Unit 2: Bayesian Learning

5. Comparing Models

10/22/2020

Comparing models

1. Both qualitative and quantitative methods can be used to distinguish between models

2. Models should be chosen on the basis of generalization, not fit

3. Some common methods for assessing both fit and generalization

What makes a model good?

A lot of you preferred the Kalman filter model to the Rescorla-Wagner model of classical conditioning.

Why?

Qualitative criteria

Assumptions check out

The assumptions of the model are plausible and consistent with other findings. They are not ad-hoc.

Explanatory adequacy

The model does more than just re-describe the data

E.g. “The power law of practice”

Interpretability

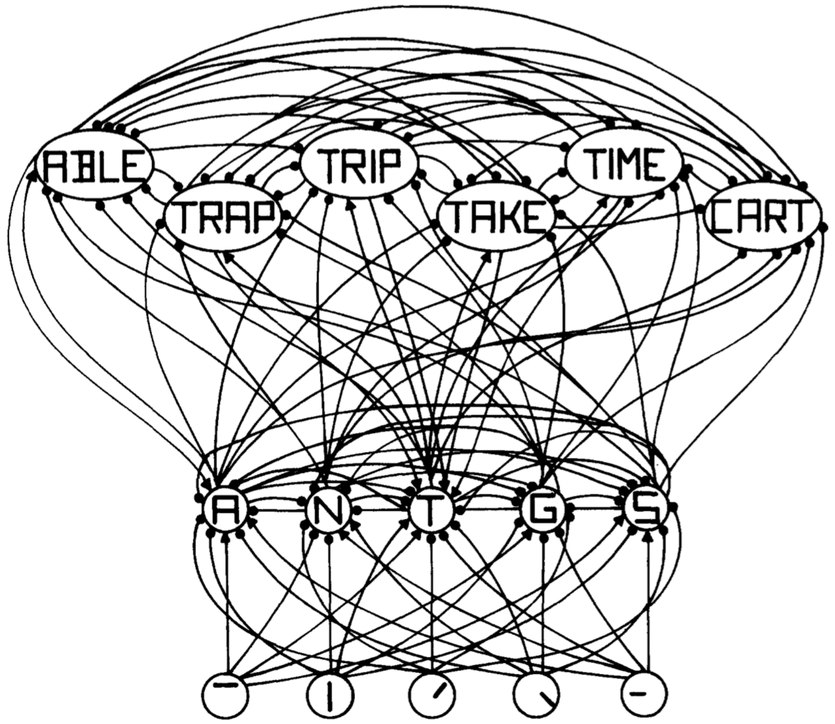
The model makes sense.

Components link to psychological or neural processes and constructs

More qualitative criteria

Stability

Results are due to core theoretical assumptions and not implementation details

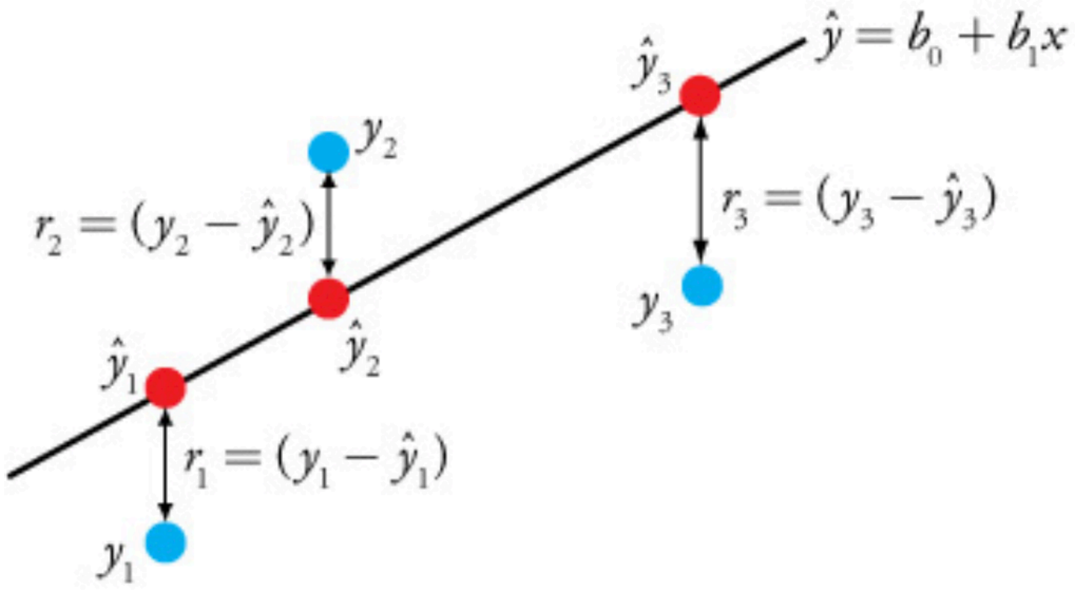


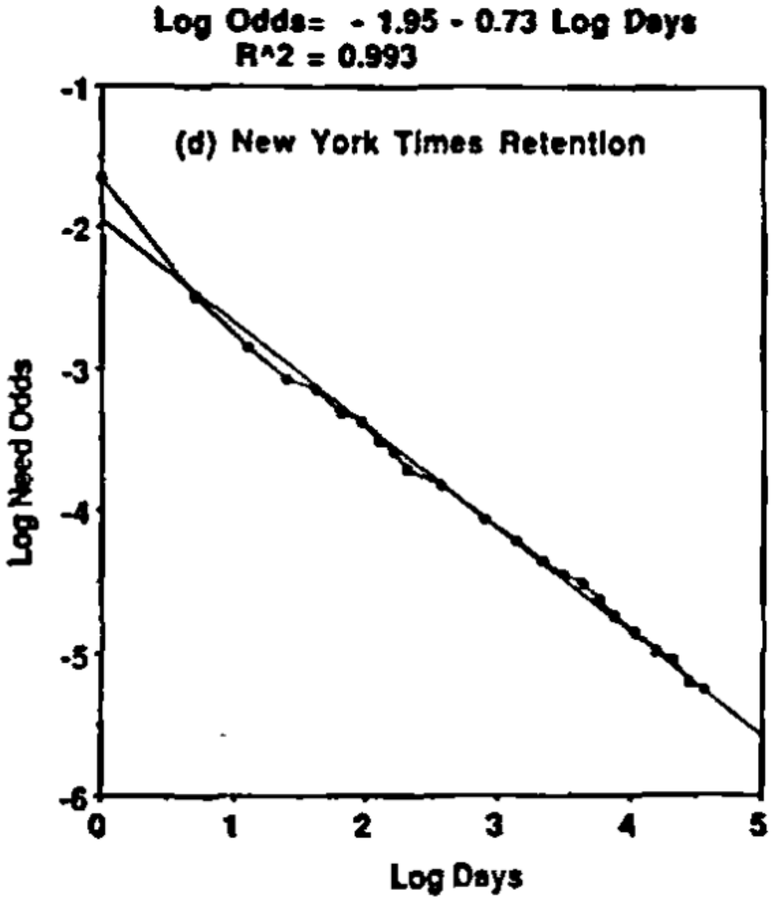
Parsimony

Simpler models are better models (Occam’s razor)

