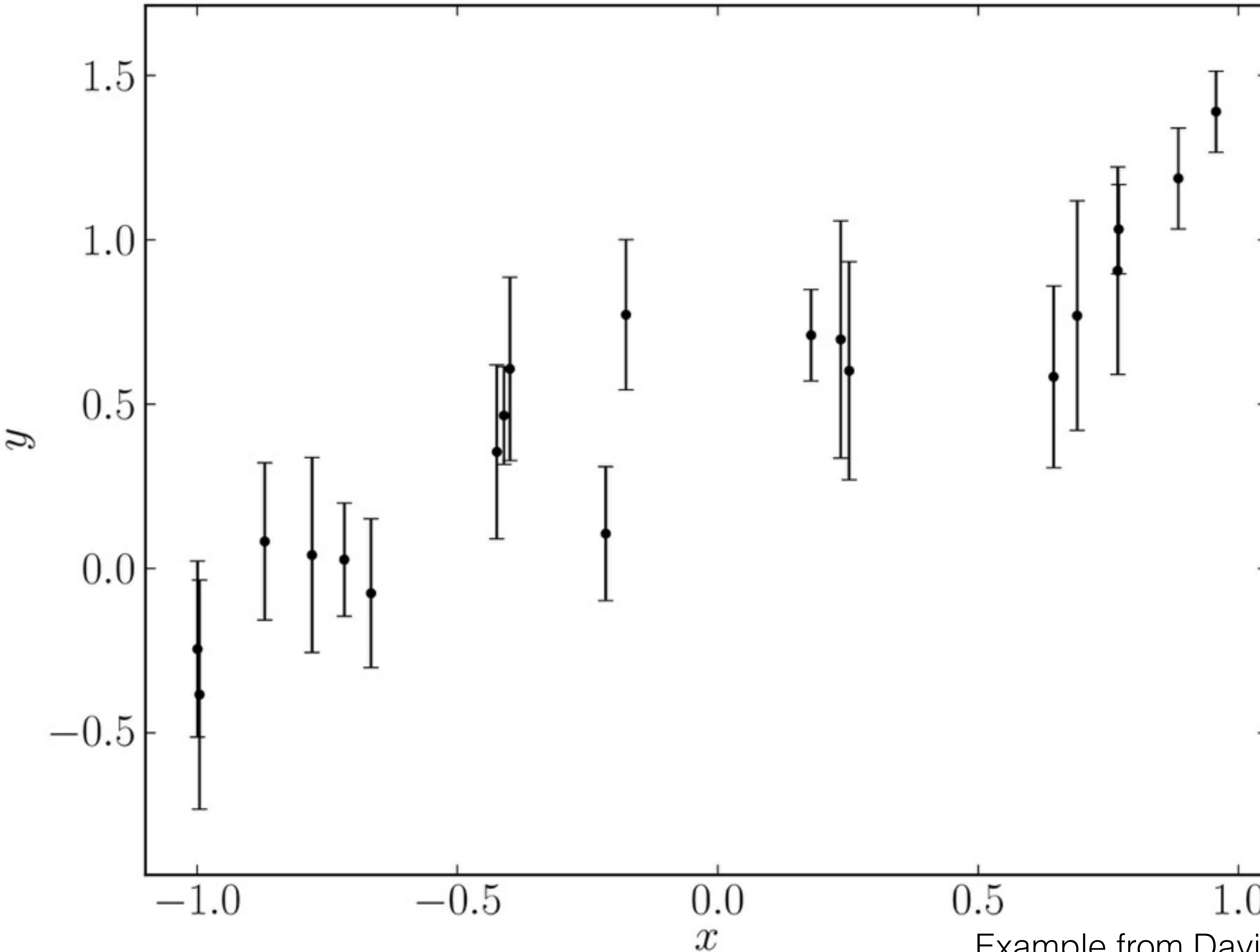
Statistics and Inference in Astrophysics

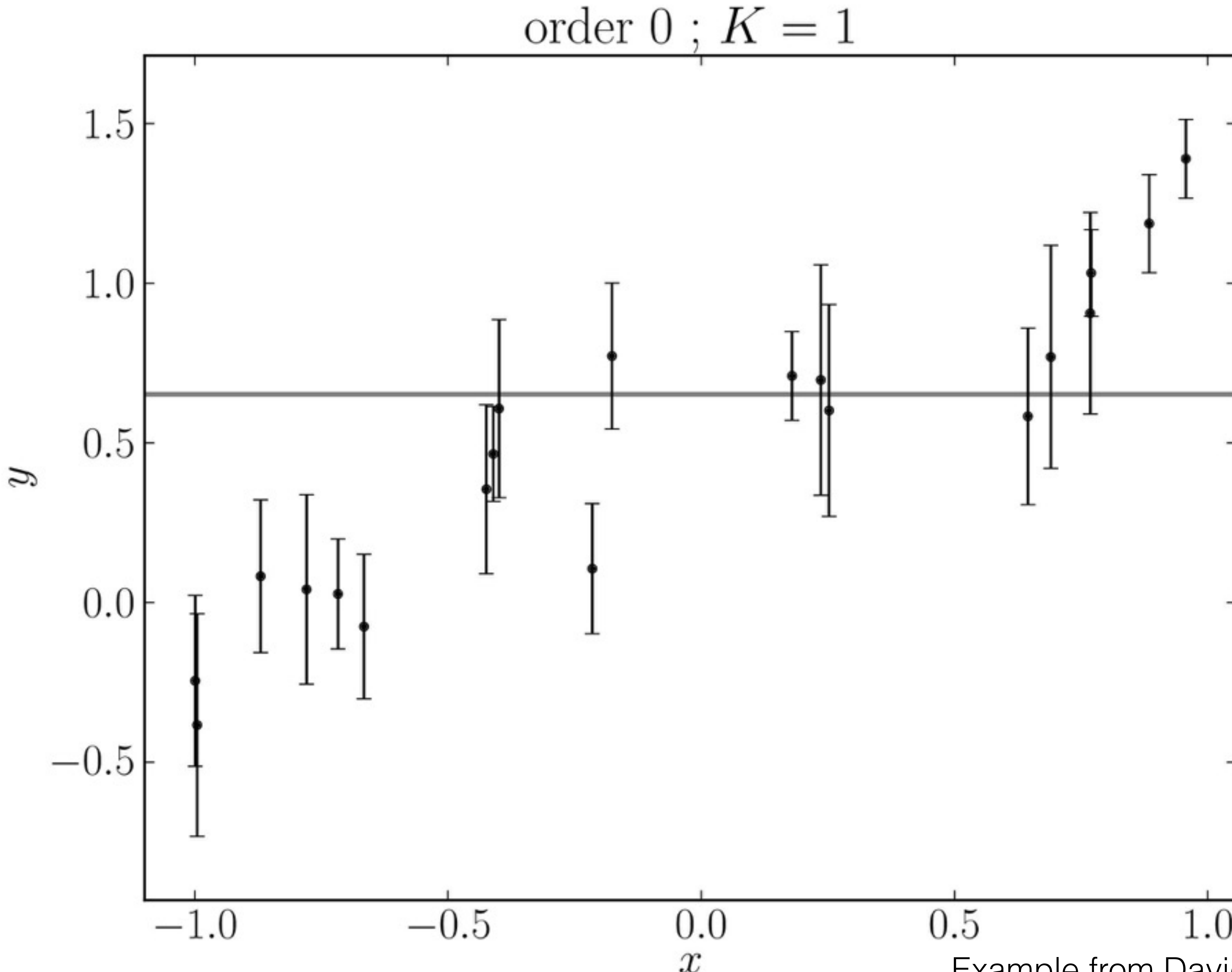
Zhu & Menard (2013)

Today

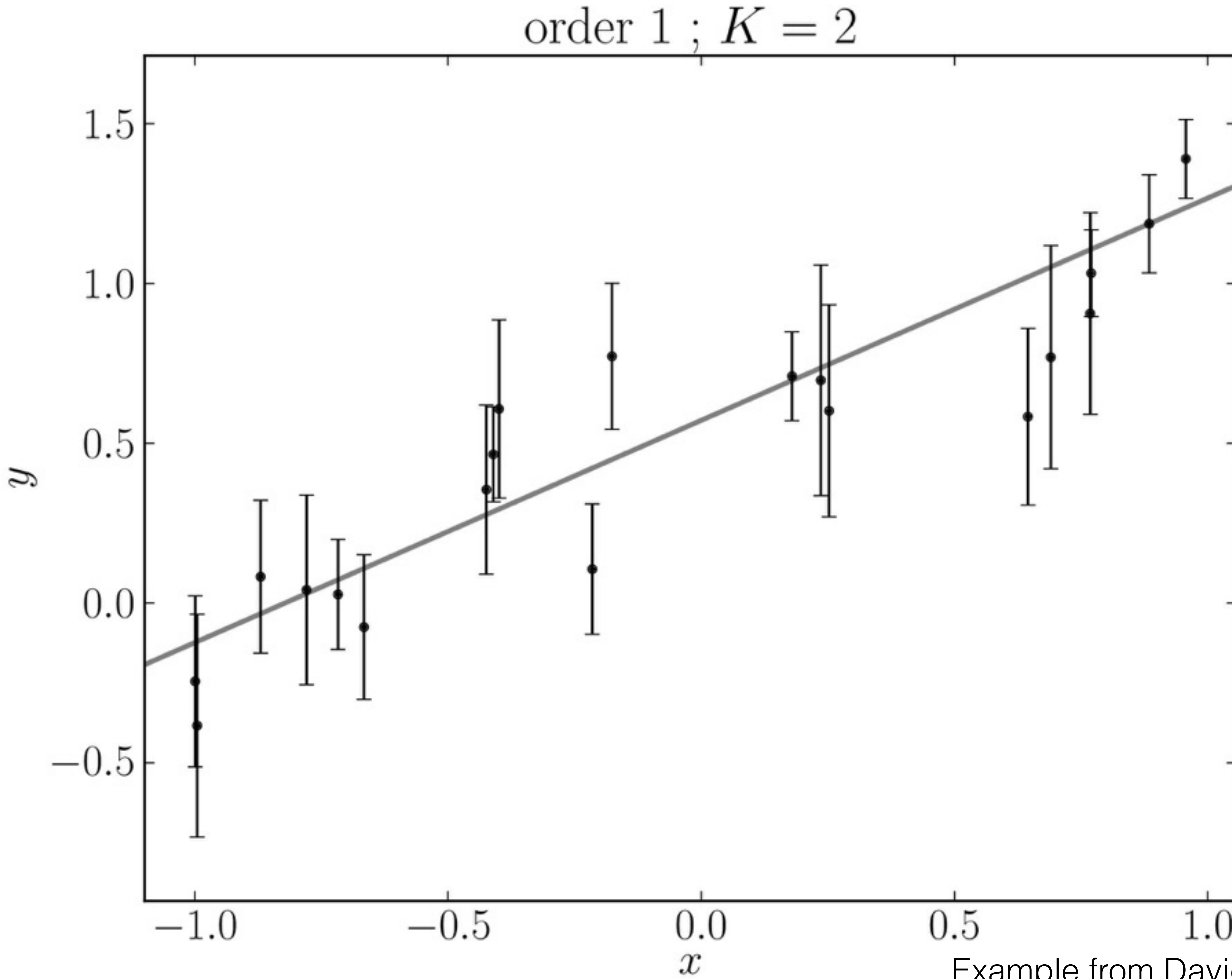
- Goodness-of-fit, model selection, cross-validation
- …



