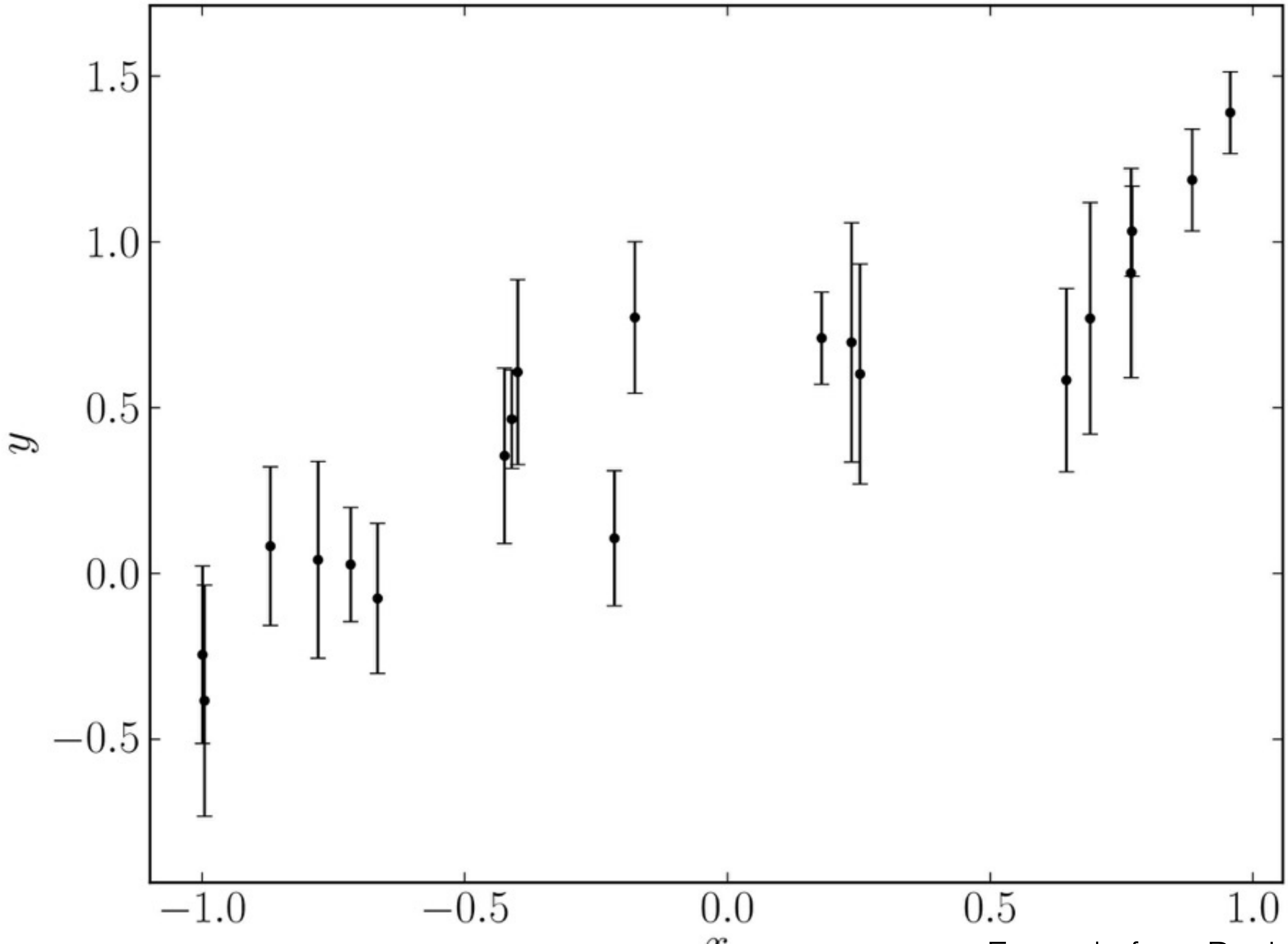
Statistics and Inference in Astrophysics

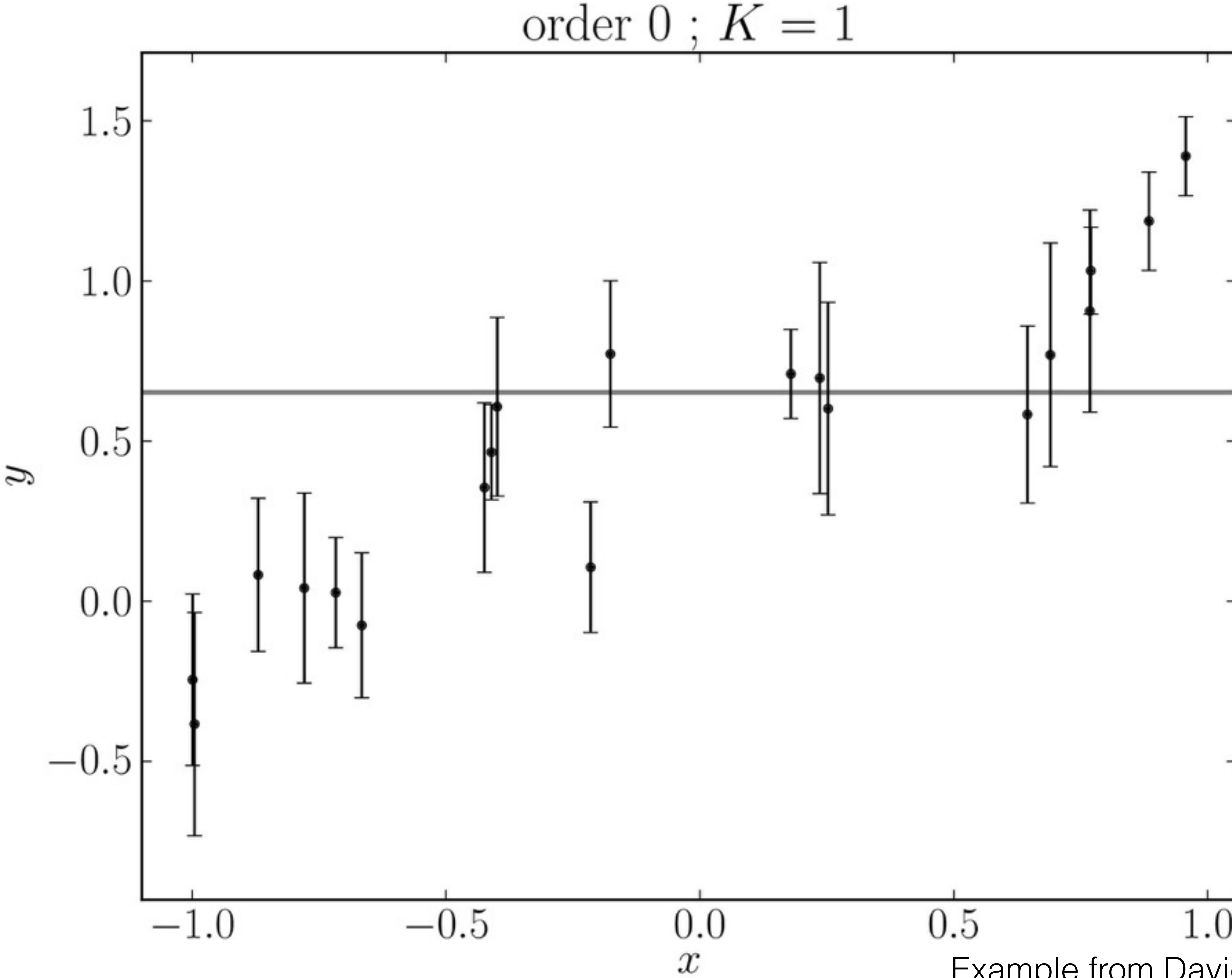
Zhu & Menard (2013)

Today

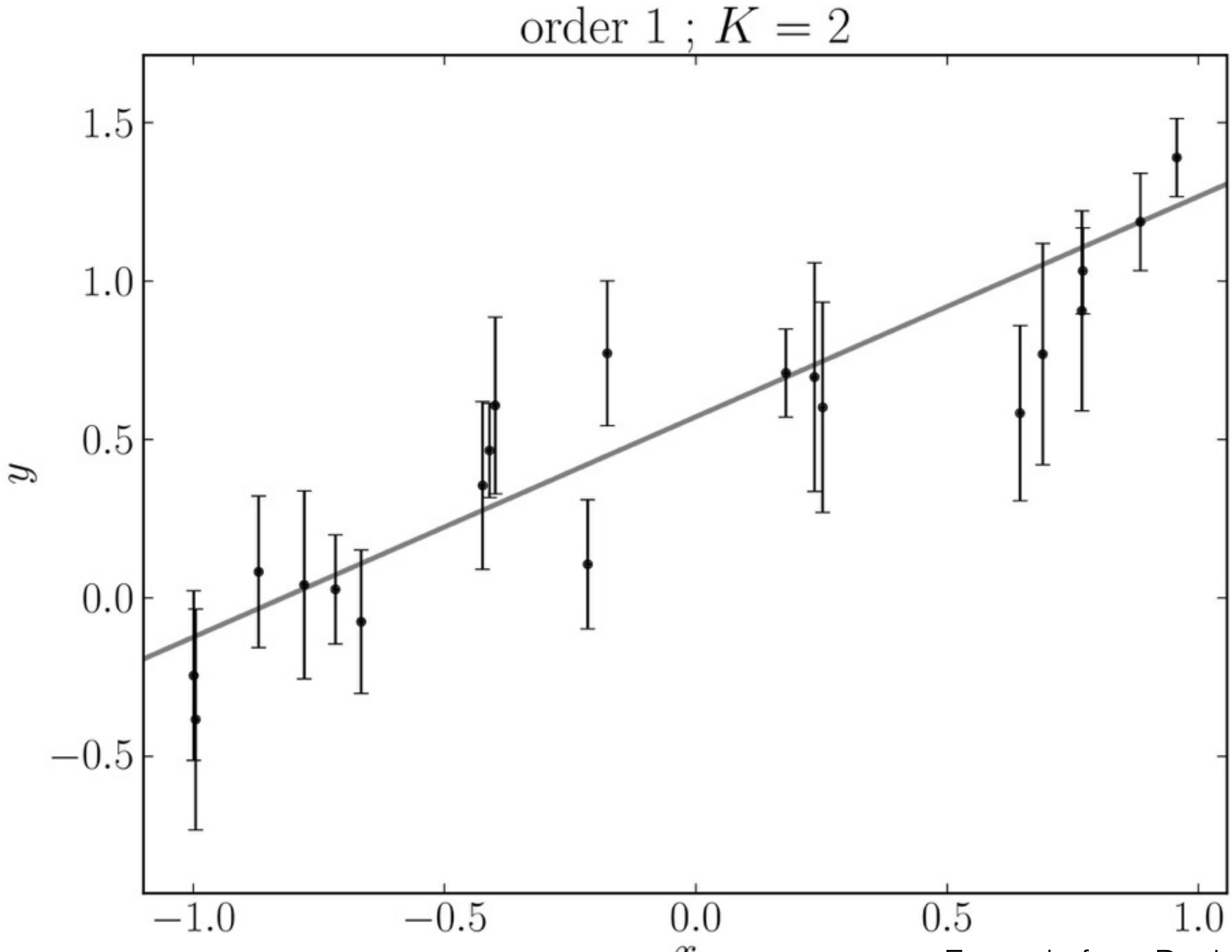
- Goodness-of-fit, model selection, cross-validation
- Outliers, robust statistics