Goodness of fit

Sum of squared errors (SSE), Log likelihood, etc





The relationship between our model and the truth

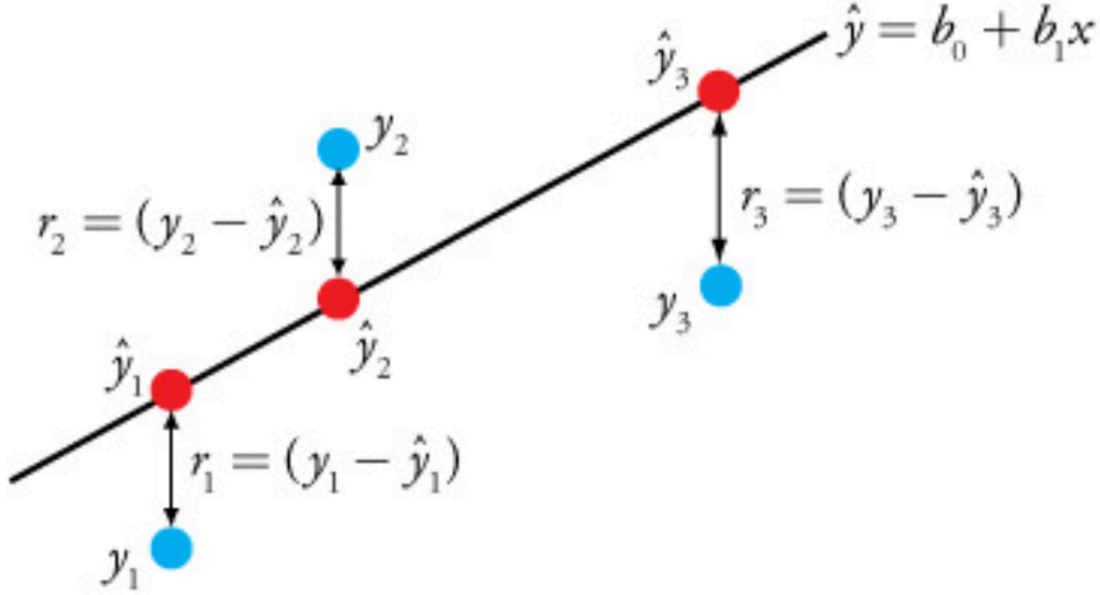
$$y = f(x, \theta) + E$$

$$\hat{y} = g(x, \theta')$$

Want to pick g , θ' such that $g \approx f, \theta' \approx \theta$

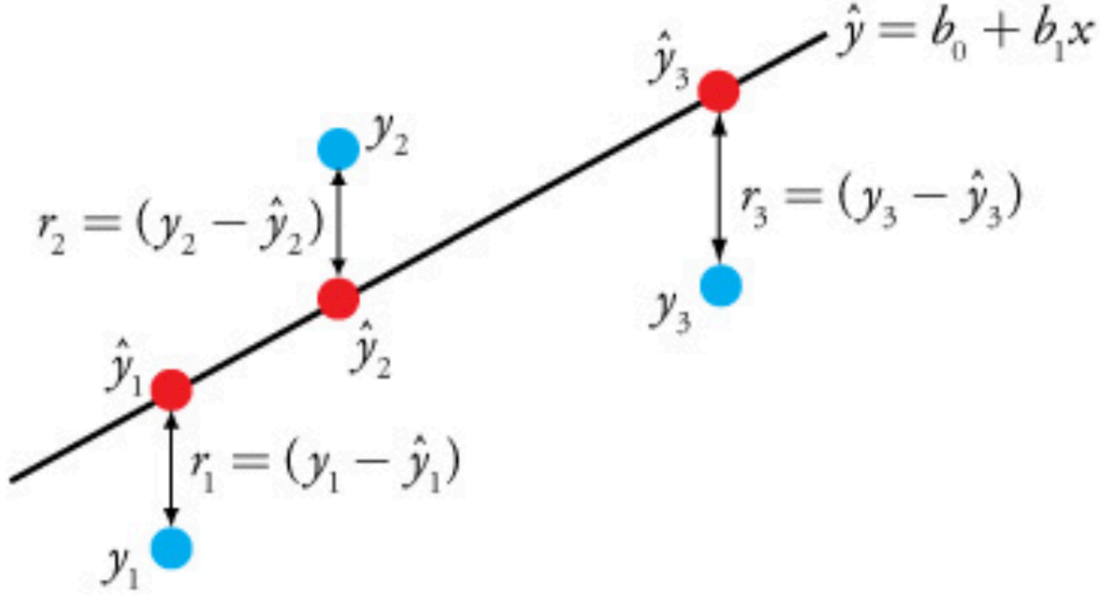
Sum of Square Errors

$$SSE = \sum (y - \hat{y})^2$$



Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{\sum (y - \hat{y})^2}{N}}$$



Percent Variance Accounted For (PVAF)

$$PVAF = \frac{SSE_{null} - SSE_{model}}{SSE_{null}}$$

$$SSE_{null} = \sum_{i} \left(y_i - \mu \right)^2$$

$$SSE_{model} = \sum_{i} \left(y_i - \hat{y}_i \right)^2$$

Likelihood

$$P\left(D \mid Model\right)$$

Note: Standard Sum of Square Errors (SSE) is equivalent to

$$y \sim \text{Normal}(f(\theta), \sigma)$$

Estimating parameters

1. Calculus
2. Grid Search
3. Optimization algorithms
4. Sampling

Using calculus to find an analytic solution

$$\hat{y}_i = \beta_0 + \beta_1 \cdot d_i$$

For some models, like linear regression, you can use calculus to find parameters that minimize your fit metric

d_i	y_i	$\hat{y}_i$
1	0.74	
2	0.59	
3	0.48	
4	0.36	

Deriving linear regression parameters

$$SSE = \sum_i \left(y_i - \hat{y}_i\right)^2$$

$$= \sum_i \left[y_i - \left(\beta_0 + \beta_1 \cdot d_i\right)\right]^2$$

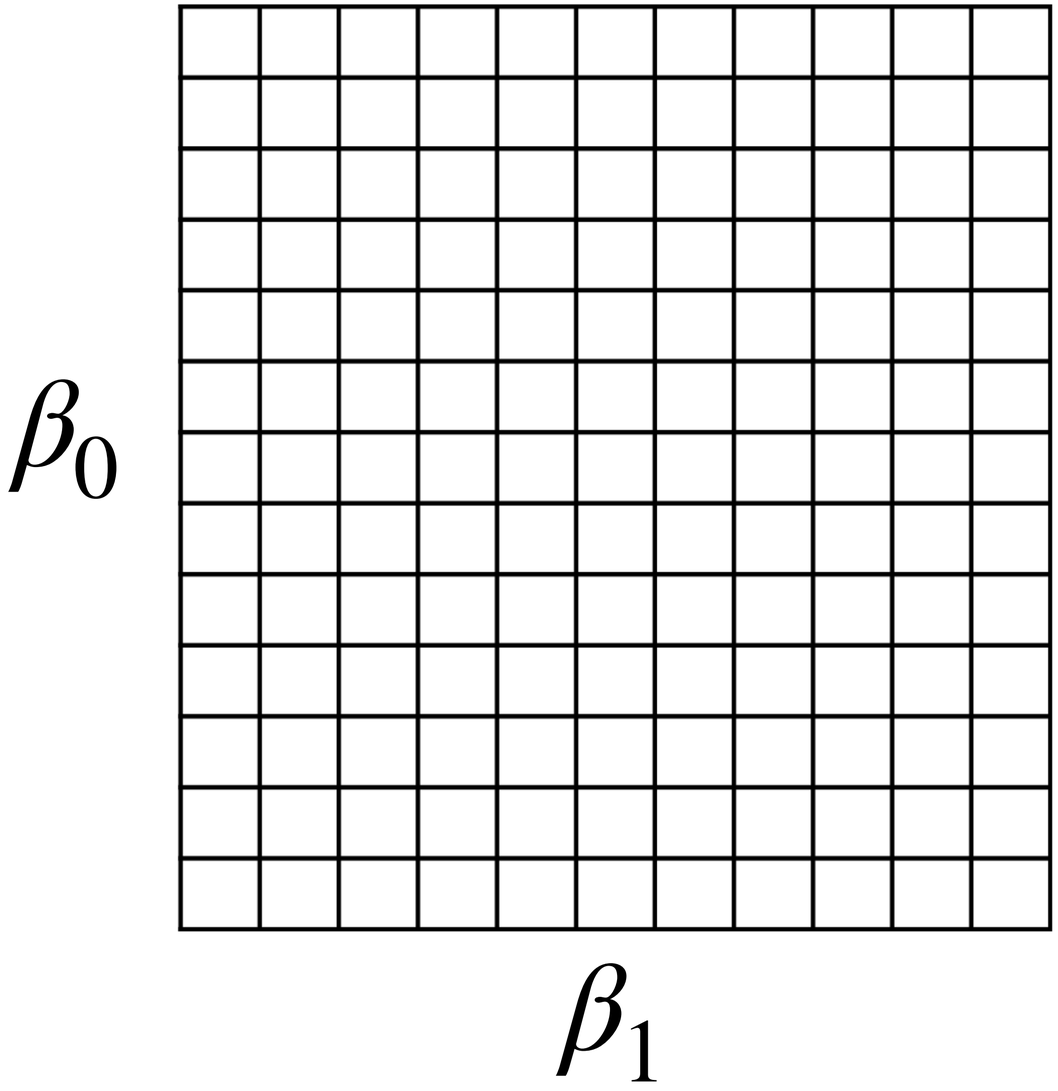
Deriving linear regression parameters

$$SSE = \sum_i \left[y_i - \left(\beta_0 + \beta_1 \cdot d_i \right) \right]^2$$

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_i \left(y_i - \beta_0 + \beta_1 \cdot d_i \right)$$