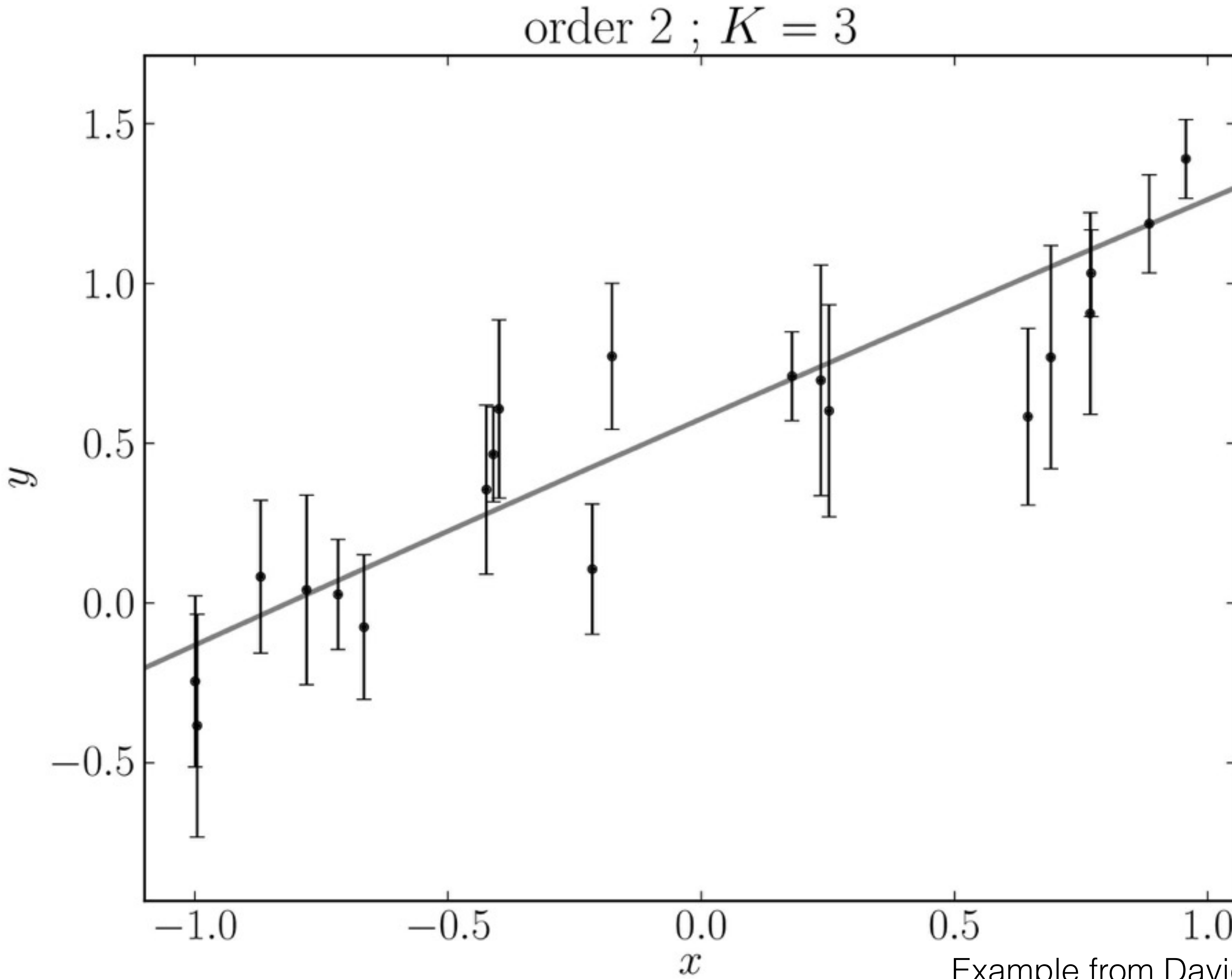
Example from David Hogg



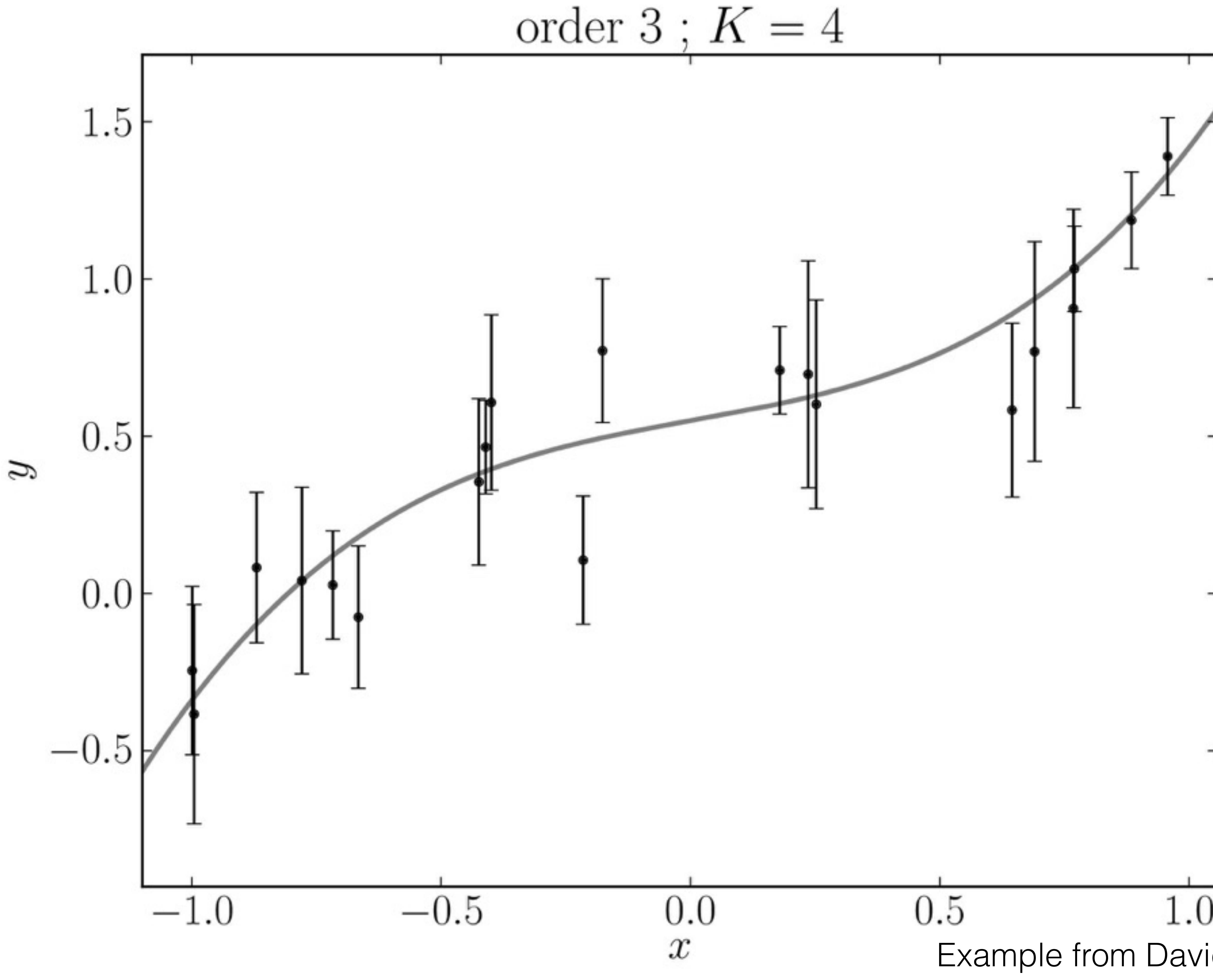
Example from David Hogg



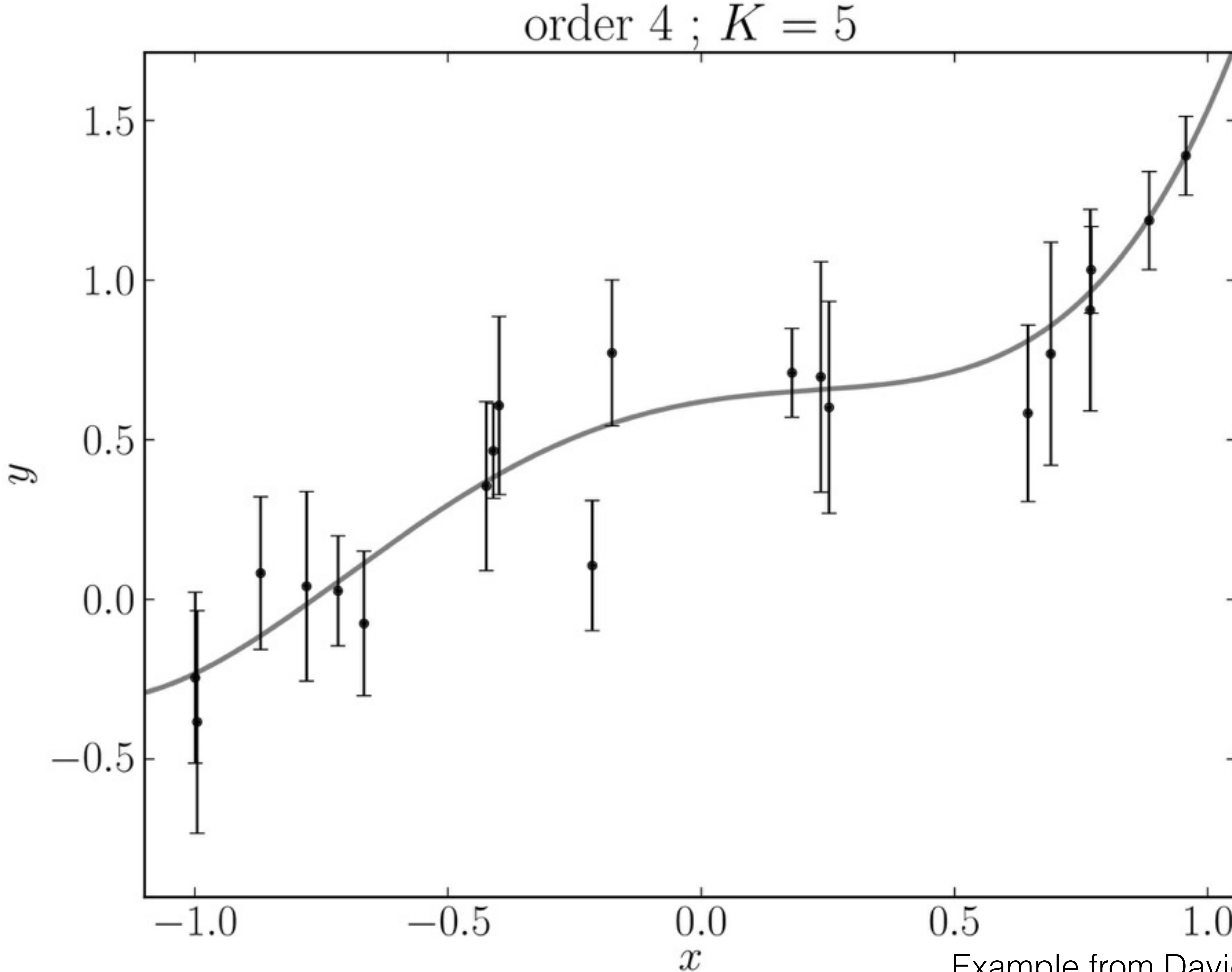
Example from David Hogg



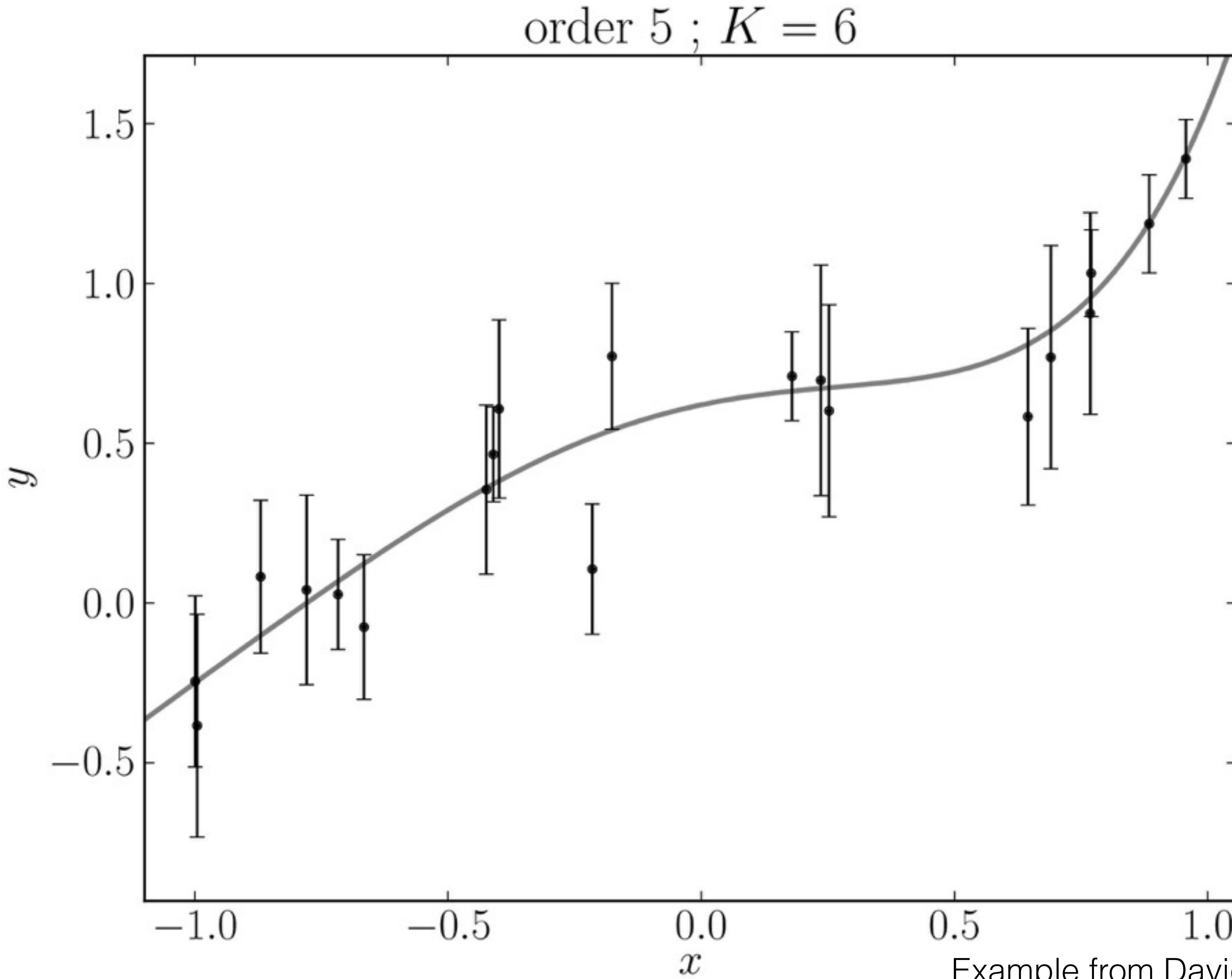
Example from David Hogg



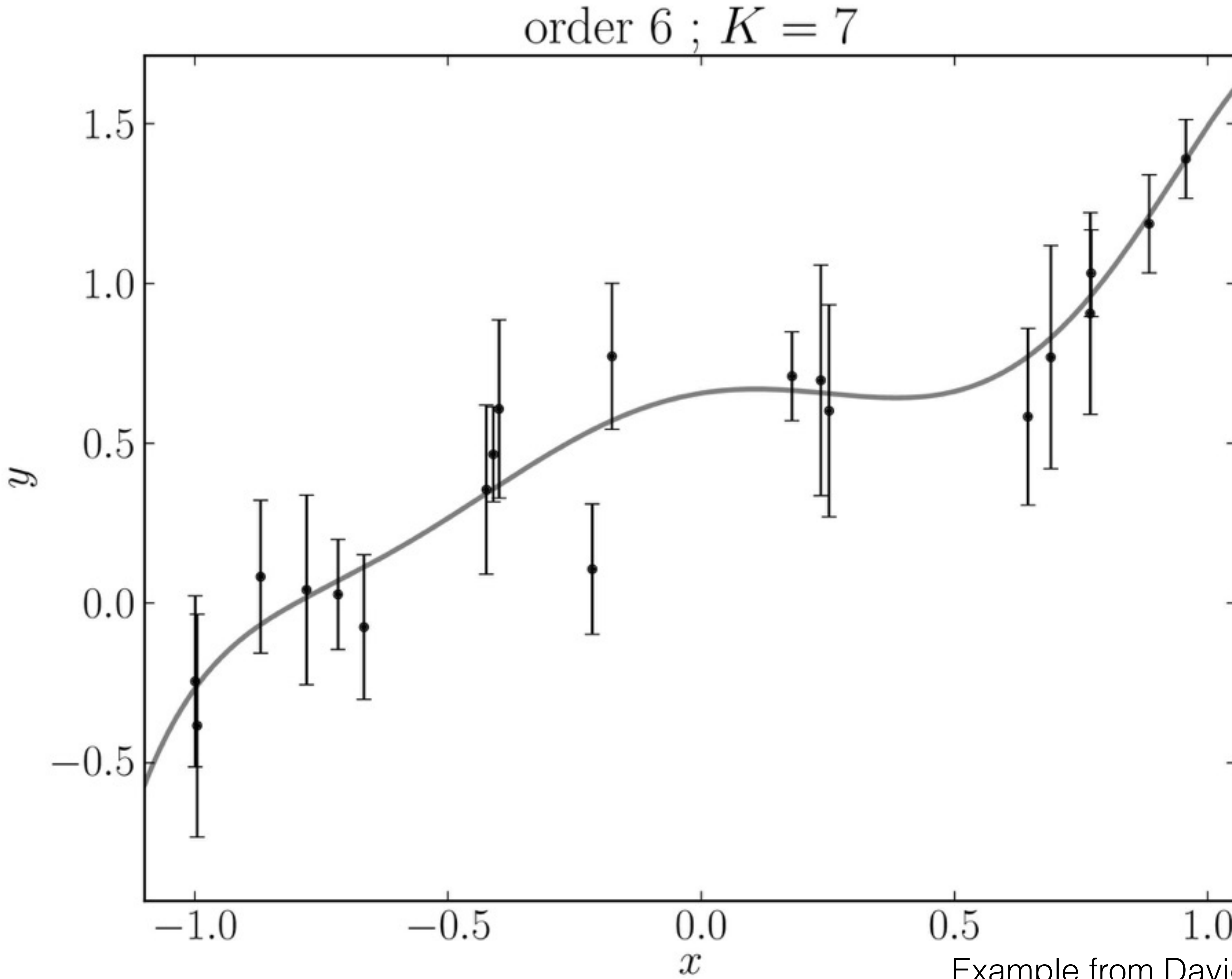
Example from David Hogg



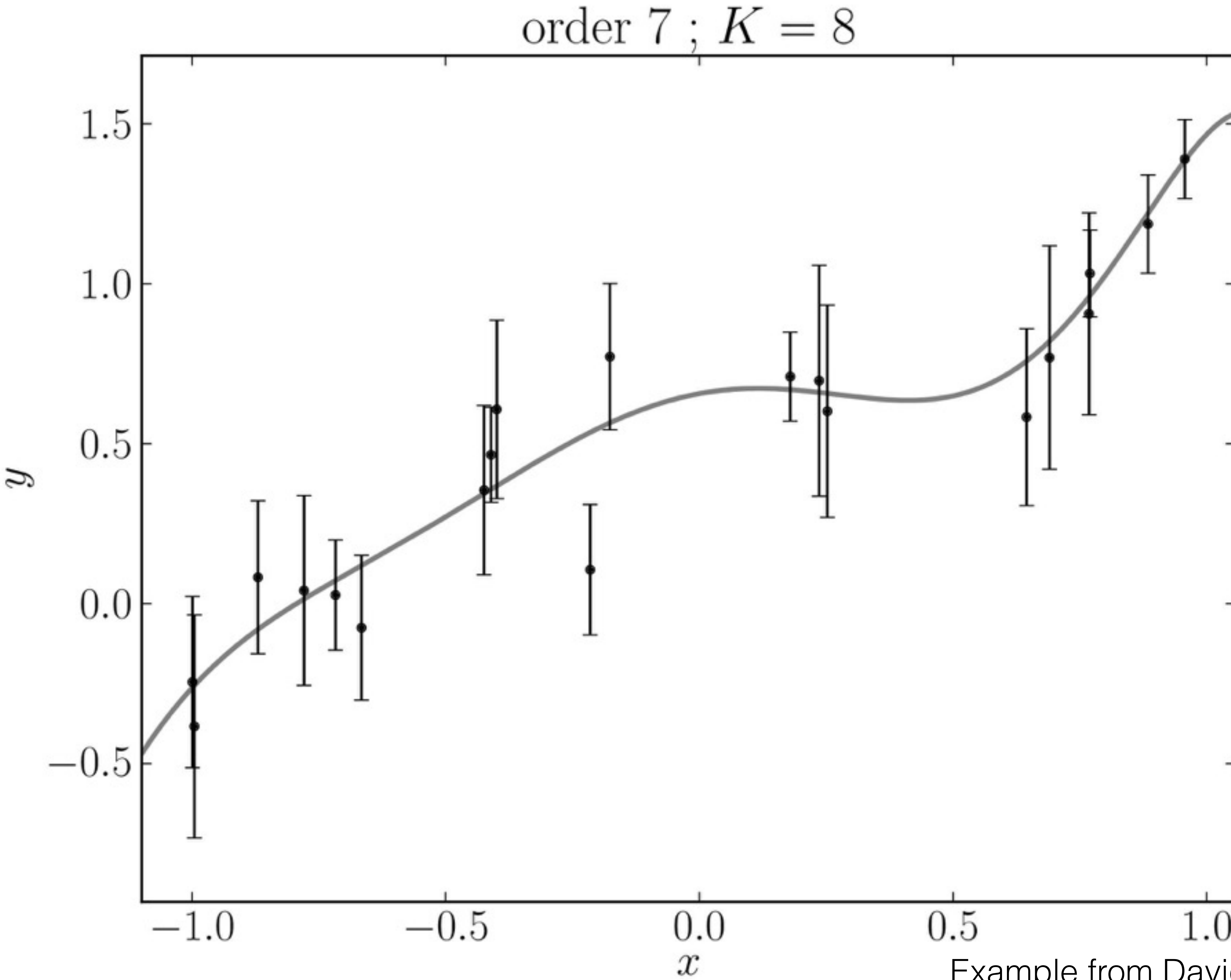
Example from David Hogg



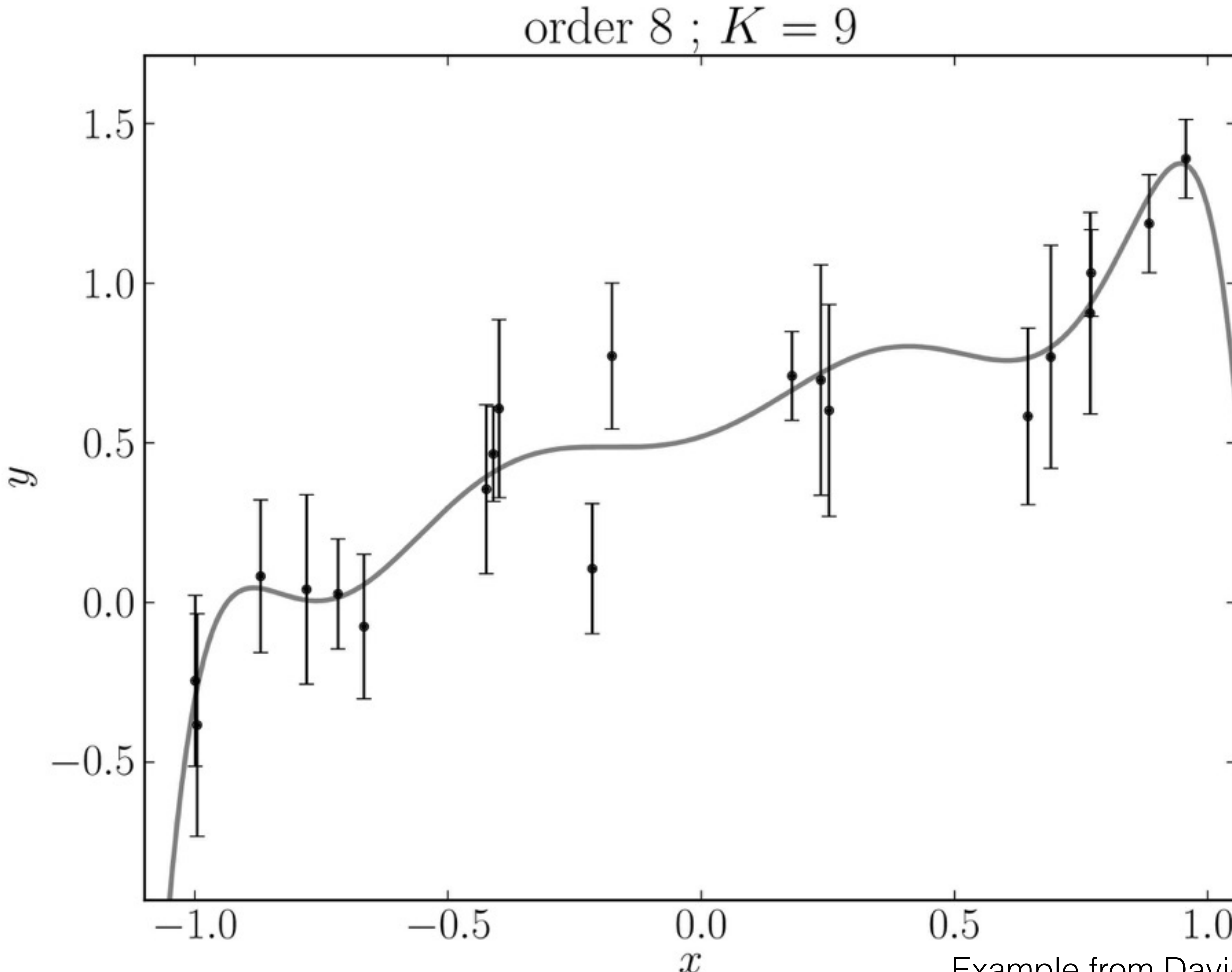
Example from David Hogg



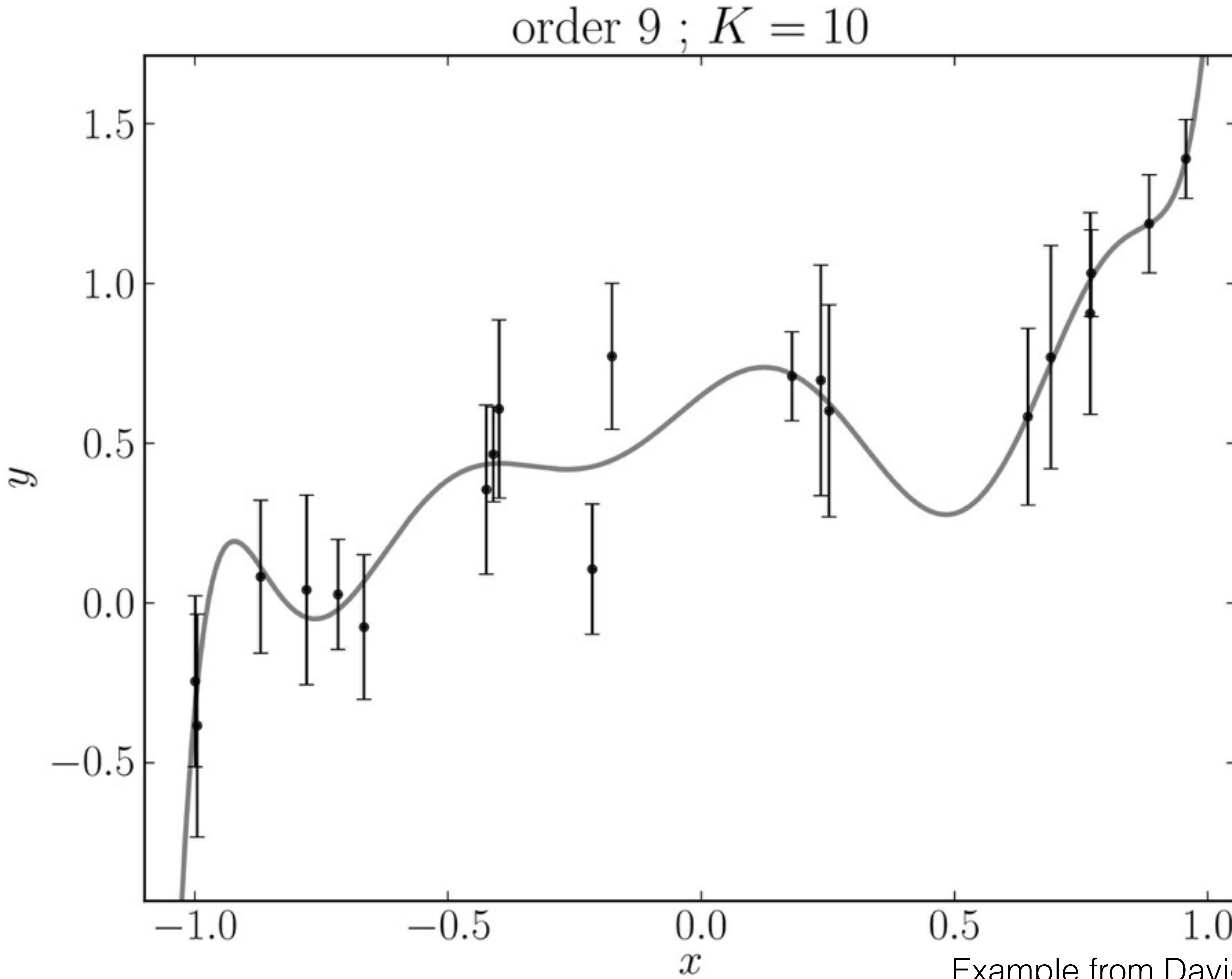
Example from David Hogg



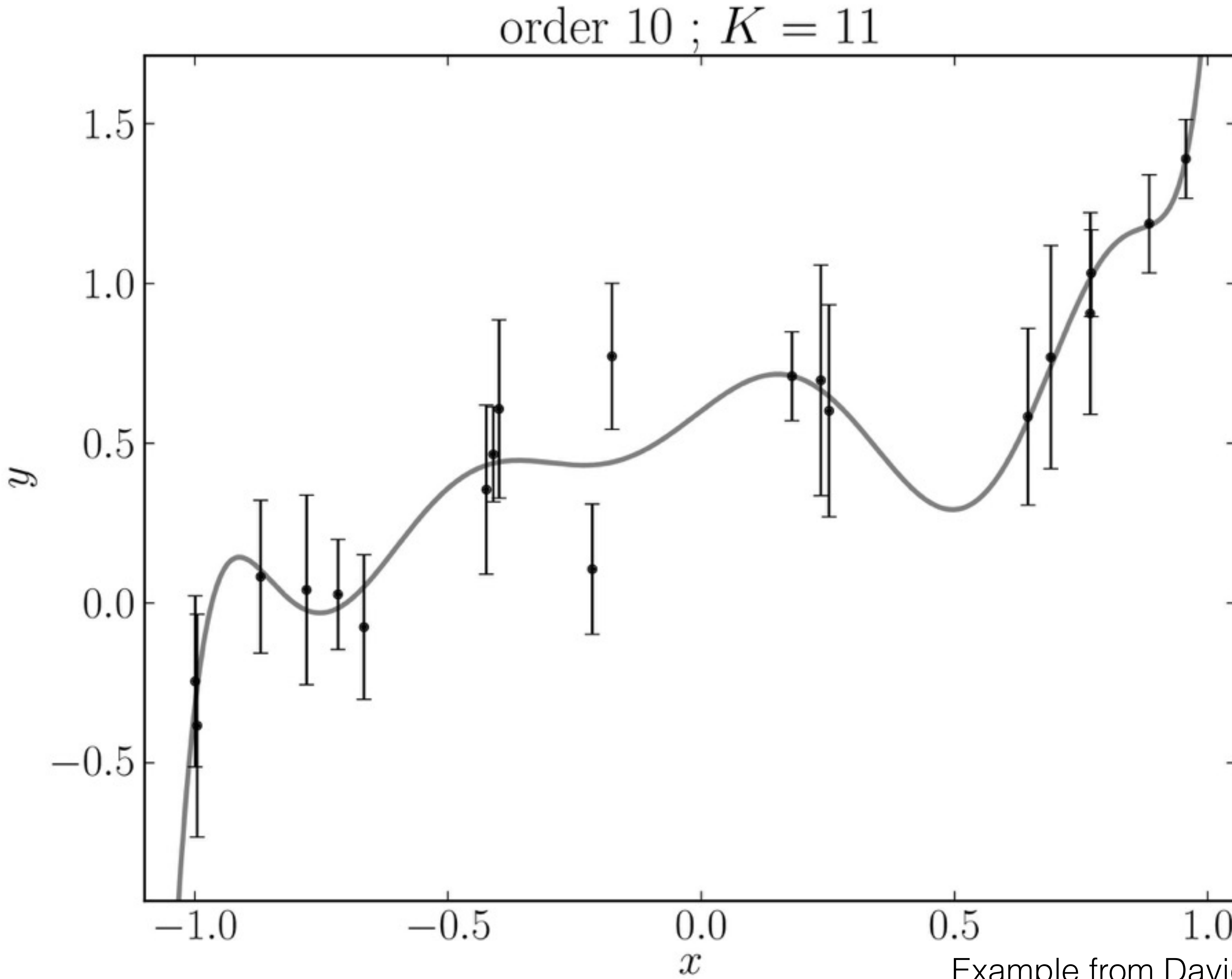
Example from David Hogg



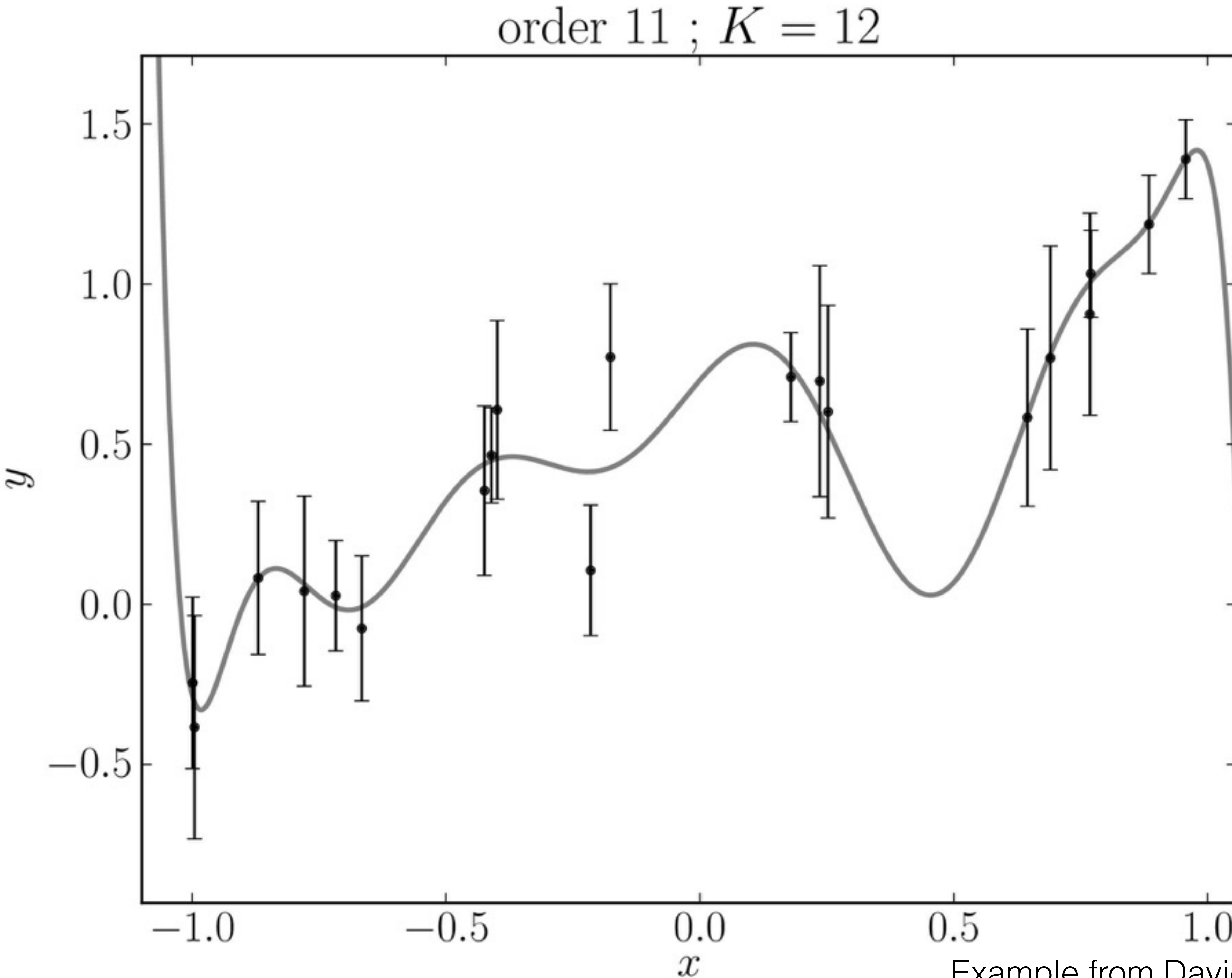
Example from David Hogg



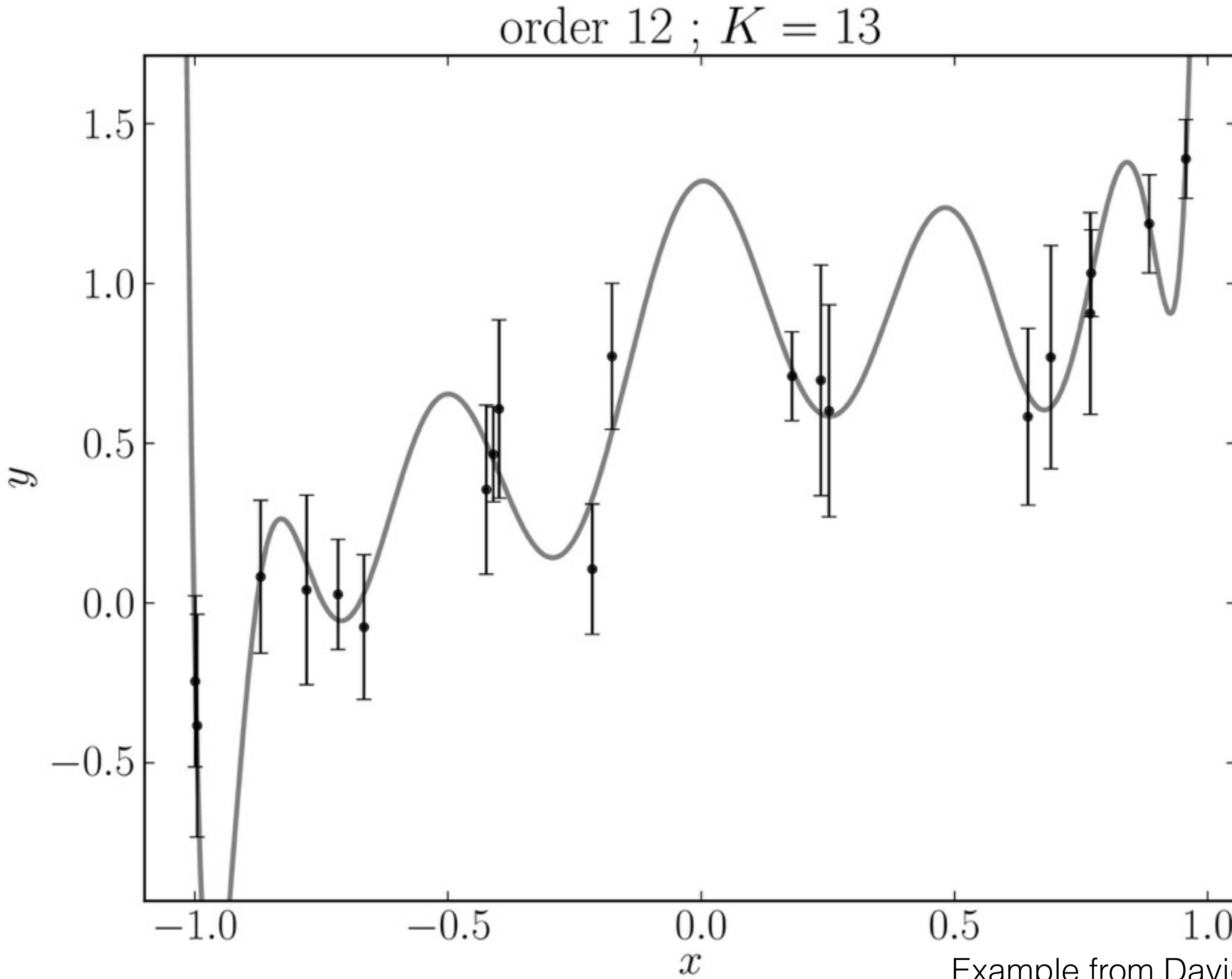
Example from David Hogg



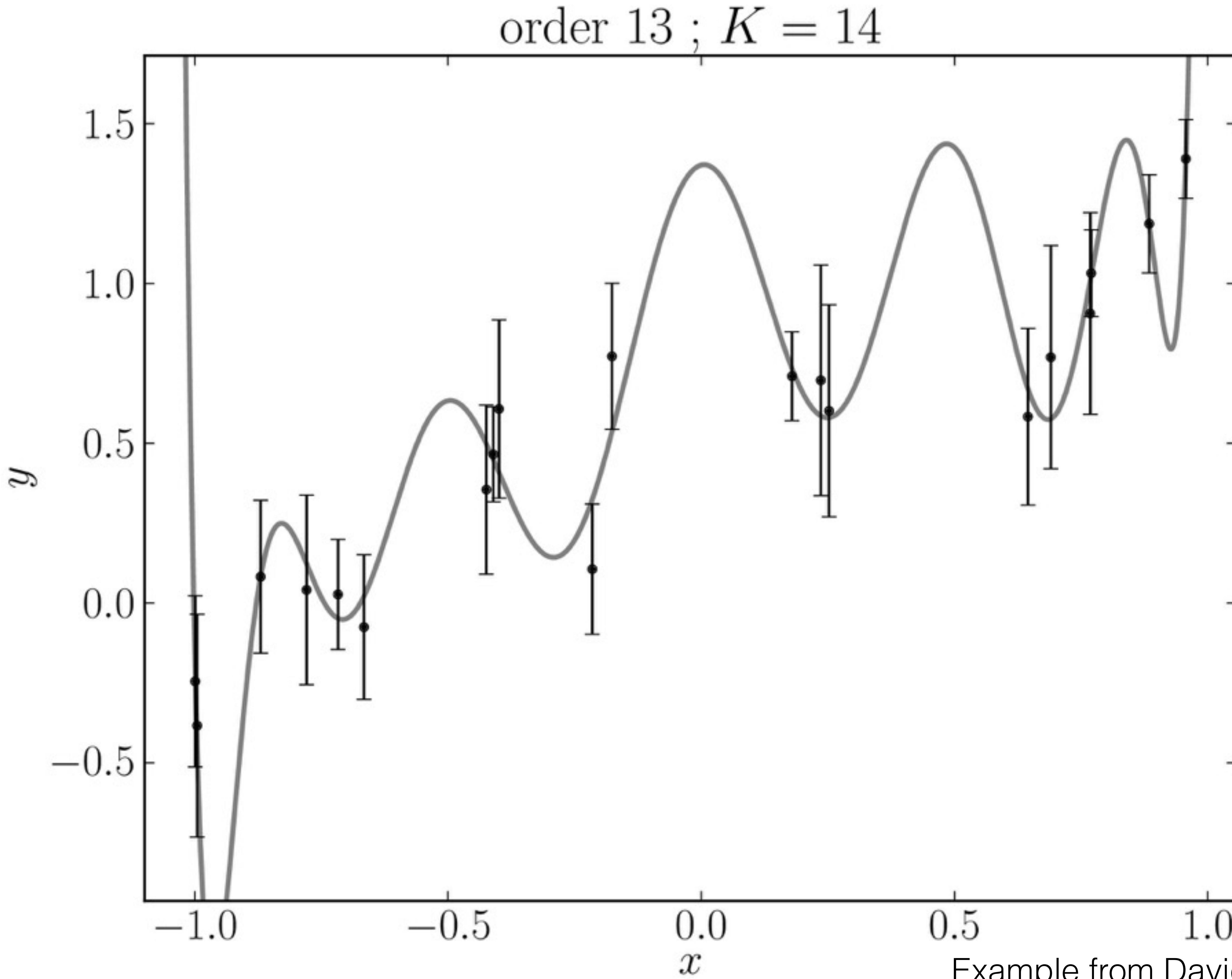
Example from David Hogg



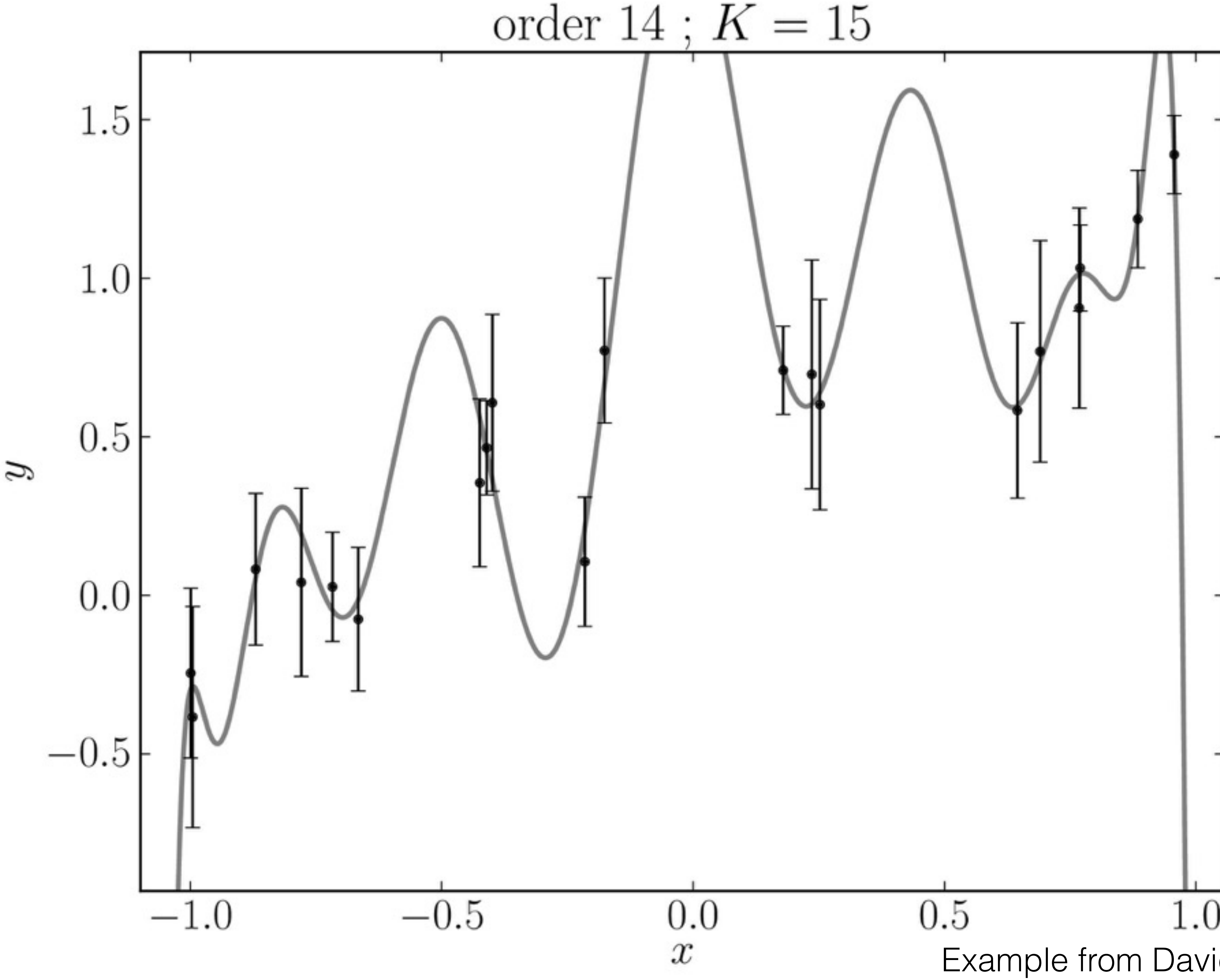
Example from David Hogg



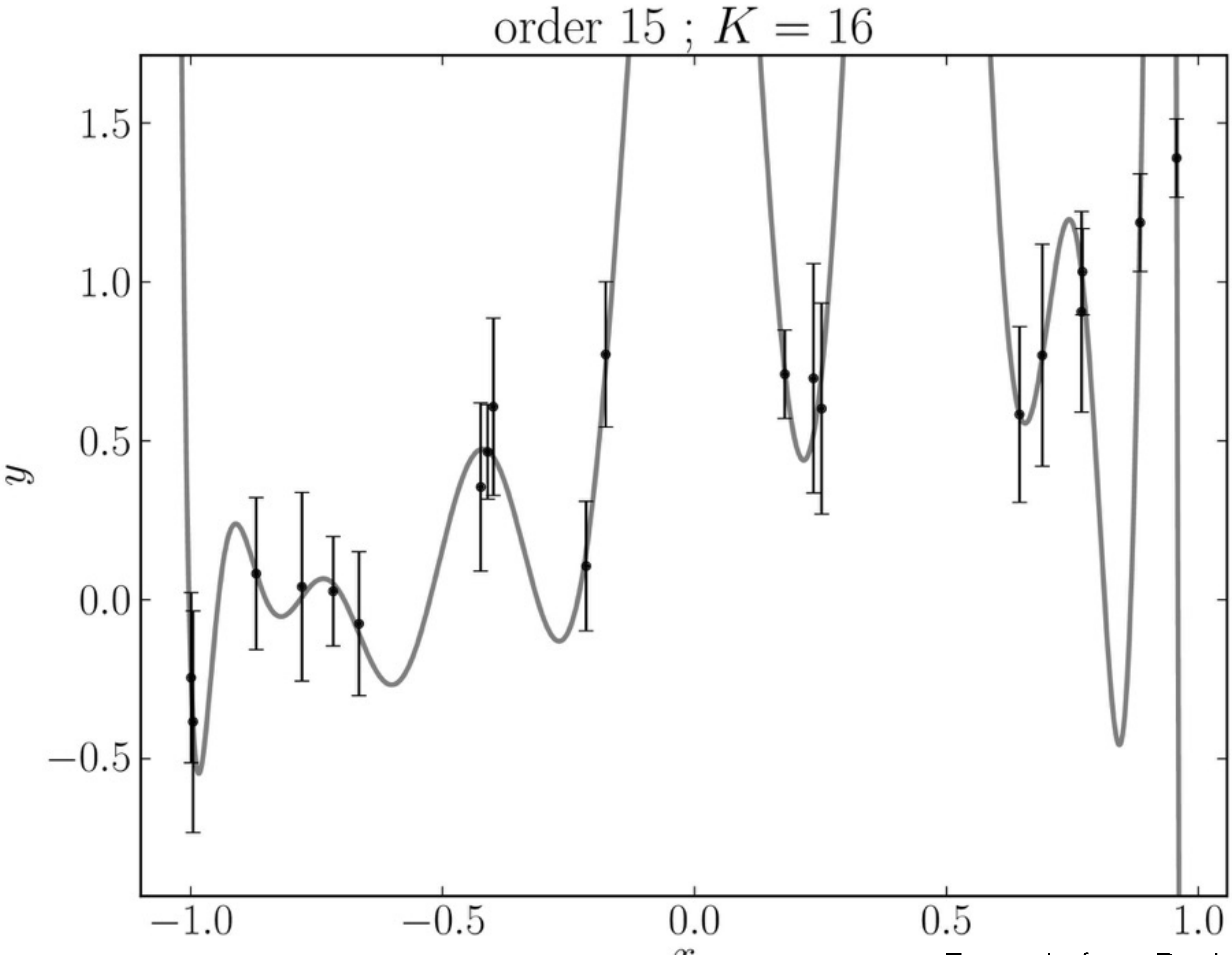
Example from David Hogg



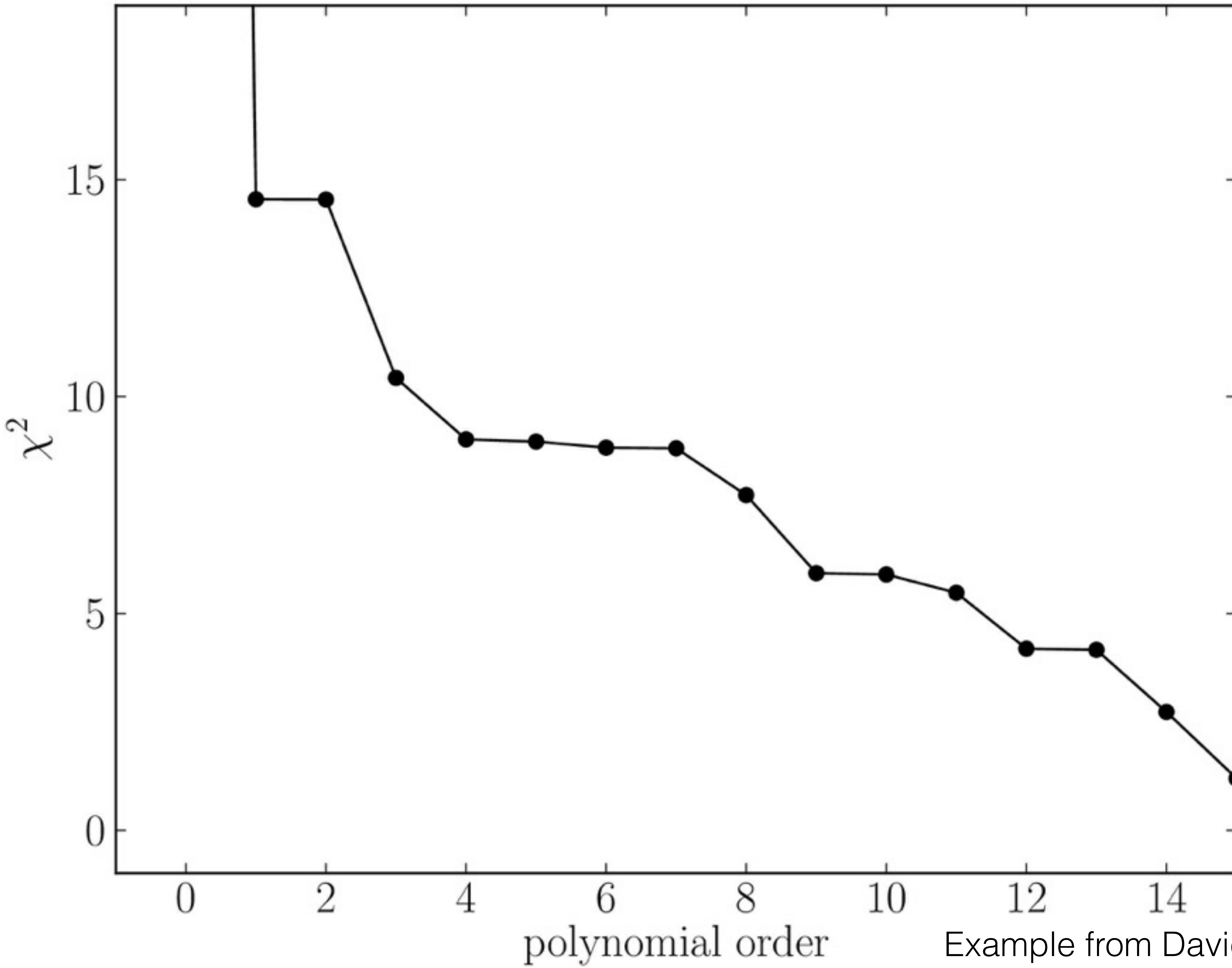
Example from David Hogg



Example from David Hogg



Example from David Hogg



Example from David Hogg

How do we decide which model is the best model?

- Chi-squared keeps improving as we increase the model's complexity

- But clear that we are overfitting!

- To determine the best model, need to figure out which model makes the *data the most probably* —> model selection

- Goodness-of-fit: is the data likely given the model?

Goodness-of-fit: General approach

- Given the model, simulate what the data would be like —> simulated data

- Compare this to the actual data that you have

- If the simulated data is very different from the actual data —> not a good fit!

- From simulated data, can reject that the data was generated by the model at X% confidence —> frequentist method at heart

Comparing simulated and actual data

- Large number of data summaries to chose from!
- Can look at plots, but for automated analysis require some low-dimensional summary
- Most popular: $\sum_i \chi_i^2$, $\sum_i |\chi_i|$, or look at distribution of χ_i

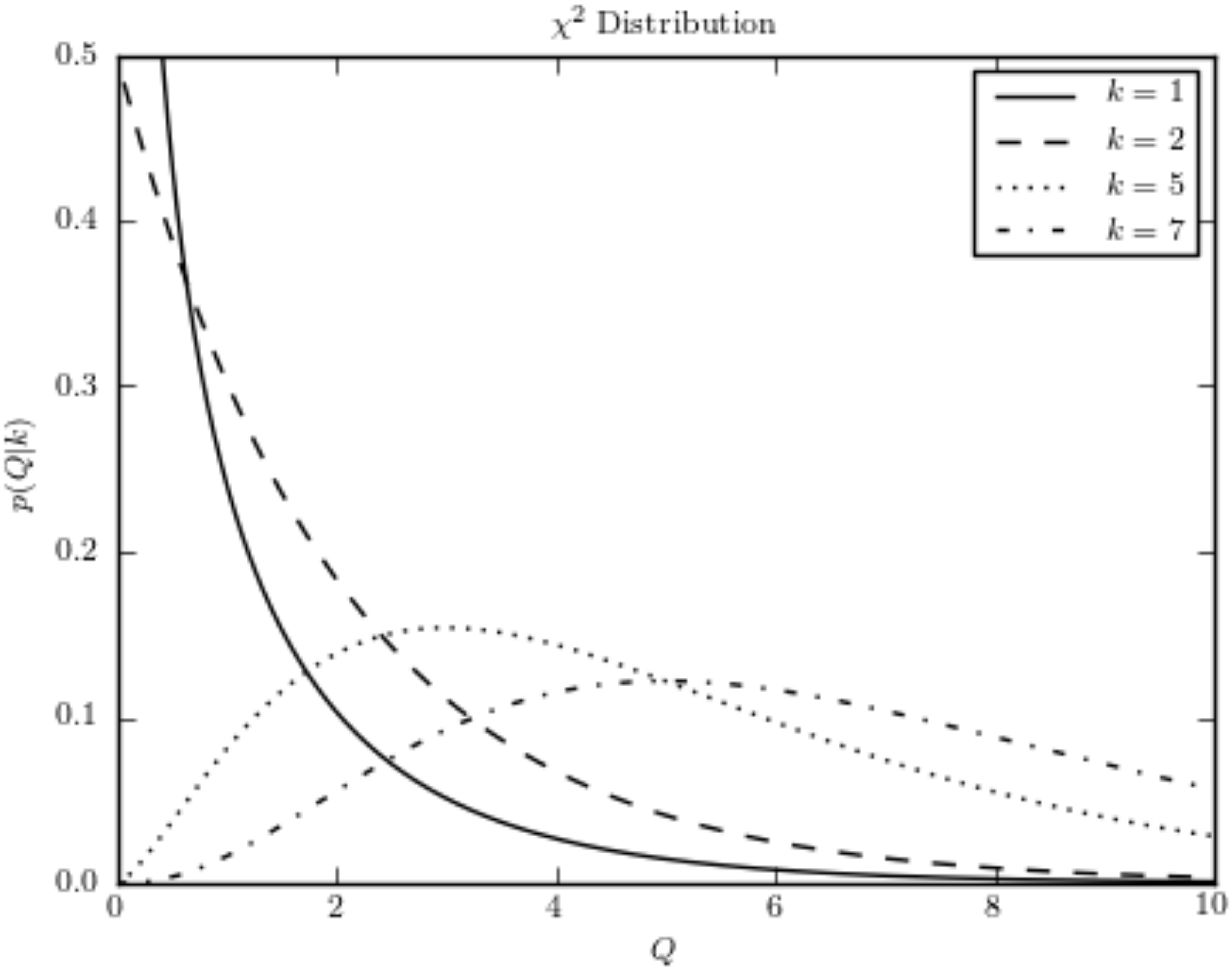
Chi-squared/ degree-of-freedom

