
Modellselektion

Dr. Dirk Hasenclever
IMISE, Leipzig

Hasenclever@IMISE.uni-Leipzig.de

Wie viele Parameter?

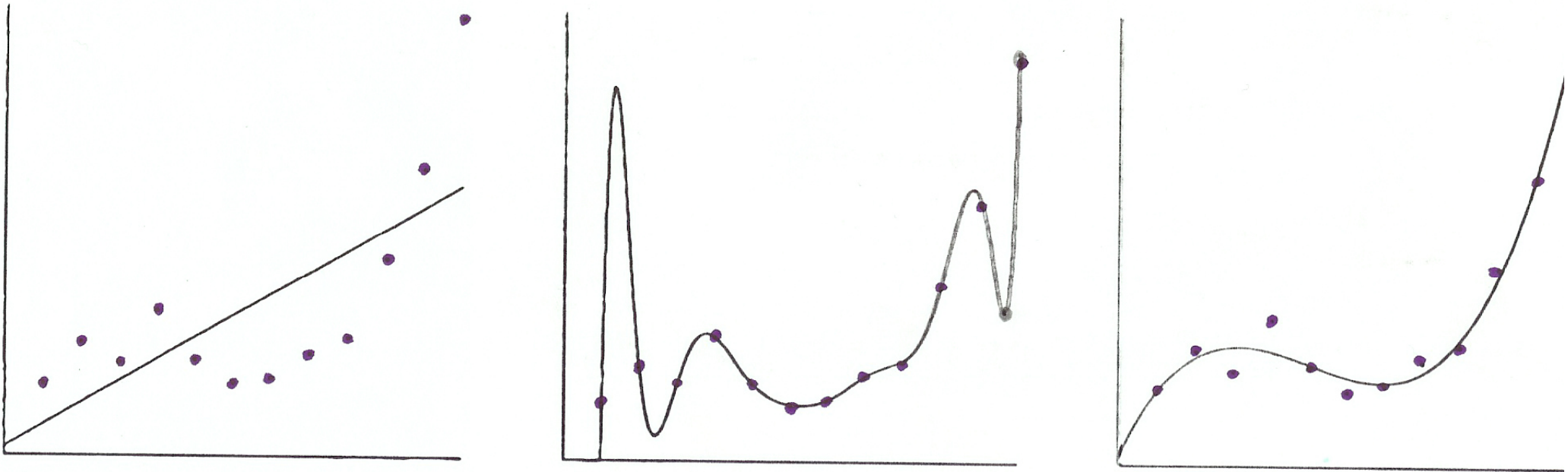


FIG. 1. A simple (a), complex (b) and a trade-off (3rd degree) (c) polynomial.