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_i d_i \left(y_i - \beta_0 + \beta_1 \cdot d_i \right)$$

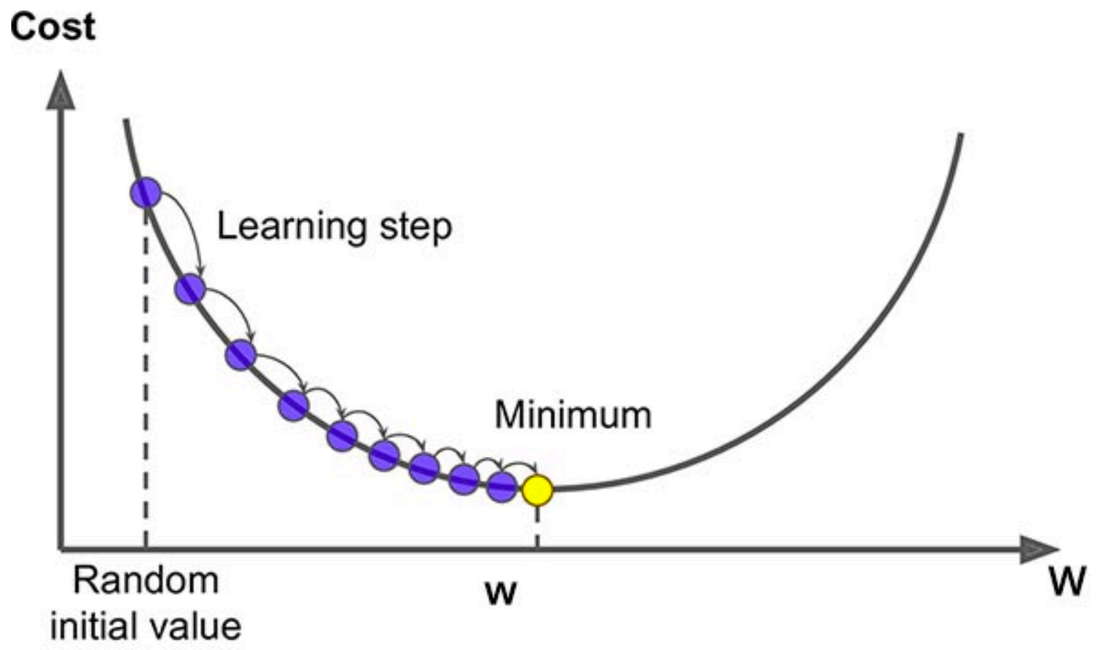
Grid search

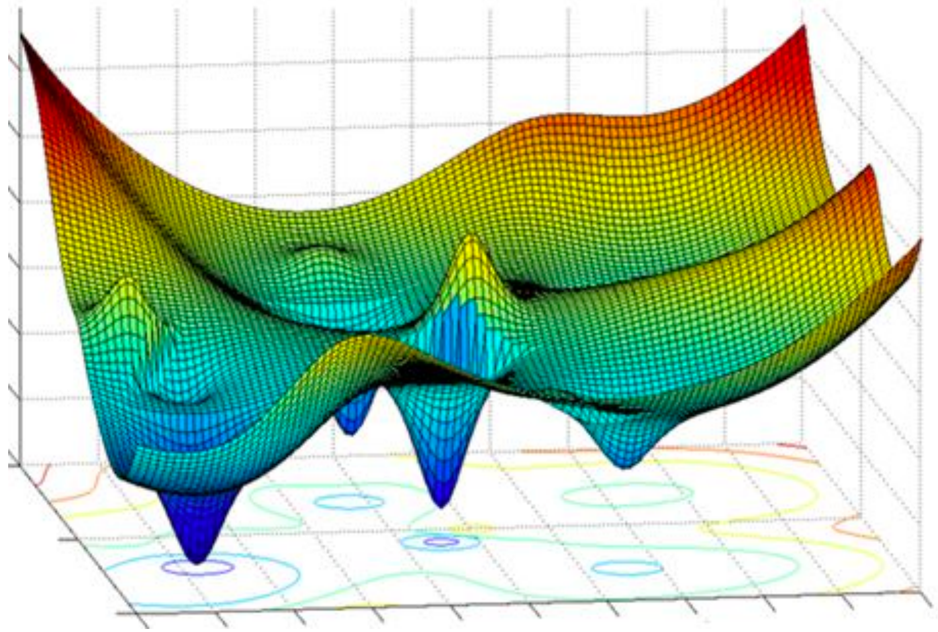


Calculate SSE for every cell and choose the best

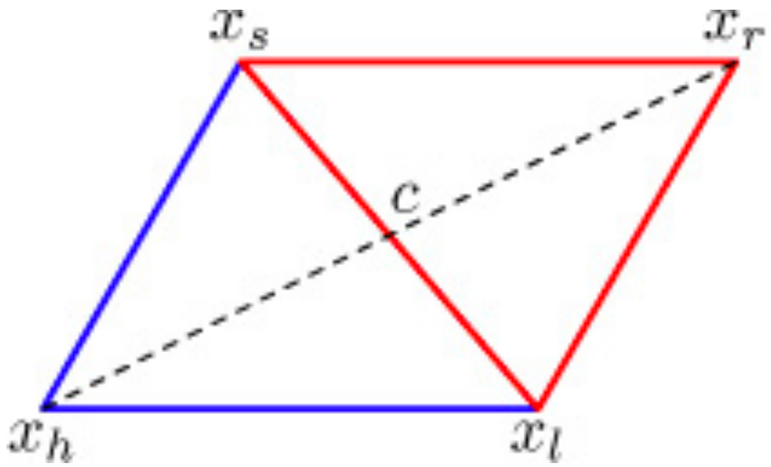
Works only for very small parameter spaces