- Most popular approach to goodness-of-fit uses chi-squared divided by the number of degrees of freedom
- Chi-squared here = $\sum_i \chi_i^2$
- Number of degrees of freedom = # data points - # of fit parameters
- Where does this come from? When does it apply?

Chi-squared distribution

- Distribution of sum of squares of *k* independent standard normal variables (those from *N*(x|0,1))
- Form: $p(x|k) = \frac{1}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)}\,x^{k/2-1}\,e^{-x/2}$
- Mean: *k*
- Variance: 2*k*
- Central limit theorem: for $k \to \infty$, p(x|*k*) → *N*(x|*k*,2*k*)

Chi-squared distribution



Ivezic et al. (2014)

Chi-squared/ degree-of-freedom

- *If* the likelihood is Gaussian: e.g.,

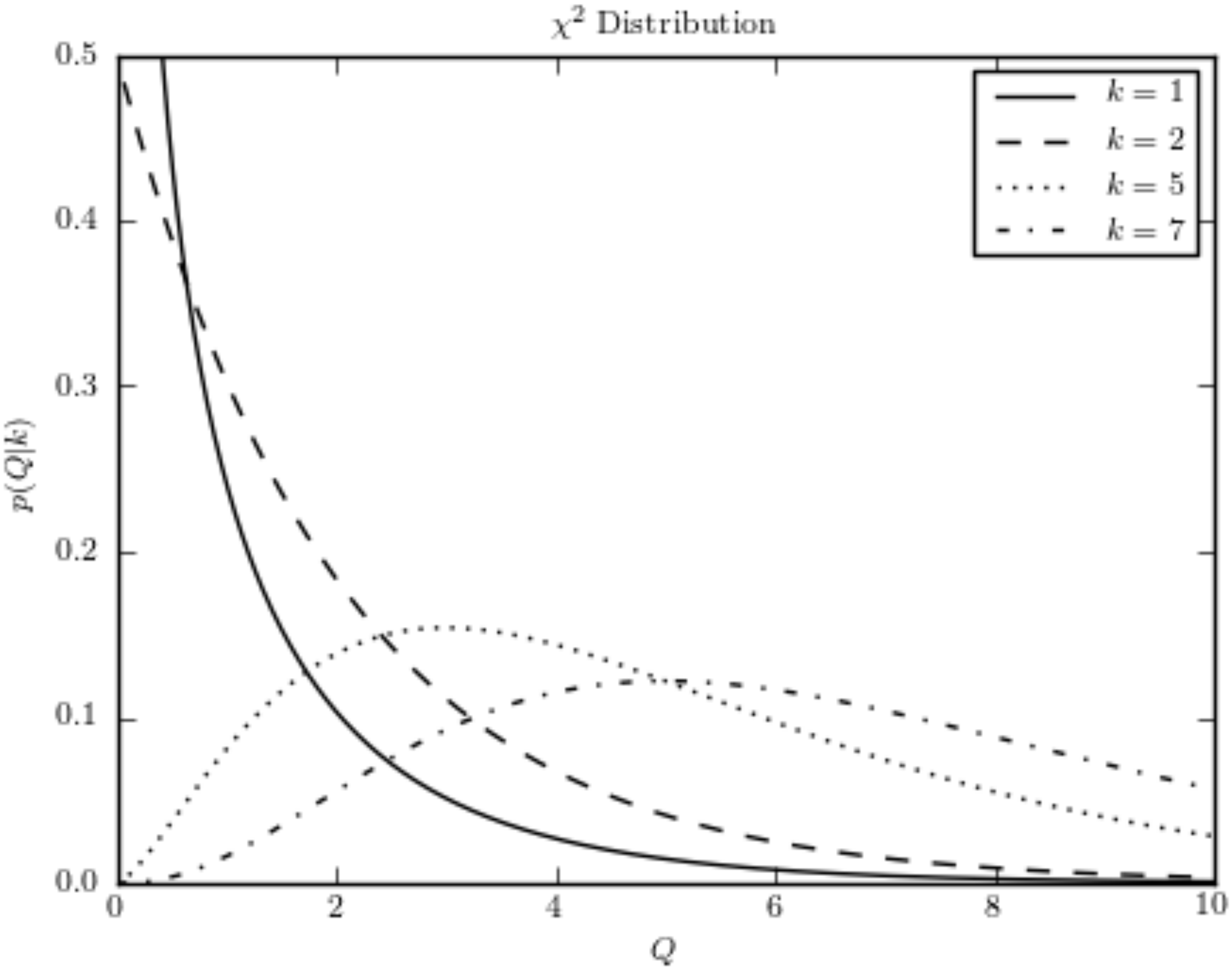
$$L = p(y_{obs} \mid \text{model}, x, \sigma) = p(y_{obs} \mid m, b, x, \sigma)$$
$$= p(y_{obs} \mid y_{true} = mx + b, \sigma)$$
$$= N(y_{obs} \mid y_{true} = mx + b, \sigma^2)$$

- Then we have: $-2 \ln L = \Sigma_i \chi_i^2$, *e.g.*, $= \Sigma_i [(y_i - mx_i - b)/\sigma_{y,i}]^2$
- *and* $\chi_i \sim N(0,1)$
- *Therefore*, $\Sigma_i \chi_i^2$ is distributed as a chi-squared distribution
- If we have fit *K* parameters to *N* data, only *N-K* of these χ_i^2 are independent —> $\Sigma_i \chi_i^2$ is chi-squared distributed with *N-K* degrees of freedom

Chi-squared distribution with *N-K* degrees of freedom

- Mean = *N-K* = dof
- Variance = 2(*N-K*) = 2 dof
- Central limit theorem:
  for *N-K* → ∞, p(x|*N-K*) → *N*(x| *N-K*,2[*N-K*])
- Therefore, expected value of χ^2/dof ~ 1
- But really should be comparing χ^2 to dof with typical scatter √[2dof]

Chi-squared distribution



Ivezic et al. (2014)

Chi-squared/ degree-of-freedom

- Assumptions one more time!
- Likelihood is Gaussian —> for Gaussian uncertainties means that the model must be *linear in the parameters* (e.g., polynomial)
- Must *believe* the uncertainties
- #dof must be large —> large data limit
- Almost never directly applies in practice! But for well-constrained parameters, any model space approx. linear near the best-fit —> widespread use of χ^2/dof
- Never just report χ^2/dof! Also report dof so variance can be determined!

