Statistik III

Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

Model Comparison

Introduction

- In the following we consider a typical situation in psychology, where data are available that are assumed to be generated by an (unknown) random process
- One is interested in finding a theory which is able to predict the data generated by this random process with sufficient precision
- In selecting appropriate theories (by *comparing models*) it is necessary to have available decision criteria to identify the "best" of competing theories
- The starting point is formed by the estimation of the free parameters contained in a theory, in order to arrive at the best possible prediction of the data at hand

Model Comparison

Introduction

- Following parameter estimation, the question arises how well a formal theory can predict the data, or which of several theories describes the data more accurately
- Model comparison has to consider at least two aspects
  - goodness of fit
  - the complexity or flexibility of the underlying theory
- An integrative account of both aspects aims at characterizing the *generalizability* of the theory
  - How well can the theory predict a new data set generated by the same random process as the present data?

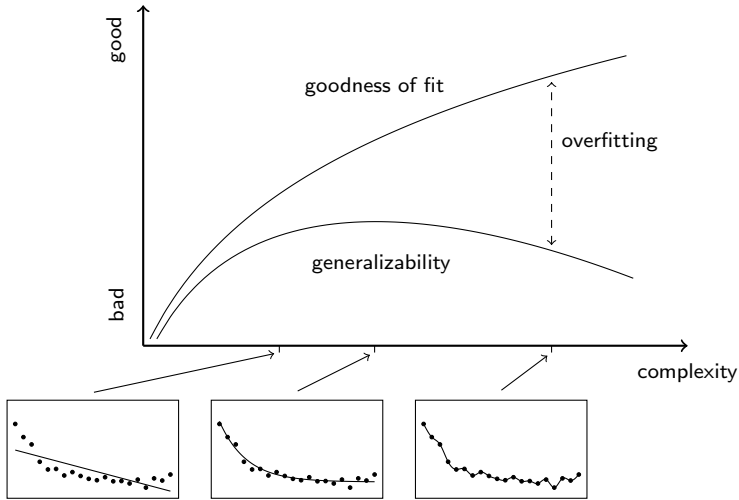
Generalizability

- As a practical guideline for deciding between alternative theories, the principle of parsimony ("Ockham's razor") suggests that of two theories having the same empirical content, the simpler one is to be preferred
- Considering the explanation of experimentally collected data by a (formal) psychological theory, its empirical content may be conceived as the collection of all the data sets that are adequately described by the theory
- The agreement between the predictions of the theory and the data is commonly referred to as *goodness of fit*
- Simplicity of the theory is inversely related to what is considered its *complexity* (or *flexibility*)

Generalizability

- The complexity of a theory depends on
  - the number of its free parameters (e. g., real constants in a formula the value of which can be freely chosen)
  - the form of the functional dependence of the predictions on the parameters of the theory
- In each case, the following figure shows the same data points predicted by
  - a linear equation
  $$y = a + b \cdot x$$
  - the exponential function
  $$y = a \cdot e^{b \cdot x}$$
  - cubic splines, i. e., piecewise functions of the type
  $$y = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$

Generalizability



Generalizability

- Goodness of fit to a given data set increases with increasing complexity
- The goal in science is
  - *not* a perfect prediction of given data
  - to predict new data (such as from a repetition of the underlying experiment) as accurately as possible, which is called *generalizability*
- Generalizability increases with goodness of fit before it starts to decrease beyond an optimal value of complexity
- The discrepancy between goodness of fit and generalizability is also called *overfitting*

Goodness of Fit

- Due to its highly desirable properties, parameter estimation in stochastic theories often follows the maximum likelihood principle
- Here, the maximum achievable likelihood is considered as a measure of the goodness of fit
- In this framework, statistics provides methods to compare theories that are related in a particular way
- The likelihood ratio test is the standard procedure for testing hierarchically nested models against each other

Likelihood Ratio Test

- We consider two theories $\mathcal{M}_g$ and $\mathcal{M}_c$, where the latter forms a constrained version of $\mathcal{M}_g$ (resulting, e. g., from fixing parameters to particular values)
- The likelihood ratio test computes the quotient of the maximum likelihood $L(\mathcal{M}_g)$ to the maximum likelihood $L(\mathcal{M}_c)$
- The statistic

$$2 \log \frac{L(\mathcal{M}_g)}{L(\mathcal{M}_c)}$$

is asymptotically χ^2-distributed with the number of degrees of freedom m equal to the difference between the number of free parameters in $\mathcal{M}_g$ and $\mathcal{M}_c$