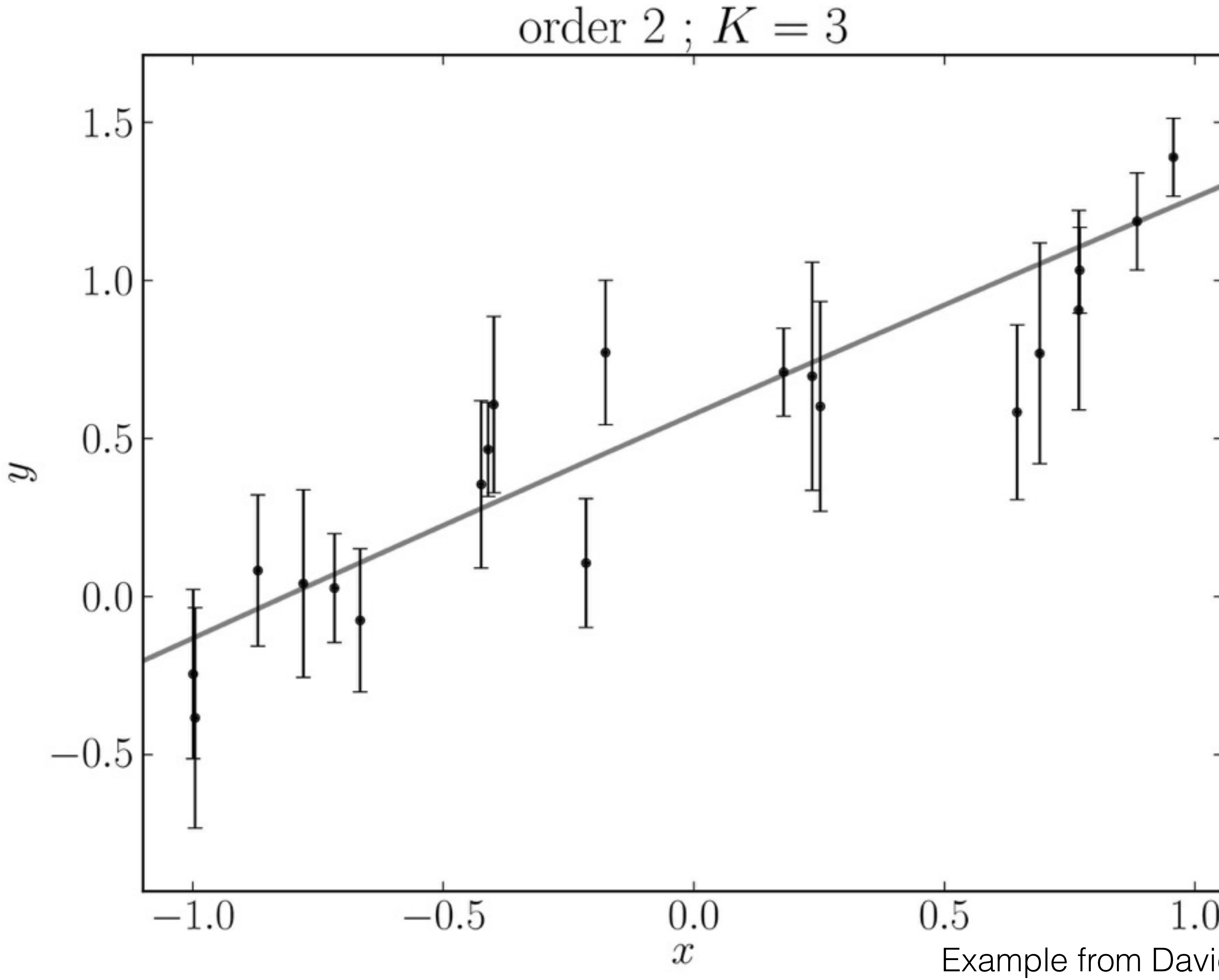
Example from David Hogg



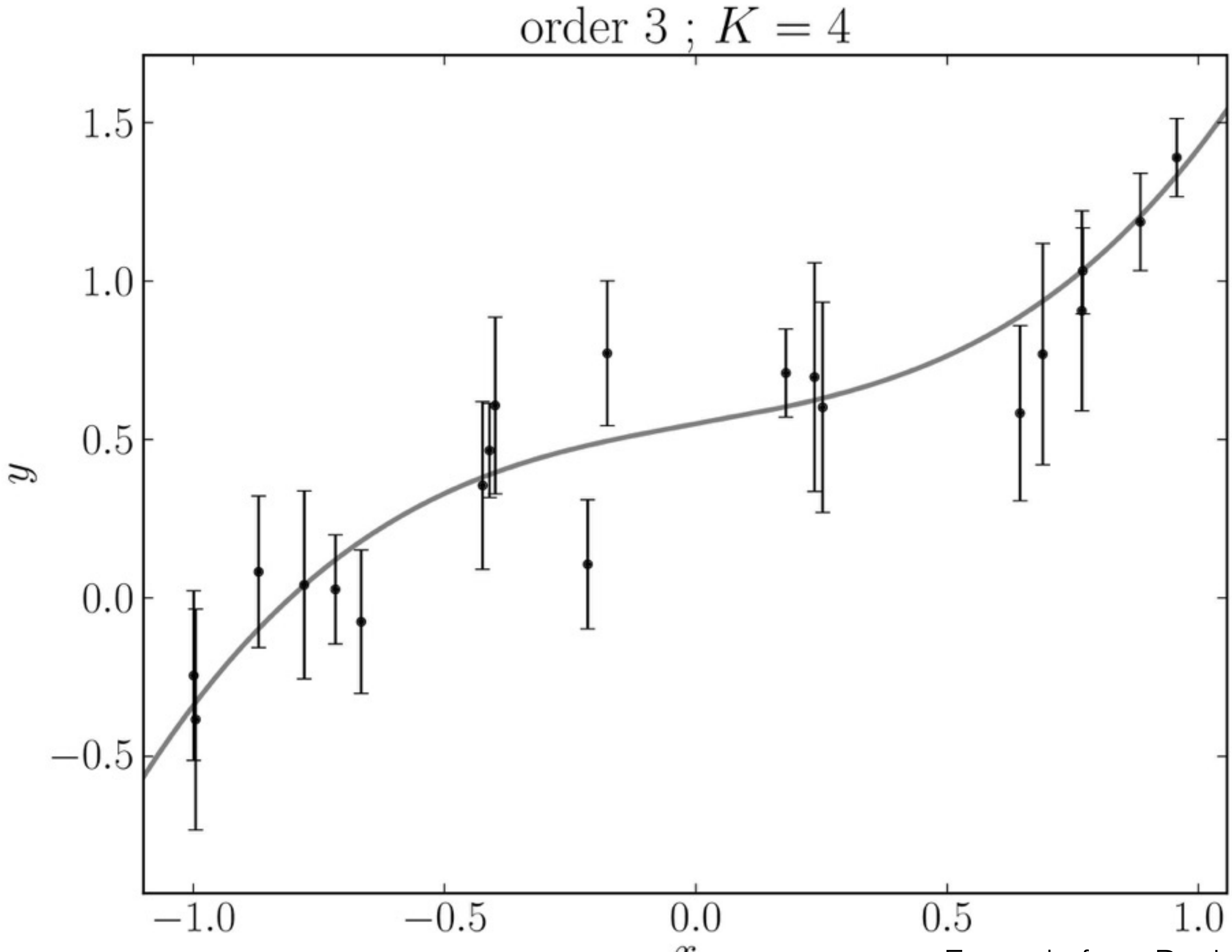
Example from David Hogg



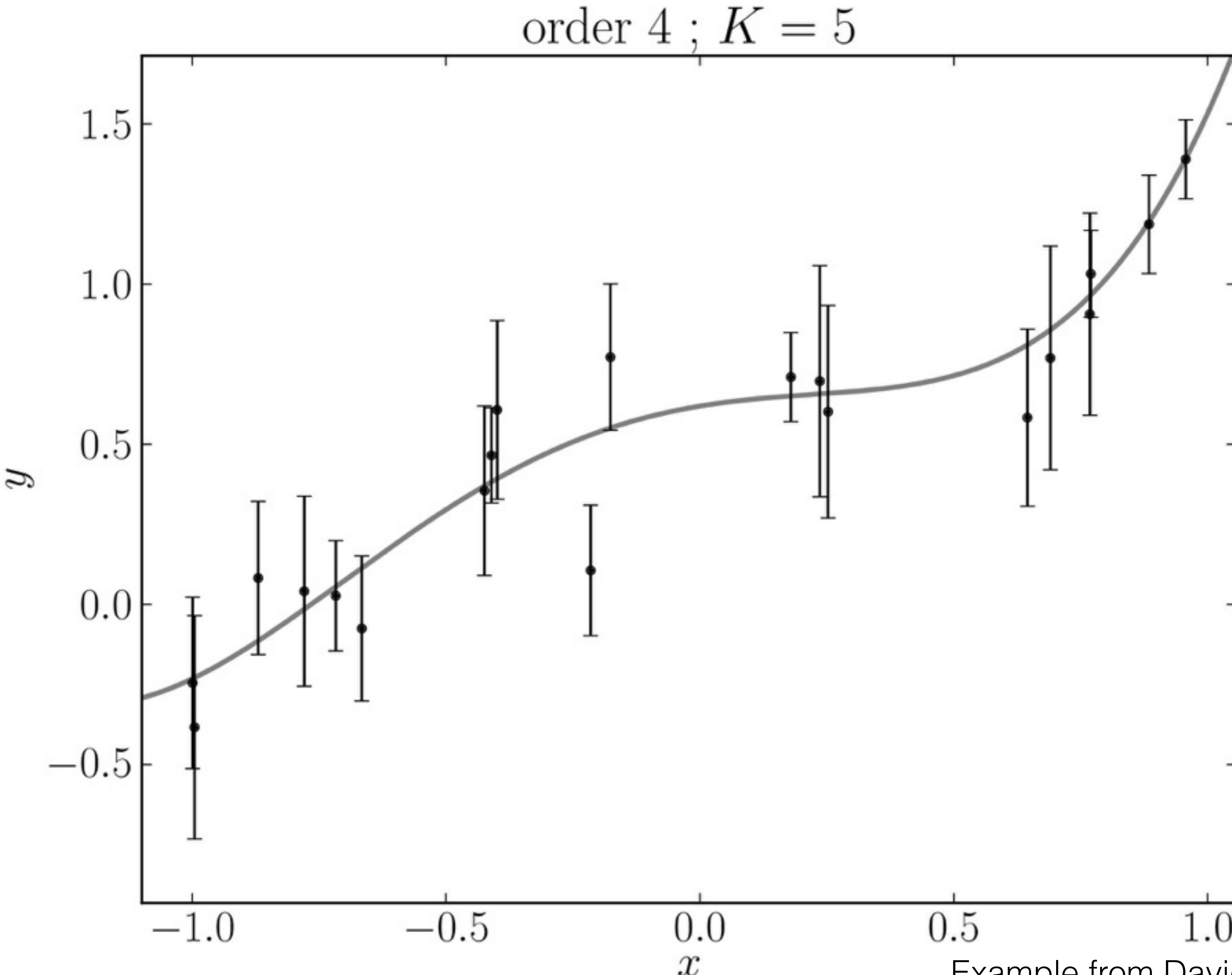
Example from David Hogg



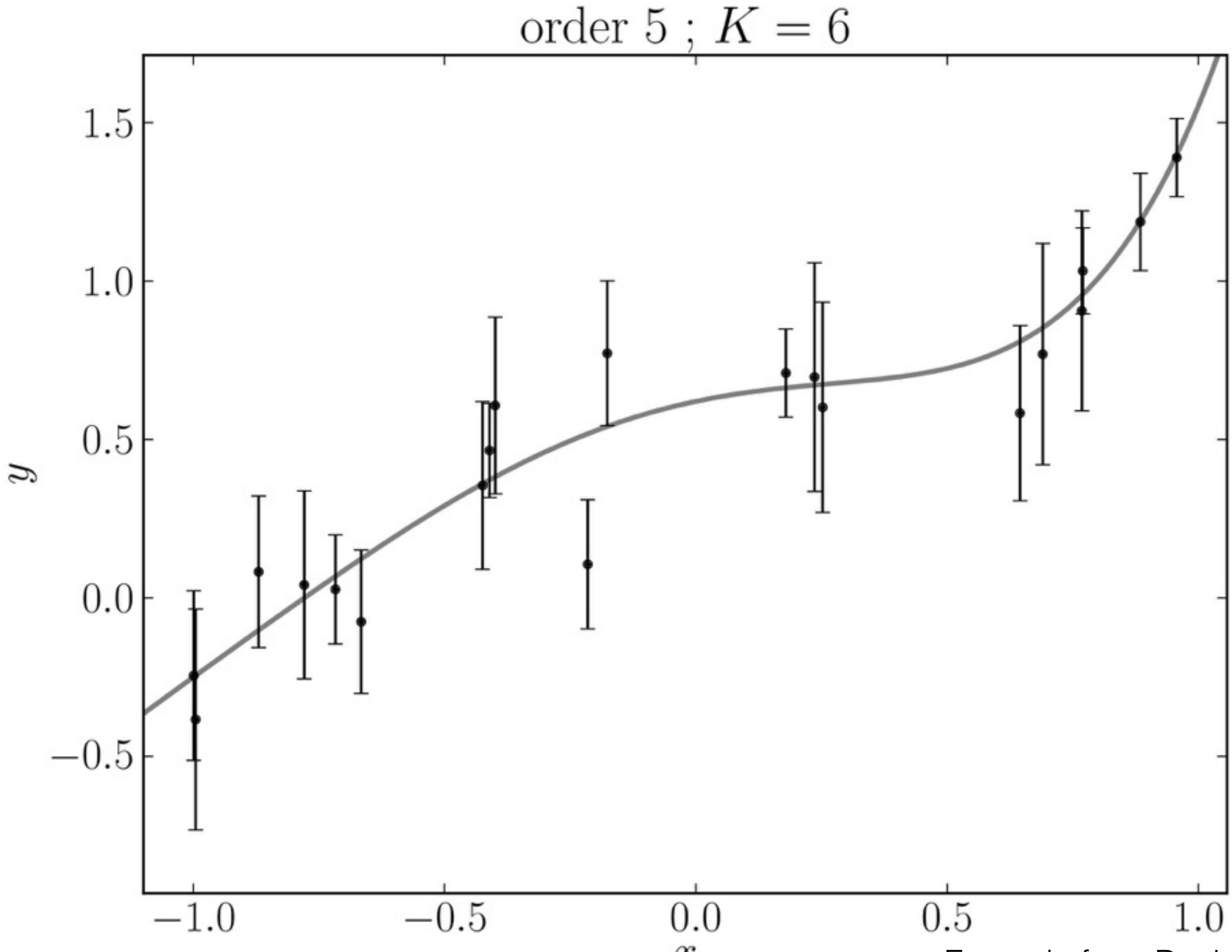
Example from David Hogg



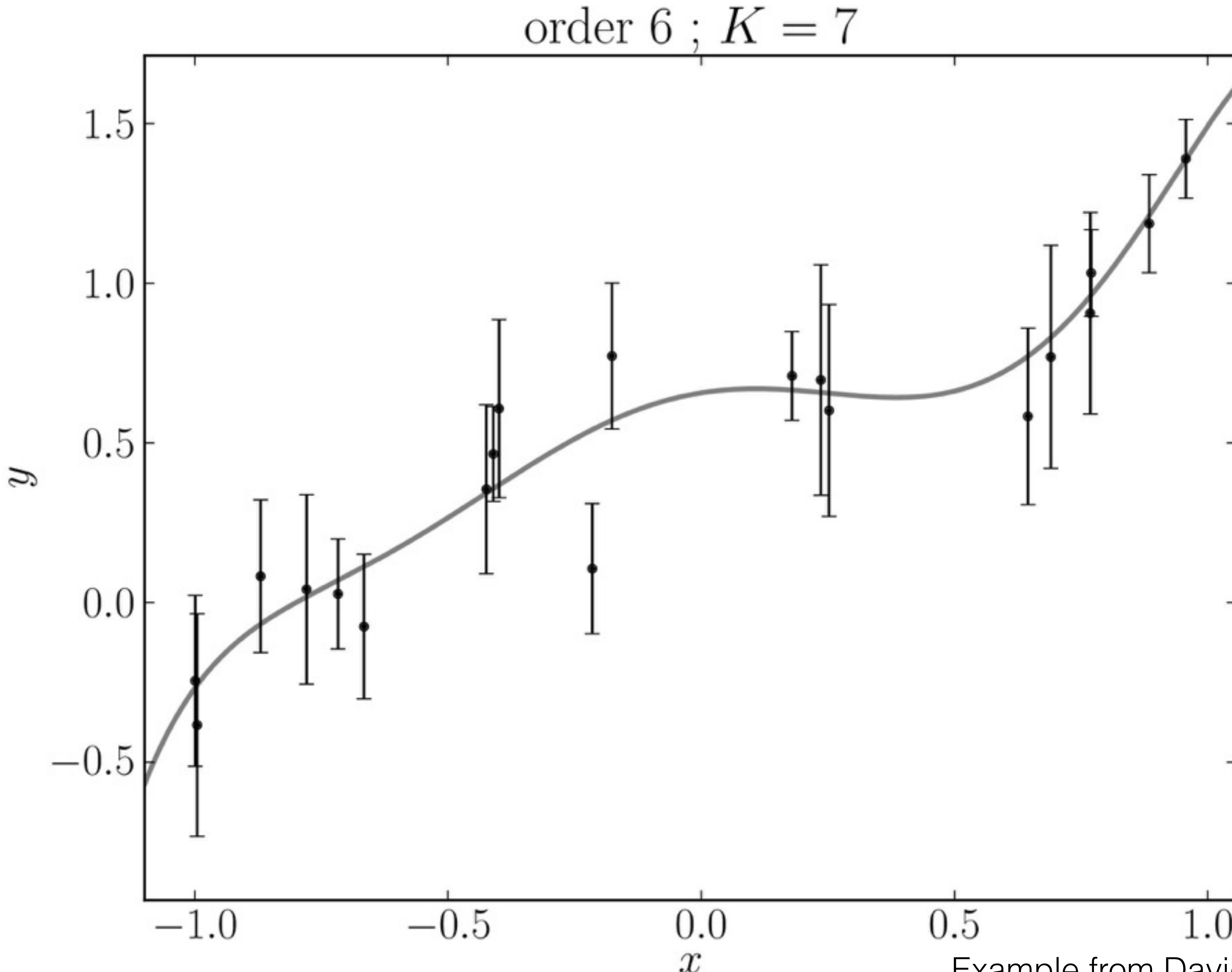
Example from David Hogg



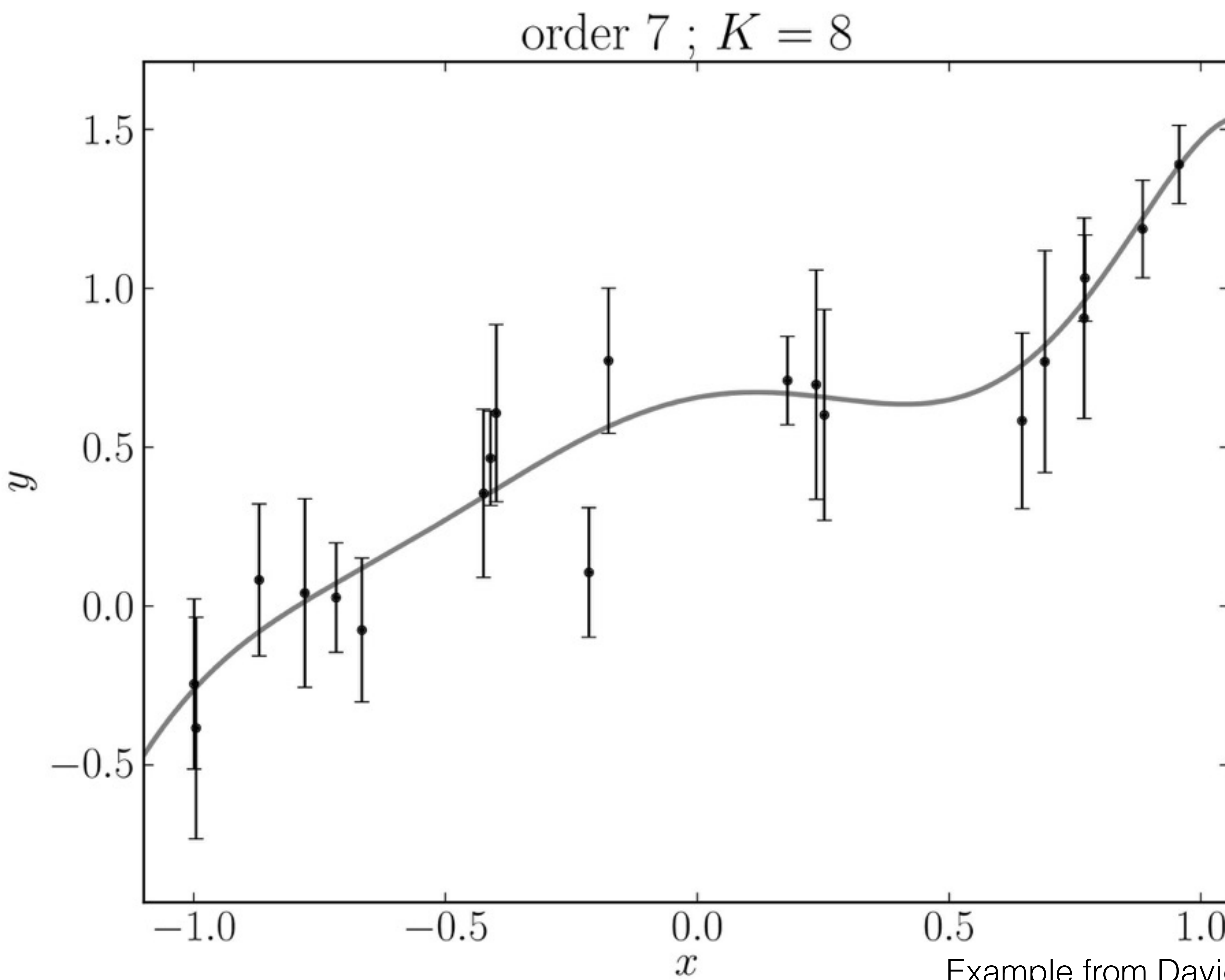
Example from David Hogg



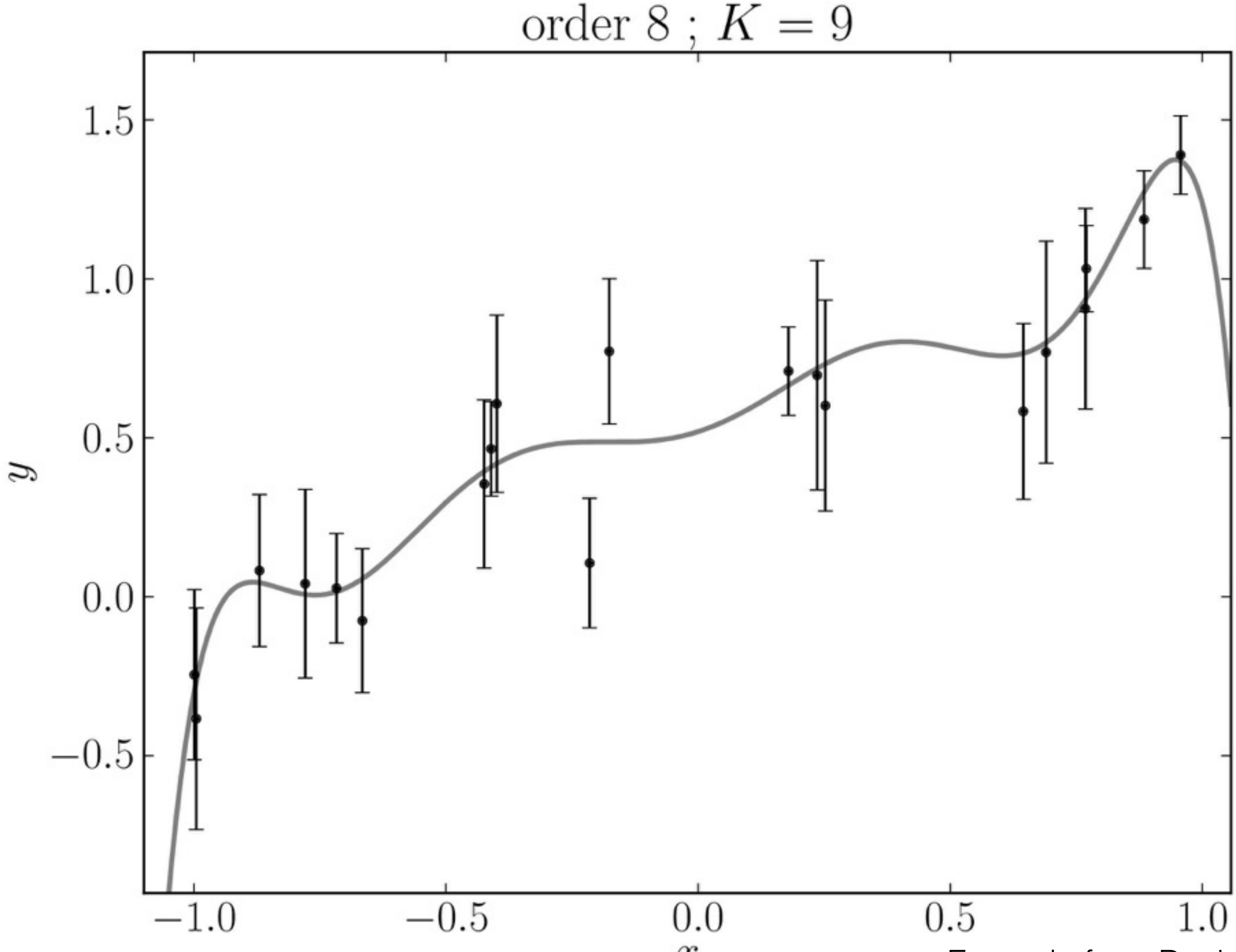
Example from David Hogg



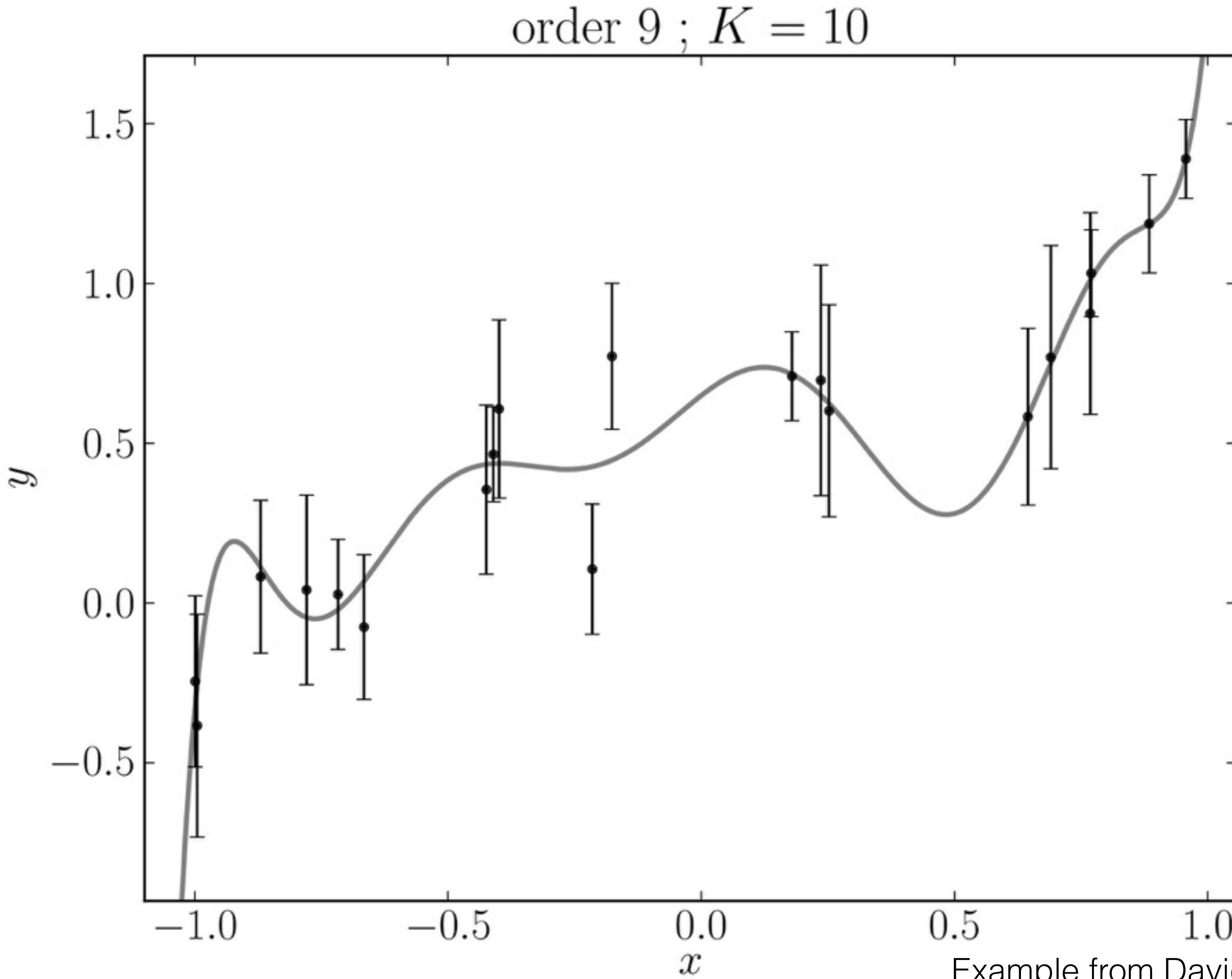
Example from David Hogg



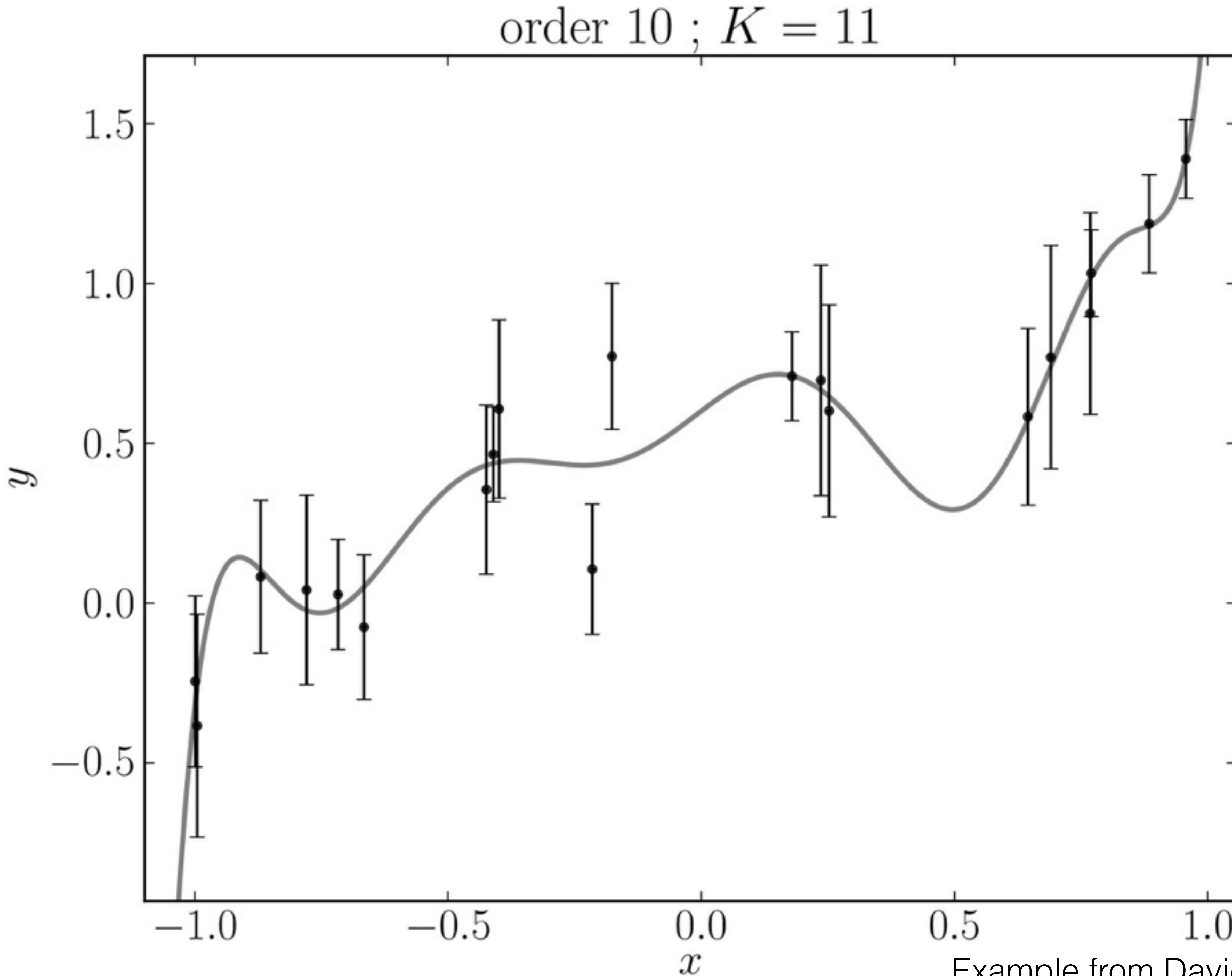
Example from David Hogg



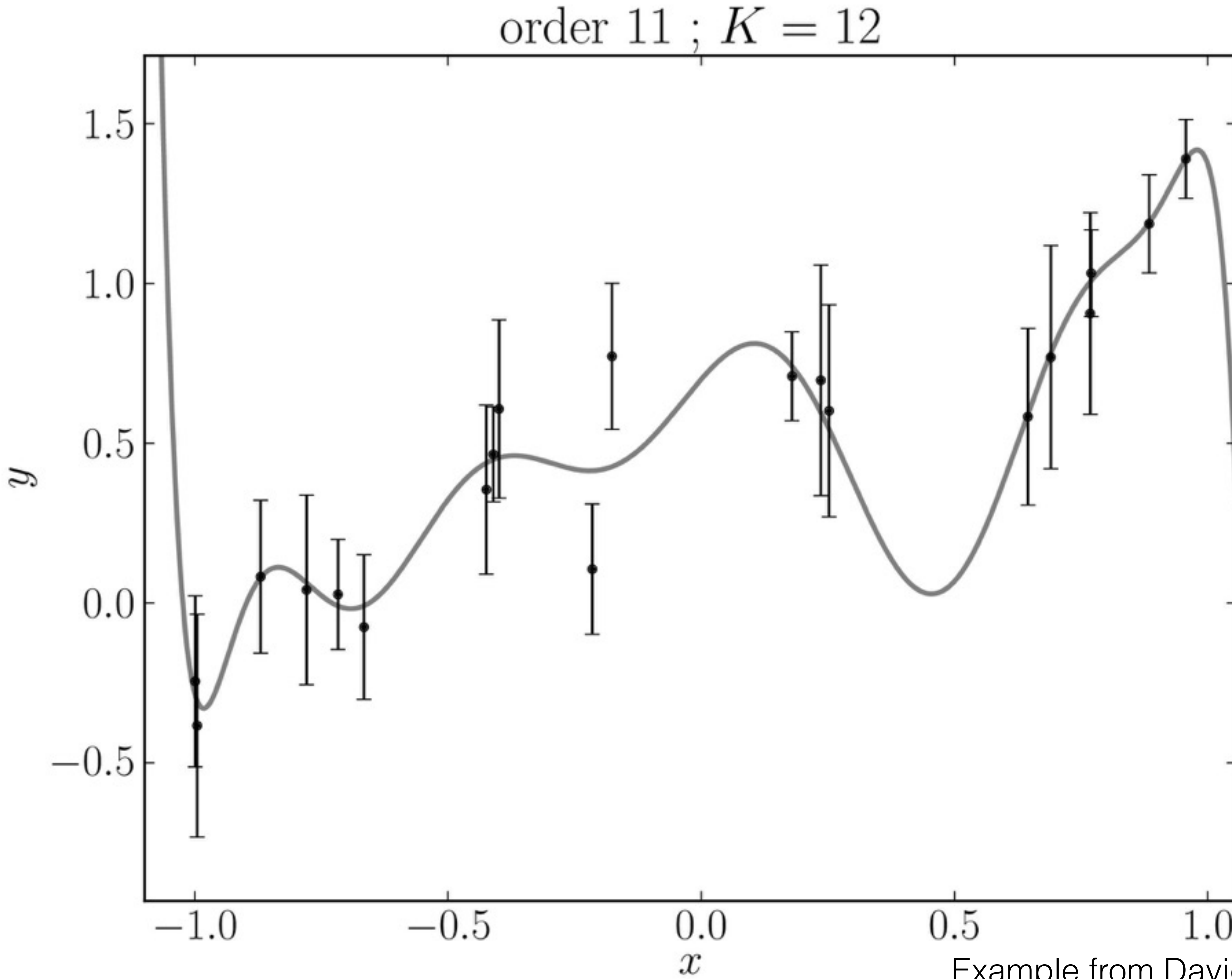
Example from David Hogg



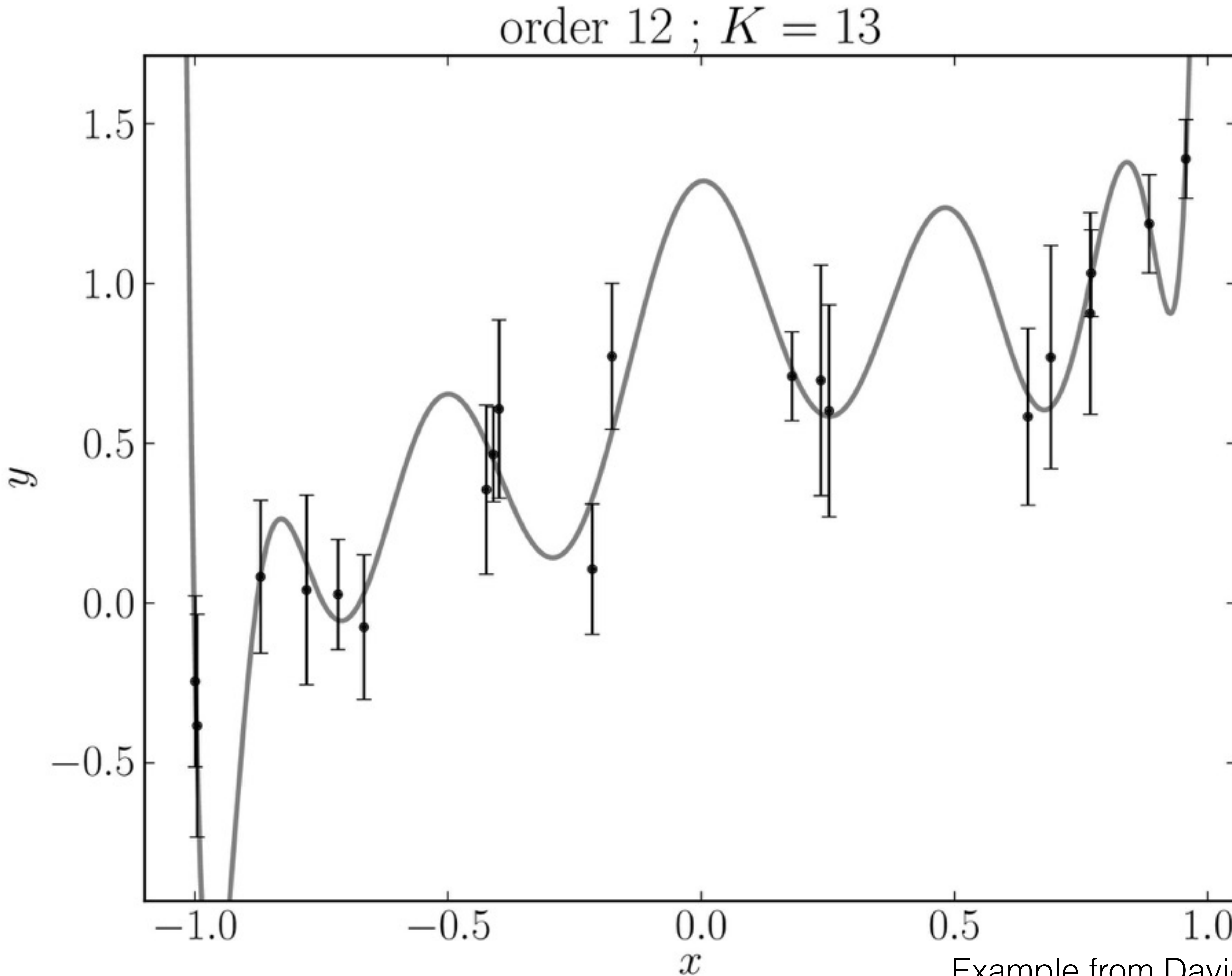
Example from David Hogg



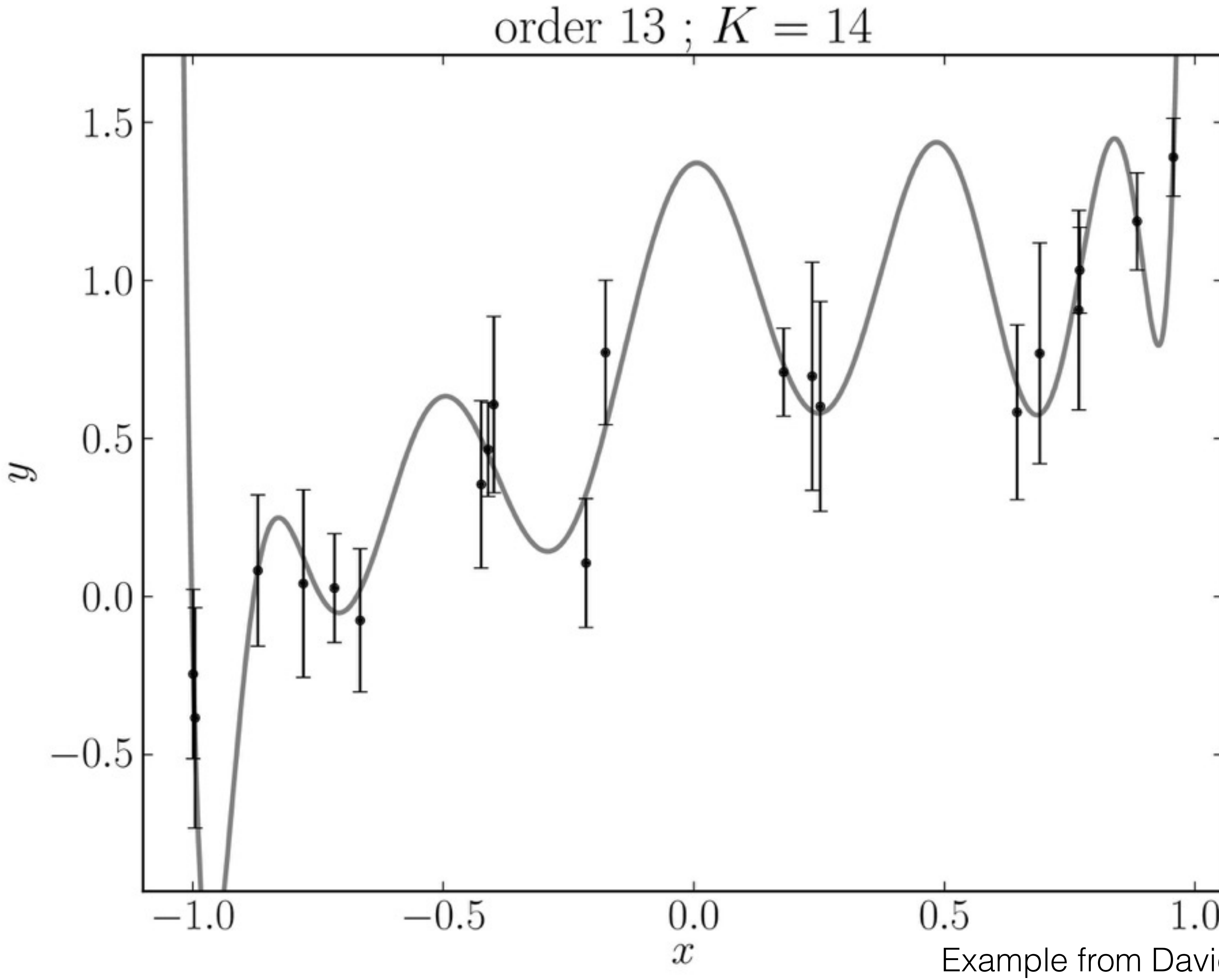
Example from David Hogg



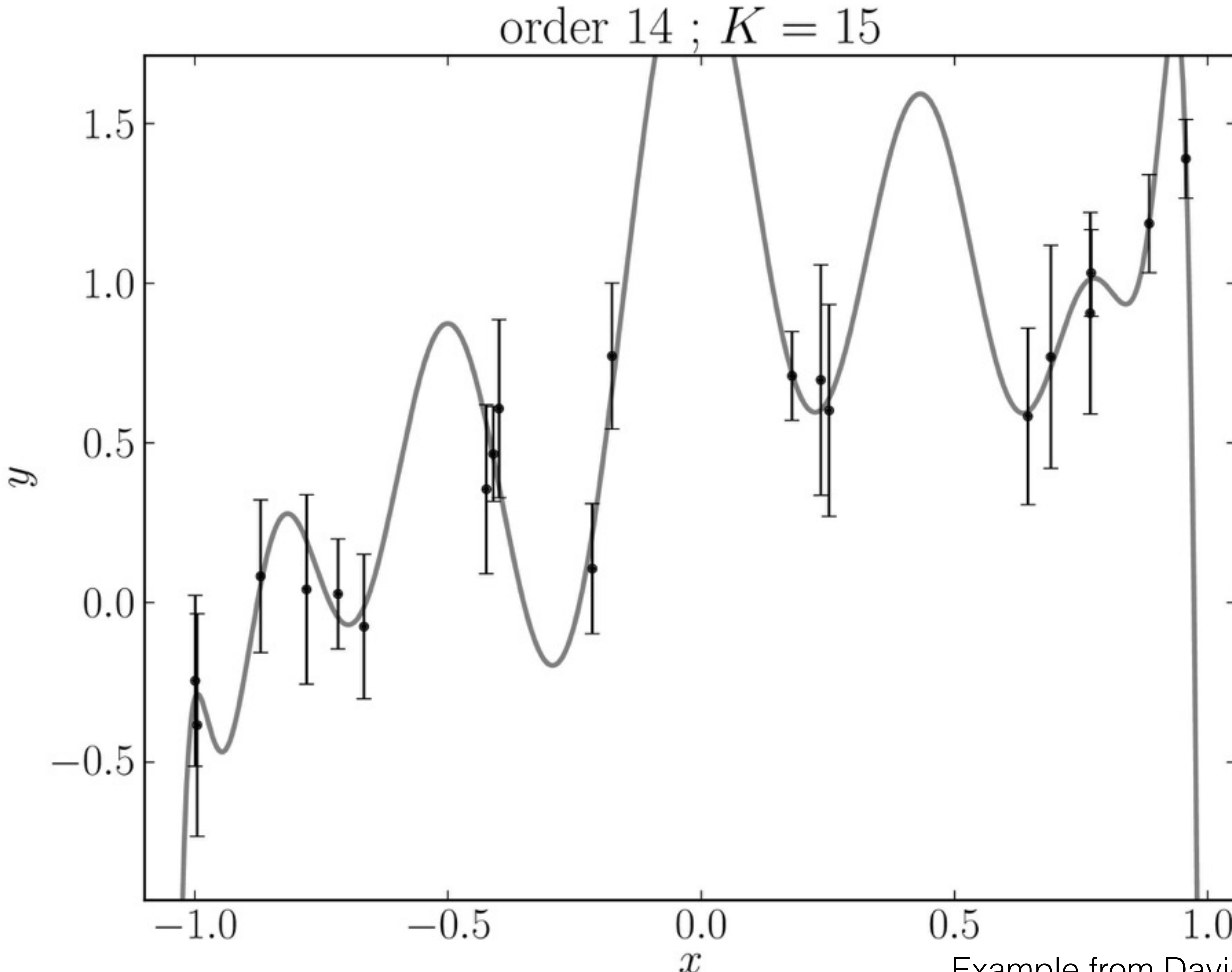