Chi-squared/ degree-of-freedom for non-linear / non-Gaussian models

- If the likelihood (data uncertainty) is not Gaussian or the model is not linear, χ^2/dof does not technically apply (except in the limit discussed on the previous slide)

- So could just directly simulate the data (e.g., linear-fit):
  1. For best fit model parameters (m,b)
  2. Simulate data: y = mx +b
  3. Draw random uncertainties and add them to y
  4. Compute χ^2
  5. Repeat to form p(χ^2)
  6. Where does χ^2 for the actual data lie in this distribution?

- Similar with other summaries (don't need to use χ^2)

Model selection using chi-squared/ degree-of-freedom

- χ^2/dof can be used to select the best model among *linear* models

- Idea is that when overfitting, χ^2 will be suspiciously close to zero

- When underfitting, χ^2 will be large

- Best model makes the data most likely —> peak of χ^2 distribution —> χ^2/dof ~ 1

Detour: difference between Delta χ^2 and χ^2/dof

- χ^2 comes up in model fitting and goodness-of-fit
- -2 ln L = χ^2 = $\sum_i \chi_i^2$
- Find best fit, can compute $\Delta\chi^2 = \chi^2 - \chi^2_{\min}$
- With uniform prior: ln p(m,b|data) = $-\Delta\chi^2/2$
- Again, chi-squared distribution, but now with *K* degrees of freedom

Detour: difference between Delta χ^2 and χ^2/dof

- Suppose you have 1 parameter (e.g., fit constant y =b rather than y = m x + b)
- p(b|data) = $\mathrm{Chi}^2(\Delta \mathrm{chi}^2, 1\ \mathrm{dof})$
- 68% confidence limit on b: that value for which $\Delta\chi^2 = 1$

```
In [7]: from scipy import stats

In [8]: stats.chi2.ppf(0.68,1)
Out[8]: 0.9889464814780228
```

Detour: difference between Delta χ^2 and χ^2/dof

- Suppose you have 2 parameters (e.g., linear fit y = m x + b)
- p(m,b|data) = Chi^2(Δchi^2,2 dof)
- 68% confidence limit on (m,b): that ellipse for which Δchi^2 = 2.3

```
In [11]: from scipy import stats

In [12]: stats.chi2.ppf(0.68,2)
Out[12]: 2.278685563767296
```

Detour: difference between Delta χ^2 and χ^2/dof

$\Delta\chi^2$ as a Function of Confidence Level p and Number of Parameters of Interest ν						
	ν					
p	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

Numerical recipes 3rd edition

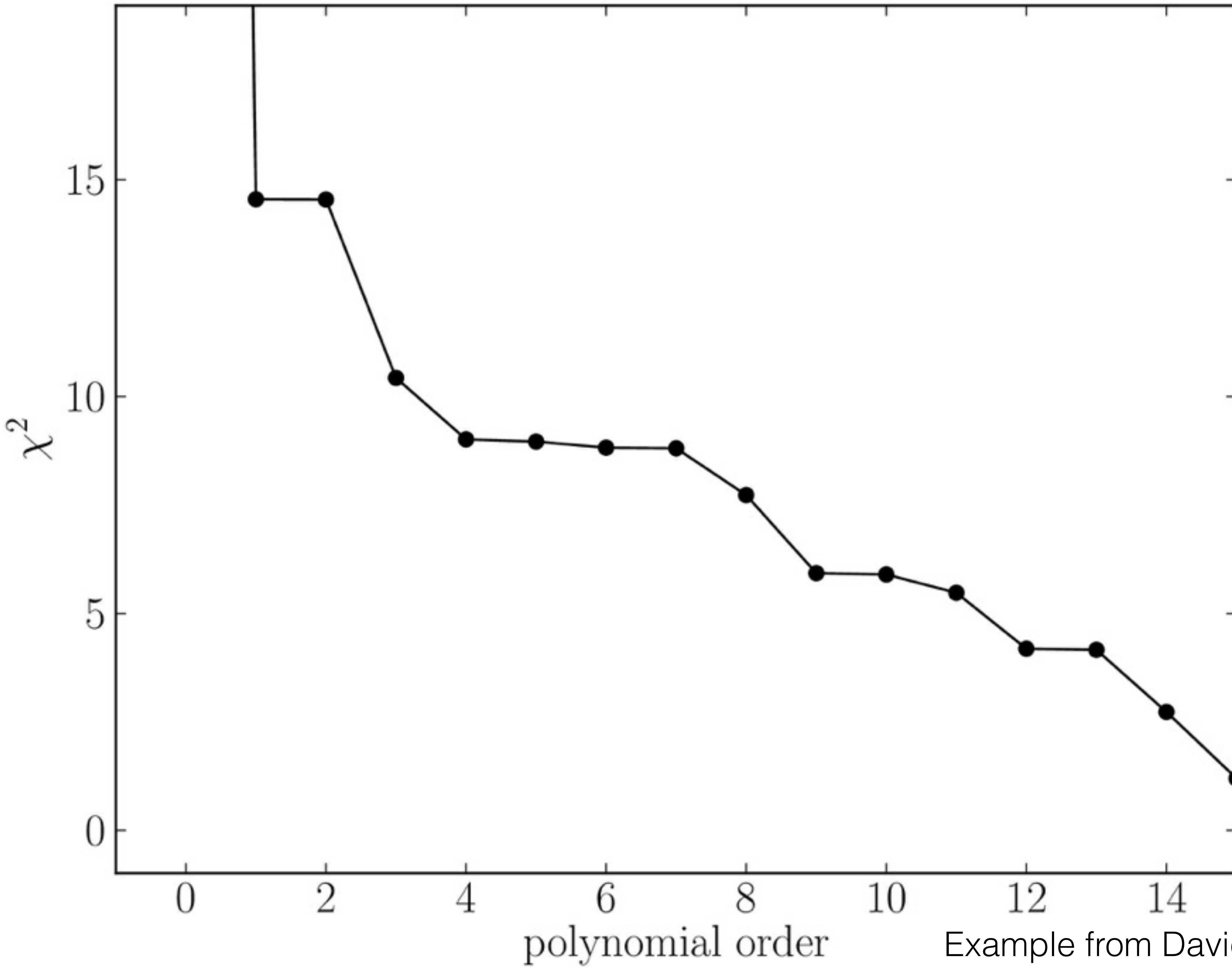
Detour: difference between Delta χ^2 and χ^2/dof

- Delta chi² used for finding uncertainty limits
- ***Don't*** use Delta χ^2 / dof for this! limits will always be much wider
- χ^2 / dof for model selection and goodness-of-fit

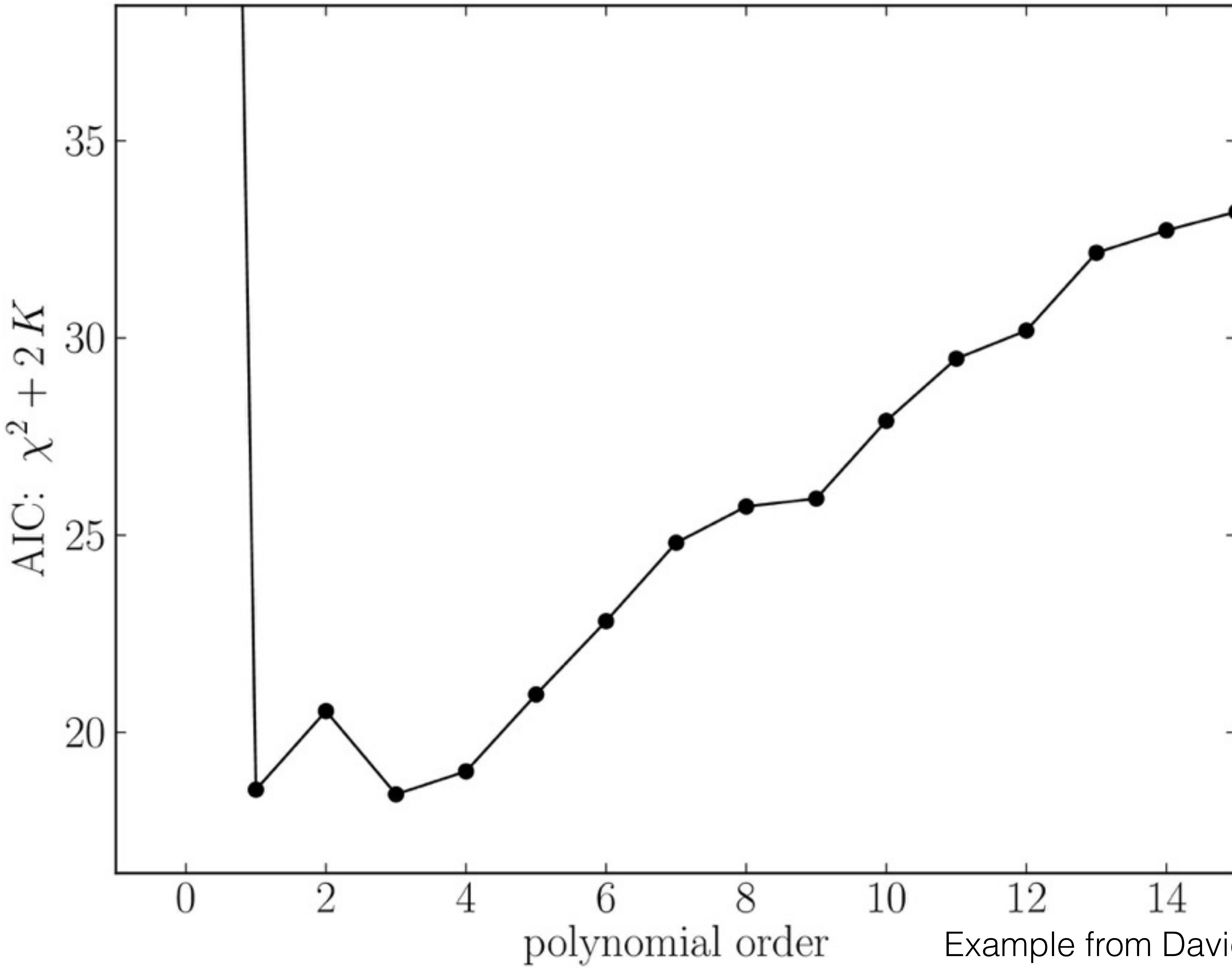
Model selection with AIC

- AIC = Akaike Information Criterion
- AIC = -2 ln L + 2 K = χ^2 + 2 K
- Delta AIC = Asymptotically the amount of information lost when using worse model G1 than better model G2
- Corrections for finite sample sizes depend on model; for 1D, linear, Gaussian model

  2K(K+1)/(N-K-1)



Example from David Hogg



Example from David Hogg

Bayesian model selection

- Bayesian methods are *very bad* at goodness-of-fit
- Application of Bayes's theorem only allows to distinguish between 2 different models, no real concept of 'good model'
- Bayes's theorem for models: model1= linear, 2=quadratic

  p(linear|data) ~ p(data|linear) p(linear)
  p(quadratic|data) ~ p(data|quadratic) p(quadratic)

- Likelihoods in these equations are marginalized over parameters of each model (c = quadratic coeff.) —> marginalized likelihood

  p(data|linear) = ∫dm d b p(data|m,b) p(m,b|linear)
  p(data|quadratic) = ∫dc dm d b p(data|m,b,c) p(m,b,c|quadratic)

  These were in the denominator of Bayes's theorem for p(m,b|linear) before!

Bayesian model selection

- Can then compute *odds ratio*:

$$\text{odds ratio} = \frac{p(\text{linear}|\text{data})}{p(\text{quadratic}|\text{data})}$$

  and select the model with the highest odds

- Requires prior over models [p(linear) and p(quadratic)]

$$\text{odds ratio} = \frac{p(\text{data}|\text{linear})}{p(\text{data}|\text{quadratic})} \times \frac{p(\text{linear})}{p(\text{quadratic})}$$

$$= \text{Bayes-factor} \times \text{prior-ratio}$$

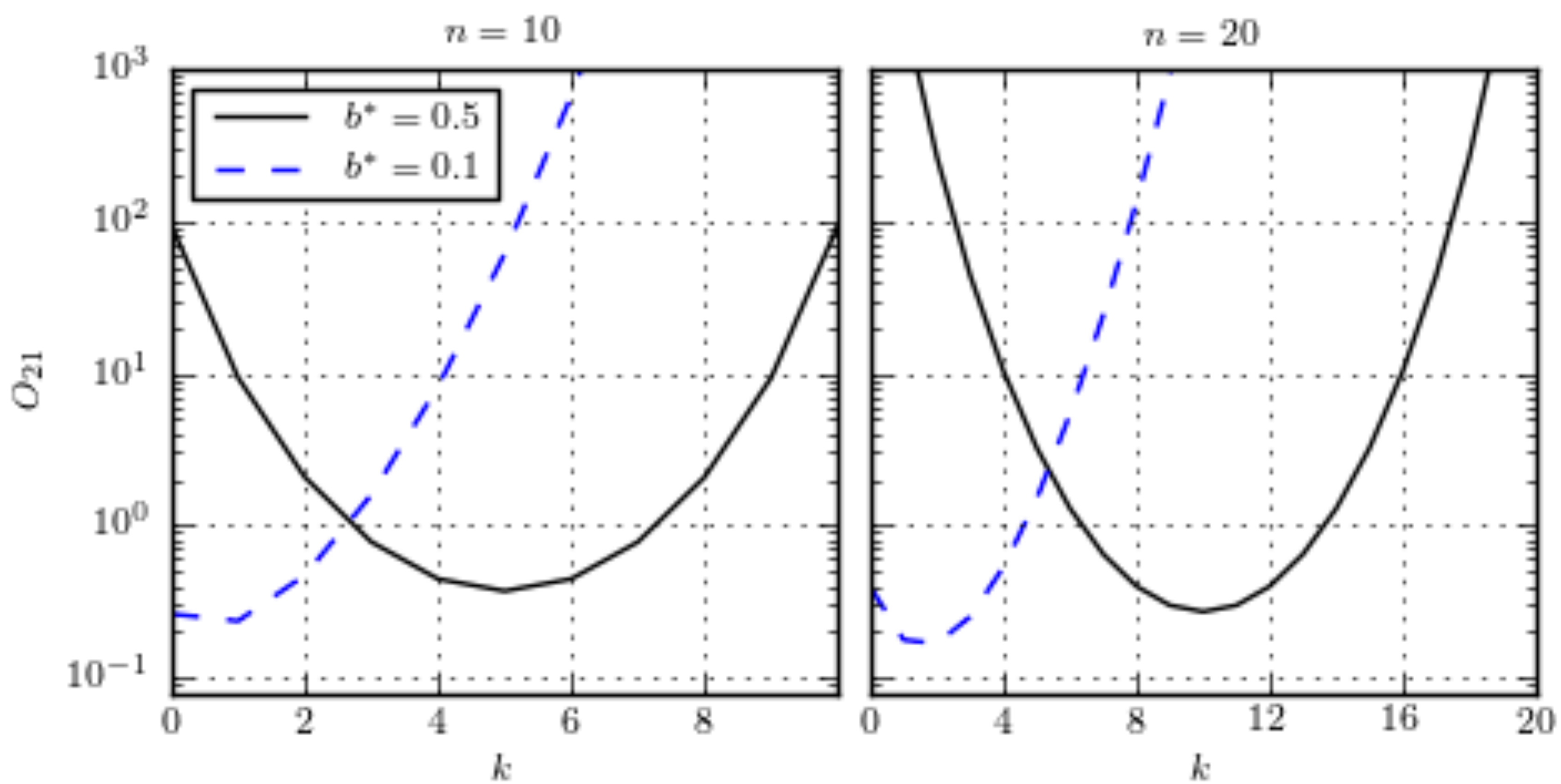
- If not strong preference, select based on Bayes-factor

Jeffreys’ scale

- What odds ratio constitutes strong evidence?
    - > 3 “substantial”
    - > 10 “strong evidence”
    - > 30 “very strong”
    - > 100 “decisive evidence”
    - < 3 “barely worth a mention”

Bayesian model selection: Example: is a coin fair?

- Flip coin n times, get k heads —> is it fair?
- Bayesian needs alternative model! Make that model: constant probability for heads that is unknown
- So two models: p=0.5 and p=unknown



Ivezic et al. (2014)

Bayesian model selection: How to compute the evidence

- Evidence = p(data|model1) = ∫dparams p(data|params) p(params|model1)

- Nested sampling: MCMC technique that returns the evidence

- If the posterior is close to Gaussian: Laplace approximation around max P*(x) with covariance matrix **A**:

$$P^*(\mathbf{x}_0)\sqrt{\frac{(2\pi)^K}{\det \mathbf{A}}}.$$

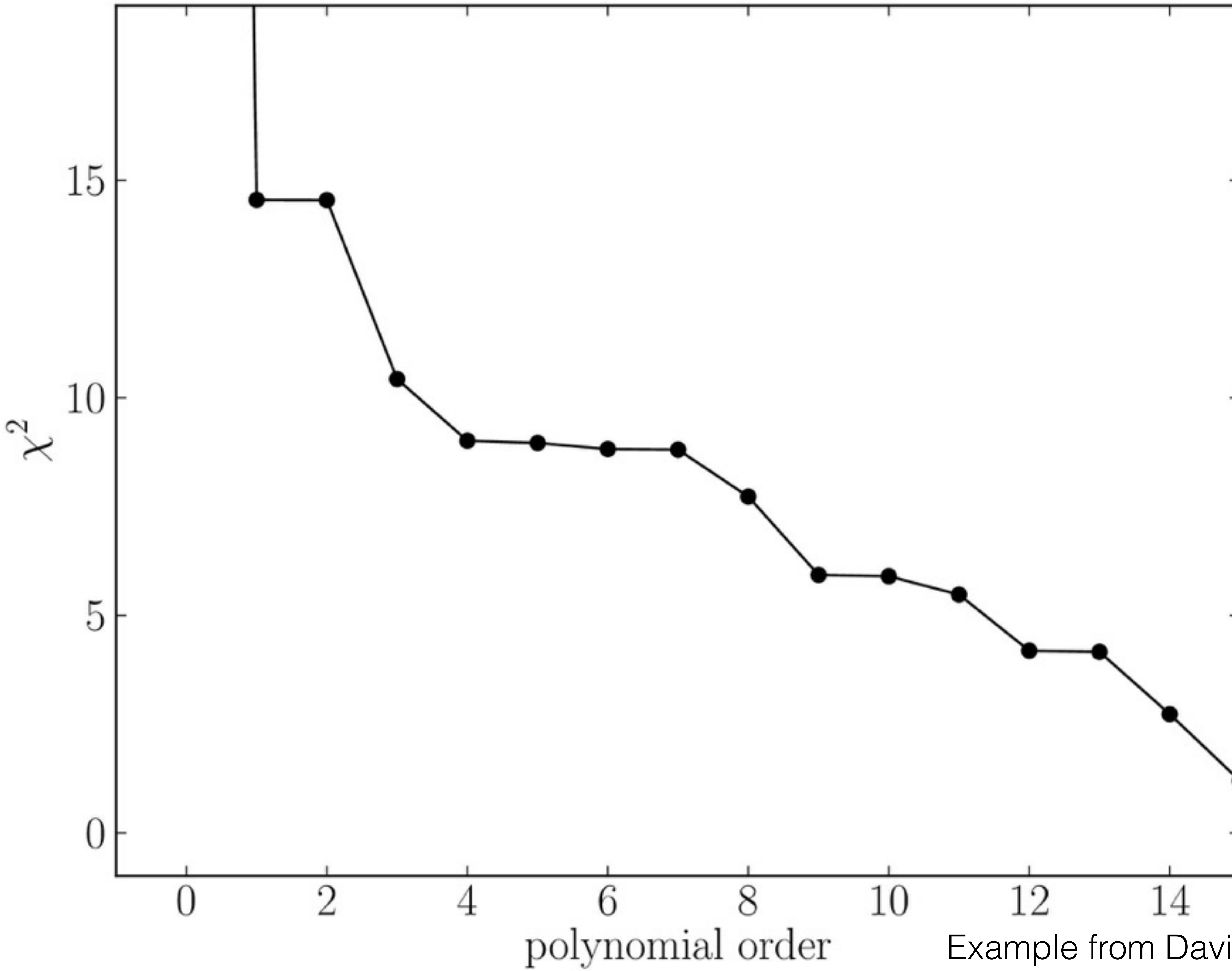
Bayesian Information Criterion (BIC)

- If the posterior is part of the exponential family of PDFs (Gaussian, chi-squared, beta, Bernoulli, …) can approximate