Optimization algorithms: Gradient Descent

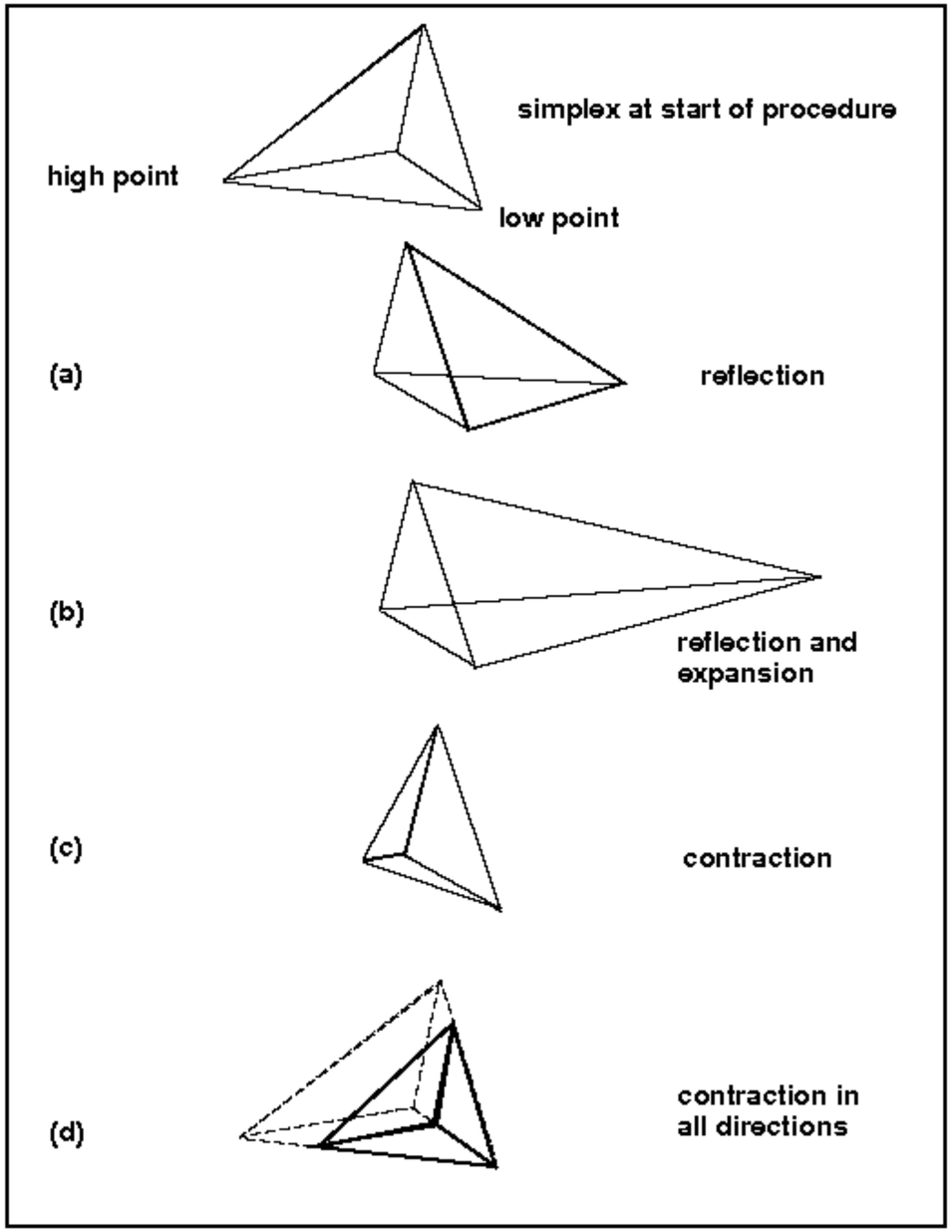




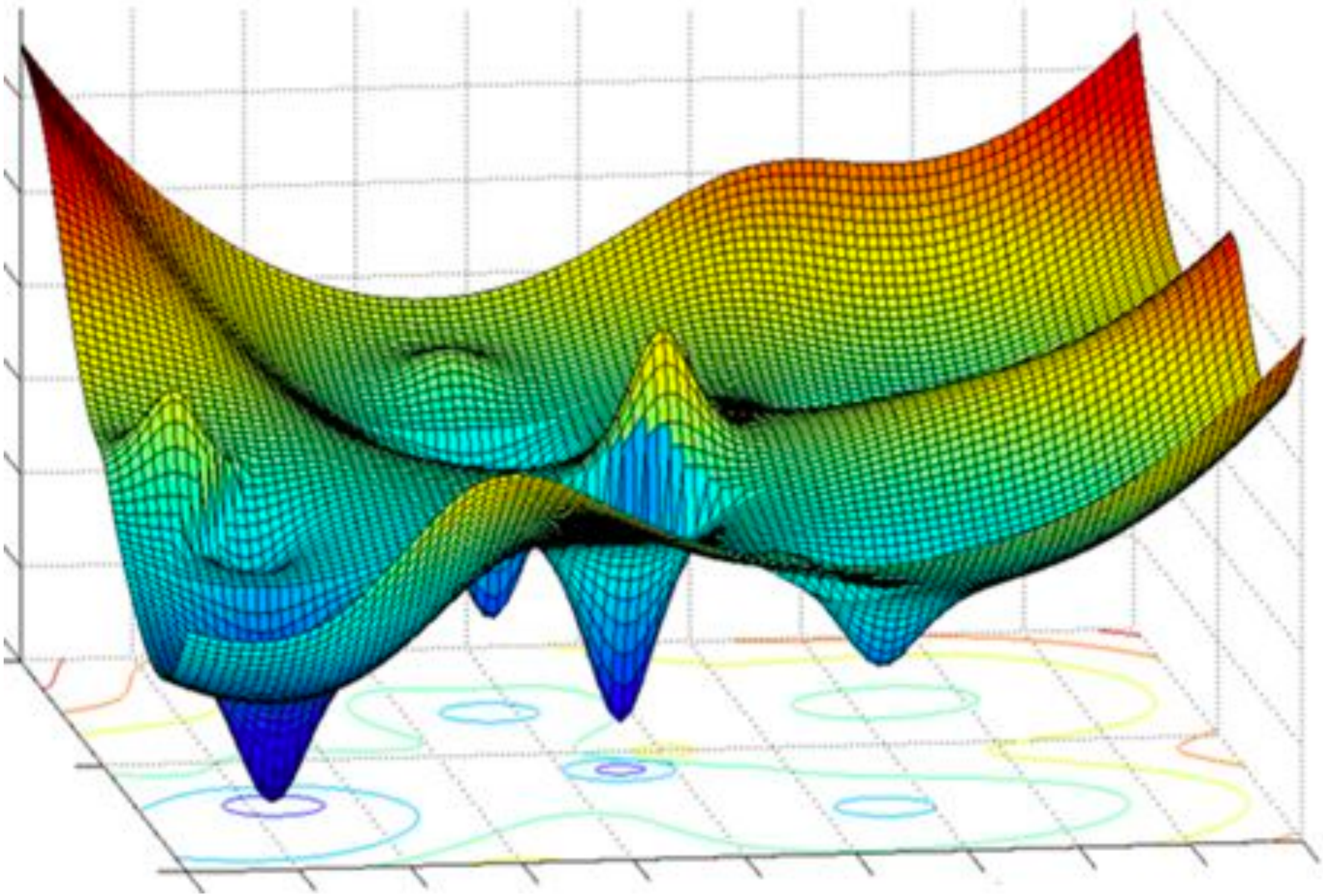
Optimization algorithms: Nelder-Mead Simplex




```
optim(initial_params, cost_function)
```



How do we solve the problem of local minima?



How do you know if your optimization procedure is working?

Run the model with known parameters to generate a simulated data set

Can you recover those parameters?

Outcomes

1. **Yes.** Great news!
2. **No.** And the fit is better with the true parameters
   you need a different search algorithm
3. **No.** But the fit is just as good as with the true parameters
   your parameters may be non-identifiable