Likelihood Ratio Test

- The decision between theories $\mathcal{M}_g$ and $\mathcal{M}_c$ via a likelihood ratio test represents a very special situation
- The theories are to be hierarchically nested and therefore have the same functional form on a part of the parameter space and everywhere else at least a highly similar form
- The difference between the complexities of the two theories $\mathcal{M}_g$ and $\mathcal{M}_c$ is captured by the difference between their numbers of free parameters
- The procedure *cannot* be used for theories that are *not* hierarchically nested
- This is a limitation, because, for example, competing cognitive theories may not be related in this way

Information Criteria

- Within a general framework of model selection, indices were developed that can account for both goodness of fit and complexity, and thus (in principle) allow for comparing theories that are not hierarchically nested
- In addition to referring to maximum likelihood as a measure of fit, these indices also capture the complexity of the respective theory
- In general, these kind of indices account for
  - the number of parameters of the theory
  - the number of observations
  - the functional form of the theory
- The objective is always to minimize the respective index
- The indices are calculated by adding to the likelihood value (representing model fit) penalty terms that increase monotonically with complexity

Akaike Information Criterion (AIC)

- This index, proposed by Akaike (1973), still constitutes the most widely used information criterion
- In the context of AIC, it is assumed that there is a "true" theory that generated the data and that is included in the class of theories under consideration (such as LMMs, GLMMs)
- When comparing two competing theories, we consider the loss of information (in terms of Kullback-Leibler divergence) which is associated with each theory's description of the data (as compared to the unknown "true" theory)
- Under the above assumptions, AIC forms an unbiased estimator of this information loss

Akaike Information Criterion (AIC)

- Formally, the index is calculated by

$$AIC = \underbrace{-2 \log L}_{\text{goodness of fit}} + \underbrace{2\,k}_{\text{complexity}}$$

  with L the maximum likelihood and k the number of parameters
- Thus, in the AIC only the number of parameters of a theory is considered as a measure of its complexity
- Rule of thumb: AIC differences > 10 provide strong evidence for the theory with the smaller value, differences ≤ 2 should rather not be interpreted (Burnham & Anderson, 2002, p. 70)

Bayesian Information Criterion (BIC)

- Schwarz (1978) proposed the BIC as an alternative to AIC
- Therefore, the term *Schwarz information criterion* (SBC, SBIC) is also used
- For large samples, BIC favors theories with the highest posterior probability
- Comparing two theories T_1, T_2 with equal prior probability and data D based on the BIC approximates model comparison according to the Bayes factor because

$$\frac{P(T_1 \mid D)}{P(T_2 \mid D)} = \frac{P(T_1)}{P(T_2)} \cdot \underbrace{\frac{P(D \mid T_1)}{P(D \mid T_2)}}_{\text{Bayes factor}}$$

Bayesian Information Criterion (BIC)

- The BIC is calculated by

$$BIC = \underbrace{-2 \log L}_{\text{goodness of fit}} + \underbrace{k \log n}_{\text{complexity}}$$

  with n the number of observations (sample size)
- Compared to AIC, weighting the number of parameters as a function of the logarithm of the number of observations will favor less complex theories (for $n \geq 8$)
- In principle, the interpretation of BIC differences is analogous to AIC (Kass & Raftery, 1995, p. 777)
- Neither AIC nor BIC take into account the functional form of the theory

Minimum Description Length (MDL)

- The minimum description length principle captures the complexity of the theory via its flexibility, i. e., by the variability of the data sets that the theory can predict
- MDL asks for ways of compressing the data:

  What is the shortest program that outputs the data set?

- The length of the shortest program that can generate the data is identified with its complexity

Minimum Description Length (MDL)

Example

- Consider the following sequences which both are 40 bits long

  0001000100010001000100010001000100010001

  1011000001010000100000011011010111011001

  and the latter resulting from an independently repeated coin toss

- The first sequence consists of the 10-fold repetition of the sequence 0001 and can be generated in R, for example, by the following program

```
for(i in 1:10) cat("0001")
```

- The second sequence may be considered random without structure
- With high probability, the shortest R program that outputs this sequence is therefore the sequence itself

```
cat("1011000001010000100000011011010111011001")
```

Minimum Description Length (MDL)

- The *Kolmogorov complexity* characterizes the regularity of the data
  - The more structure, the smaller the Kolmogorov complexity
- This formulation of MDL provides a very general theory which, however, is not practically applicable
  - There is no algorithm providing the shortest-length program to generate given data ("uncomputability")
  - The invariance theorem (which states that results do not depend on the programming language used) only applies to infinitely large data sets
- The following indices are approximations to the principle of minimum description length