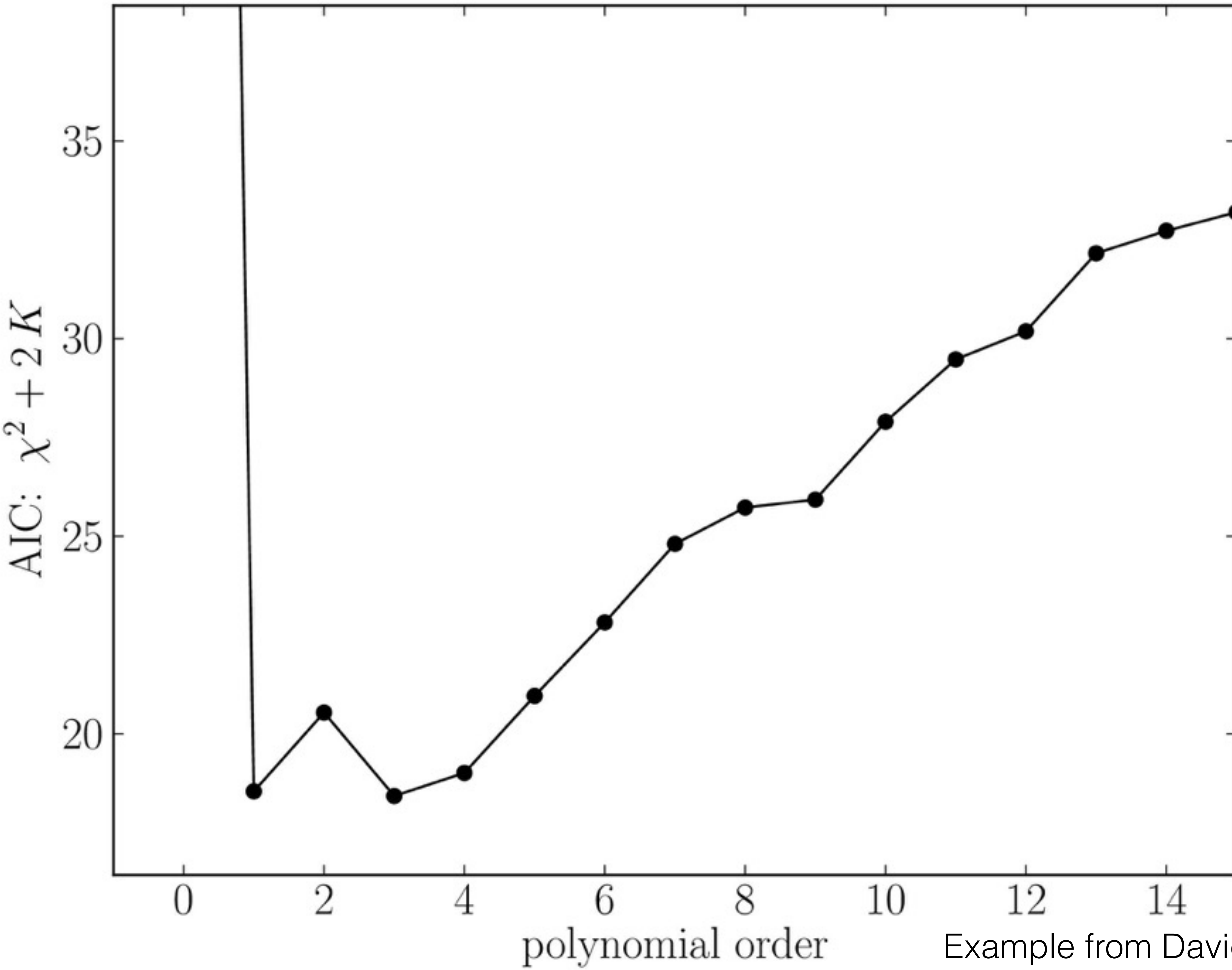
  Evidence ~ BIC = -2 ln L + K ln[N]

  for K parameters and N data points

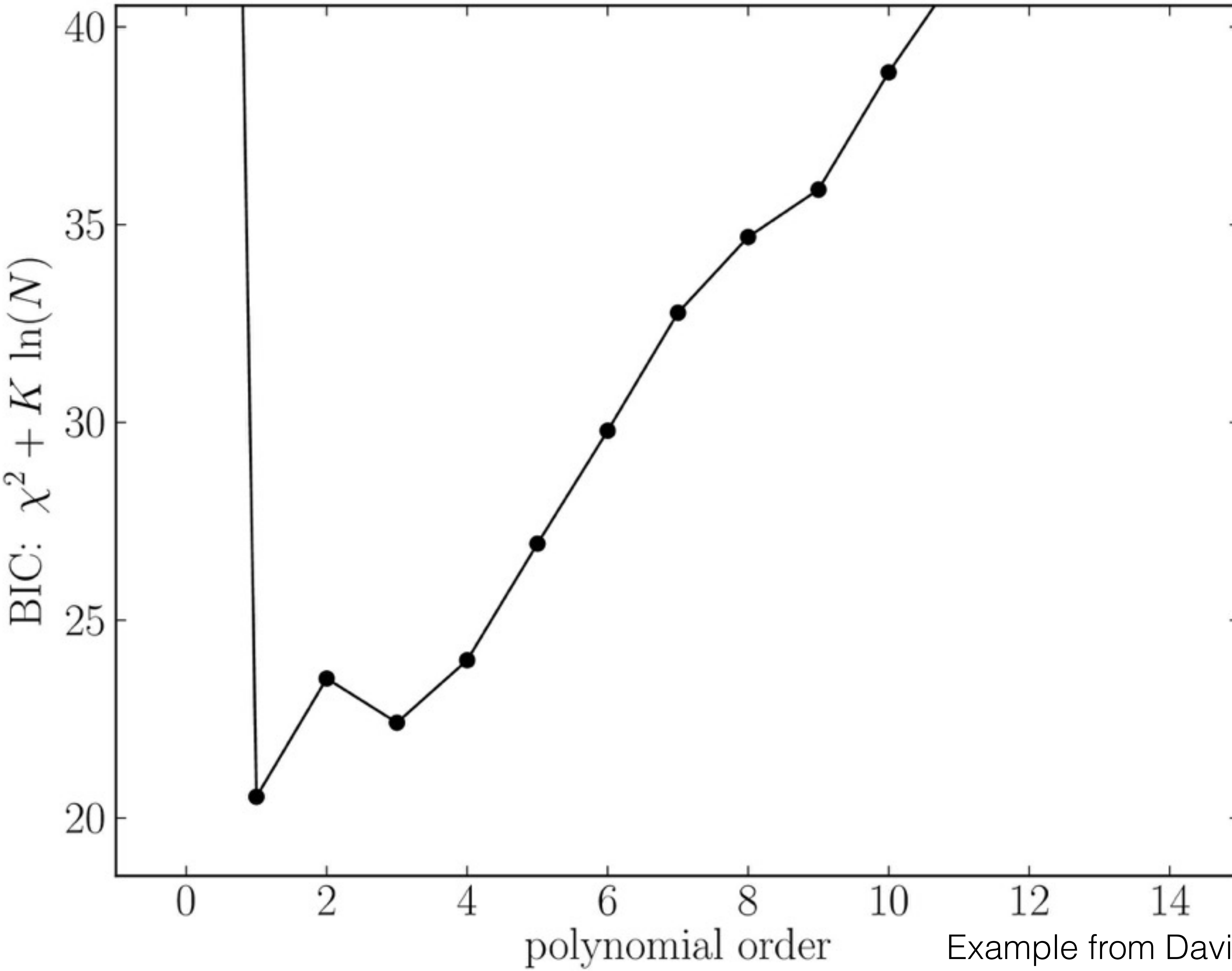
- Similar to AIC, but Bayesian :-)



Example from David Hogg



Example from David Hogg



Example from David Hogg

Cross-validation

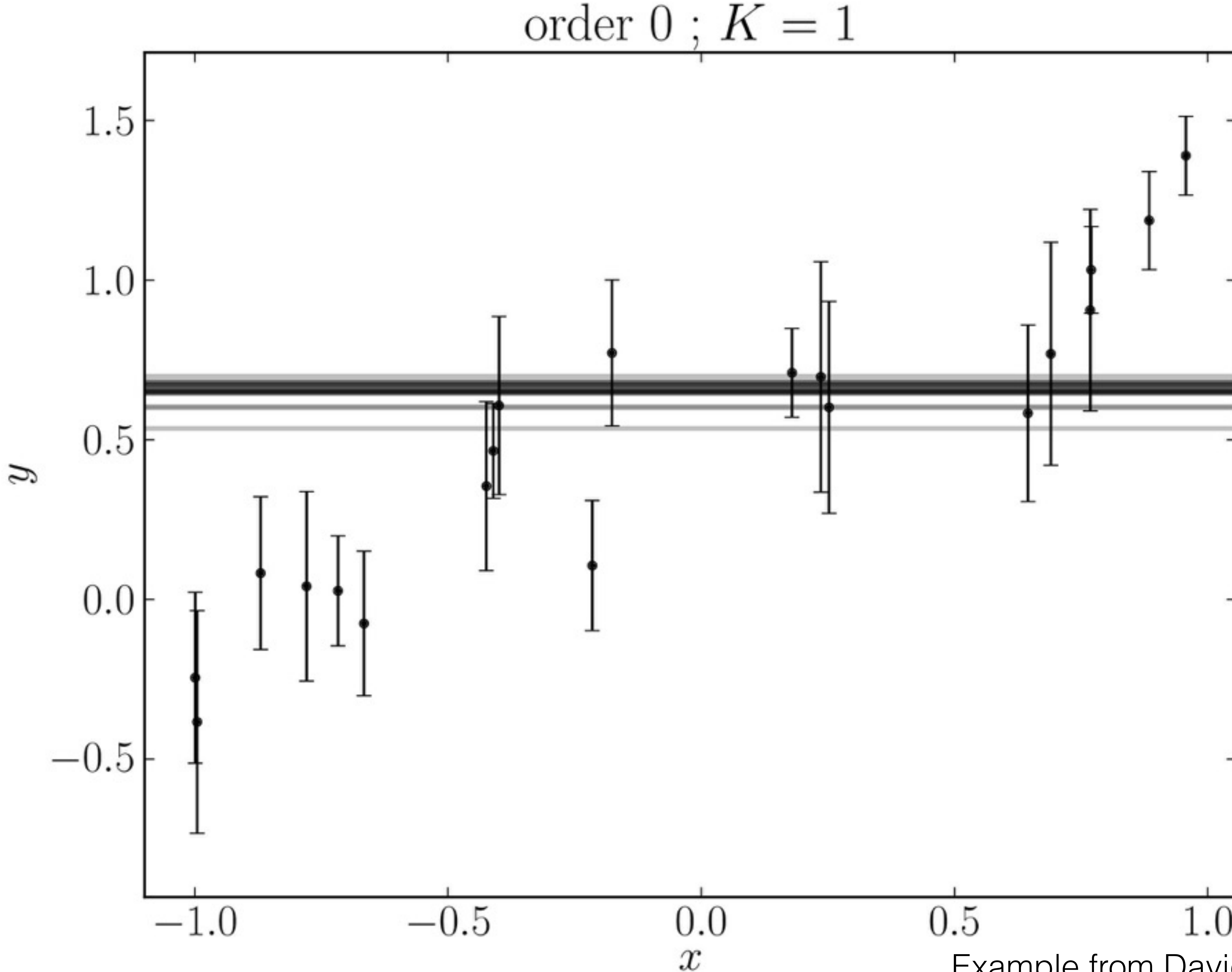
- All previous methods require many assumptions
- Take approach similar to bootstrap and jackknife before, using the data themselves to generate 'new' data
- Cross-validation is very similar to jackknife:

  1. Generate *N* data sets that leave out 1 data point at a time
  $[\{x_1,x_2,x_3,\ldots\}, \{x_0,x_2,x_3,\ldots\}, \{x_0,x_1,x_3,\ldots\}, \ldots]$

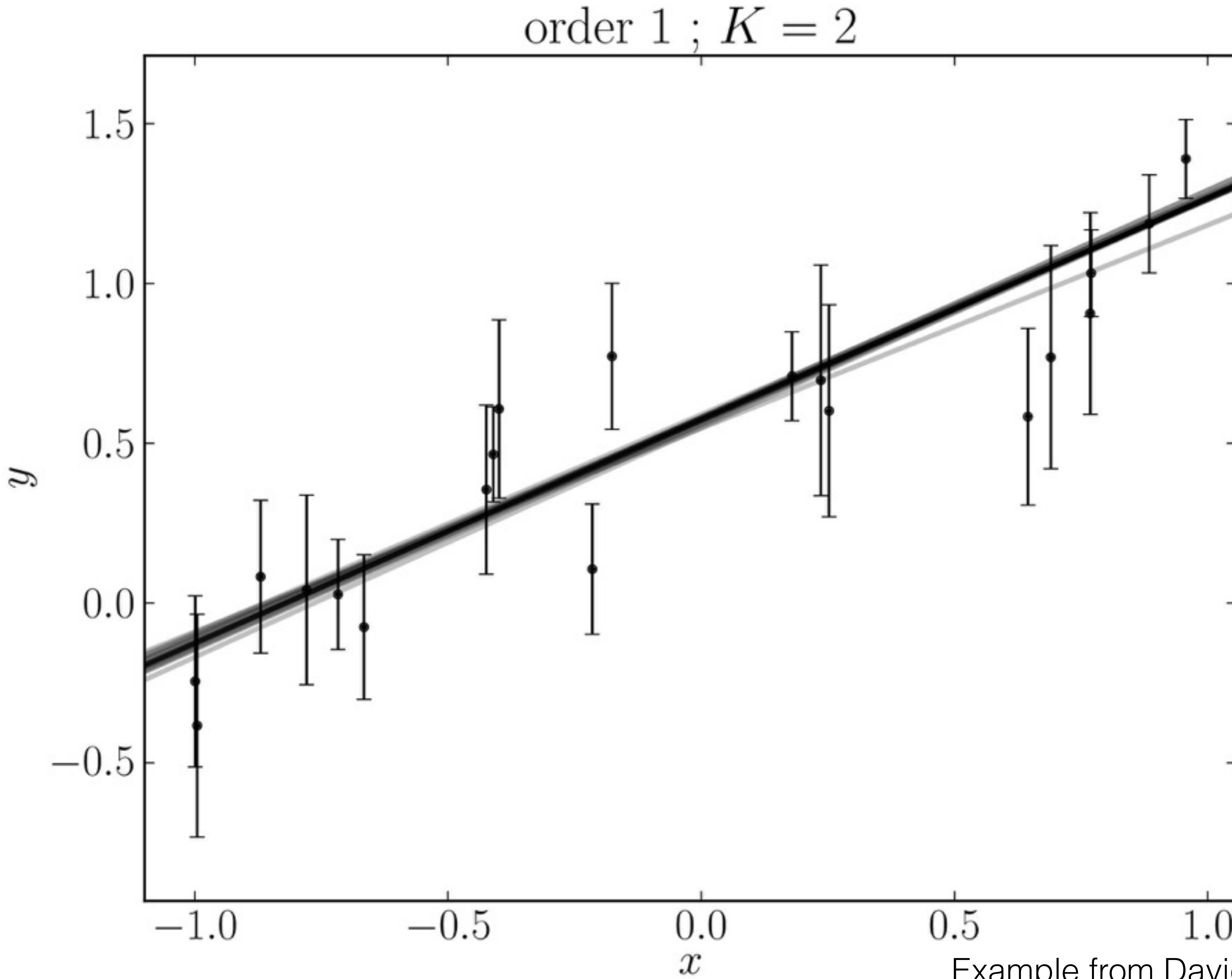
  2. Fit the model to each data set

  3. Compute the likelihood of the data point that was left out: L_i

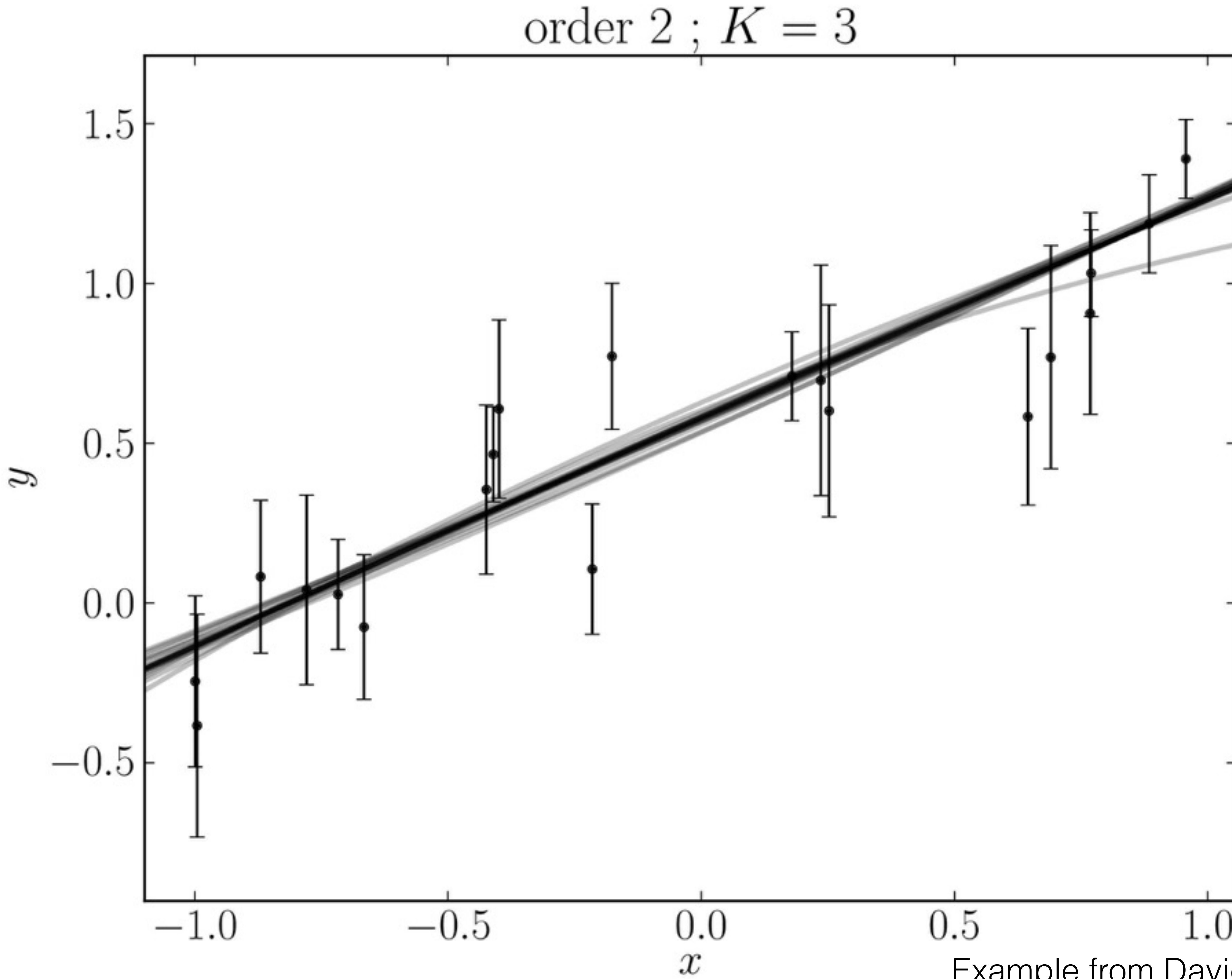
  4. Cross validation likelihood $L_{cval} = Prod_i\ L_i$