Schlechter Fit

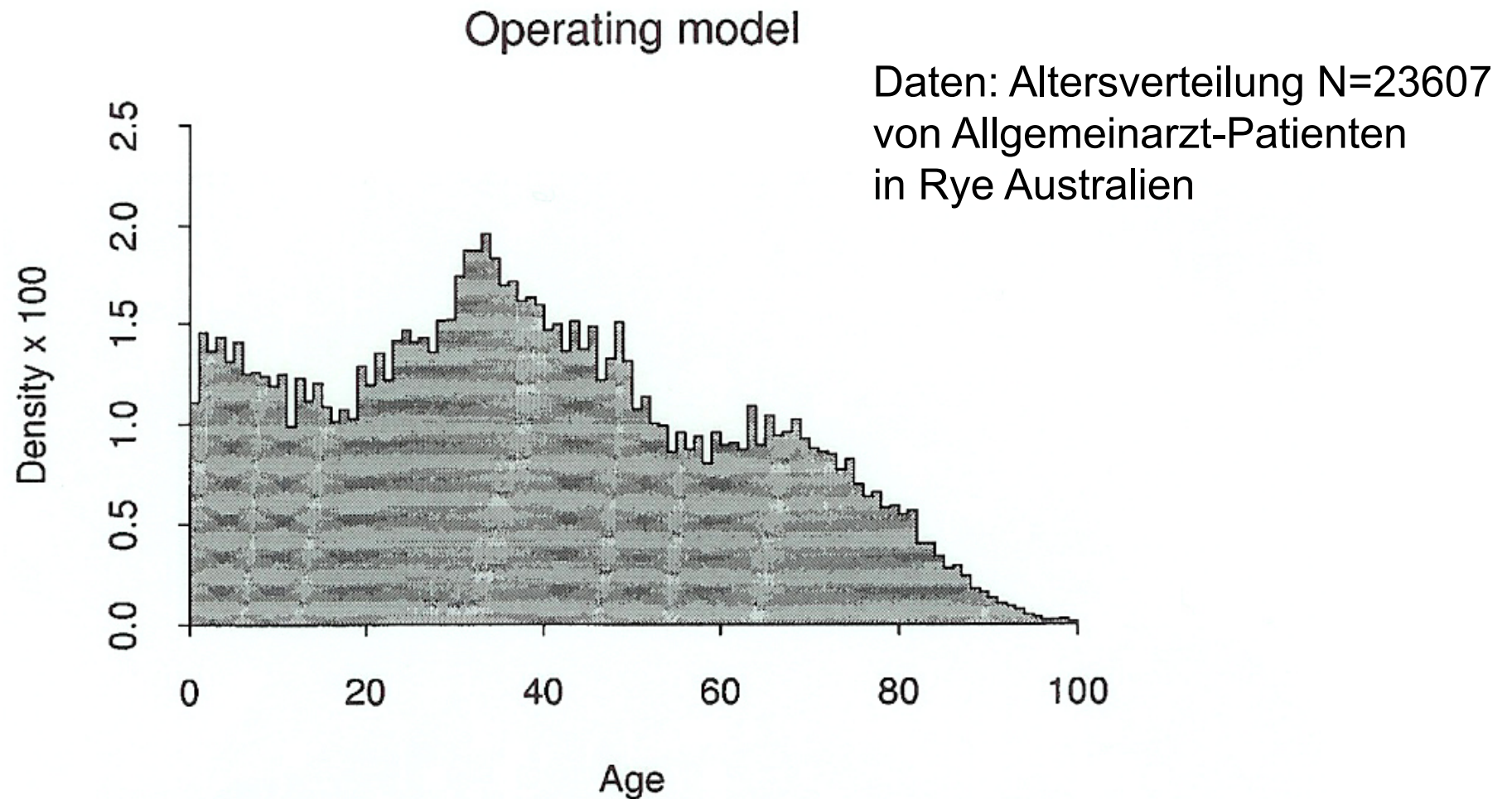
Too good to be true

angemessen

Das Modell-Selektions Problem

- Gegeben: Daten, die durch Modelle mit unterschiedlich viele Parameter beschrieben werden. Z.B. Regression: Viele Kovariate.
- Aufgabe: Wähle „**bestes**“ Modell
- Vorsicht: **Overfitting**
Es ist trivial, dass man mit mehr Parametern die gegebenen Daten auch besser beschreiben kann. (Vgl. Bsp: Zucchini et al.)