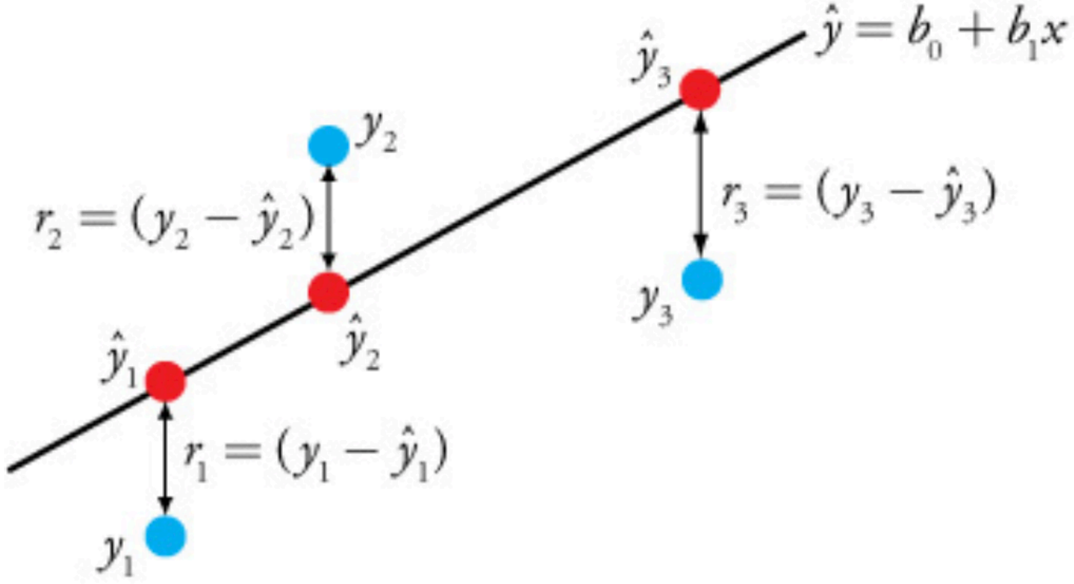
Quantitative criteria

Goodness of fit

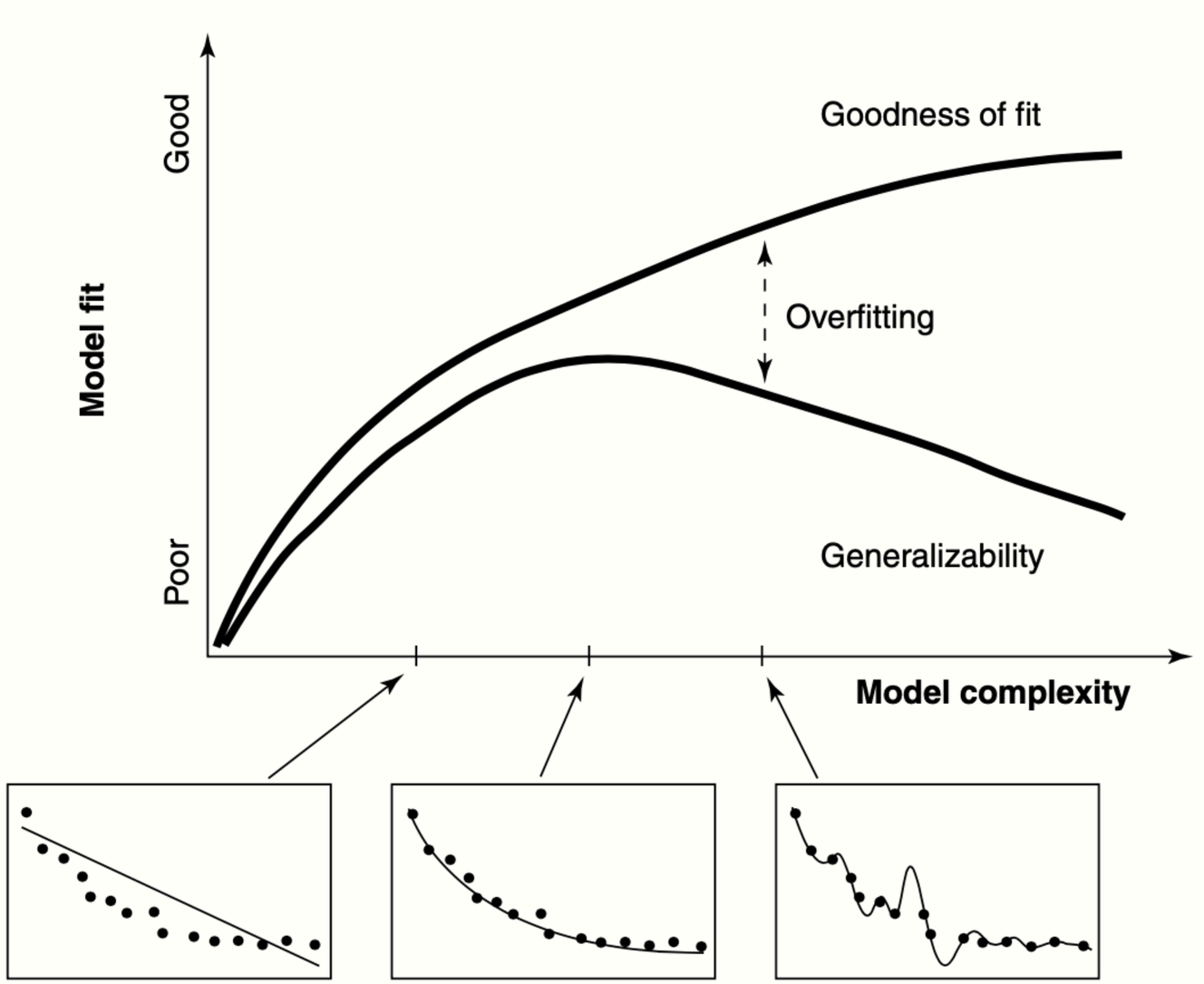
Sum of squared errors (SSE), Log likelihood, etc

A good fit is important
but not sufficient.

Why?



Too much flexibility leads to overfitting (Pitt & Myung, 2002)



Can we recover the true model?