Normalized Maximum Likelihood (NML)

- The index is calculated by

$$NML = \underbrace{-\log L}_{\text{goodness of fit}} \underbrace{+\log \int_{x \in \Omega} L(x)dx}_{\text{complexity}}$$

- Integration is to be performed over all possible data sets $x \in \Omega$
- The respective term becomes large when the theory is flexible enough to predict many different data sets (i. e., many data sets have a substantial likelihood)
- Numerical estimation techniques such as Monte Carlo (MC) integration are often too time intensive for practical purposes

Fisher Information Approximation (FIA)

- FIA is asymptotically equivalent to NML and calculated by

$$FIA = \underbrace{-\log L}_{\text{goodness of fit}} + \underbrace{\frac{k}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det I(\theta)}d\theta}_{\text{complexity}}$$

  with $I(\theta)$ the Fisher information matrix, i. e., the covariance matrix of the derivative of the log-likelihood (score function), and $\det I(\theta)$ denoting its determinant
- The last term characterizes the functional form of the theory (in the unbiased case the inverse of the information matrix forms the covariance matrix of the ML estimators)
- The Fisher information matrix is often available in closed form, and the integral can be estimated via MC integration

Minimum Description Length (MDL)

Several factors influence the complexity of a theory (as approximated, e. g., by FIA)

- Number of parameters (complexity increases with the increasing number)
- Functional form
- Number of observations (complexity increases with sample size)
- Changes of the parameter space (complexity grows with expansion)
- Type of likelihood function or error distribution
- Experimental design (e. g., how independent variables are defined)

Example: Stevens's and Fechner's Laws

(Pitt, Myung & Zhang, 2002)

Model equations

- Fechner's law

$$Y = a \cdot \log(X + b) + \varepsilon$$

- Stevens's power law

$$Y = a \cdot X^b + \varepsilon$$

with error $\varepsilon \sim N(0, \sigma^2)$

Simulation

- 1000 datasets are simulated with $x = 1, 2, \ldots, 6$, $\sigma = 0.3$, and parameter values $a = 2$, $b = 2$ for Fechner and $a = 2$, $b = 5$ for Stevens

Example: Stevens's and Fechner's Laws

(Pitt, Myung & Zhang, 2002)

Results

	Data generated with	
Theory	Stevens	Fechner
AIC		
Stevens	12.42 (100%)	11.92 (47%)
Fechner	52.22 (0%)	11.86 (53%)
MDL (FIA)		
Stevens	10.71 (100%)	10.46 (0%)
Fechner	25.40 (0%)	5.21 (100%)

Mean value of the information criterion (percentage of selecting the model)

Example: Stevens's and Fechner's Laws

(Pitt, Myung & Zhang, 2002)

Results

- Fechner's law

$$FIA_F = -\log L + 2.48$$

- Stevens's power law

$$FIA_S = -\log L + 8.00$$

Discussion

- Since both theories have the same number of parameters, the selection according to AIC is based on the maximum likelihood only
- Due to its greater flexibility, Stevens's power law can achieve a greater likelihood even for data generated by Fechner's law
- The MDL criterion (FIA) can reliably distinguish between the two theories

Example: Latent Trait Modeling

(De Boeck et al., 2011)

- For the `VerbAgg` data set in the `lme4` package on a questionnaire on verbal aggression, we considered the following models

```
## Rasch model (participants as random effects)
m1 <- glmer(r2 ~ 0 + item + (1 | id), VerbAgg, binomial)

## LLTM
m2 <- glmer(r2 ~ 0 + btype + mode + situ + (1 | id),
            VerbAgg, binomial)

## LLTM with homoscedastic error
m3 <- glmer(r2 ~ 0 + btype + mode + situ + (1 | id) + (1 | item),
            VerbAgg, binomial)
```

Example: Latent Trait Modeling

(De Boeck et al., 2011)

- Number of parameters
  - Model `m1`: 25 parameters

	24	item parameters
+	1	variance across persons

  - Model `m2`: 6 parameters

	3	behaviors
+	1	mode (2 modes, 1 reference category)
+	1	situation (2 situations, 1 reference category)
+	1	variance across persons

  - Model `m3`: 7 parameters

	3	behaviors
+	1	mode (2 modes, 1 reference category)
+	1	situation (2 situations, 1 reference category)
+	2	variance across persons and items

Example: Latent Trait Modeling

(De Boeck et al., 2011)