Problem: Overfitting



Zucchini 2000

Problem: Overfitting II

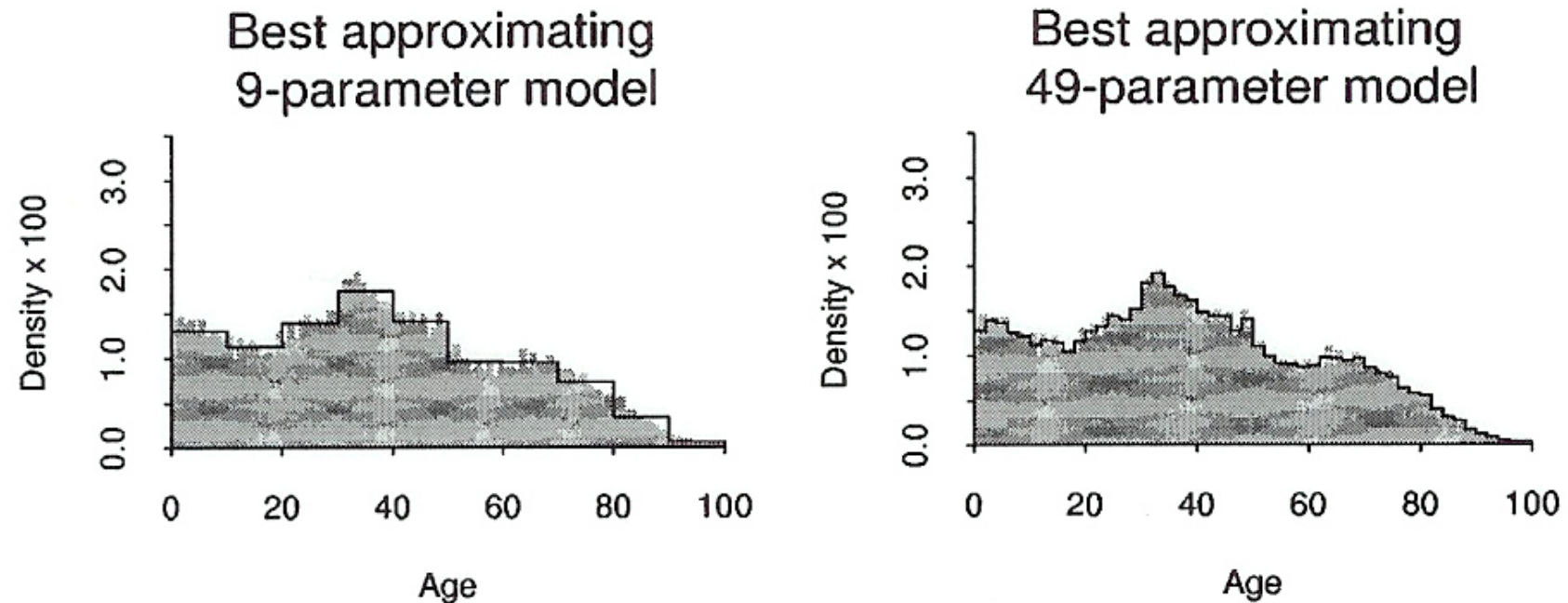


FIG. 3. The operating model for the age distribution of the Ryde population (shaded) and the best approximating models with 9 and with 49 parameters (dark lines).

Zucchini 2000

Problem: Overfitting III

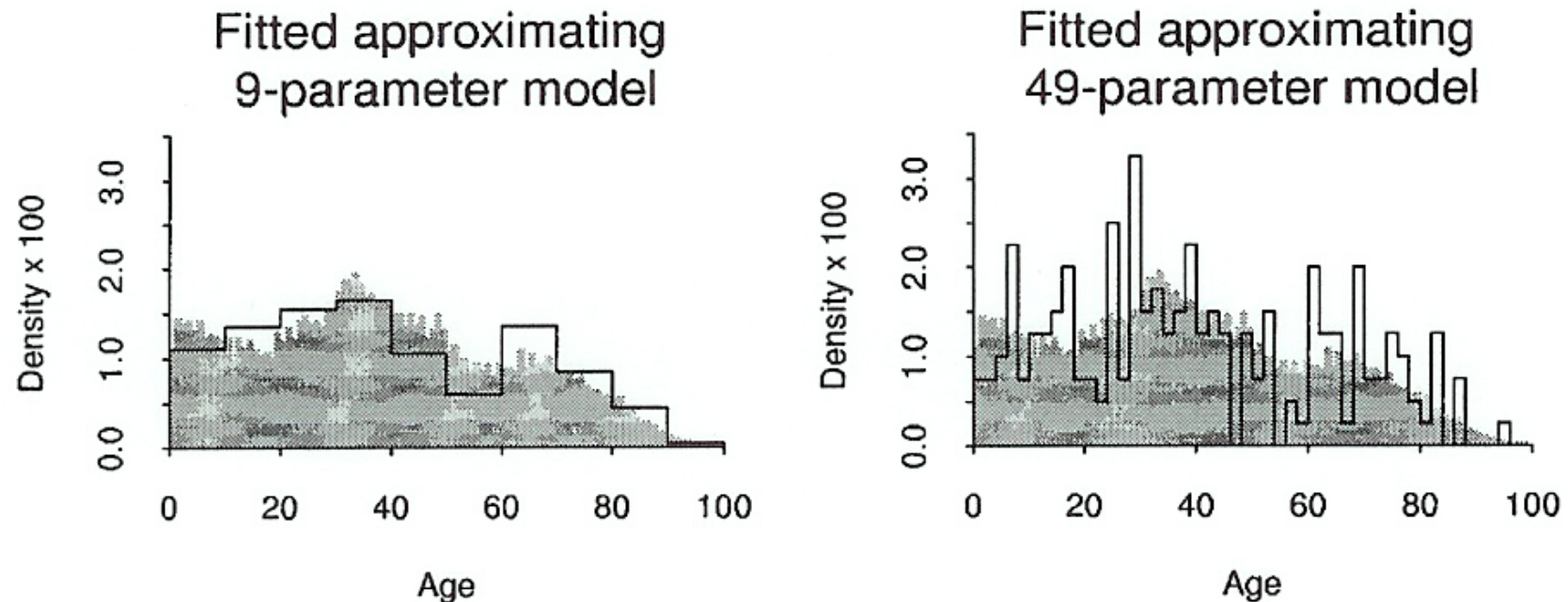


FIG. 4. The operating model (shaded) and the fitted models with 9 and with 49 parameters for the sample of 200 Ryde residents (dark lines).