Example from David Hogg



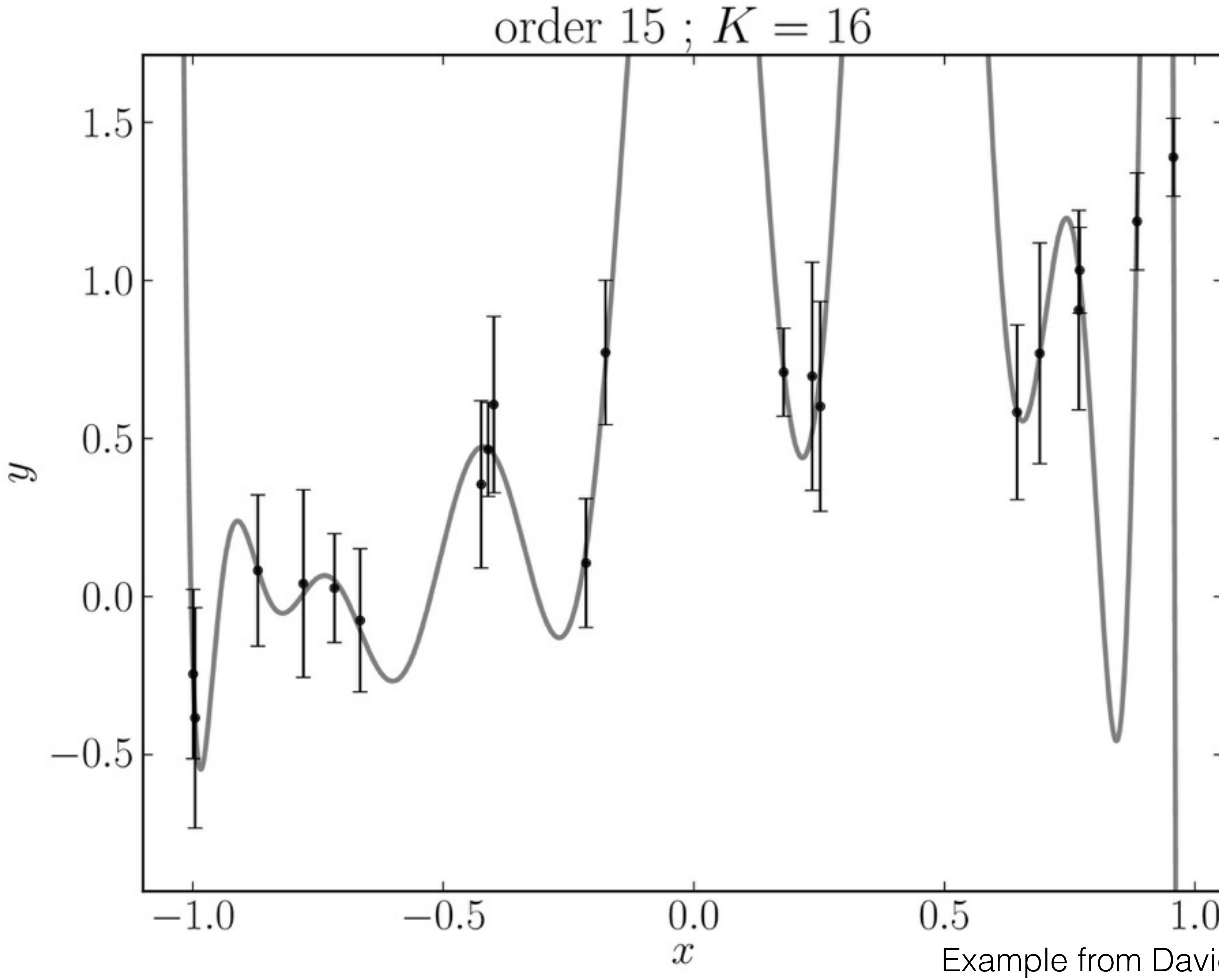
Example from David Hogg



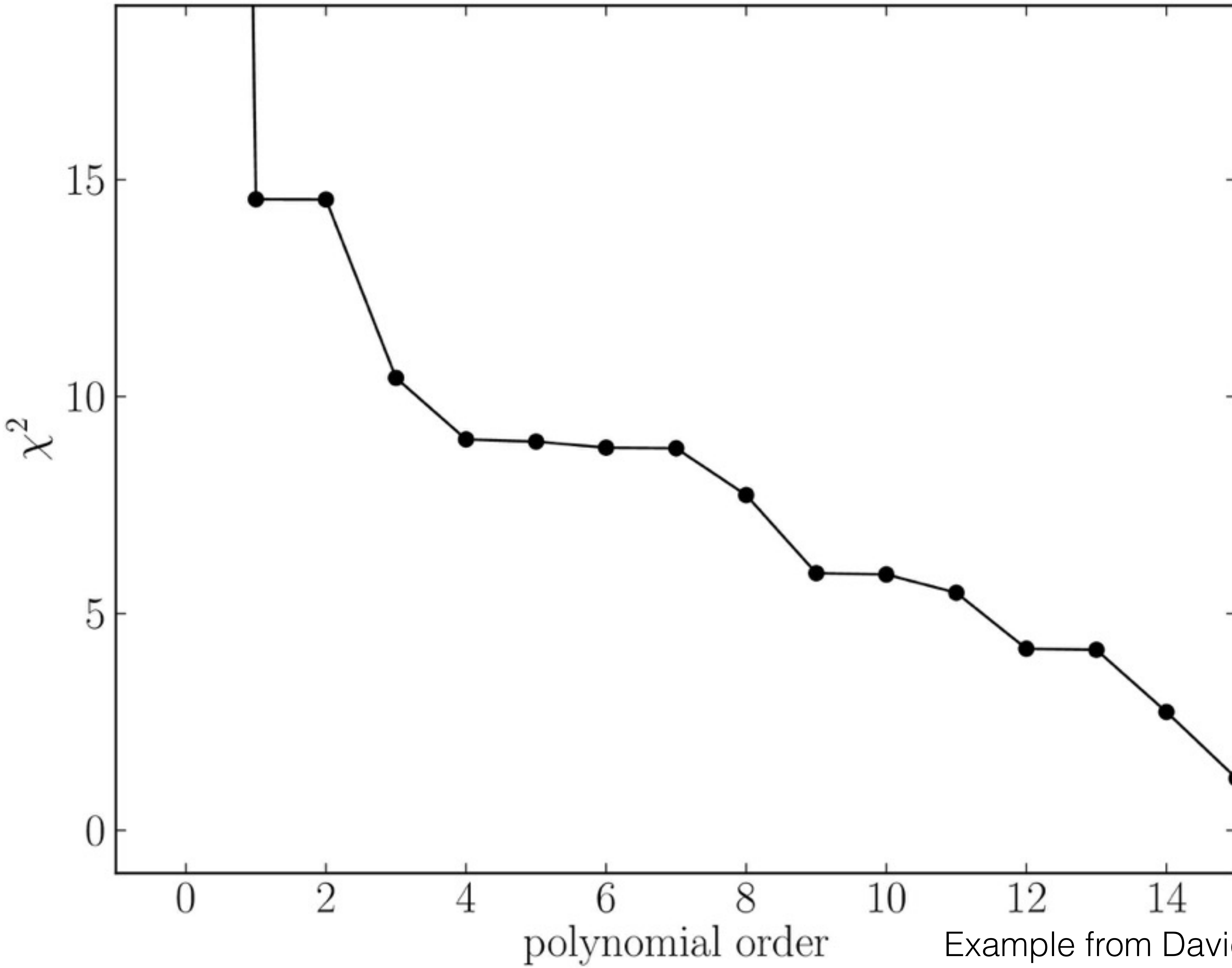
Example from David Hogg



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 = $\sum_i \chi_i^2$

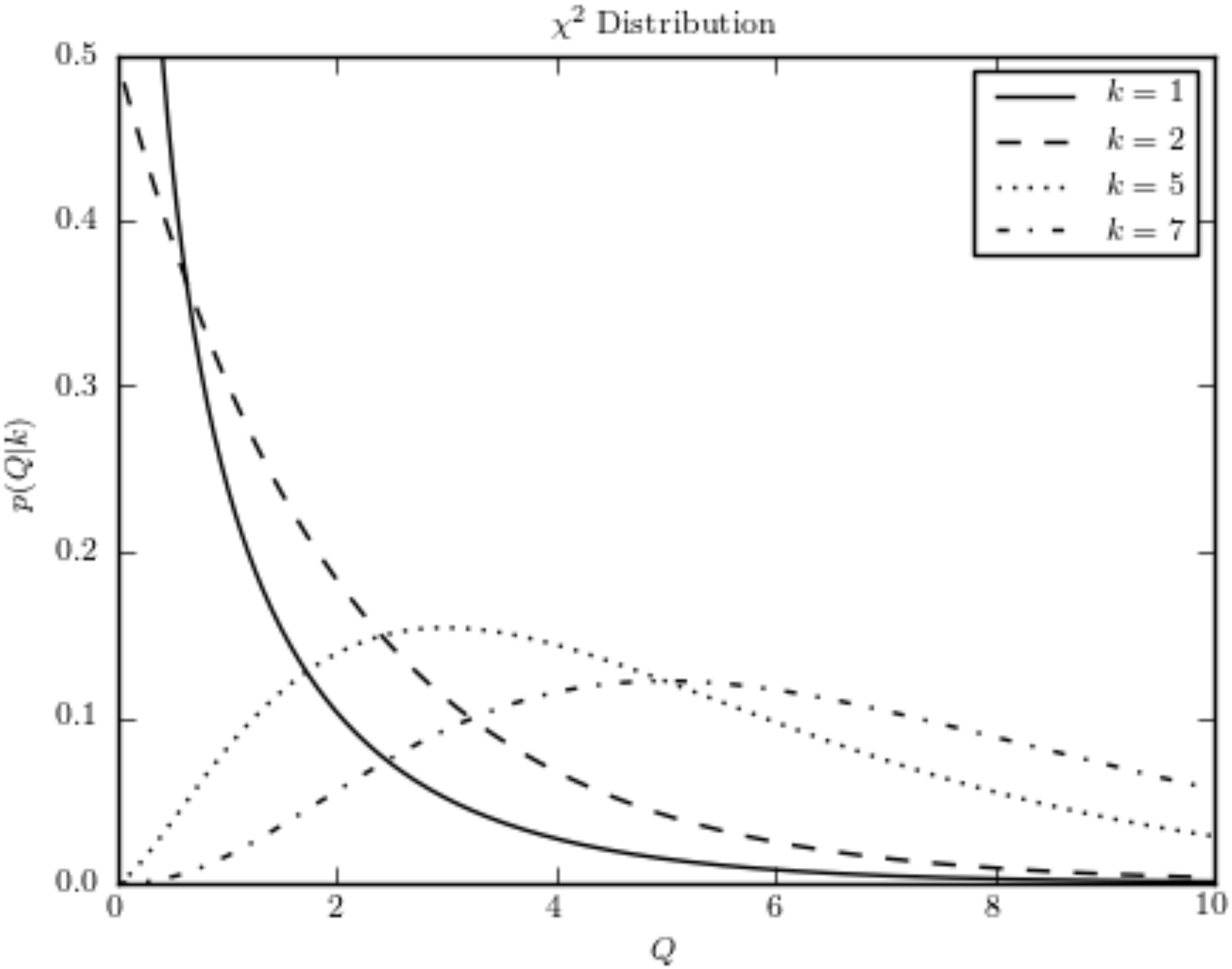
- Number of degrees of freedom = # data points - # of fit parameters

- 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*
- Basis for *chi-squared-per-degree-of-freedom* goodness-of-fit
- Central limit theorem: for $k \to \infty$, $p(x|k) \to N(x|k,2k)$

Chi-squared distribution



Ivezic et al. (2014)

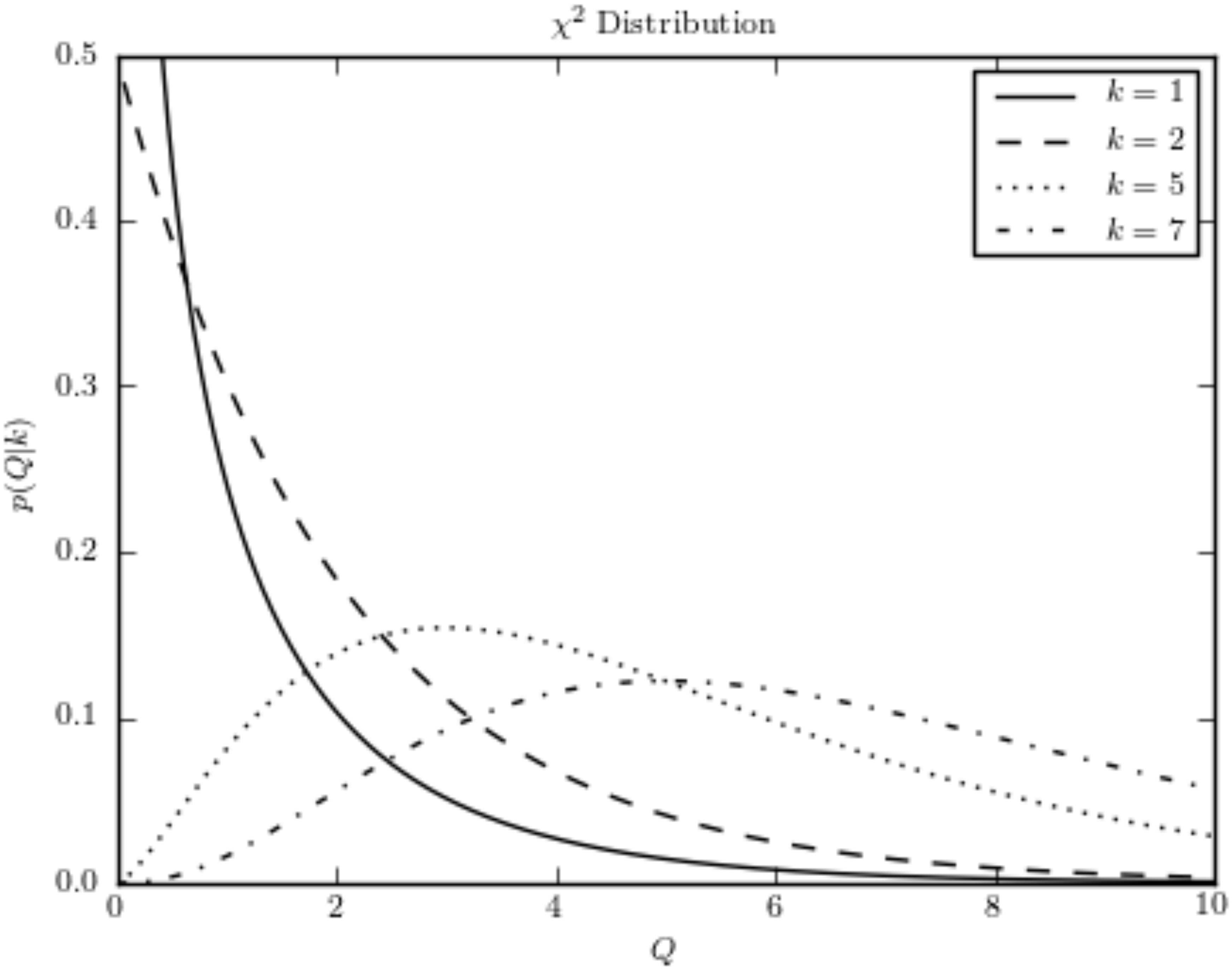
Chi-squared/ degree-of-freedom

- *If* the likelihood is Gaussian: e.g., $p(y_i|m,b,x_i,\sigma_{y,i}) = N(y_i|mx_i+b,\sigma^2_{y,i})$
- 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*, Σ_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 chi$_i$^2 are independent —> Σ_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 $\rightarrow \infty$, p(x|N-K) $\rightarrow$ N(x| N-K,2[N-K])
- Therefore, expected value of chi^2/dof ~ 1
- But really should be comparing chi^2 to dof with typical scatter $\sqrt{2 dof}$

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* (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 chi^2/dof

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, chi^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 chi^2
  5. Repeat to form p(chi^2)
  6. Where does chi^2 for the actual data lie in this distribution?

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

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

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

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

- When underfitting, chi^2 will be large

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

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

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

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

- Suppose you have 1 parameter (e.g., fit constant y =b rather than y = m x + b)

- p(b|data) = Chi^2(Δchi^2,1 dof)

- 68% confidence limit on b: that value for which Δchi^2 = 1

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

In [8]: stats.chi2.ppf(0.68,1)
Out[8]: 0.98894648147802289
```

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

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

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

In [12]: stats.chi2.ppf(0.68,2)
Out[12]: 2.278685663767296
```

