



PSA 2016

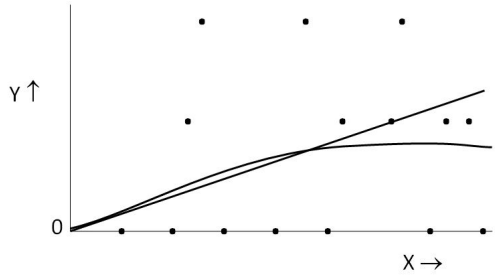
# Inherent Complexity

## A problem for Statistical Model Evaluation

Jan-Willem Romeijn
University of Groningen

# Curve fitting

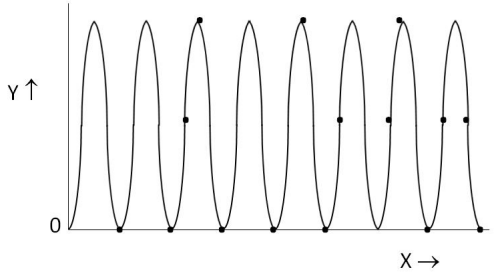
A prime example of model selection is fitting a curve to a scatter plot.




From what family of curves should we choose the best fitting one? Linear, quadratic, or otherwise?

# Perfect fit

What to make of this curve? It uses only two parameters but it fits the scatter plot perfectly.




This paper diagnoses what is wrong with the sine curves, and then draws a number of general lessons.

# Contents


**1 Statistical model evaluation** **5**

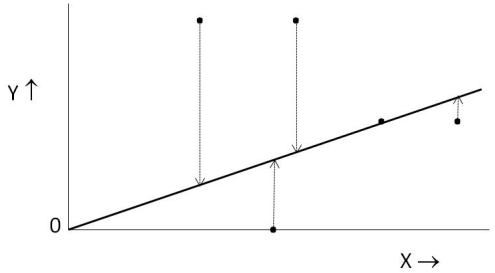
**2 Cheap and almost perfect fit** **8**

**3 Diagnosis of the problem** **12**

**4 Conclusions** **15**

# 1 Statistical model evaluation

Curve-fitting is statistical inference. Each curve determines a probability for possible scatter plots, $P_\theta(S_{xy})$.




We choose a curve from a family, e.g., linear curves, by maximizing the likelihood of the curve for a given scatter plot.

**Evaluating families of curves**

Curve fitting also concerns the evaluation of families of curves, or models. A model $\mathcal{M}$ can be scored on a variety of performance measures:

$$\text{AIC}(\mathcal{M}) = 2\log\left(P_{\hat{\theta}}(S_{xy})\right) - 2\dim(\mathcal{M}),$$
$$\text{BIC}(\mathcal{M}) = 2\log\left(P_{\hat{\theta}}(S_{xy})\right) - \log(m)\dim(\mathcal{M}).$$

These so-called information criteria are based on different good-making features but they all involve a trade-off between fit and simplicity.

**Dimension of the model**

Moreover, simplicity often shows up as the number of model parameters, $\dim(\mathcal{M})$.

- AIC employs the number of parameters as a result of approximating the expected Kullback-Leibler divergence to the true hypothesis.
- For the BIC a similar penalty drops out of an approximation of the predictive performance, or marginal likelihood, of the model.

The example with the sine curves shows that this expression of simplicity does not cover everything that is salient.

## 2 Cheap and almost perfect fit

We fit the data with a model based on trigonometric functions. For every pair of points $\langle X, Y \rangle$ we have

$$P_\alpha(\langle X, Y \rangle) = N(\mu(X), \sigma).$$

But instead of a polynomial function we now employ a trigonometric one:

$$\mu(X) = \alpha_1 - \alpha_1 \cos(\alpha_2 X).$$

This function gives the mode of the distribution over $Y$ for values of $X$. No distribution is imposed over the latter.

**Really a perfect fit?**

Say that the scatter plot $S_{xy}$ is contained in $X < L$ and $Y < H$. Then choose

$$\alpha_1 = \frac{H}{2}.$$

Consider the curves with $\alpha_2 = {}^{L}\!/_{t}$ for increasing $t$, and look at the specific point $\langle x, y \rangle$. For what values of $t$ do we have that

$$y = \frac{H}{2} - \frac{H}{2}\cos\left(\frac{Lx}{t}\right) \quad ?$$

Or rather, how close does the curve have to be in the $X$-direction, for the curve to be close enough in the $Y$-direction?