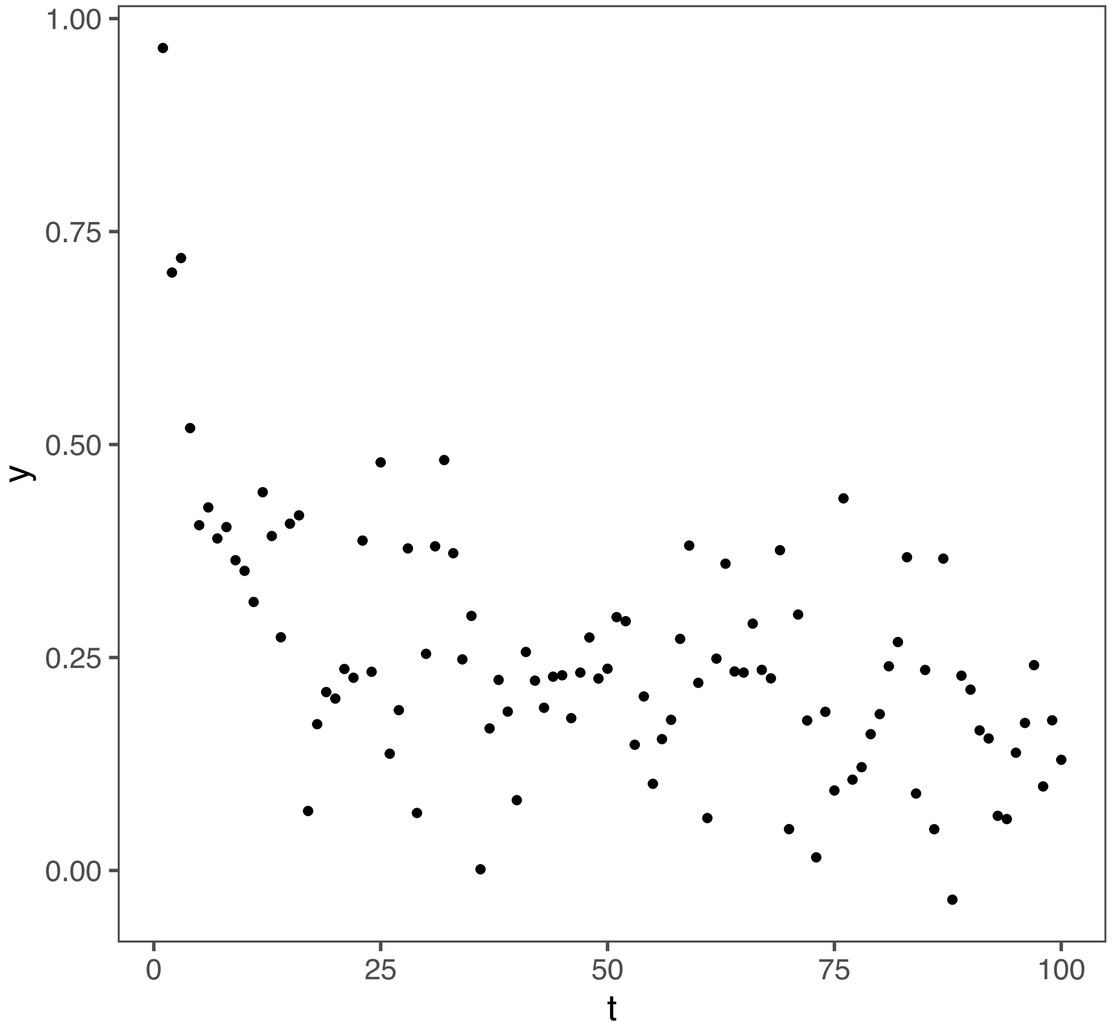
$$M_a : y = (1 + t)^{-a}$$

$$M_b : y = \left(b + ct\right)^{-a}$$

Generate data from

$$M_a + \mathrm{N}(0,.1)$$

$a = .4$



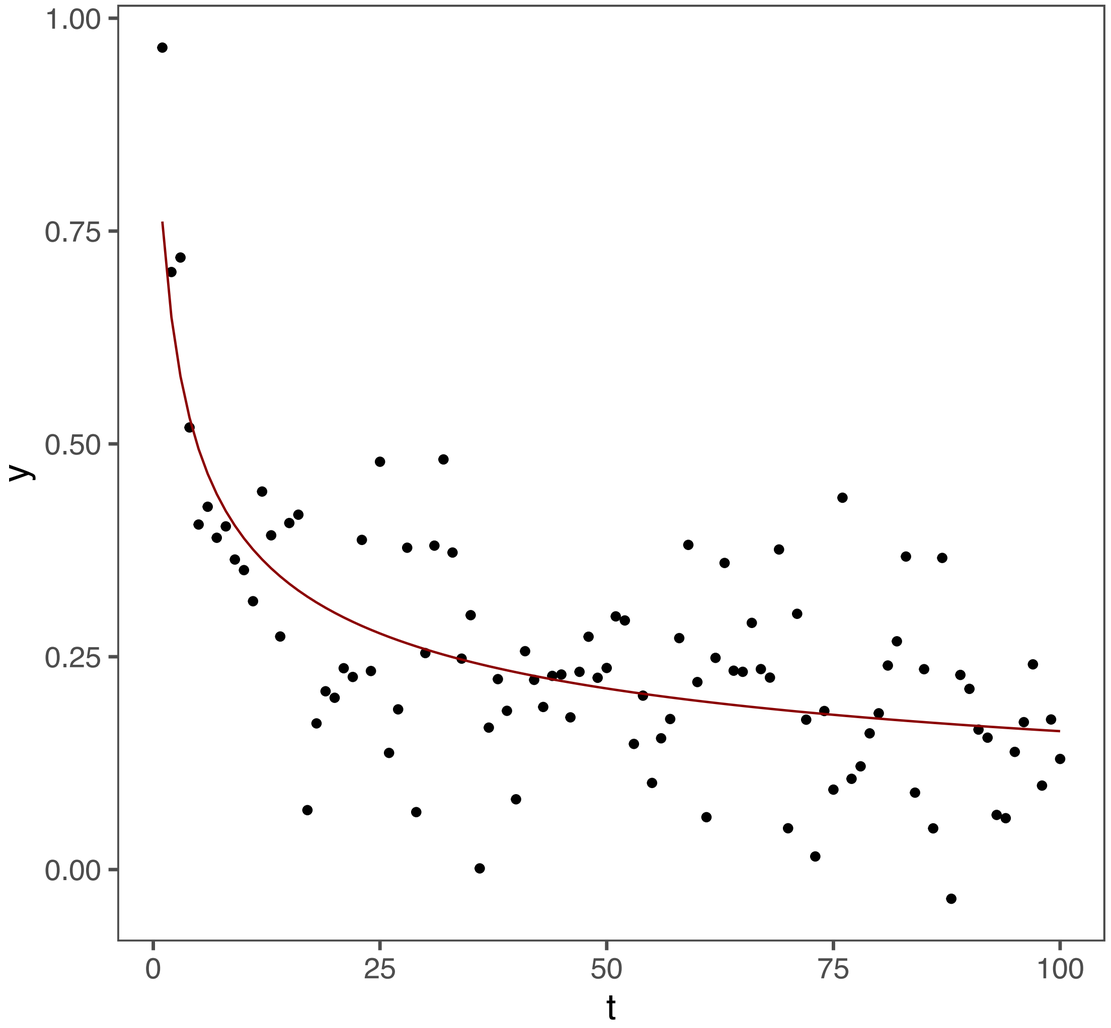
Can we recover the true model?

$$M_a : y = (1 + t)^{-a}$$

$$M_b : y = \left(b + ct\right)^{-a}$$

```
a_opt <- optim(.3,
               fn = loss_a)
```

$$a = .39$$



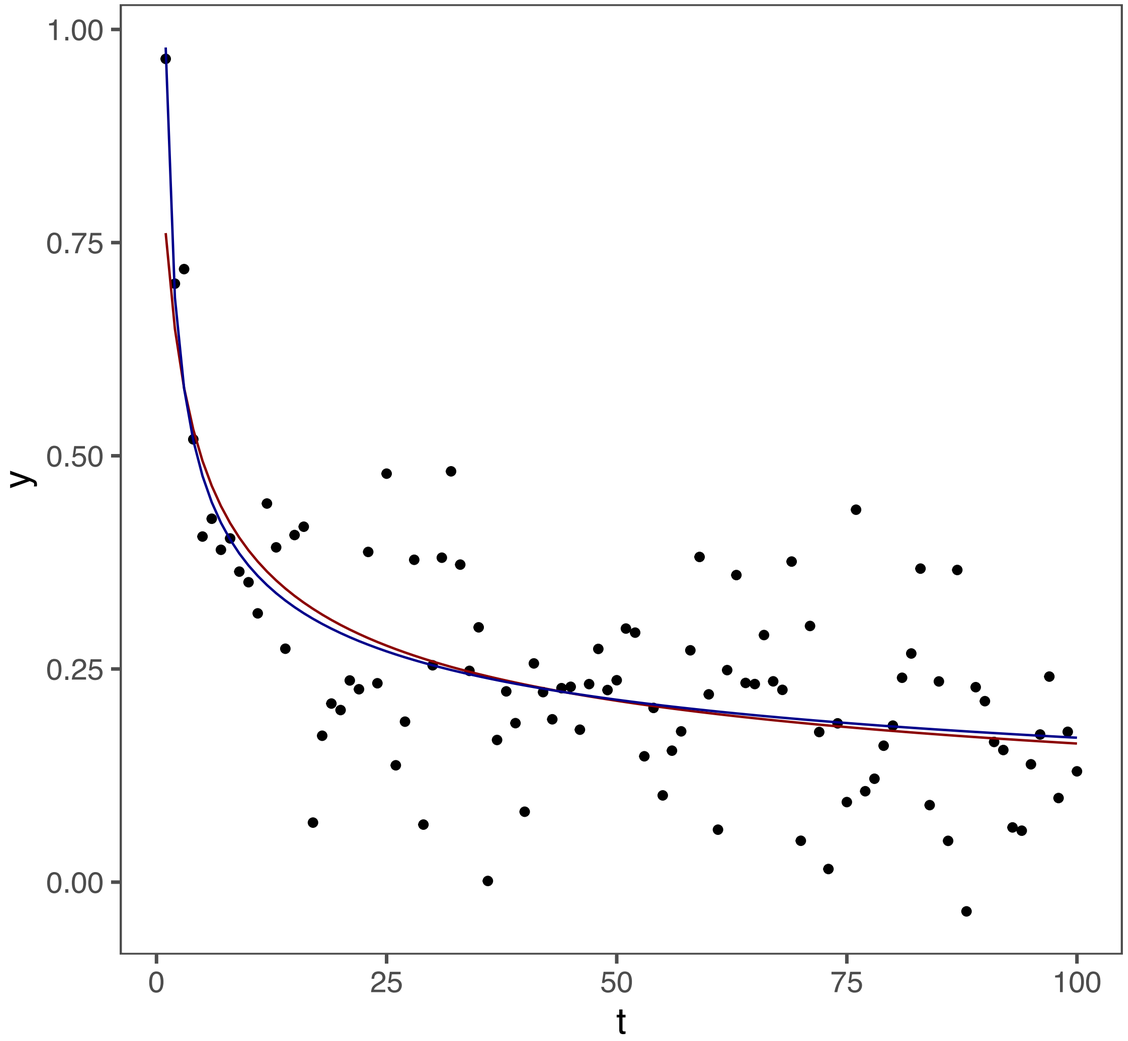
The complex model is preferred

$$M_a : y = (1 + t)^{-a}$$

$$M_b : y = \left(b + ct\right)^{-a}$$

```
a_opt <- optim(.3, fn = loss_a)

b_opt <- optim(c(.3, .2, .1),
               fn = loss_b)
```



$$a = .33, b = -.96, c = 2.02$$

Too much flexibility leads to overfitting (Pitt & Myung, 2002)

Table I. Results of a model recovery simulation in which a GOF measure (RMSE) was used to discriminate models when the source of the error was varied.