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

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

Model selection with AIC

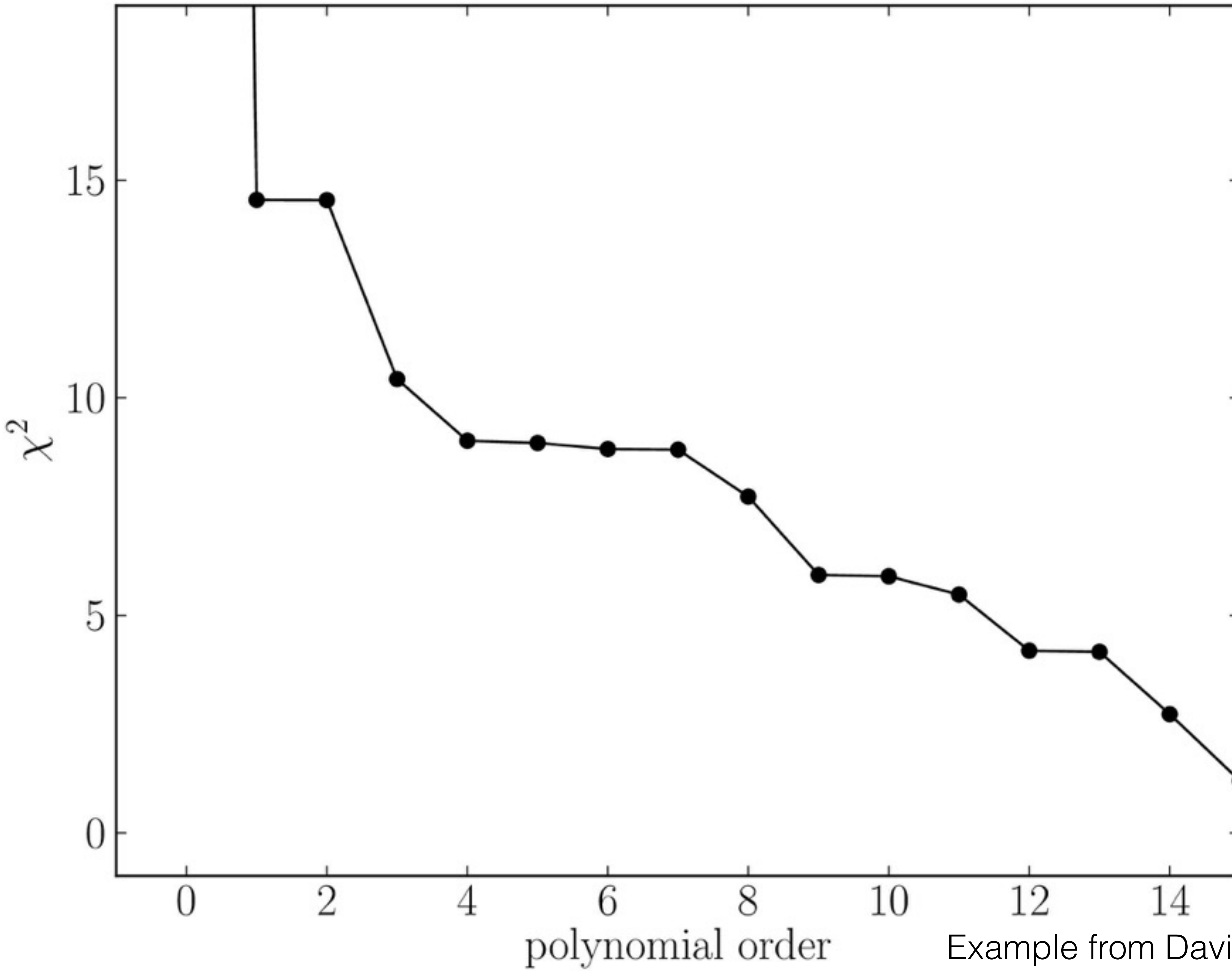
- AIC = Akaike Information Criterion

- AIC = -2 ln L + 2 K = chi^2 + 2 K

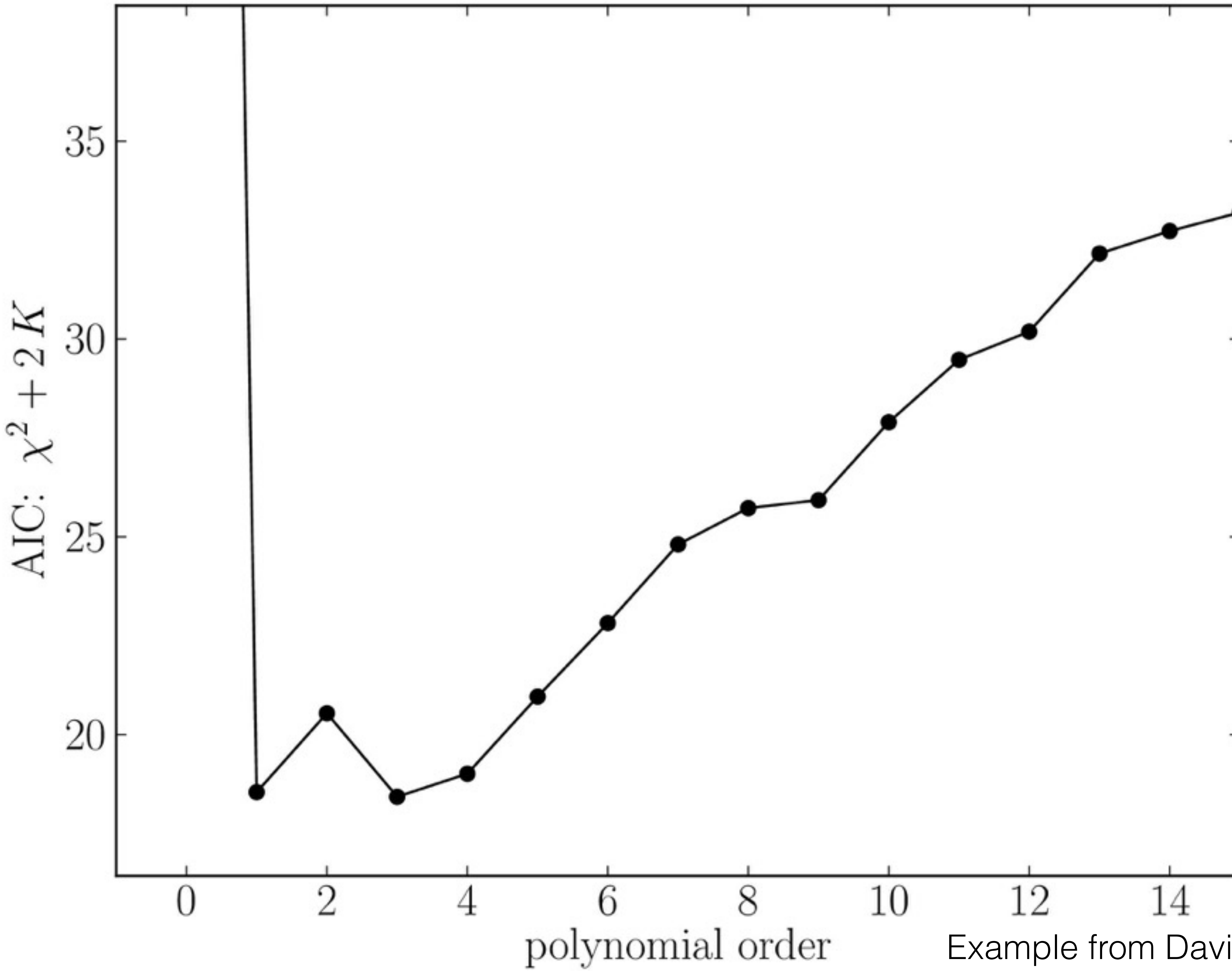
- Delta AIC = Asymptotically the amount of information lost when using worse model G1 than better model G2

- 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 method 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)]

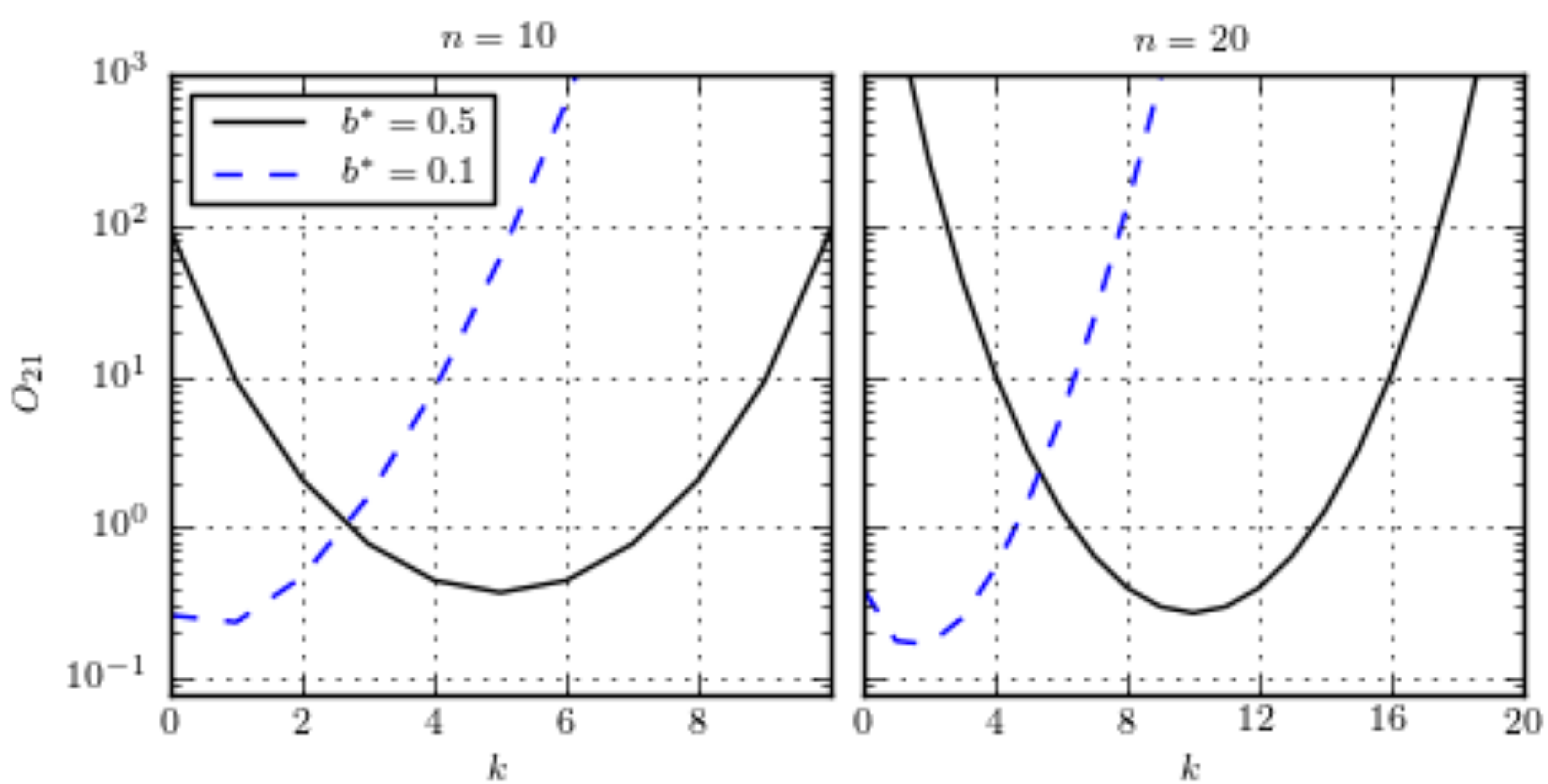
$$\text{odds ratio} = \frac{p(\text{data}|\text{linear})}{p(\text{data}|\text{quadratic})} \times \frac{p(\text{linear})}{p(\text{quadratic})}$$

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

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

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**:

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

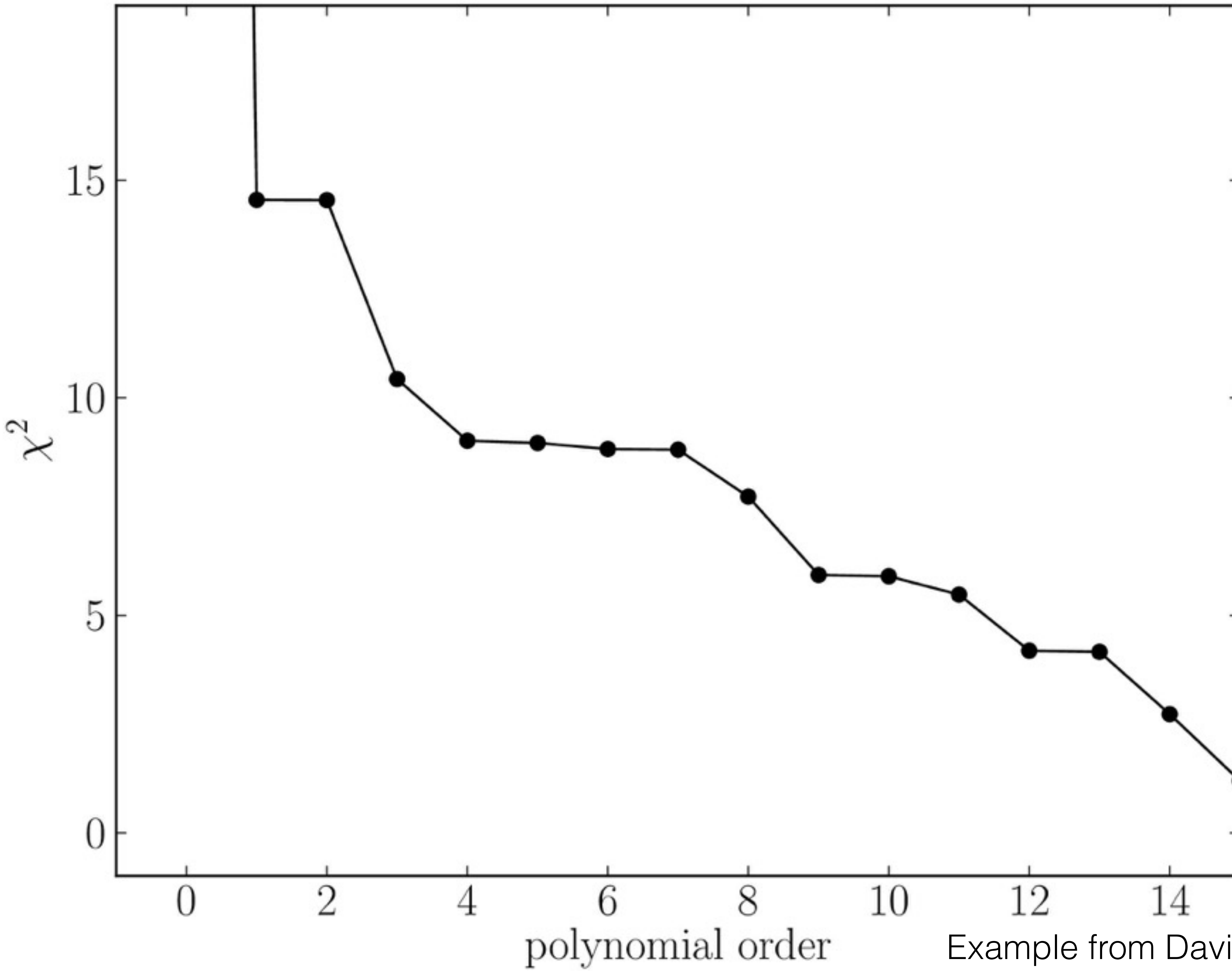
Bayesian Information Criterion (BIC)

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

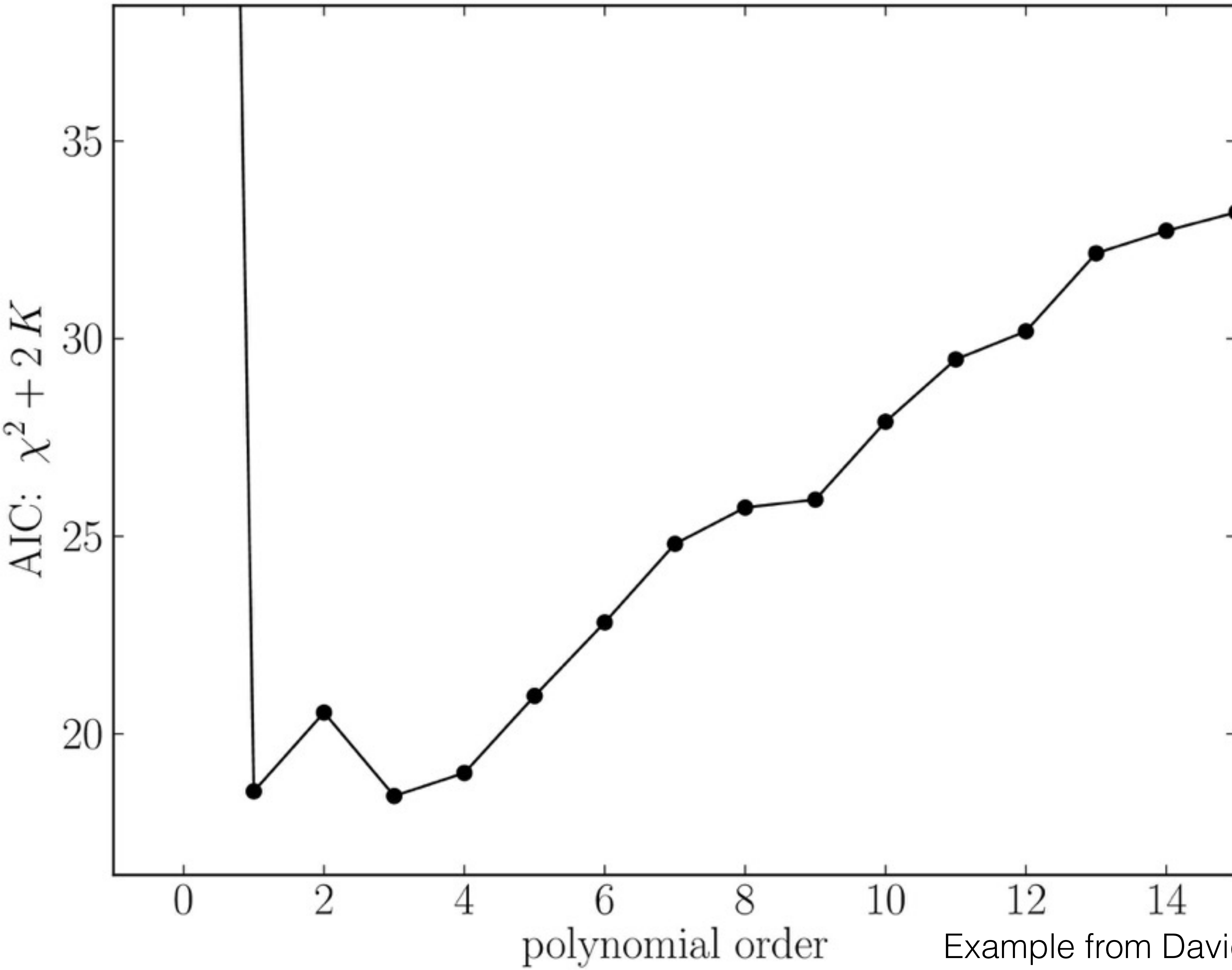
  Evidence ~ BIC = -2 ln L + K ln[N]

  for K parameters and N data points

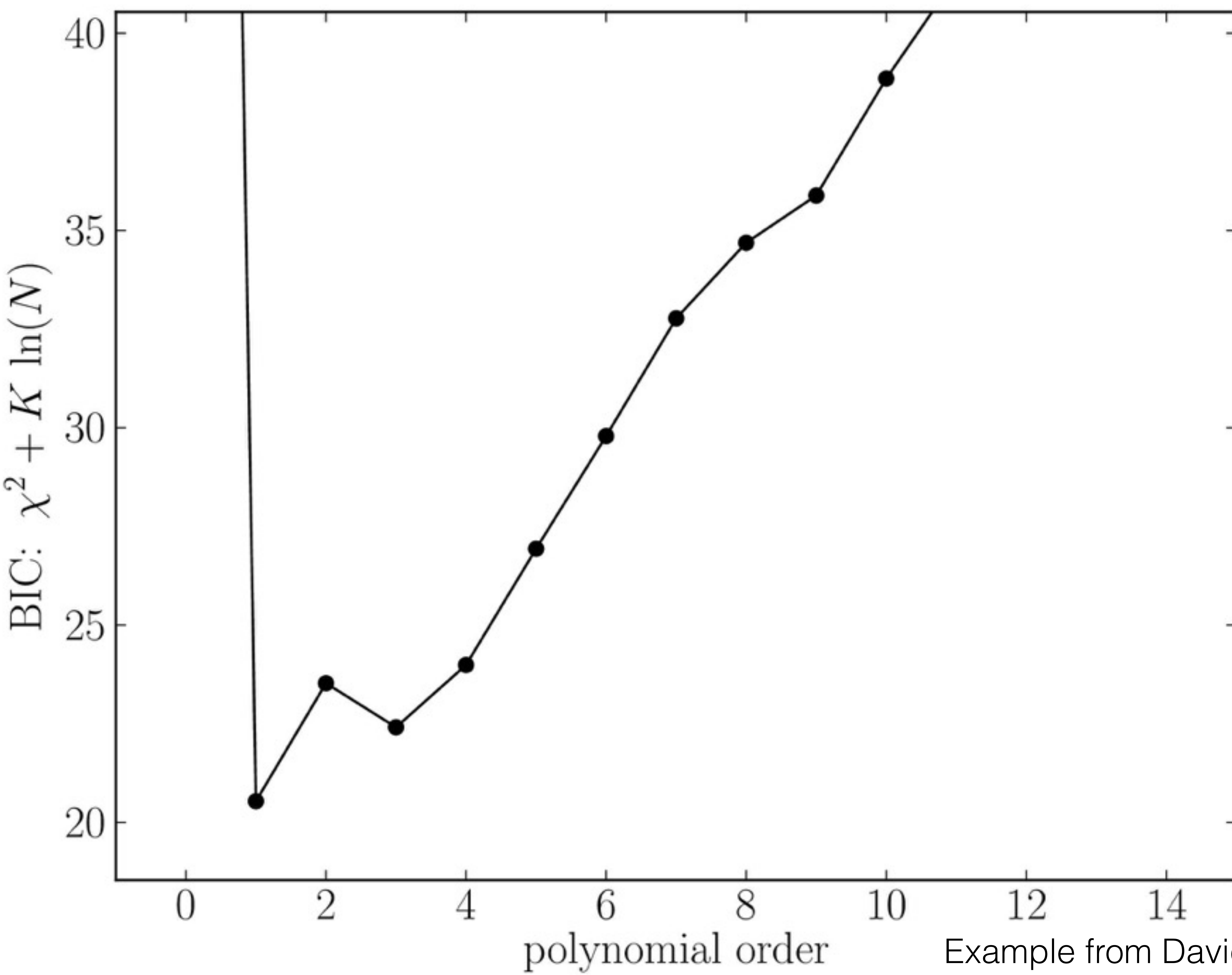
- Similar to AIC, but Bayesian :-)