**An almost perfect fit**

We set the error in the $Y$-direction to $\epsilon$. For some $k$ we have $x \in [{}^{kL}/_{t}, {}^{(k+1)L}/_{t}]$, in which the curve covers the range $[0, H]$ twice. Then $x$ is close enough if

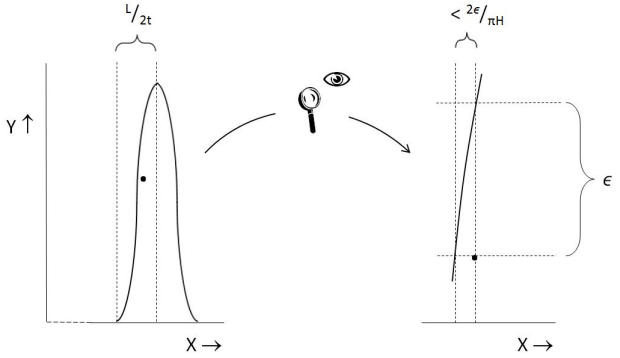
$$\frac{L}{\pi t}\cos^{-1}(1-{}^{2(y-\epsilon)}/_{H}) \quad < \quad x-\frac{kL}{t} \quad < \quad \frac{L}{\pi t}\cos^{-1}(1-{}^{2(y+\epsilon)}/_{H}),$$

and similarly for the second half of the interval. Owing to the slope of the cosines, we have

$$\left[\frac{L}{\pi t}\cos^{-1}(1-{}^{2y}/_{H})\right]-\frac{2\epsilon L}{\pi t H} \quad < \quad x-\frac{kL}{t} \quad < \quad \left[\frac{L}{\pi t}\cos^{-1}(1-{}^{2y}/_{H})\right]+\frac{2\epsilon L}{\pi t H}$$

**In a picture**

For every value of $t$ we can check if $x$ is included in the designated interval of length $4\epsilon L/\pi tH$ within $[kL/t, (k+1)L/t]$. Both intervals shrink at the same rate.




Assuming $x$ to be a random number, the frequency of $x$ being included will tend to $4\epsilon/\pi H$. For $m$ points, a fraction $(4\epsilon/\pi H)^m$ of curves fits with error $\epsilon$.

# 3 Diagnosis of the problem

For any curve fitting problem, there are *infinitely many* sine curves that fit almost any scatter plot almost perfectly. However*...*

- The sine model is far from robust: small nudges to the data lead to completely different curves.
- The dimensions of the sine model are deceptive: the space of distributions is very densely packed.
- In the space of distributions, the set of well fitting sine curves is erratic and disjointed.

In short, the sine model seems wrong.

**Model evaluation criteria**

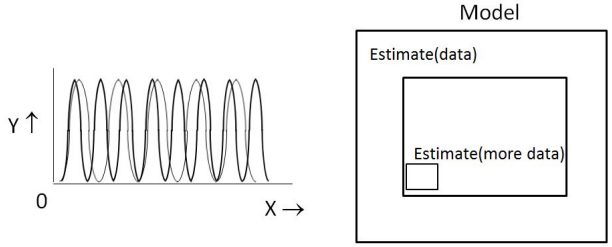
Do the extant model evaluation tools, like the AIC and BIC, give the intuitively correct answer?

- AIC does not return a verdict because it is not defined for unidentified models.
- Superficially the BIC gets it wrong because it simply counts the number of adjustable parameters.
- That also holds for other model evaluation tools like MDL and the DIC, but here the situation is more nuanced.

Several criteria can be refined and adapted to offer the correct verdict.

**Solution key**

With every new data point, the set of well fitting curves diminishes very rapidly. The marginal likelihood of the model brings this out.




To express this diminishing of the solution set, we need a measure over the model. In particular, we need a prior to compute the marginal likelihood.

# 4 Conclusions

There are several general lessons to take away.

- Do not mistake the number of parameters in a model for its actual complexity.
- Regularly return to the deeper motivations for the model evaluation tools, namely the good-making features of models.
- One of these features relates to model size: allowing for fewer possible data patterns is preferable.

But... can we objectively determine how many possible data patterns are packed together in a model?

# Thank you

The slides for this talk will be available at http://www.philos.rug.nl/ romeyn.
For comments and questions, email j.w.romeijn@rug.nl.