- While `m2` is nested in `m3`, both models and `m1` are non-nested
- AIC favors `m1`, and BIC favors `m3`

```
Formula: r2 ~ 0 + item + (1 | id)
     AIC      BIC   logLik deviance df.resid
  8128.5   8301.9  -4039.3   8078.5     7559

Formula: r2 ~ 0 + btype + mode + situ + (1 | id)
     AIC      BIC   logLik deviance df.resid
  8249.9   8291.5  -4119.0   8237.9     7578

Formula: r2 ~ 0 + btype + mode + situ + (1 | id) + (1 | item)
     AIC      BIC   logLik deviance df.resid
  8163.8   8212.4  -4074.9   8149.8     7577
```

Cross-Validation

Basic idea

- The methods covered so far estimate generalizability via an equation that explicitly specifies the trade-off between goodness-of-fit and complexity
- Cross-validation estimates generalizability directly from the observed data by simulating data-generation and model-prediction without explicitly specifying complexity
- In general, cross-validation is a three-step procedure
  1. The data sample is split into non-overlapping subsamples, the calibration sample y_{cal} and the validation sample y_{val}
  2. The model is fit to the calibration sample y_{cal} to compute the model's ML estimate $\theta_{MLE}(y_{cal})$
  3. The goodness-of-fit of the model for the validation sample y_{val} with parameter values fixed to the ML estimate $\theta_{MLE}(y_{cal})$ is taken as the model's generalizability estimate

Cross-Validation

Leave-one-out (LOO) cross-validation

- The data sample is split into a calibration sample of $n-1$ observations and a validation sample with the remaining one observation ($y_{\neq i}$ vs. y_i)
- This $(n-1)$-vs-1 split process is repeated for all possible n splits
- The model's final generalizability results from aggregating goodness-of-fit over all splits

$$LOOCV = -\sum_i \log L(y_i \mid \theta_{MLE}(y_{\neq i}))$$

- Advantages
  - Ease of computation (unlike FIA and NML, no high dimensional integration involved)
  - Implicitly takes into account all aspects of model complexity (in contrast to AIC and BIC)

Bayesian Model Selection

- Given data D, Bayesian Model Selection (BMS) compares theories T_1, T_2 via the Bayes factor

$$\frac{P(D \mid T_1)}{P(D \mid T_2)}$$

  with $P(D \mid T_i)$ the marginal likelihood of theory T_i
- The marginal likelihood results from averaging likelihood across all parameter values, weighted by the prior distribution on the parameters
- BMS selects the theory with the largest *average* likelihood, not the one with the largest *maximal* likelihood
- Averaging the likelihood accounts for all the aspects of the complexity of the theory

Bayesian Model Selection

- Connection to information criteria
  - BIC asymptotically corresponds with model selection by the Bayes factor
  - An asymptotic approximation of BMS turns out to be just the same as FIA
- These formal equivalences are quite surprising due the fact that they are rooted in divergent theoretical and philosophical foundations
- Computation of BMS also involves numerical integration of high dimensions and is plagued by similar computational difficulties as FIA and NML

Conclusions

Pitt & Myung (2002)

> *Methods like AIC and MDL can give the impression that model selection can be automated and require minimal thought. . . . Selection methods should be viewed as a tool that researchers can use to gain additional information about the models under consideration. These tools, however, are ignorant of other properties of the models, such as their plausibility or the quality of the data, so it is inappropriate to decide between them solely on what is learned from a test of generalizability. (p. 425)*

References

Myung, J.I., Cavagnaro, D., & Pitt, M.A. (2016). Model evaluation and selection. In W.H. Batchelder, H. Colonius, E.N. Dzhafarov, & J.I. Myung (Eds.), New handbook of mathematical psychology: Foundations and methodology (pp. 552–598). Cambridge University Press.

Myung, J.I., Navarro, D.J., & Pitt, M.A. (2006). Model selection by normalized maximum likelihood. Journal of Mathematical Psychology, 50, 167-179.

Pitt, M.A., & Myung, I.J. (2002). When a good fit can be bad. Trends in Cognitive Sciences, 6(10), 421-425.

Pitt, M.A., Myung, I.J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. Psychological Review, 109, 472-491.

Su, Y., Myung, I.J., & Pitt, M.A. (2005). Minimum description length and cognitive modeling. In P. Grünwald, I.J. Myung, & M.A. Pitt (Eds.) Advances in Minimum Description Length: Theory and Applications (pp. 411-433). Cambridge, MA: MIT Press.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E. (2015). Model comparison and the principle of parsimony. Oxford Handbook of Computational and Mathematical Psychology. (pp. 300-317) `http://cidlab.com/prints/vandekerckhove2014model.pdf`