Condition (sources of variation)	Model the data were generated from			Model fitted	
	M_A $a = 0.4$	M_A $a = 0.6$	M_B	M_A	M_B
(1) Sampling error	100	–	–	0.040 (0%)	0.029 (100%)
(2) Sampling error + individual differences	50	50	–	0.041 (0%)	0.029 (100%)
(3) Different models	–	50	50	0.075 (0%)	0.029 (100%)
(4) Sampling error	–	–	100	0.079 (0%)	0.029 (100%)

Quantitative criteria

Goodness of fit

Sum of squared errors (SSE), Log likelihood, etc

A good fit qualifies the model as one of the candidate models for further consideration... necessary but not sufficient.

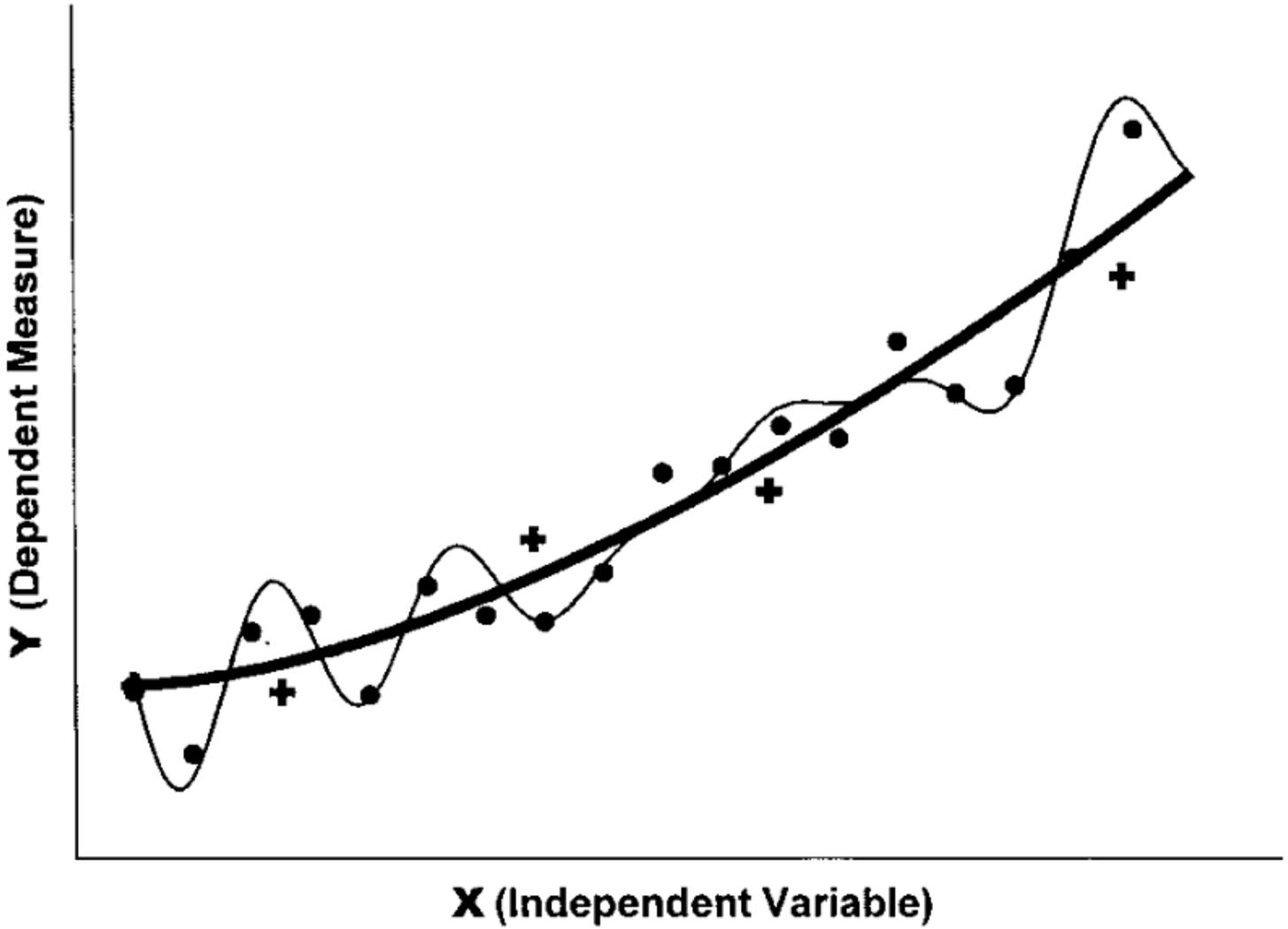
Parsimony

The simplest model that does not fit significantly worse than the most complex model

Generalizability

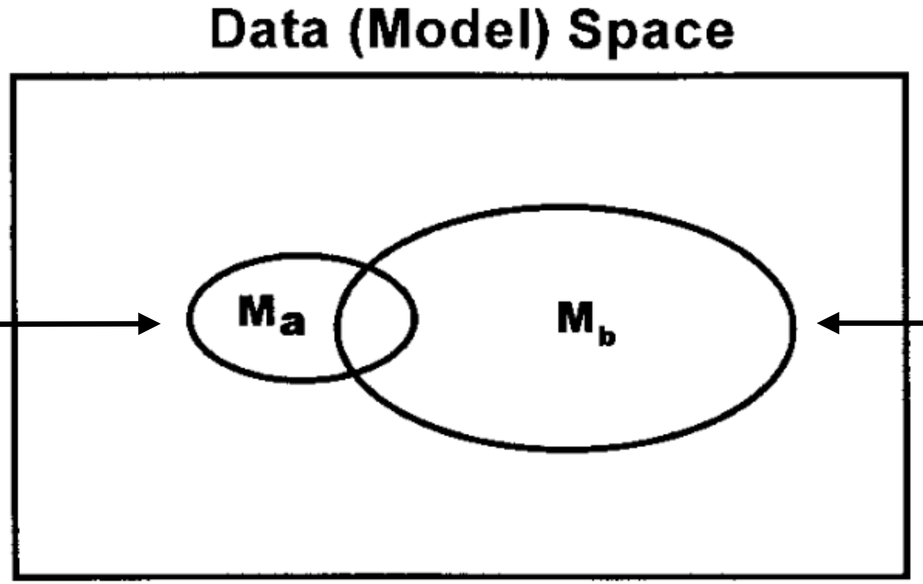
Can the model predict new data?

Fit and generalization can trade off (Pitt, Myung, & Zhang, 2002)

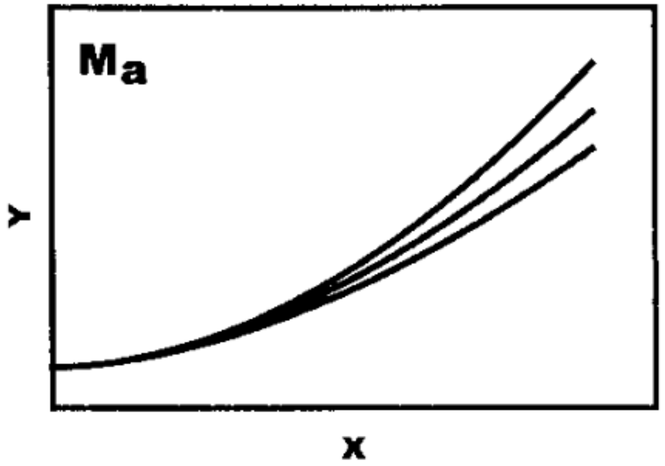


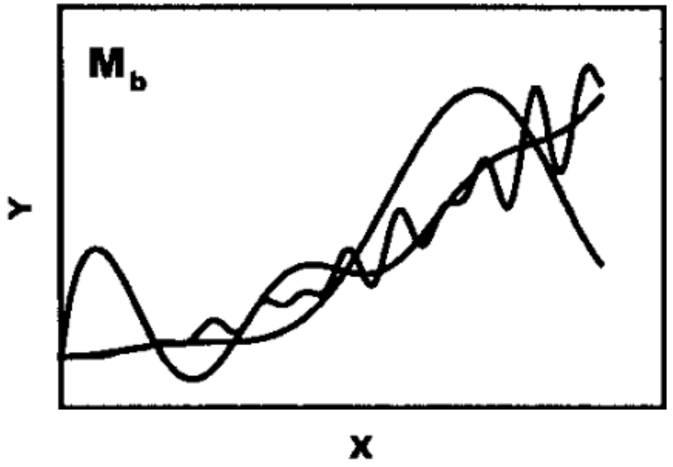
Complex models can predict a lot of different patterns of data

Simple model
(Easy to falsify)



Complex model
(Harder to falsify)





Comparing models

If more complex models can always fit the data better,
how do we compare simple and complex models?