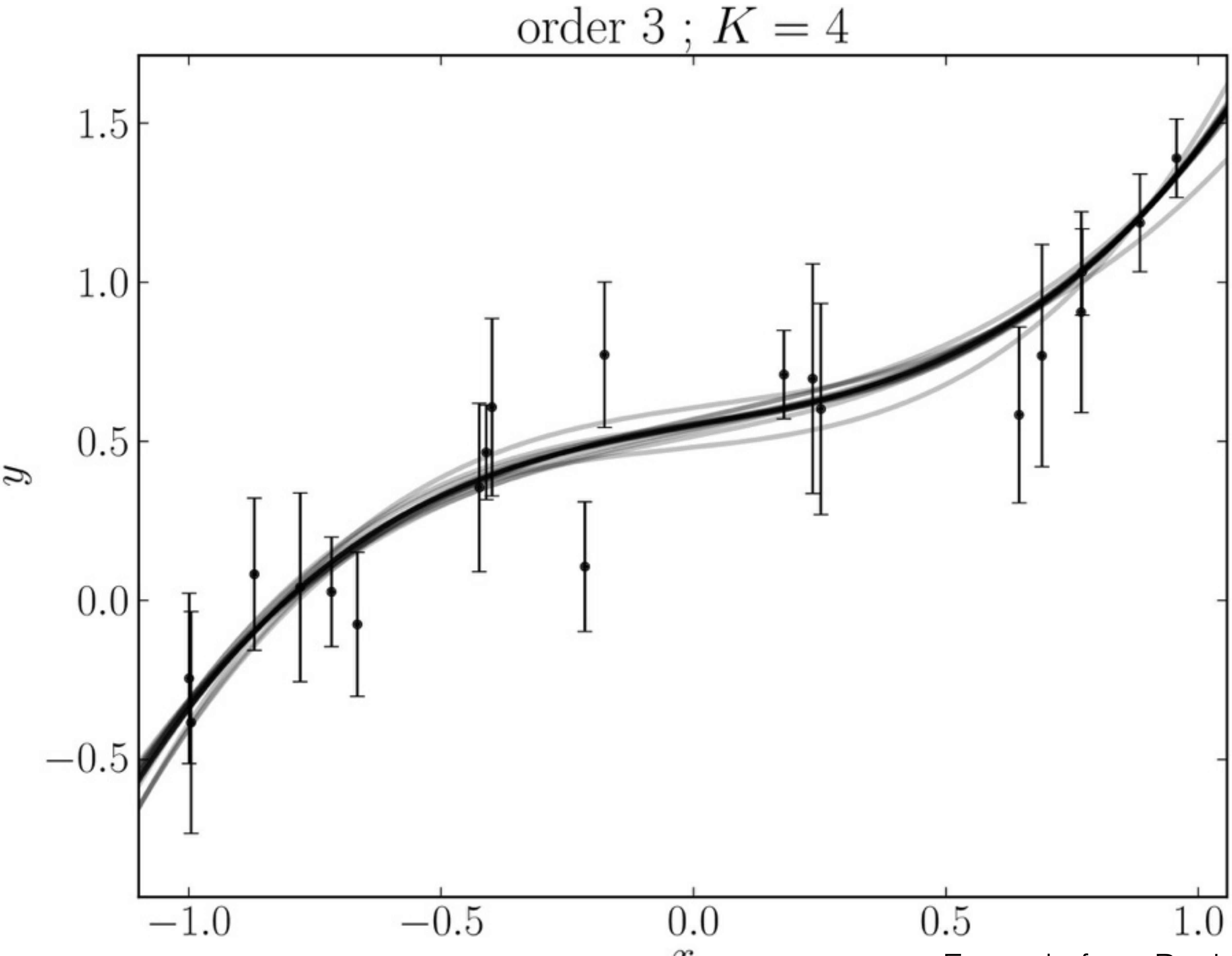
Example from David Hogg



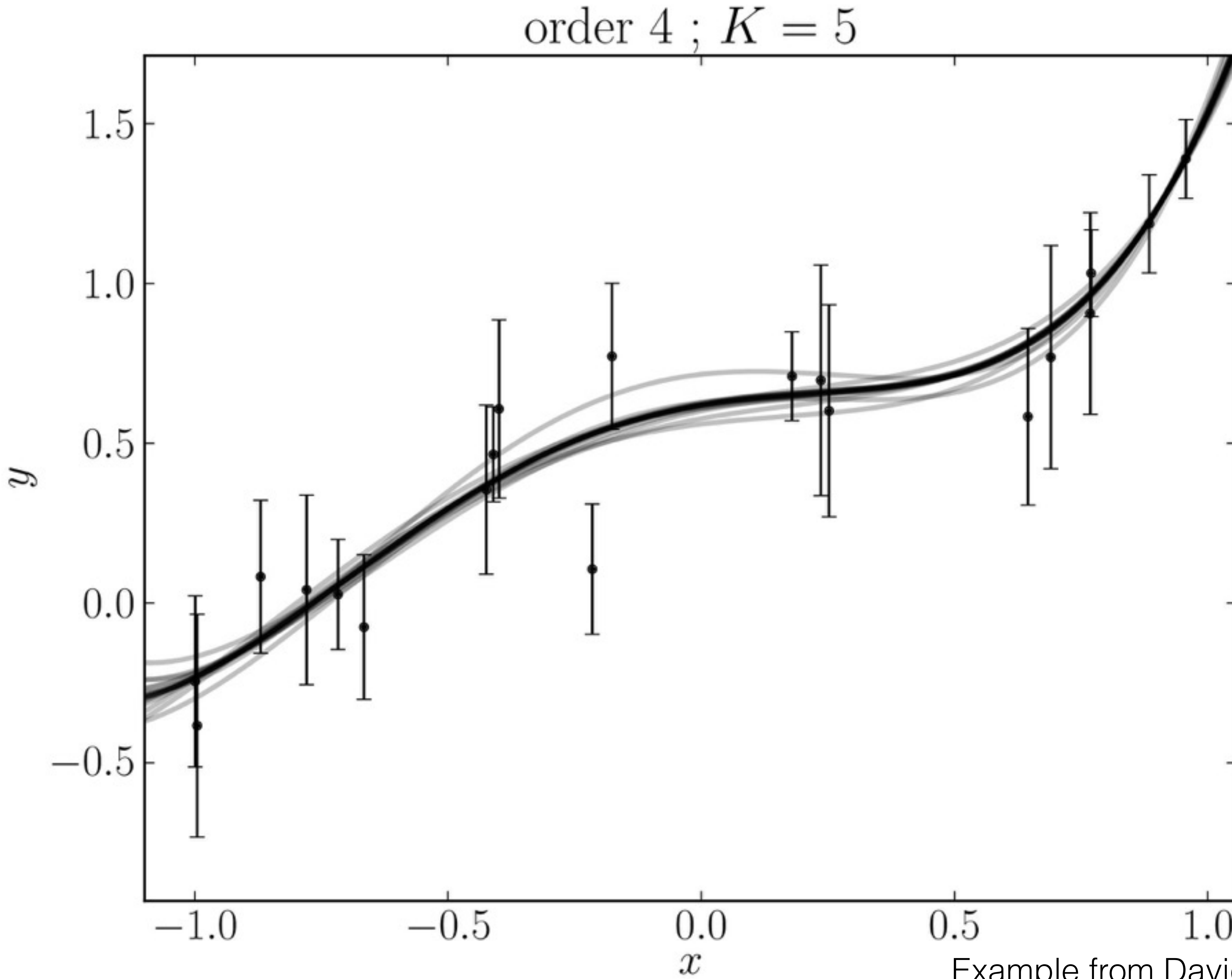
Example from David Hogg



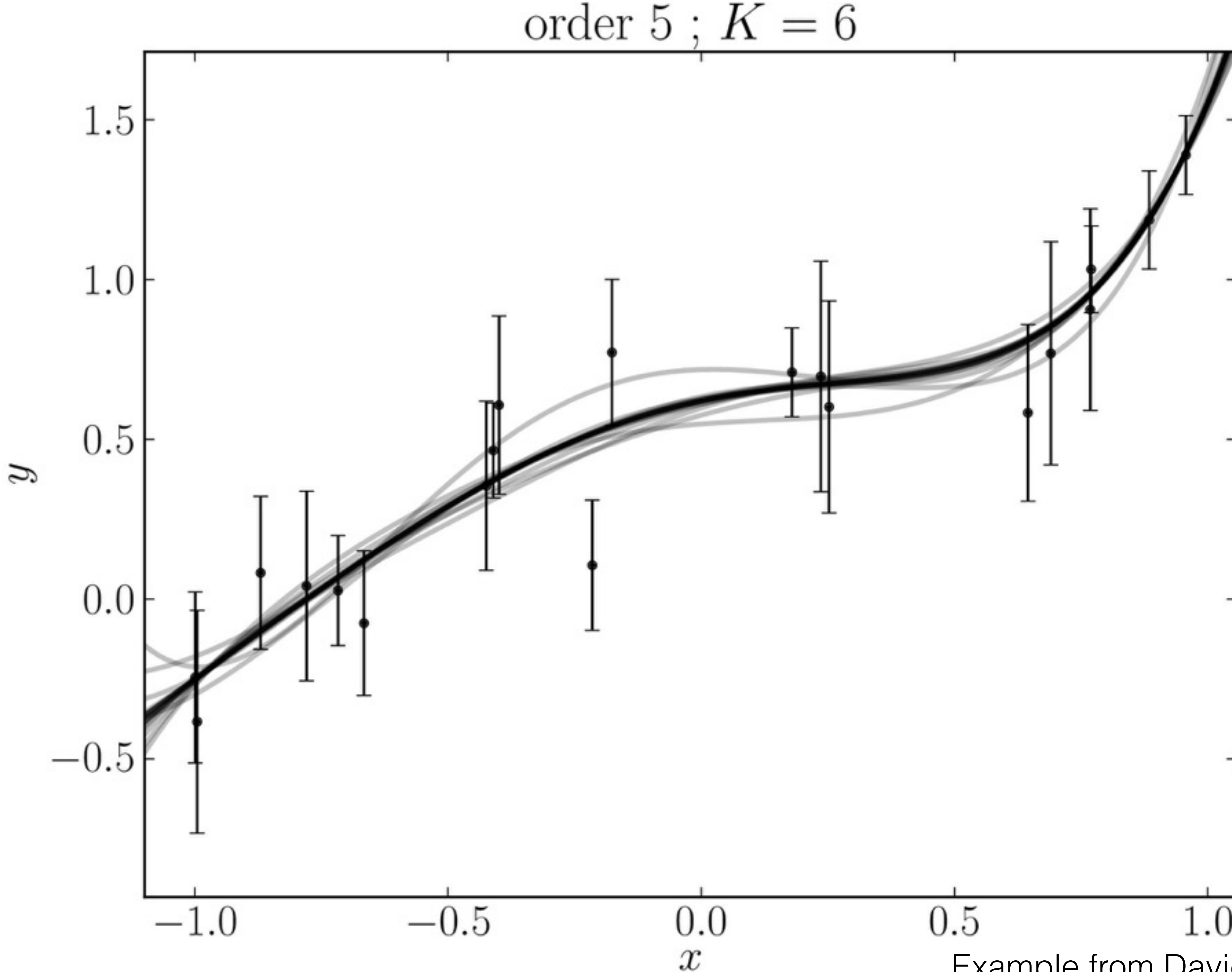
Example from David Hogg



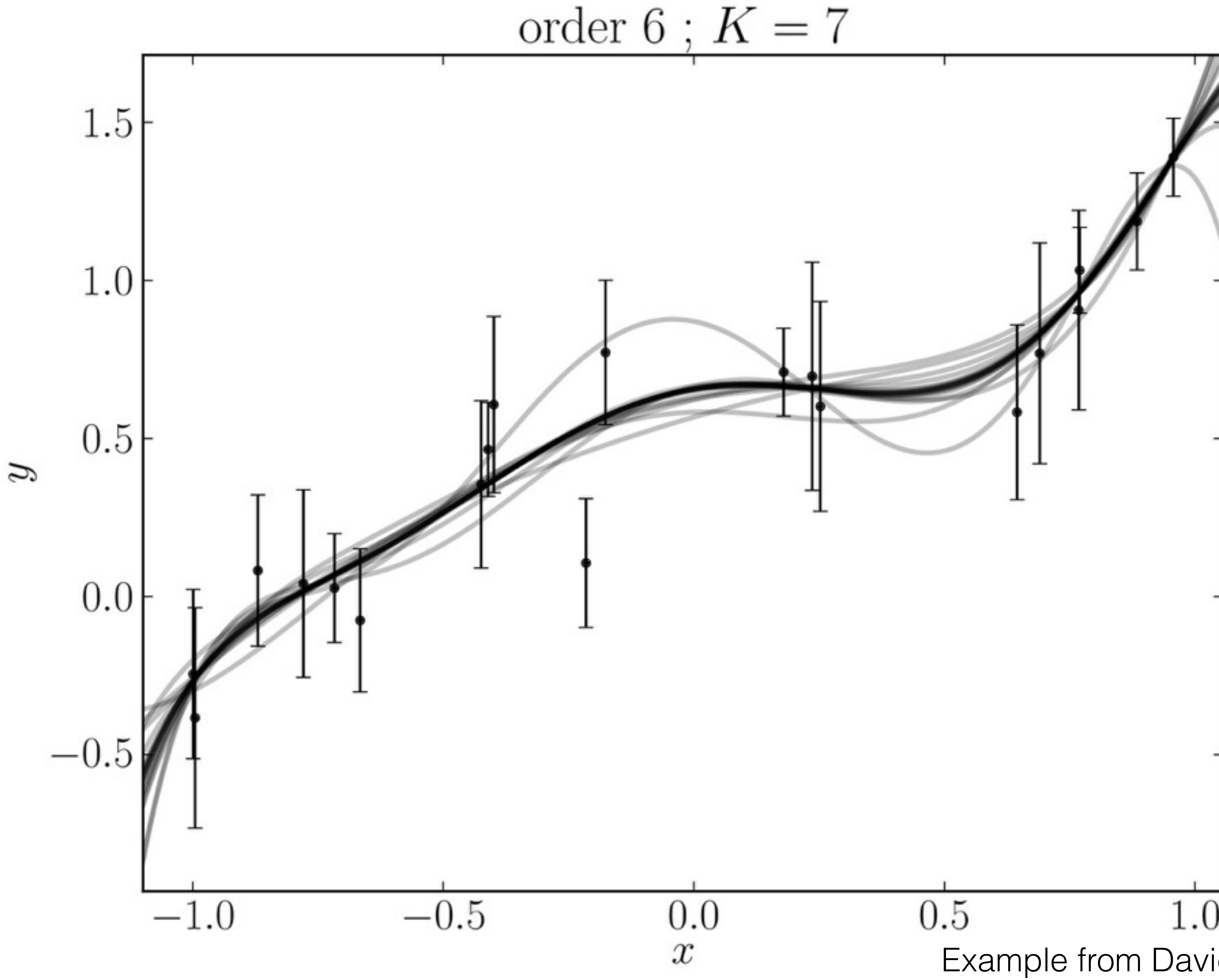
Example from David Hogg



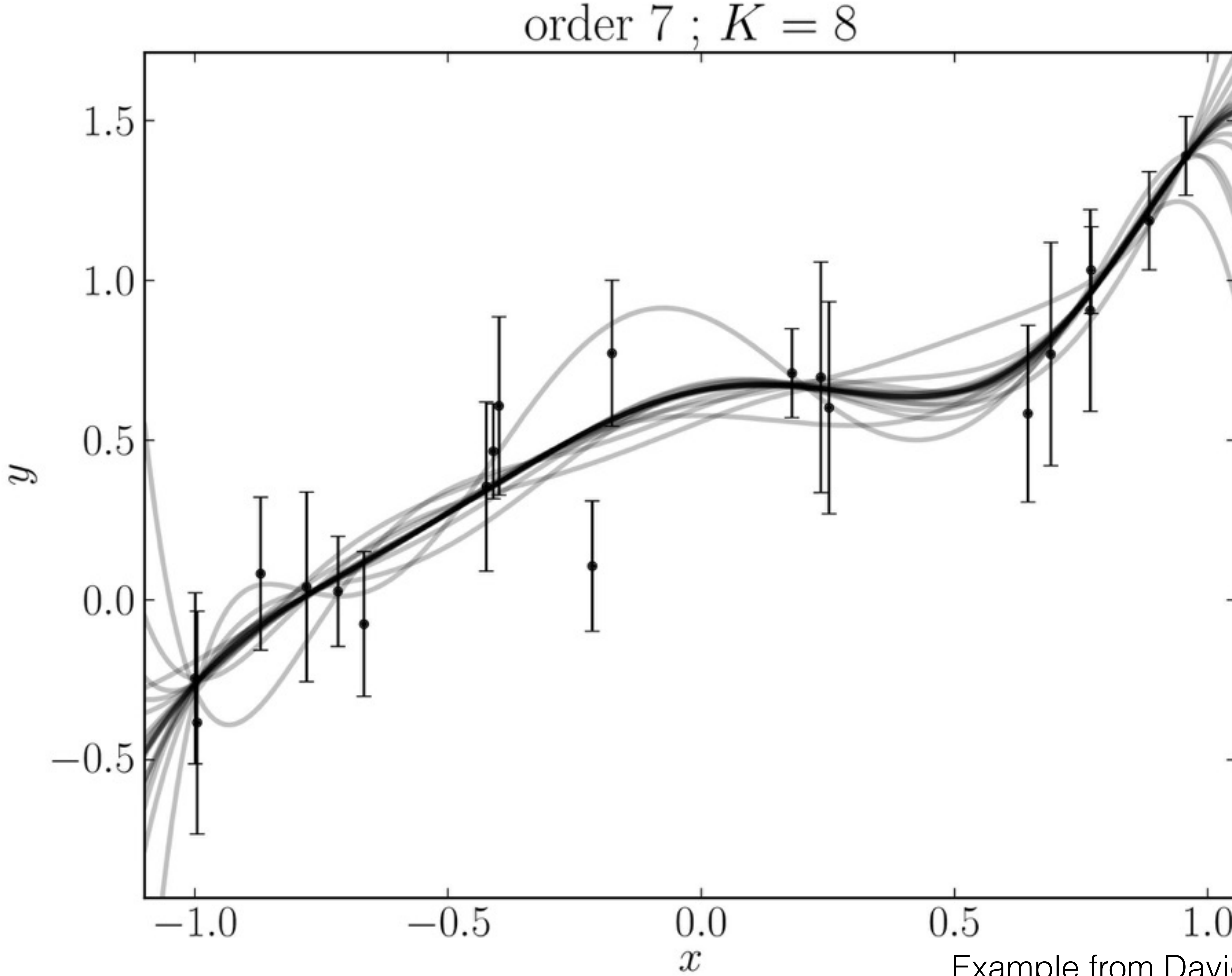
Example from David Hogg



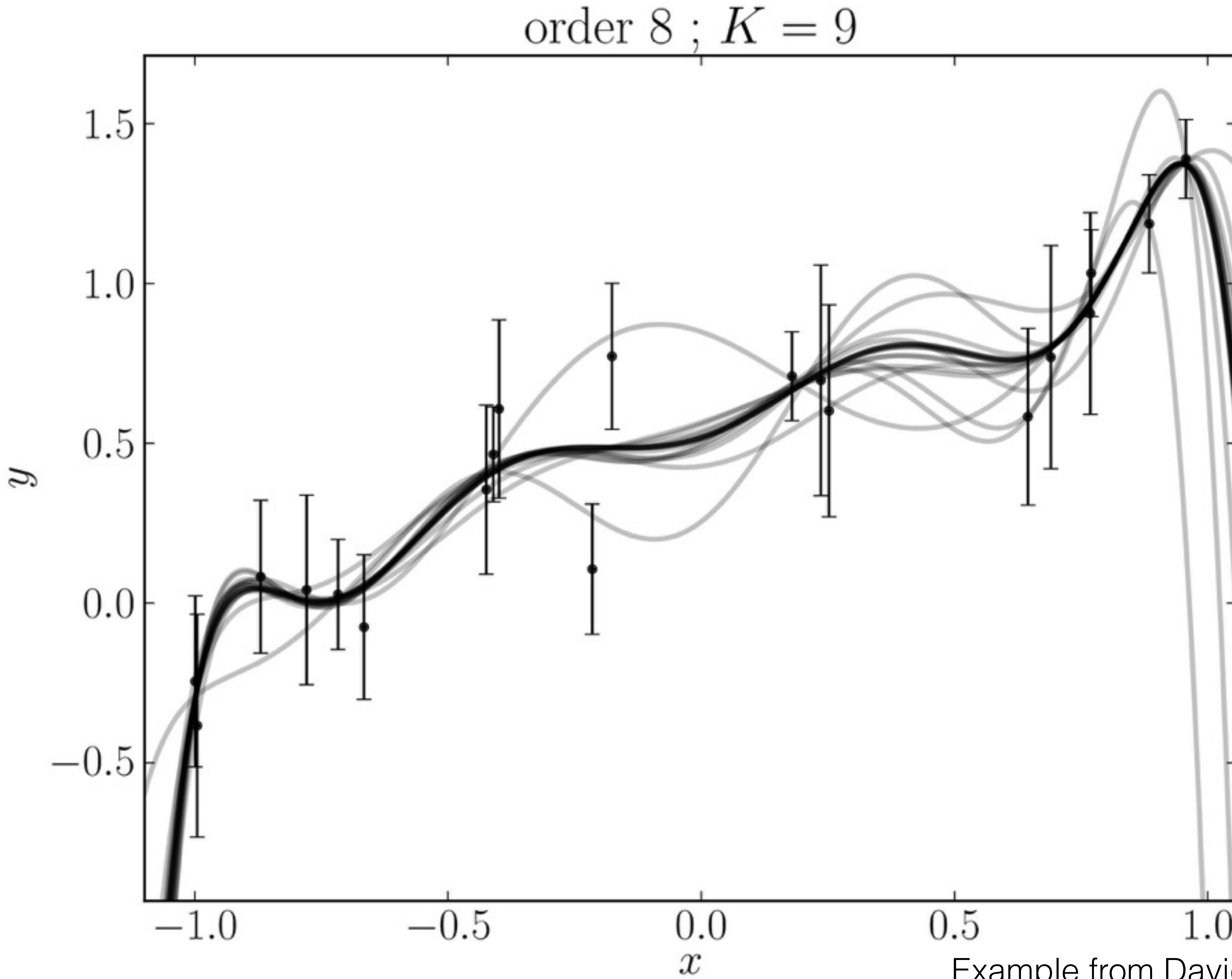
Example from David Hogg



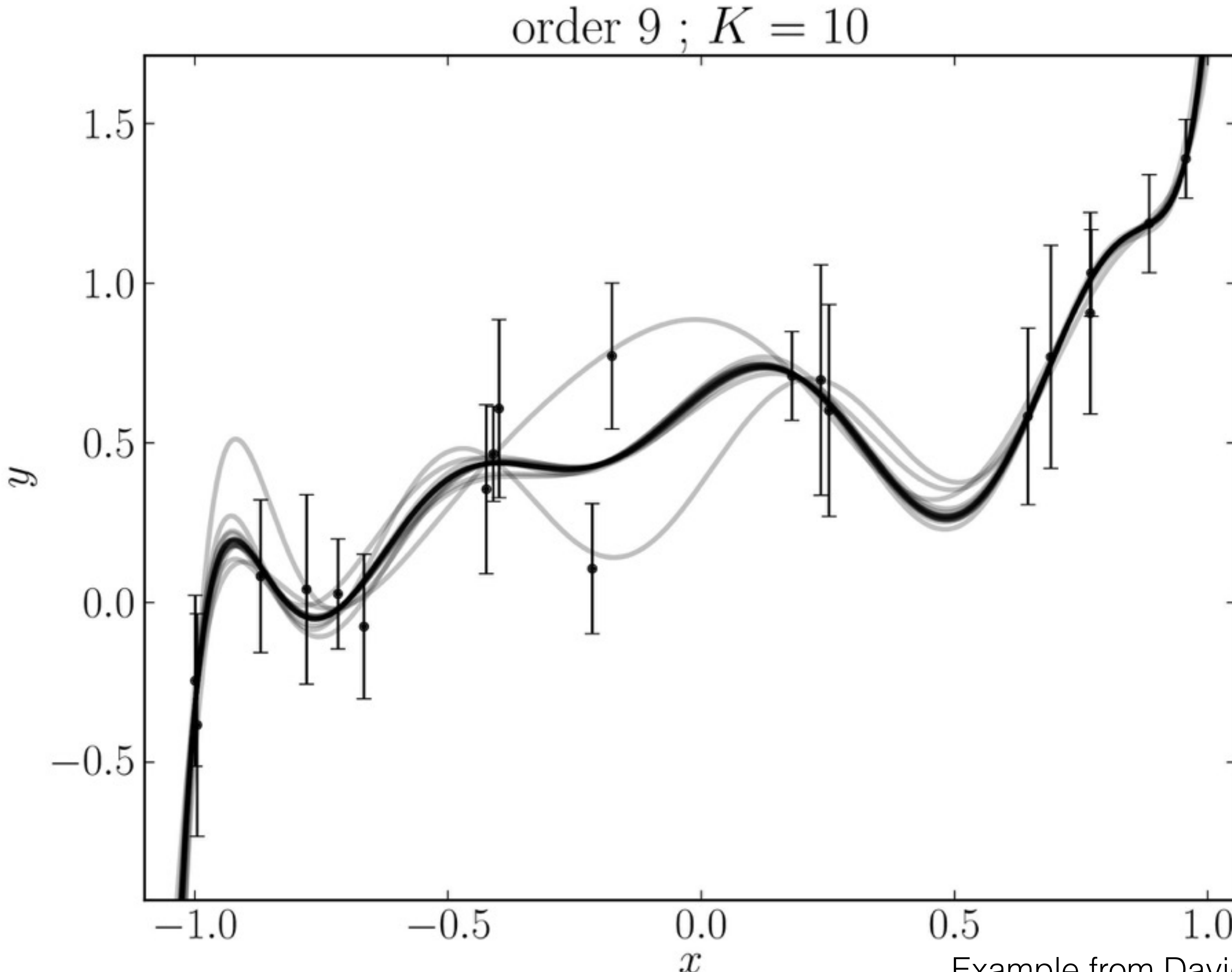
Example from David Hogg



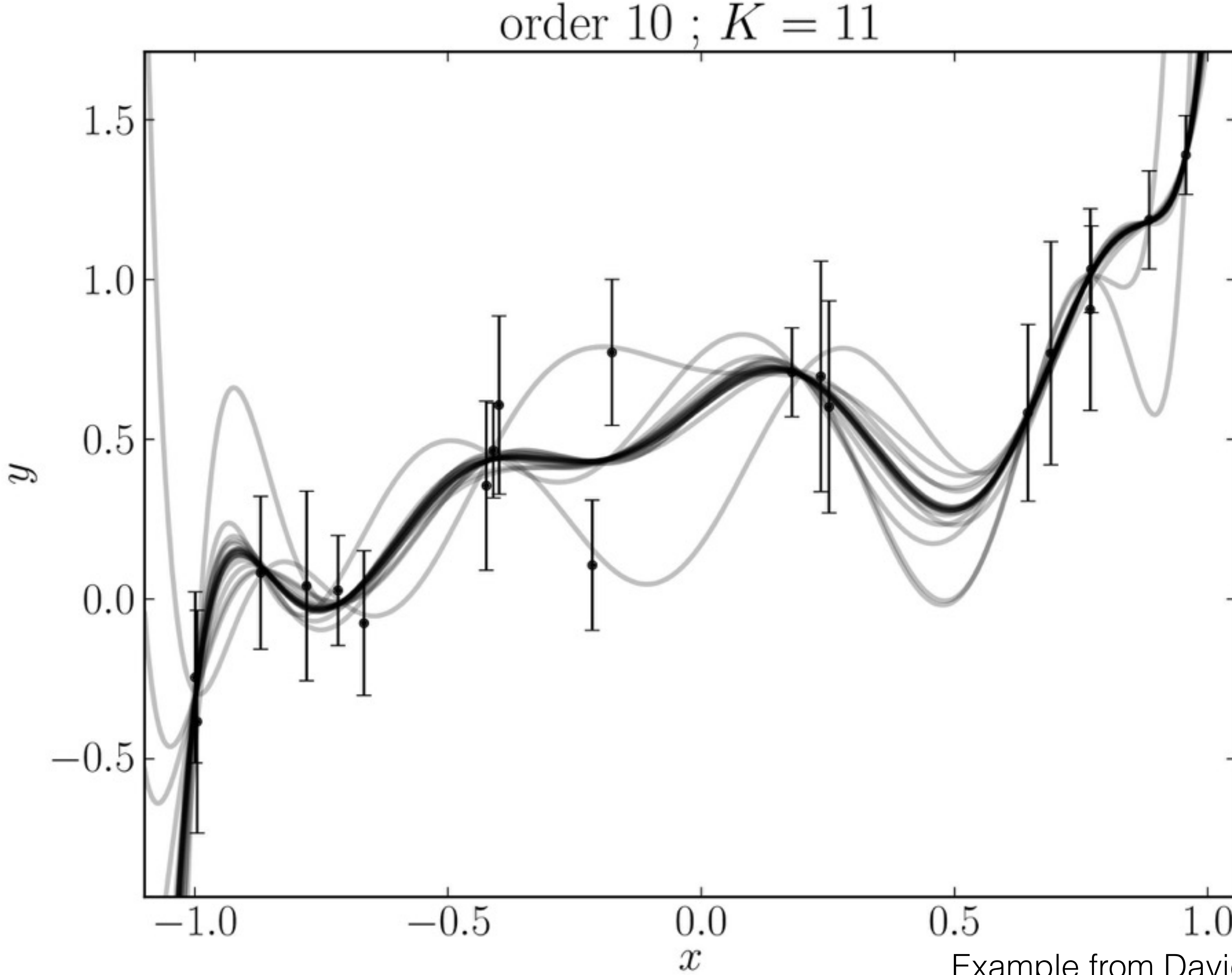
Example from David Hogg



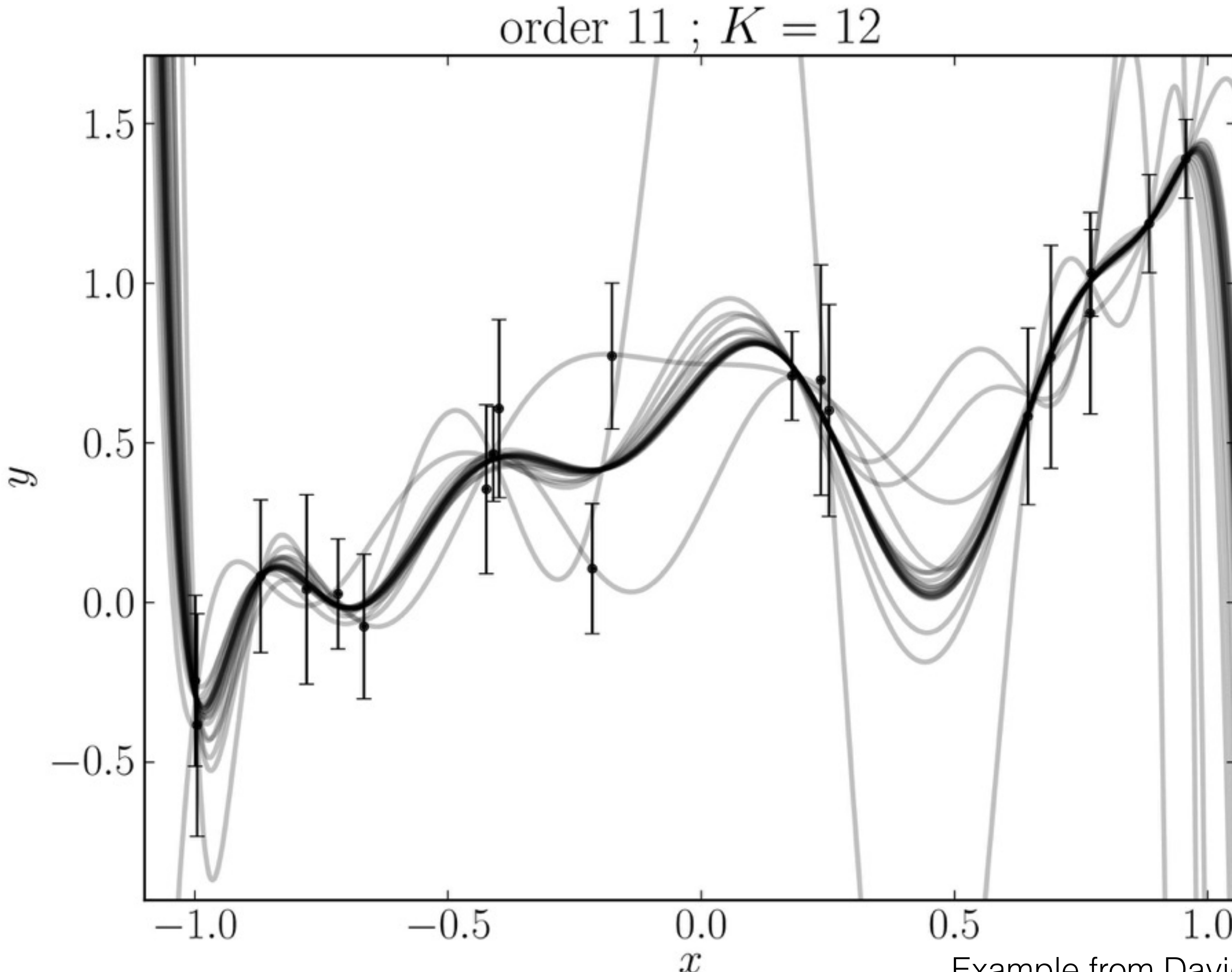
Example from David Hogg



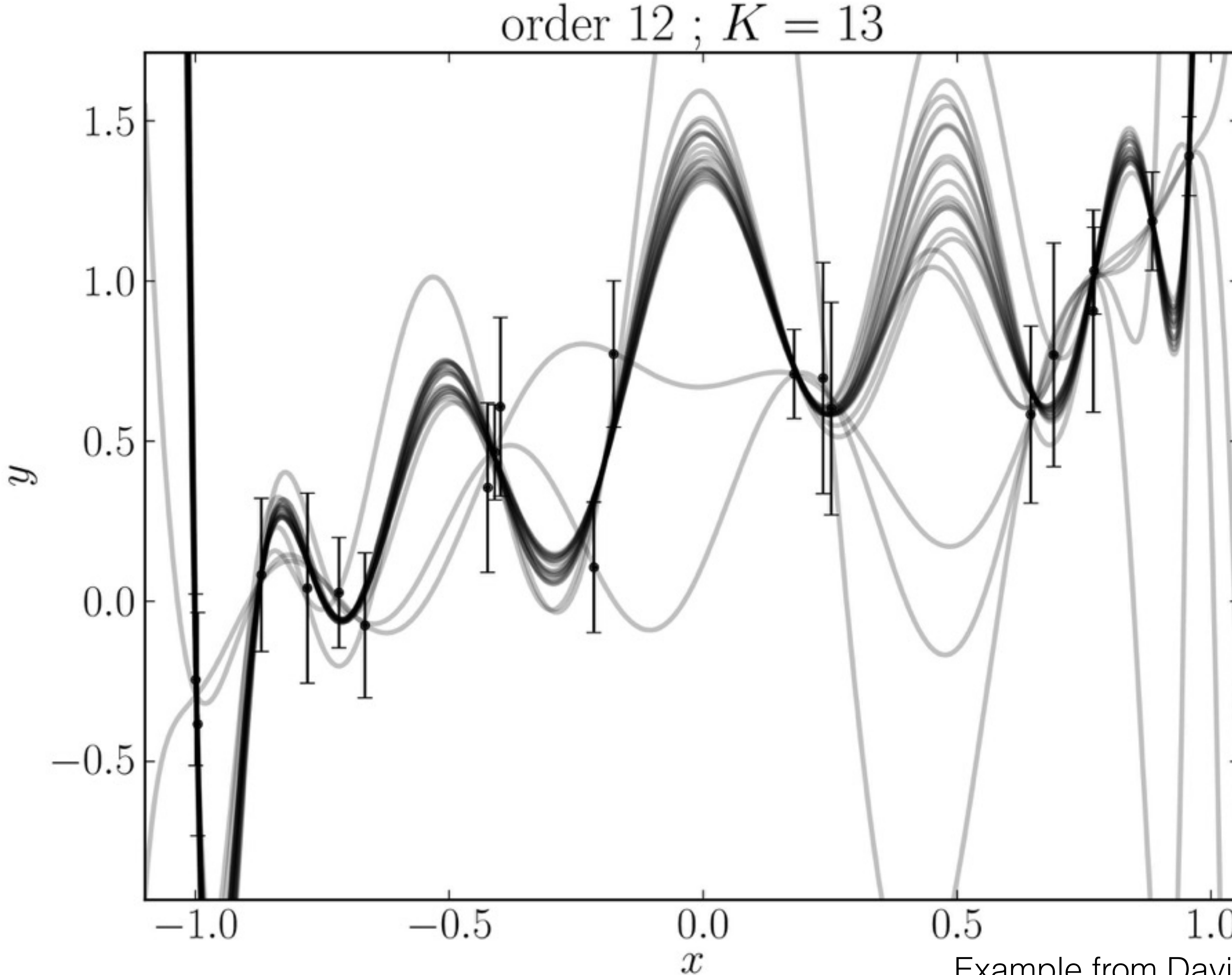