Simuliere Stichproben à 200

Zucchini 2000

Problem: Overfitting IV

Modelle geschätzt
mit zu vielen
Parameter werden
- obwohl im Trainings-
Sample besser –
bei Prädiktion
unabhängiger Daten
schlechter als
einfachere Modelle.

Distribution of the overall discrepancy
= L_2 Norm Abstand zur wahren Dichte

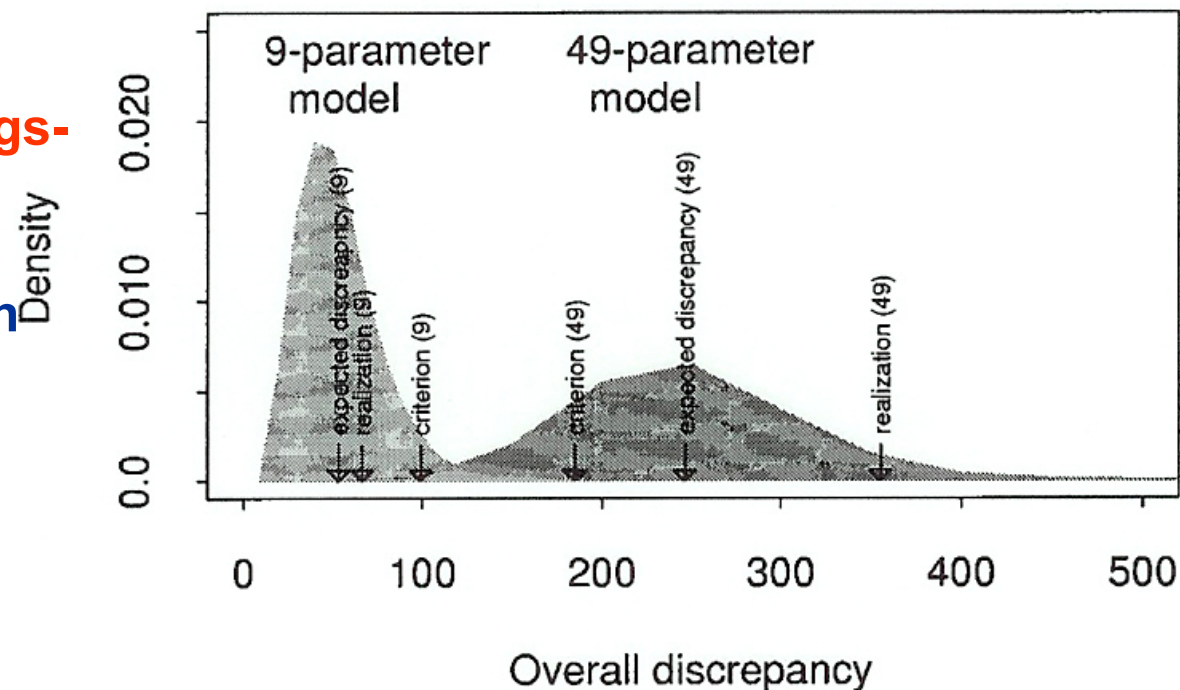


FIG. 5. The distributions of the overall discrepancies for $n = 200$ and the values of the overall discrepancy that were realized for the sample. The (entire) criterion, which is explained later in the paper, is also shown.