Example from David Hogg



Example from David Hogg



Example from David Hogg

Cross-validation

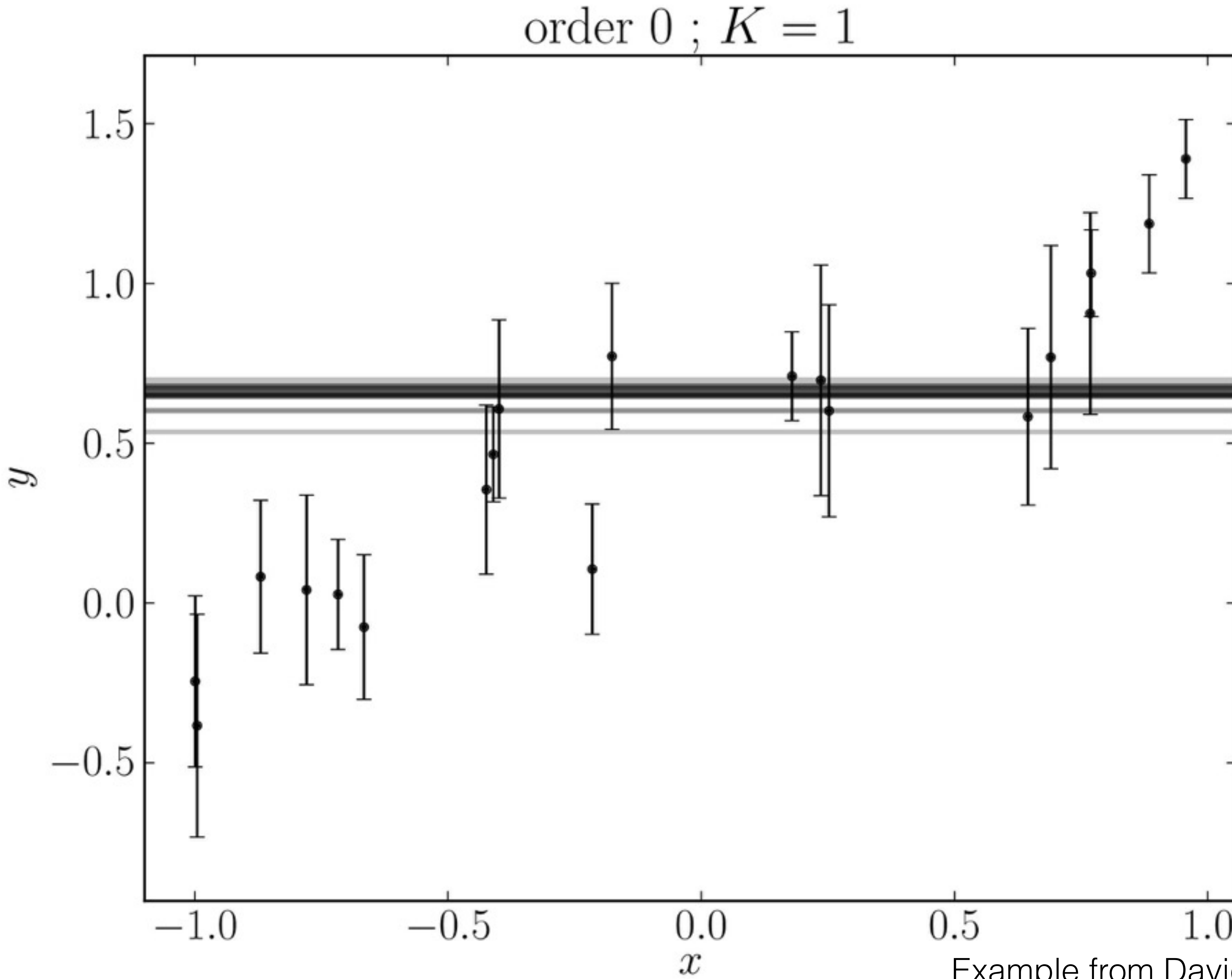
- All previous methods require many assumptions
- Take approach similar to bootstrap and jackknife before, using the data themselves to generate 'new' data
- Cross-validation is very similar to jackknife:

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

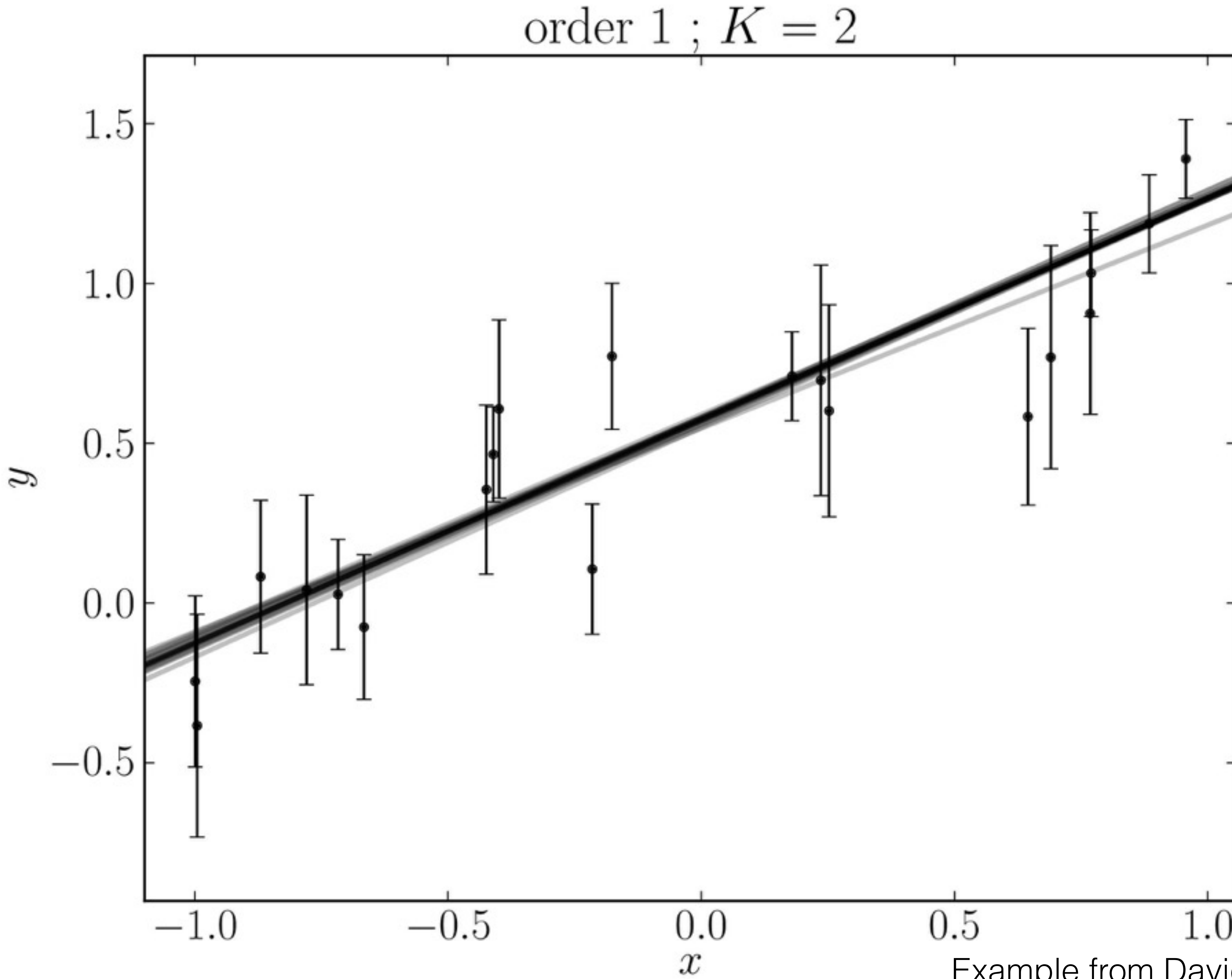
  2. Fit the model to each data set

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

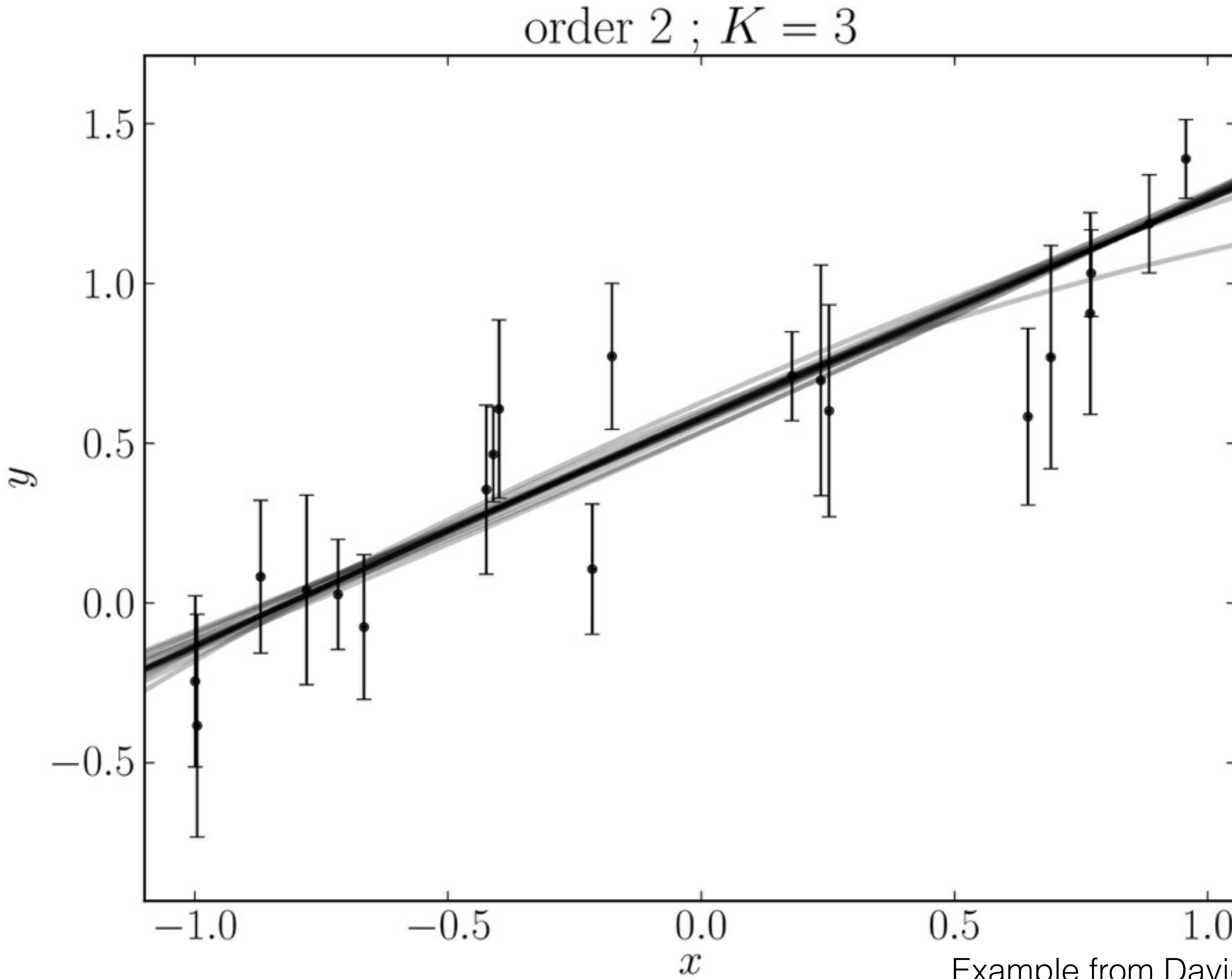
  4. Cross validation likelihood $L_{cval} = \mathrm{Prod}_i\, L_i$