Example from David Hogg



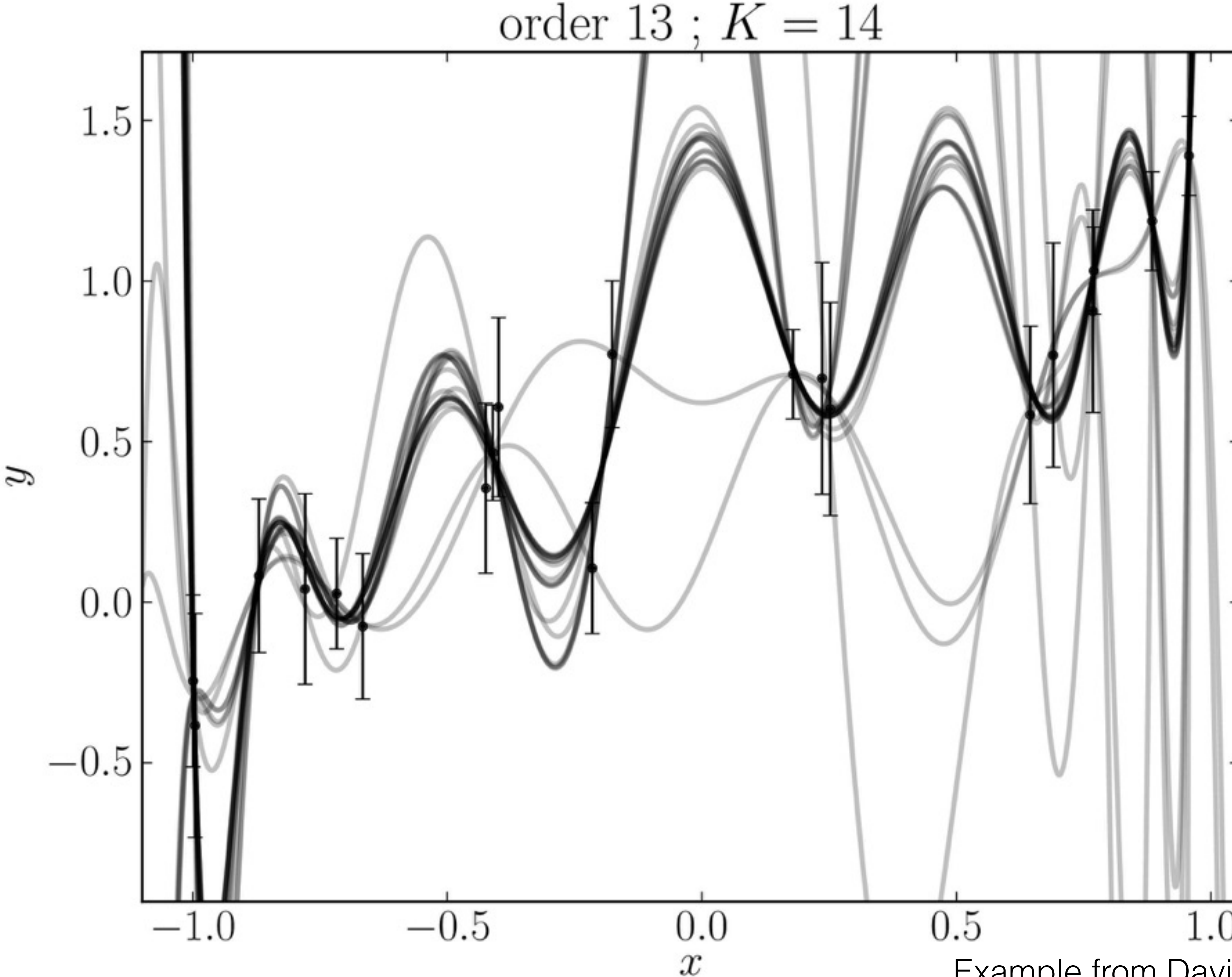
Example from David Hogg



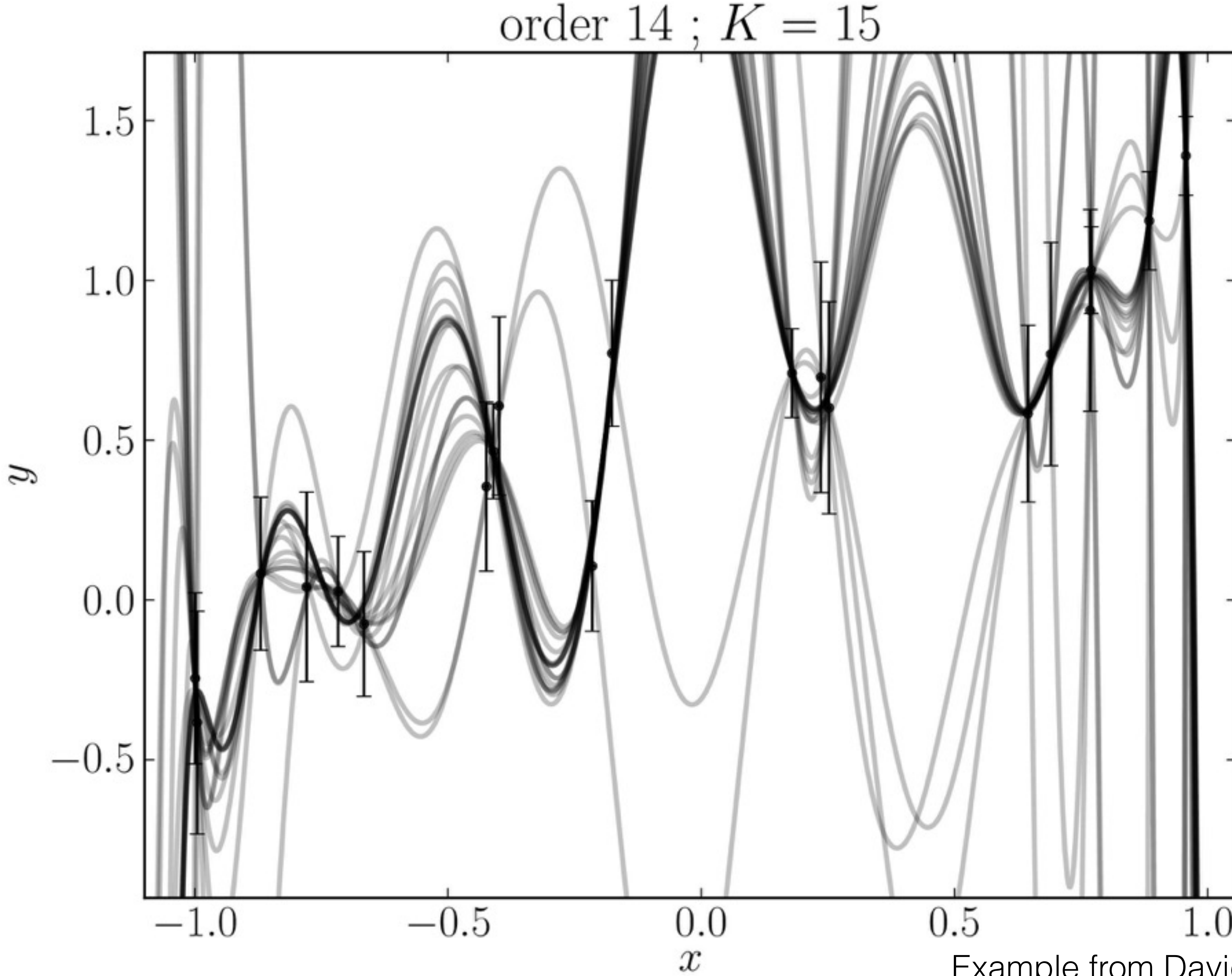
Example from David Hogg



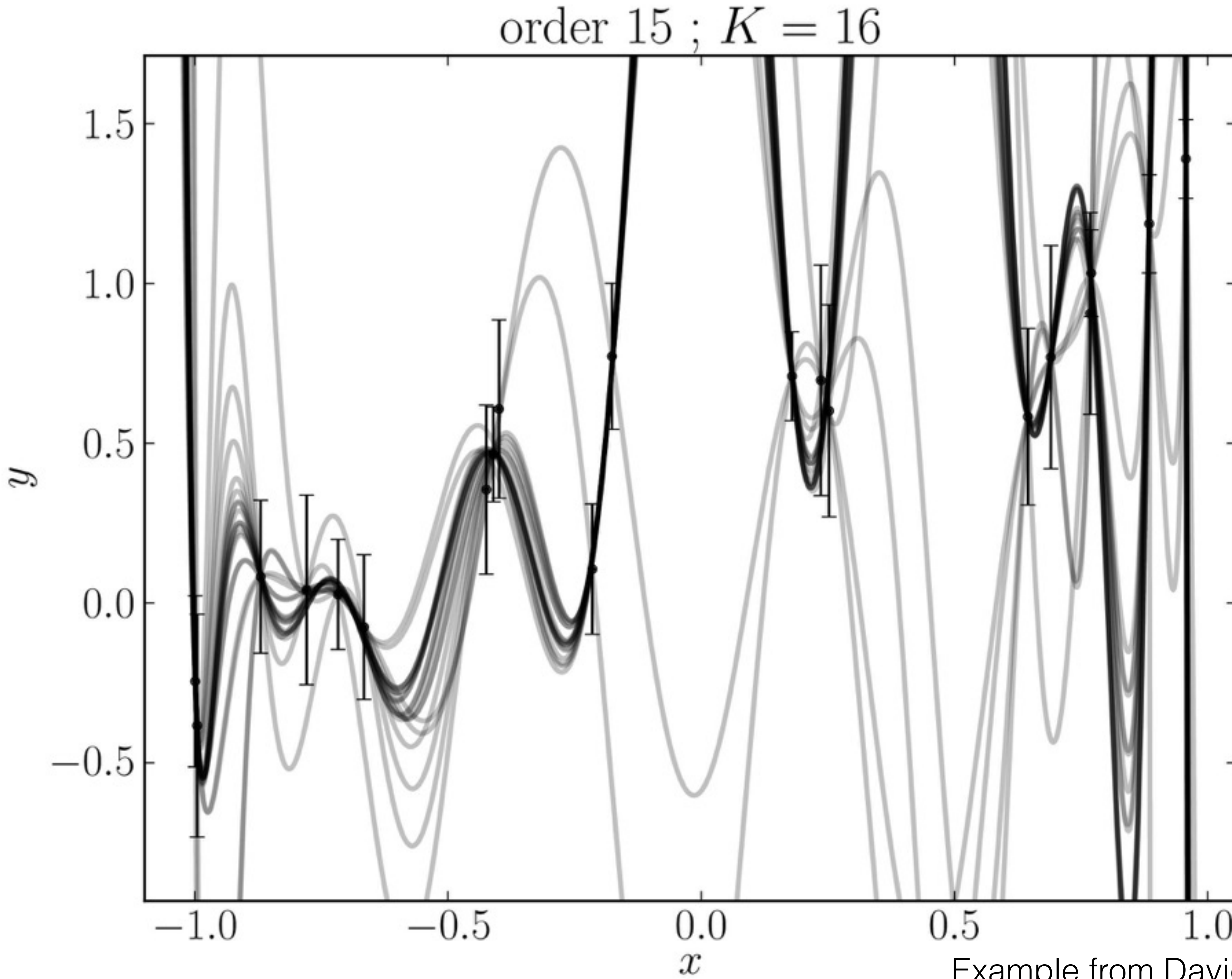
Example from David Hogg



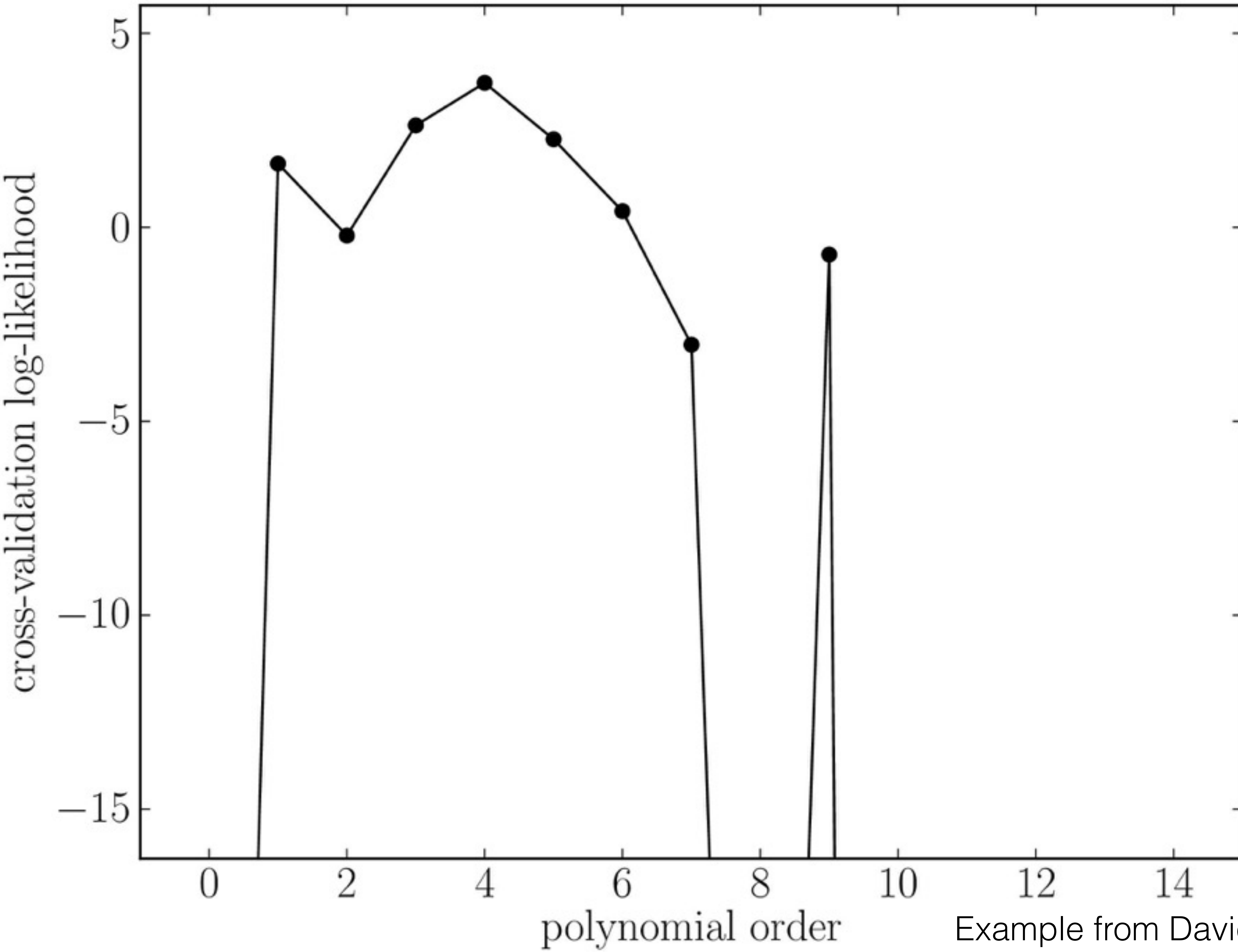
Example from David Hogg



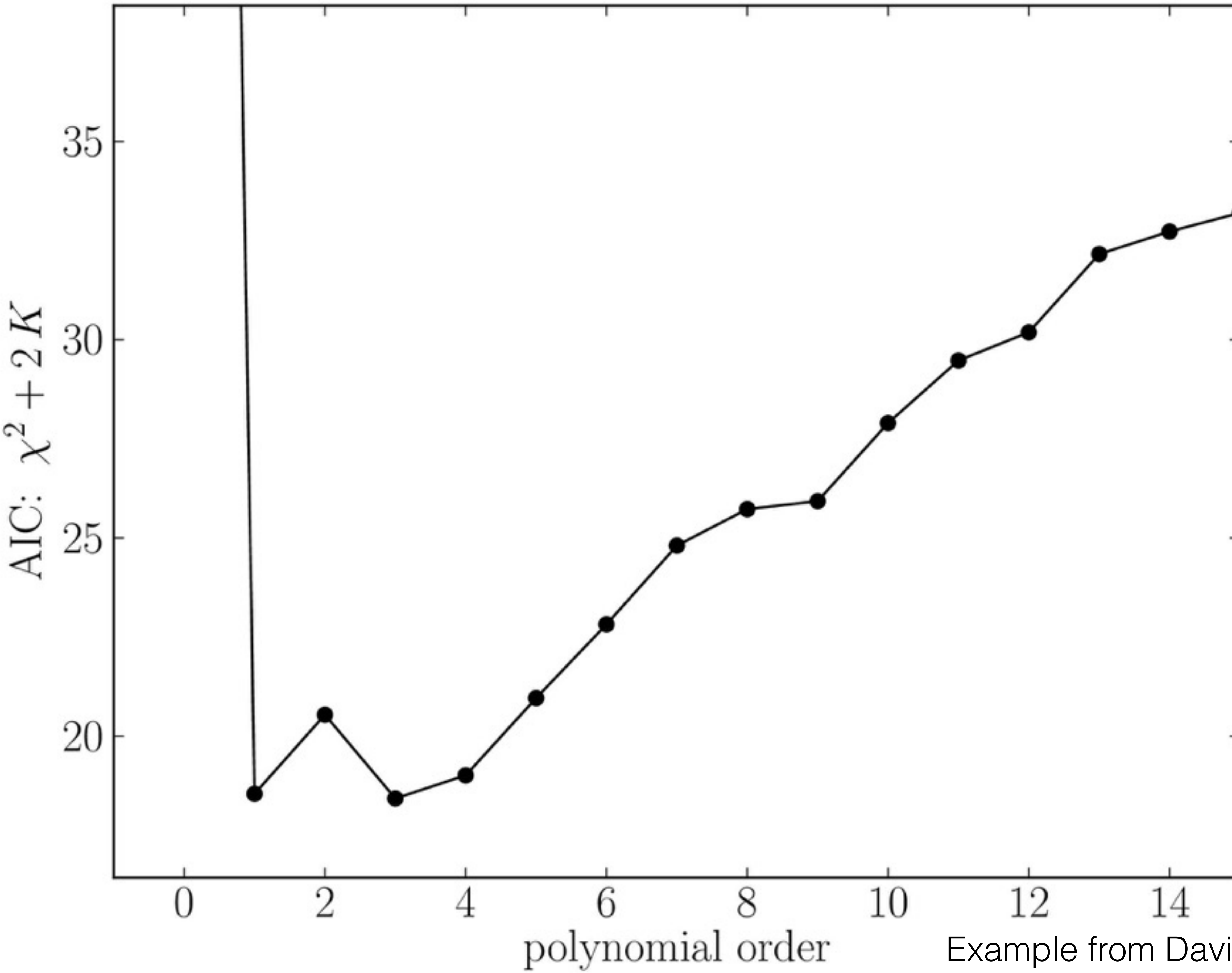
Example from David Hogg



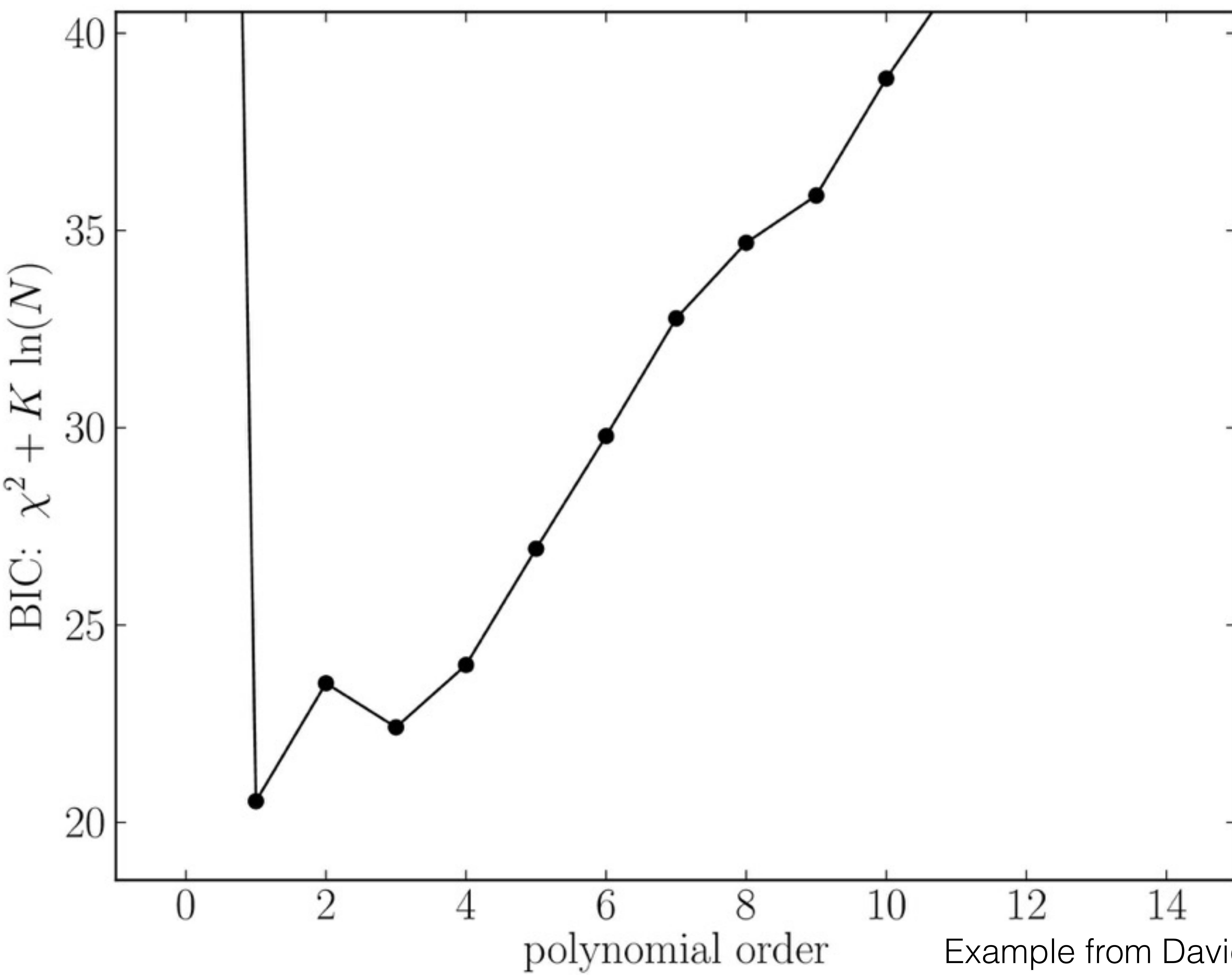
Example from David Hogg



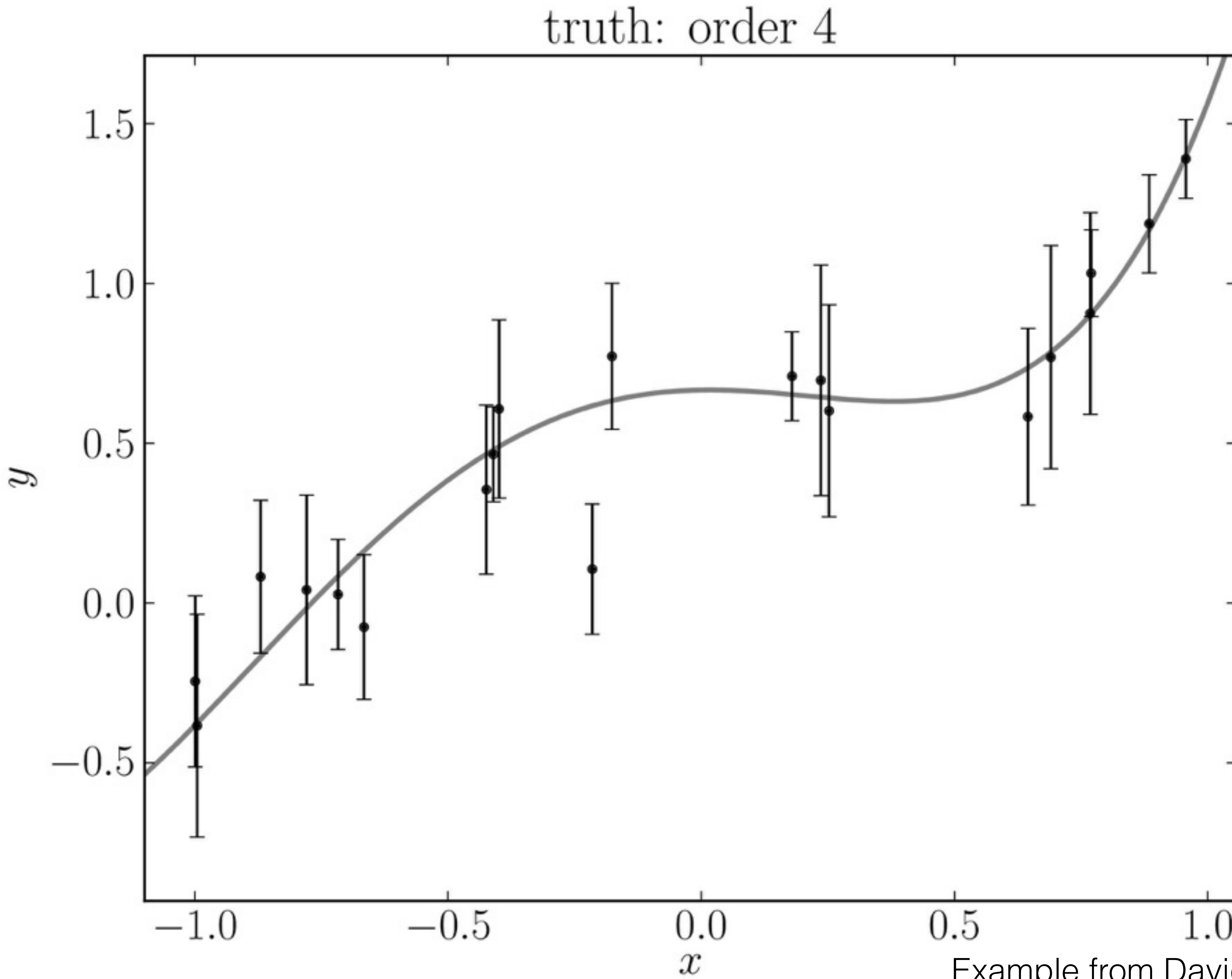
Example from David Hogg



Example from David Hogg



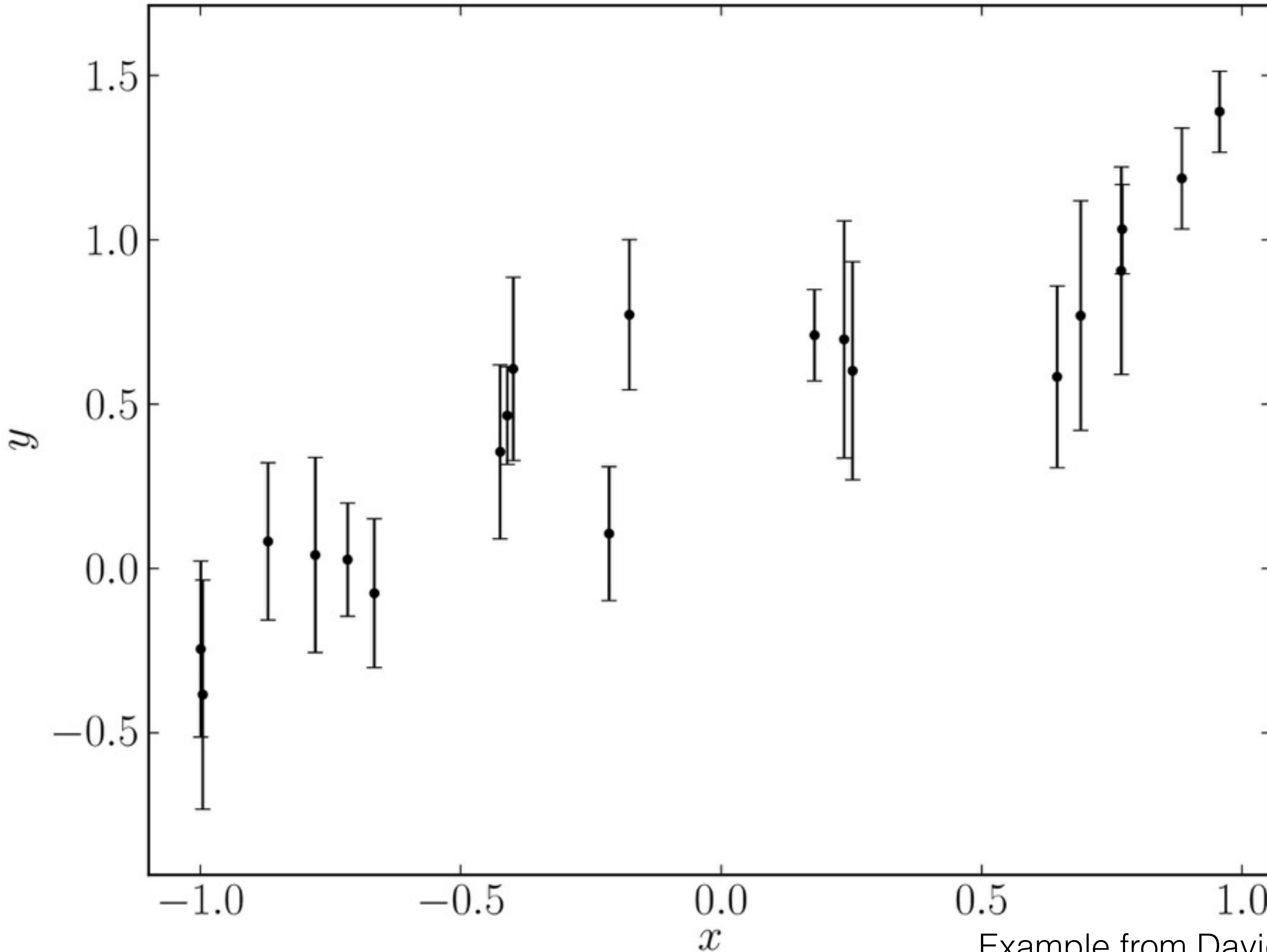
Example from David Hogg



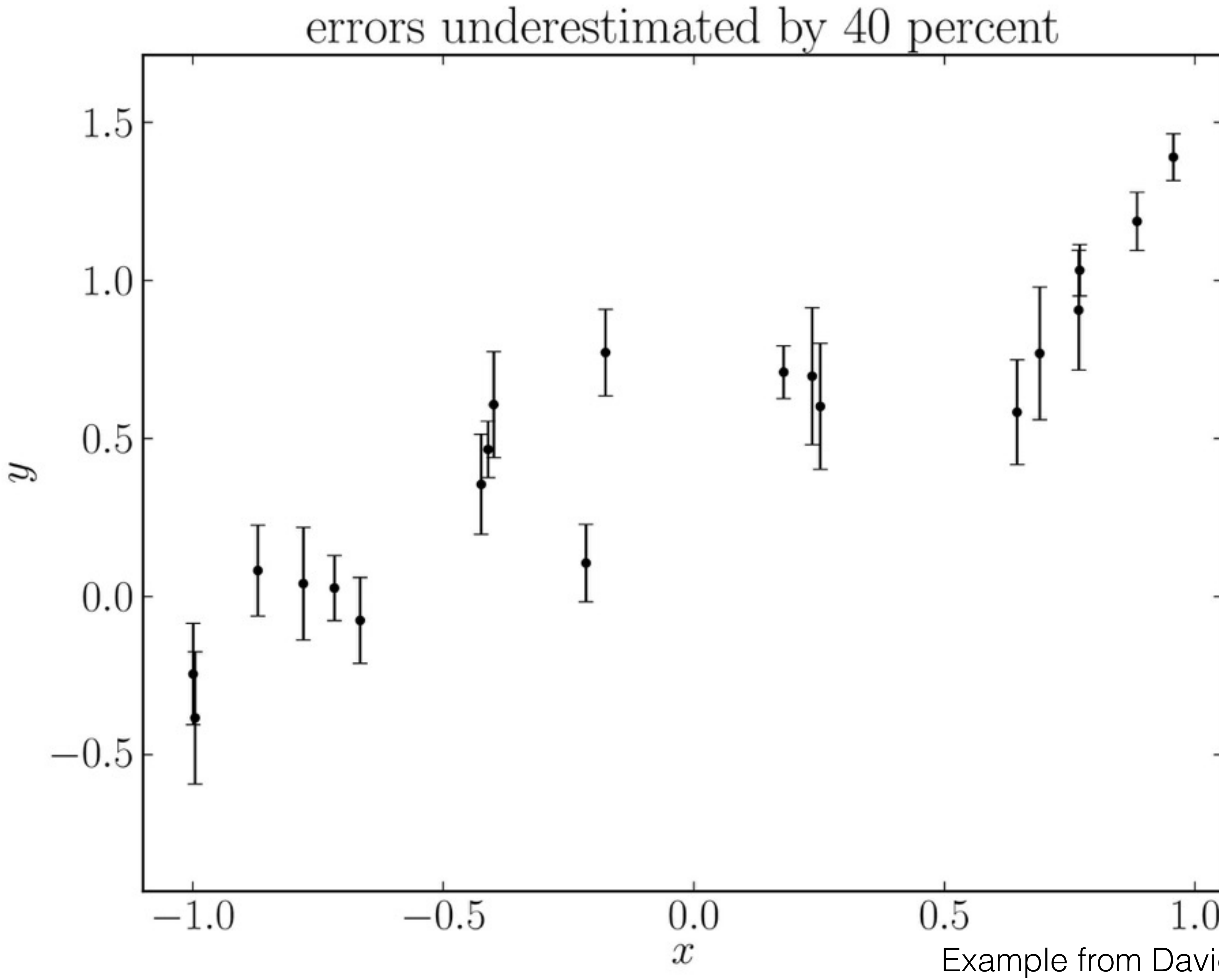
Example from David Hogg