Zucchini 2000

Deviance und ihre Verteilung

- Definiere die **Deviance eines Modells** zu gegebenen Daten als

$$-2 * \log\text{likelihood}(\text{MLE der Parameter})$$

- Modell werden verglichen aufgrund der Differenz in ihrer Deviances (=umskalierte Likelihood Ratio)
- Die Deviance eines Modells ist eine **Zufallsvariable**

Deviance und ihre Verteilung II

- Für die Differenz von Deviances von **verschachtelten Modellen** mit unterschiedlich vielen Parametern gilt asymptotisch:
- Unter der Nullhypothese H_0 : zusätzliche Parameter sind überflüssig $=0$

ist die **Differenz der Deviances χ^2 verteilt**:

- $$-2 (\text{loglike}(\text{modell}_p) - \text{loglike}(\text{modell}_q))$$
$$\sim \chi^2(p-q)$$

df= Differenz der Parameterzahlen der Modelle

$$= p-q$$

Frequentistische Modellselection

- (Nur bei geschachtelten Modellen):
 - Serie von Anpassungstests z.B. über Deviance
 - Step-up
 - Step down
 - ...
- Konzeptuelle Probleme
 - Inkompatibilität (siehe Bsp.)
 - Fehler 1. Art nicht klar definiert, bei bedingten Serien von Tests
 - Keine explizite Kontrolle des Overfitting

Kompatibilitätsproblem

Angenommen man macht gleichzeitig zwei unabhängige Versuche, und betrachte jeweils Modelle mit 0 oder einem Parameter

a) Separat :

$$- 2 (l_1(\psi_1) - l_1(0)) = 3.0 < 3.84 = \chi^2_{1; 5\%}$$

$$- 2 (l_2(\psi_2) - l_2(0)) = 3.2 < 3.84 = \chi^2_{1; 5\%}$$

⇒ beide Parameter nicht "signifikant" von 0 verschieden.

b) Simultan für Gesamtexperiment :

$$-2(l_1(\psi_1) + l_2(\psi_2) - l_1(0) - l_2(0)) = 3.0 + 3.2 = 6.2 > 5.99 = \chi^2_{2; 5\%}$$

⇒ Modell mit zwei Parametern besser; d.h. mindestens ein Parameter von 0 verschieden.

?!

William of Ockham (1321)

- **Pluralitas non est ponenda - sine neccesitate!**
("Ockham's Rasiermesser")
- Wähle das einfachste Modell, welches mit den Daten, dem relevanten gesicherten Vorwissen und der Fragestellung verträglich ist.
- Problem:
 - Finde "angemessenen" quantitativen Trade-off zwischen
 - Güte der Modellanpassung an gegebene Daten und
 - Einfachheit (= # Parameter) des Modells

Modellselektion im Likelihood Paradigma I

- Problem: Likelihood mit mehr Parametern wird trivialerweise besser durch Overfitting.
- Idee: Ziehe den trivialerweise zu erwartenden Gewinn an LogLikelihood pro zusätzlichen Parameter von Log-Likelihood ab.

Modellselektion im Likelihood Paradigma II

- Ansatz **Crossvalidierung**:
Simuliere unabhängige Prädiktion, um
Overfitting herauszurechnen
 - Nehme ein Element heraus und fitte Modell an $n-1$ Datenpunkte
 - Berechne Likelihood des herausgenommenen Datenpunkts bzgl. dieses Modells
 - Berechne crossvalidierte Likelihood als Produkt dieser Beiträge.
 - Wähle Modell mit der maximalen crossvalidierten Likelihood

AIC

- Weniger rechenintensiv:
- **Akaike Informationskriterium (AIC)**
 - Crossvalidierte loglikelihood $\approx$
loglikelihood – q (q = # Parameter)
(im Erwartungswert asymptotisch.
- Vorgehen: Vergleiche
$$E(\chi^2_{(df=q)}) = q$$
$$AIC = -2 \loglikelihood(\text{Modell}(q)) + 2 \cdot q$$

Wähle Modell mit minimalem AIC

Dieses Modell lässt beste Prädiktion zukünftiger Beobachtungen erwarten.

Bayesianische Modellselektion

- **Falls** einige Modelle gegeben, von denen eines das wahre ist, wähle hierarchischen Prior

- Prior welches Modell
- Prior über Modellparameter gegeben Modell

- Entscheide Modellwahl aufgrund des Posterior-Bayes-Faktors

- Falls wenig Vorinformation:

(Bayes'sches Informationskriterium)

$$\mathbf{BIC} = -2 \log\text{likelihood}(\text{Modell}(q)) + q \cdot \ln(n) \quad (\text{Schwarz})$$

Differenz in BIC approximiert den Bayesfaktor zwischen Modellen.

BIC favorisiert Modelle mit weniger Parametern