on crops

Summary

- An effective method of scientific research is Platt's strong inference which is an deductive-inductive methods using experiments to test deducted hypotheses
- Statistical methods can be used to compare competing hypotheses using sample data
- Generally, simpler models are to be preferred, but trade-off in ability for postdiction and prediction needs to be clarified
- Parsimony is an important principle in science
- Parsimonious models use data better
- Complete/Non-parsimonious Models: Danger of overfitting
- But be careful: Theories/models sometimes have to be complex

Further Reading

- ?, The Scientific Method at Work, Chapter 8

References i