Cross-validation: advantages

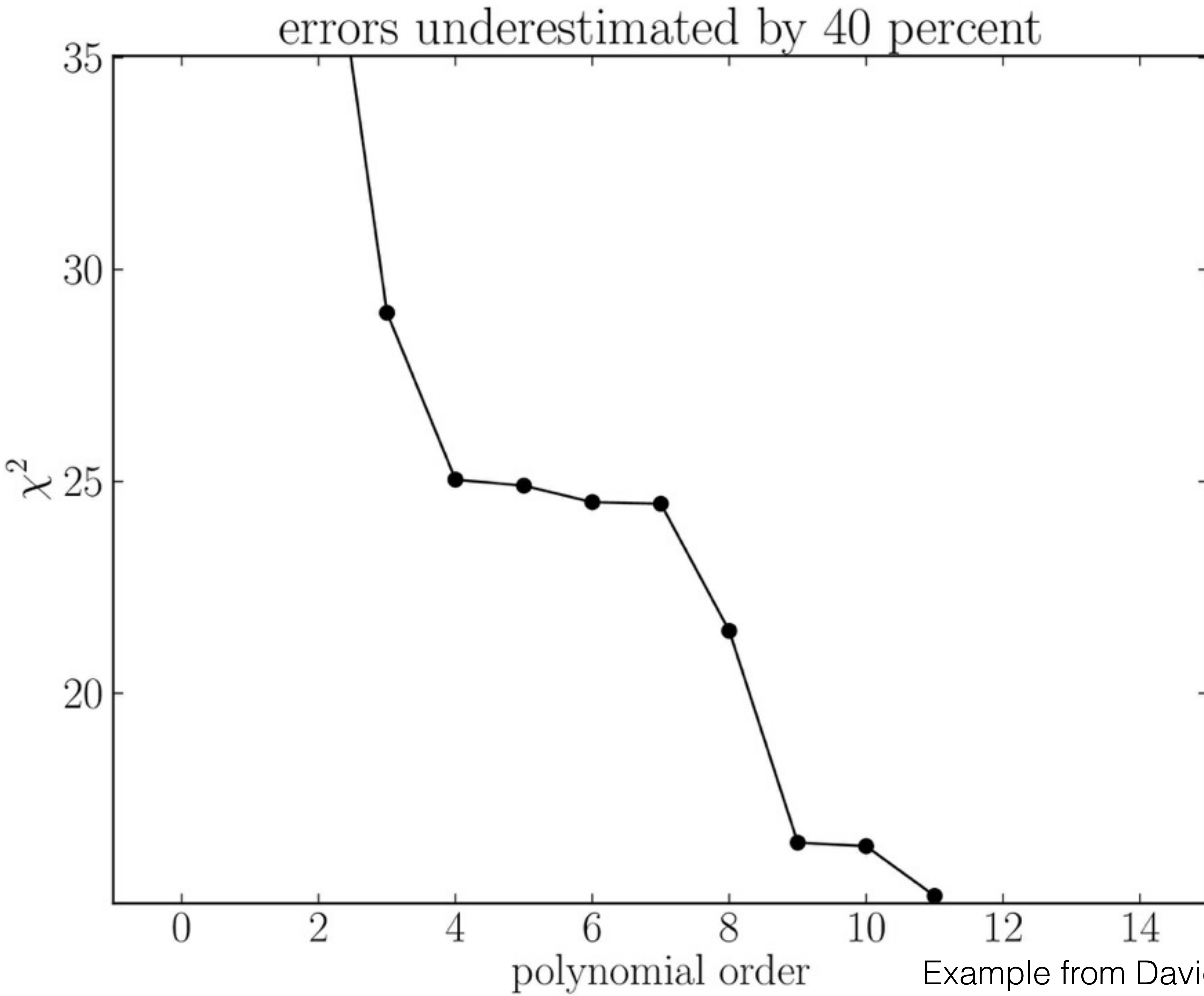
- Makes sense and does not make strong assumptions (just that all data points are typical)
- Easy to implement, but can be expensive if each fit is expensive (can use leave-N-out instead)
- Robust against certain types of under/overestimates of the uncertainties (similar to bootstrap)



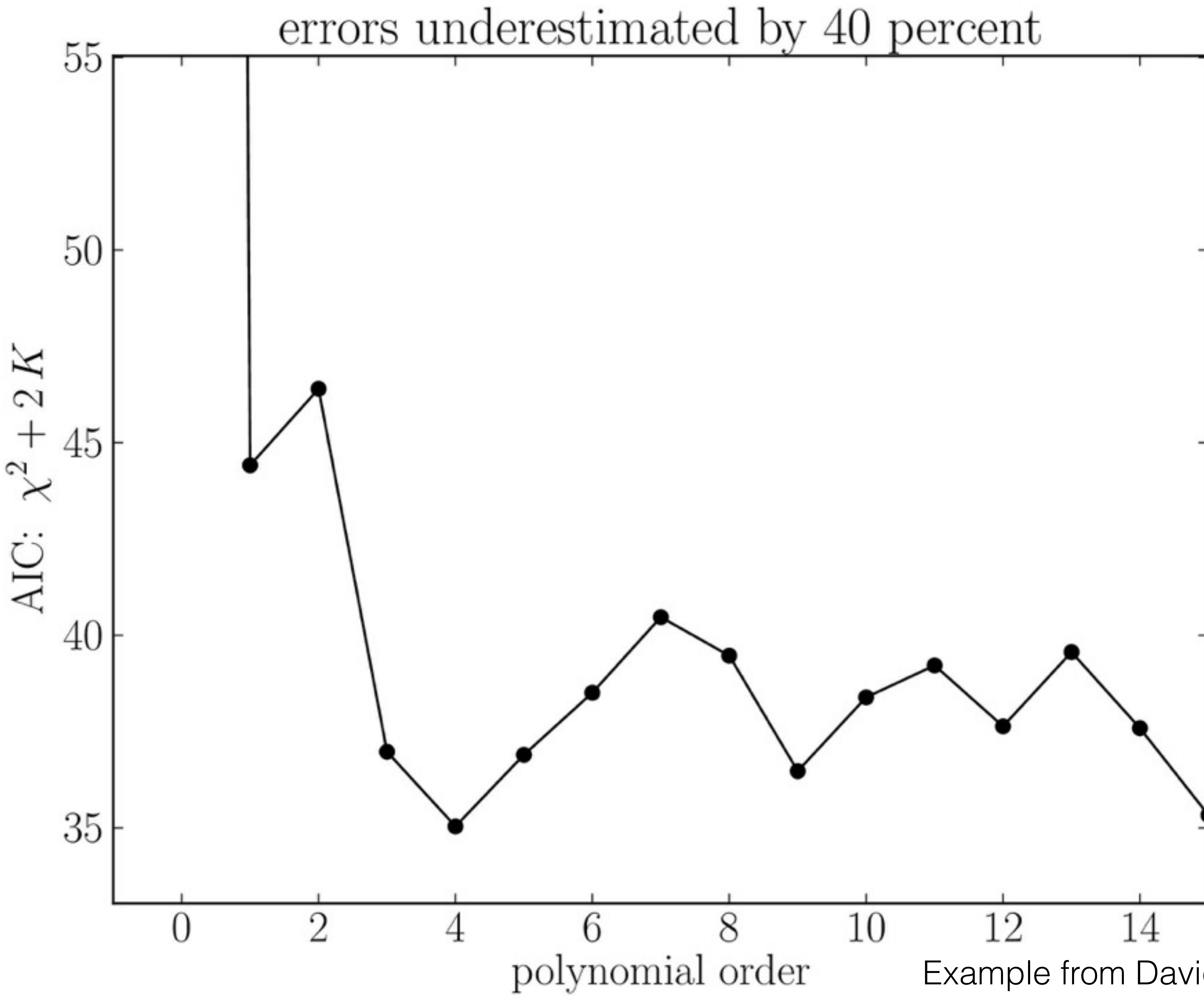
Example from David Hogg



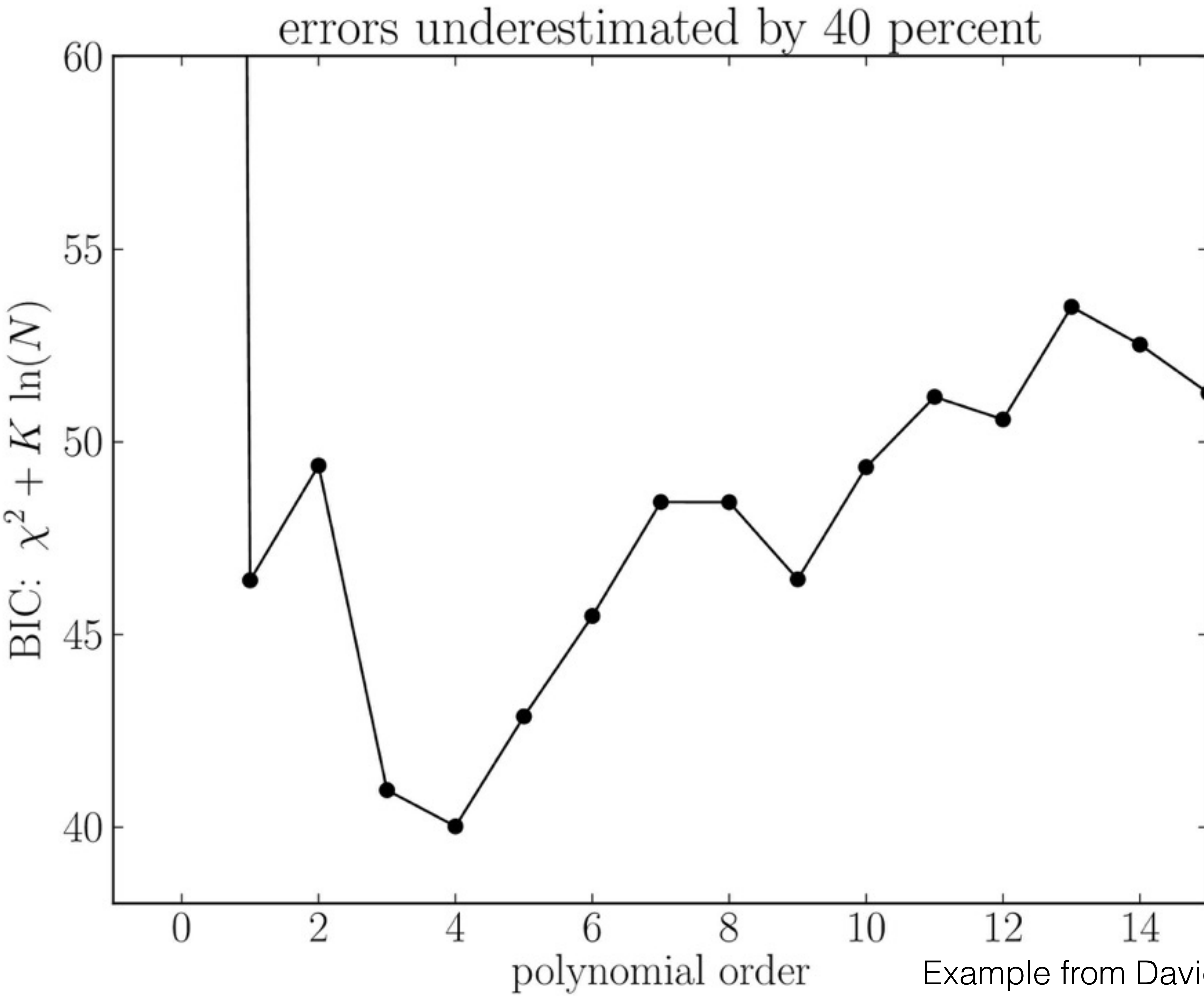
Example from David Hogg



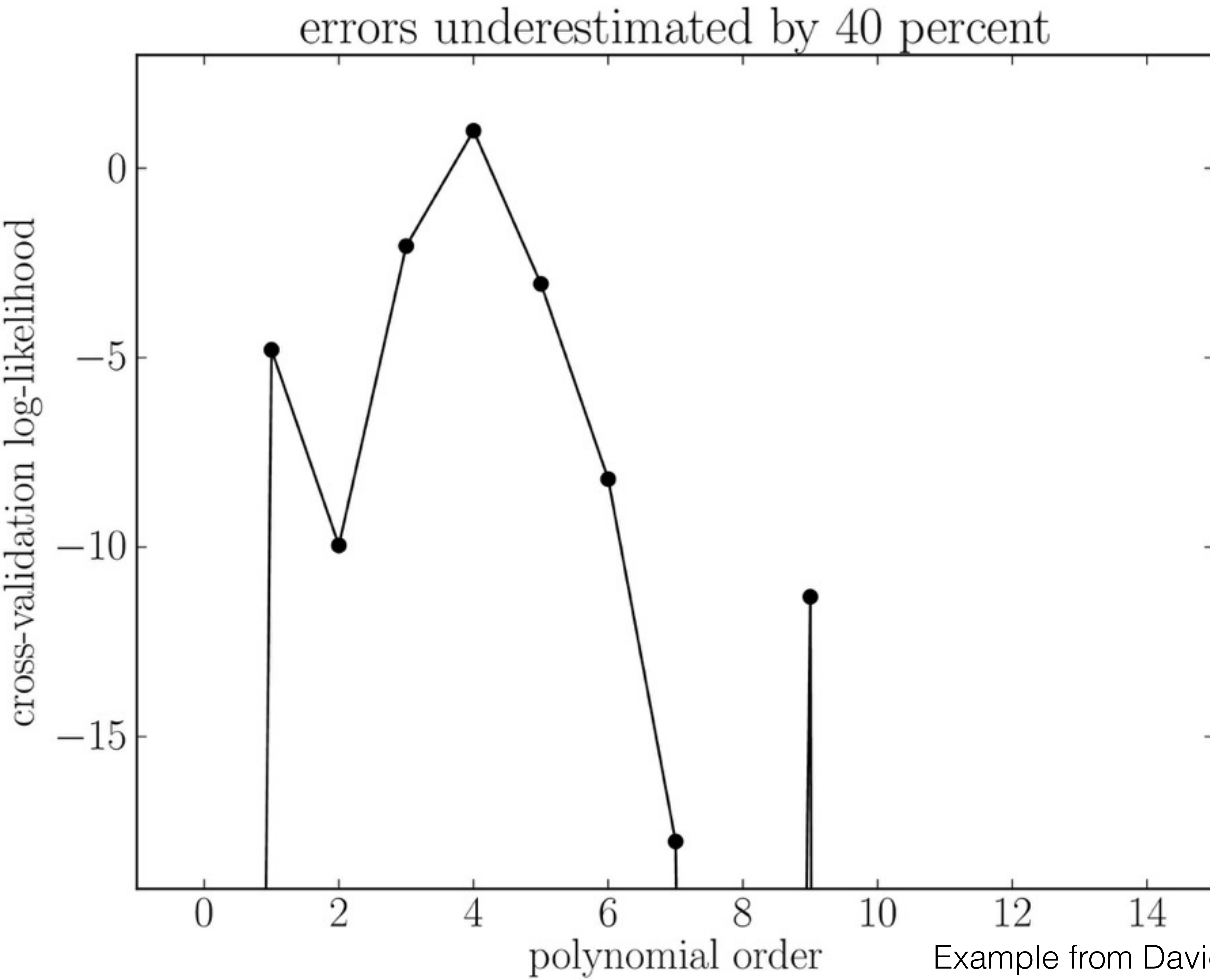
Example from David Hogg



Example from David Hogg



Example from David Hogg



Example from David Hogg