Intuition: Penalize complex models for their complexity

But how?

Measures of fit and generalizability

Selection method	Criterion equation	Dimensions of complexity considered
Root Mean Squared Error	$RMSE = (SSE/N)^{1/2}$	None
Percent Variance Accounted For	$PVAF=100(1-SSE/SST)$	None
Akaike Information Criterion	$AIC = -2\ ln(f(y\|\theta_0)) + 2k$	Number of parameters
Bayesian Information Criterion	$BIC = -2\ ln(f(y\|\theta_0)) + k \cdot ln(n)$	Number of parameters, sample size
Bayesian Model Selection	$BMS=-ln \int f(y\|\theta)\pi(\theta)d\theta$	Number of parameters, sample size, functional form
Minimum Description Length	$MDL=-ln\ (f(y\|\theta_0)) + (k/2)ln(n/2\pi)+ln \int \sqrt{det(I(\theta))}d\theta$	Number of parameters, sample size, functional form

Solution 1: Penalize for number of parameters and sample size

Akaike's Information Criterion

$$AIC = -2\log L + 2K$$

K parameters

N data points

Bayesian Information Criterion

$$BIC = -2\log L + K\log N$$

Akaikie's information criterion (AIC)

$$AIC = -2\log L + 2K$$

K parameters

Better fit

More complexity

$$AIC_b - AIC_a$$

> 0 model a is better

= 0 models are equivalent

< 0 model b is better

Bayesian information criterion (BIC)

$$BIC = -2\log L + K\log N$$

Related to “Bayes Factor”

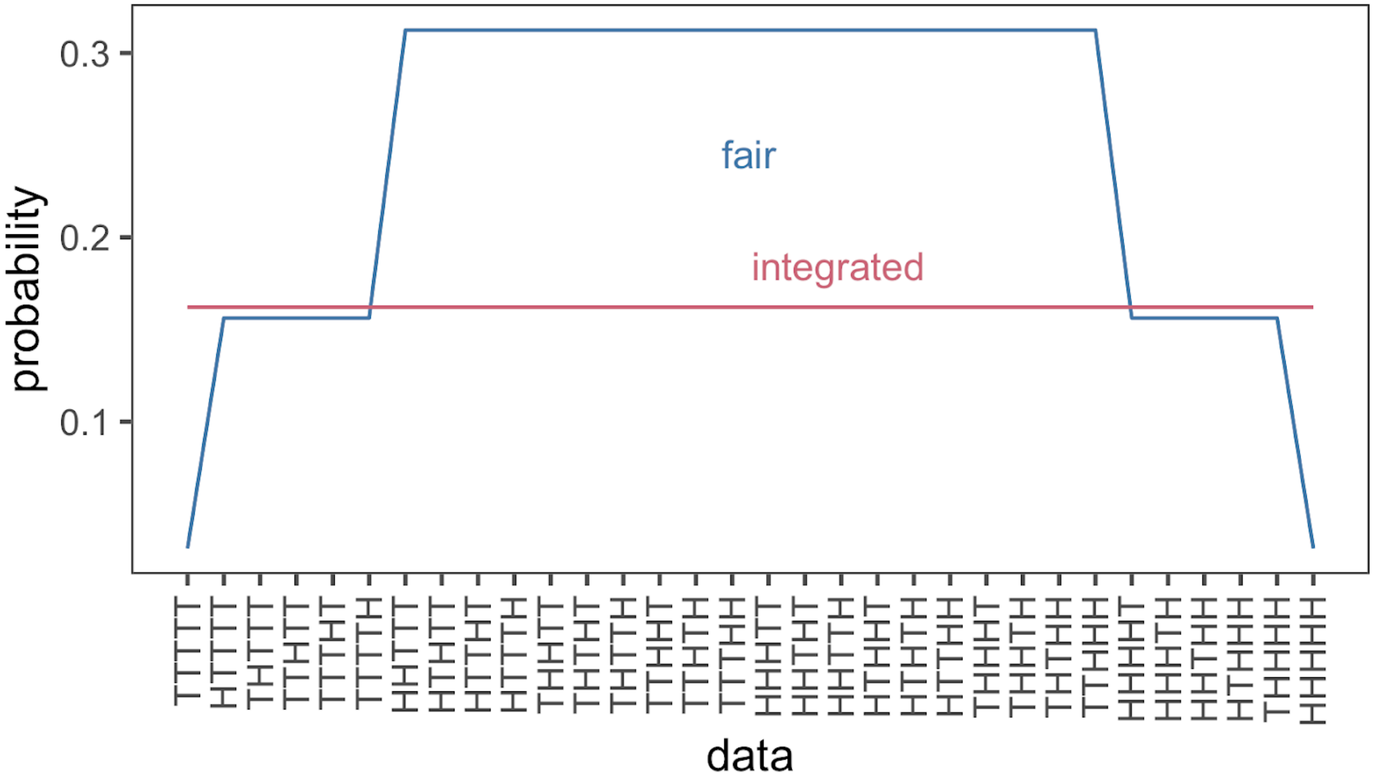
$$B = \frac{p\left(M_1 \mid y\right)}{p\left(M_2 \mid y\right)} = e^{-\frac{1}{2}\Delta BIC}$$

The problem with AIC and BIC: Is complexity the same as parameters?

$$M_a : y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$

$$M_b : y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2$$

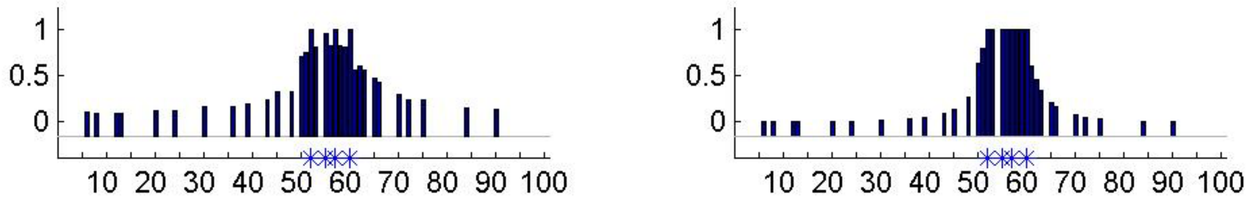
Solutions: Be a Bayesian and penalize the model for functional form as well



$$P\left(M \mid D\right) = \int_{\theta} P\left(D \mid \theta\right) p\left(\theta\right)$$

What else is missing?

Predictions from the number game model:



60 52 57 55

Model has one free parameter: λ

$$P\left(M \mid D\right) = \int_{\lambda} P\left(D \mid \lambda\right) p\left(\lambda\right)$$

Be a frequentist! Use cross-validation

d_i	y_i	$\hat{y}_i$
1	0.74	
2	0.59	
3	0.48	
4	0.36	

Test Set

Training Set

Estimate parameters on the Training set.

Pick models based on the test set

Comparing models

1. Both qualitative and quantitative methods can be used to distinguish between models

2. Models should be chosen on the basis of generalization, not fit

3. Some common methods for assessing both fit and generalization