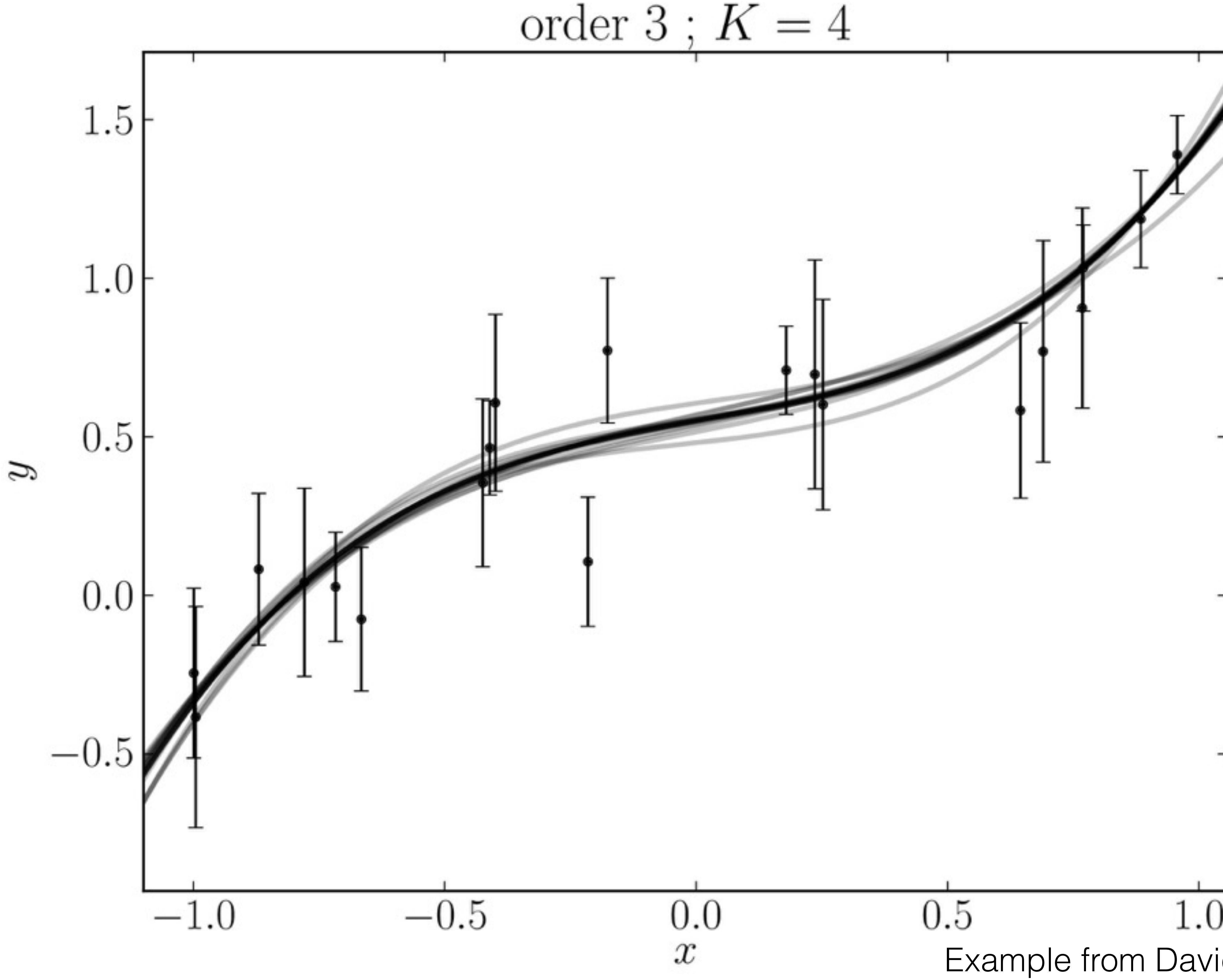
Example from David Hogg



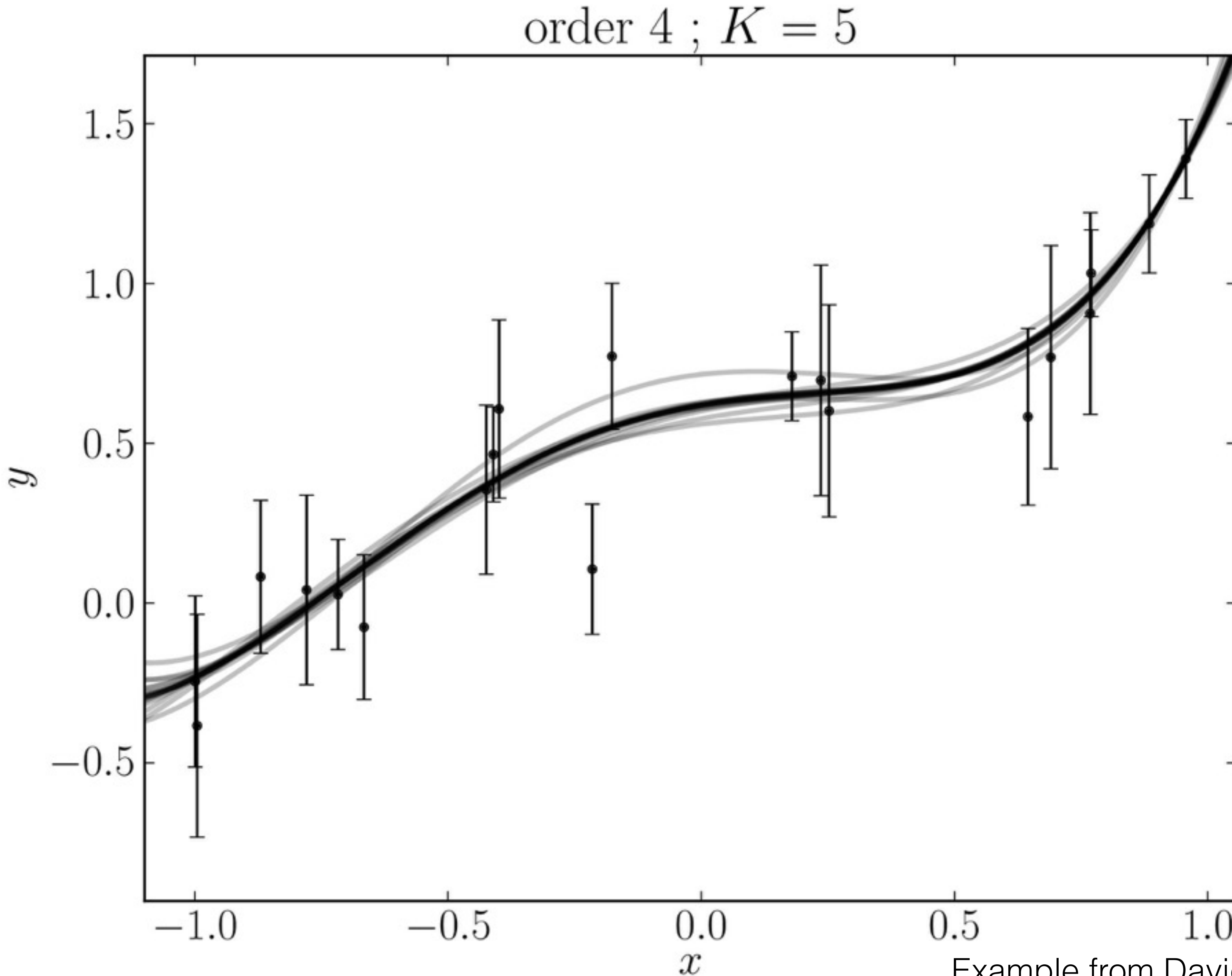
Example from David Hogg



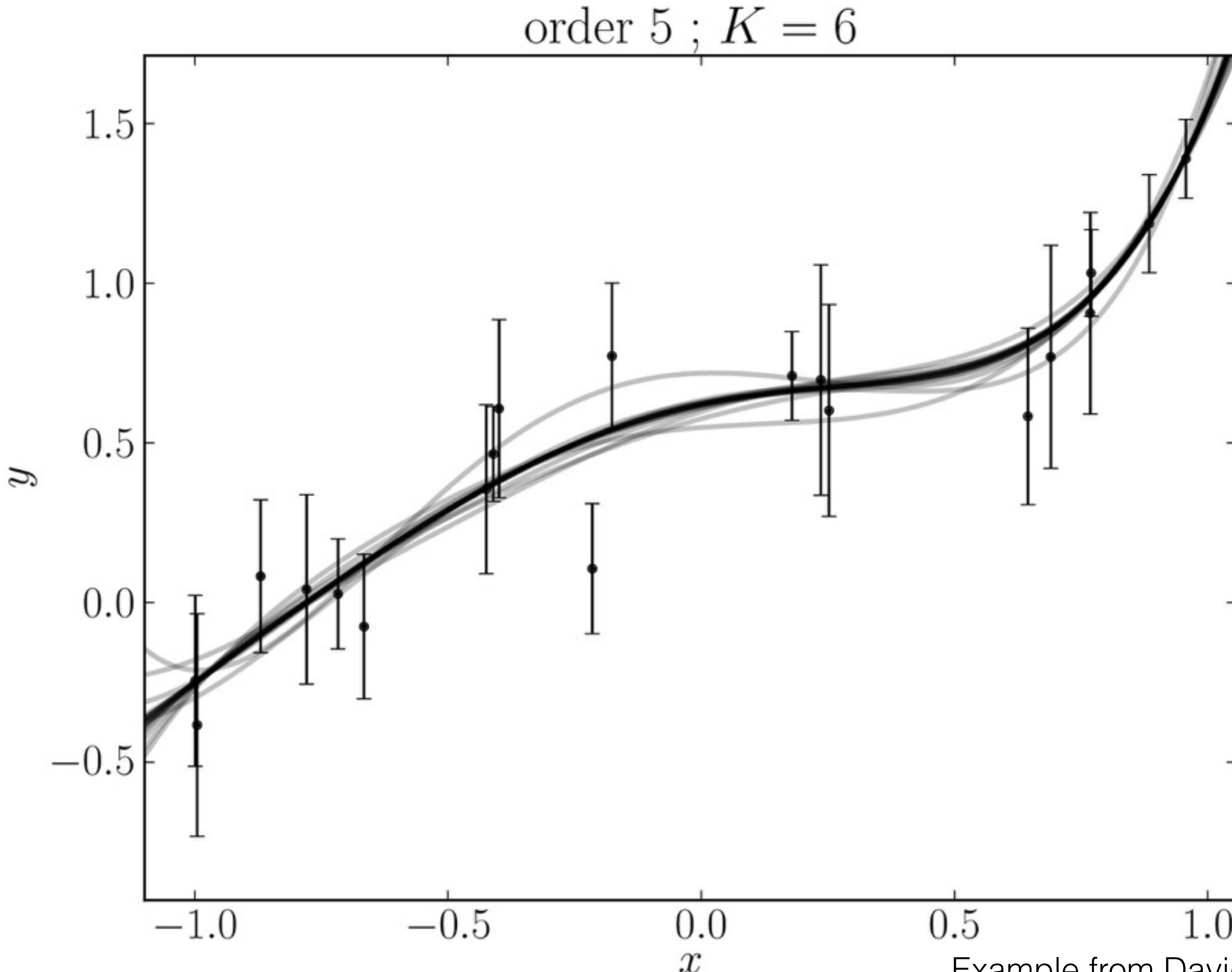
Example from David Hogg



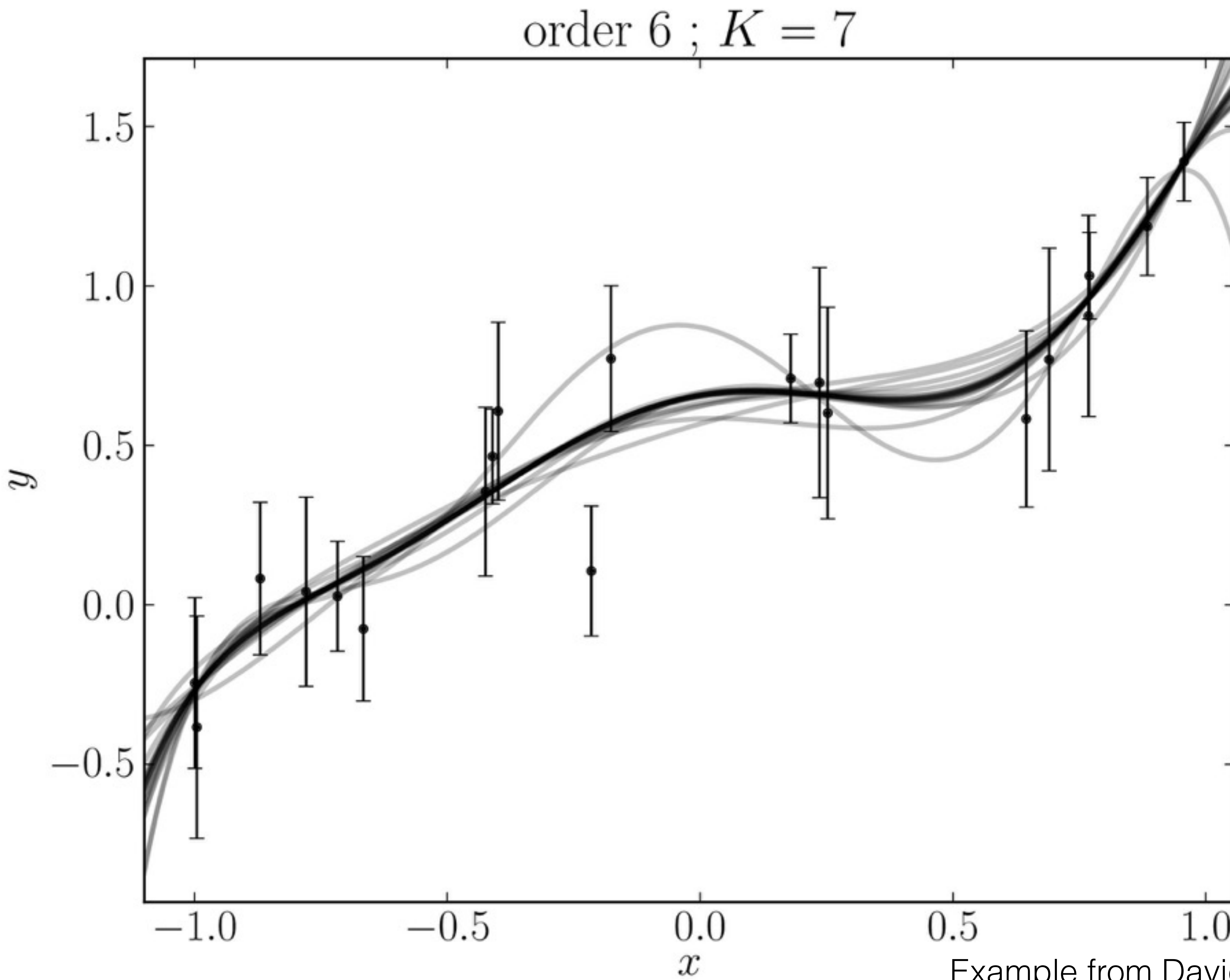
Example from David Hogg



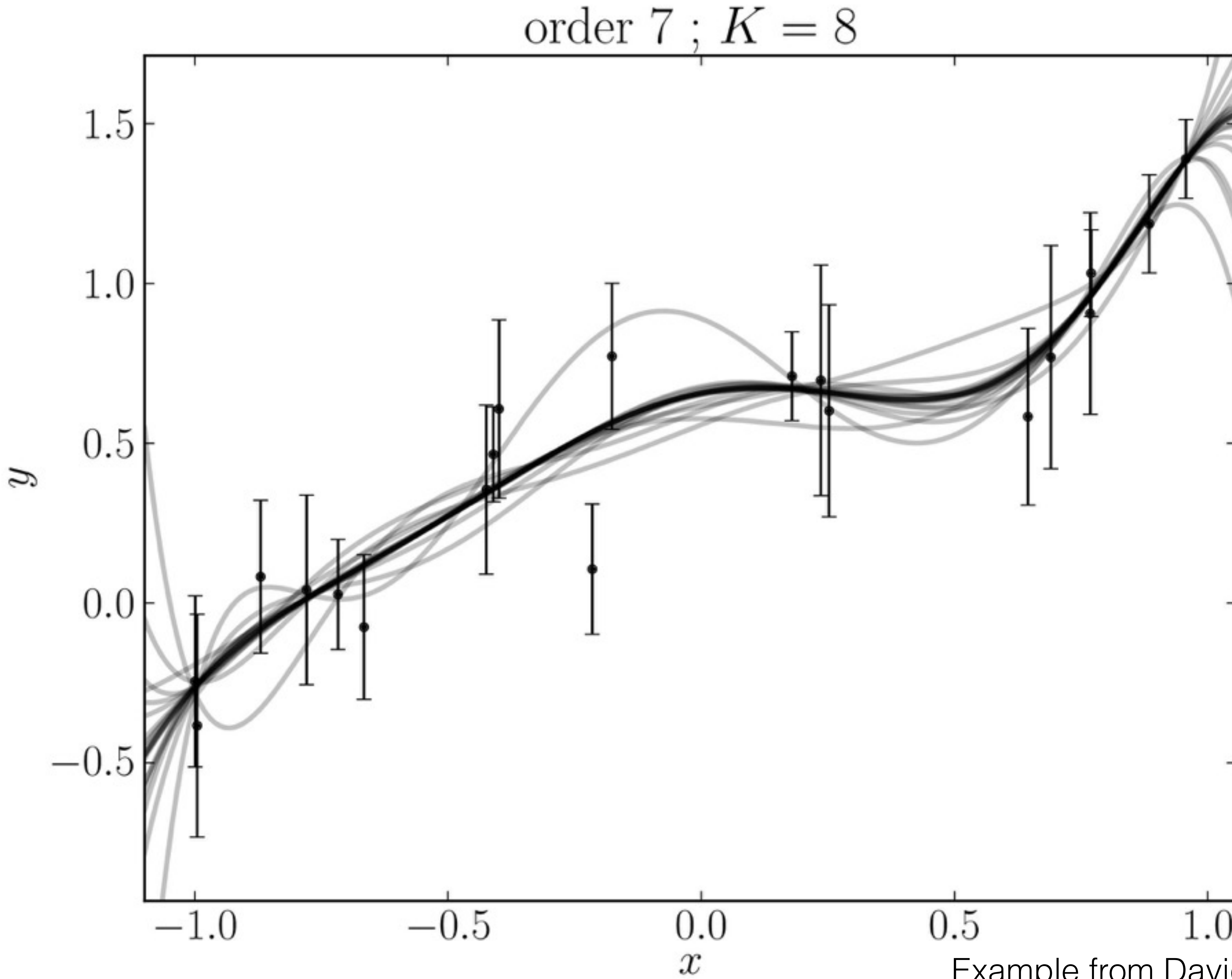
Example from David Hogg



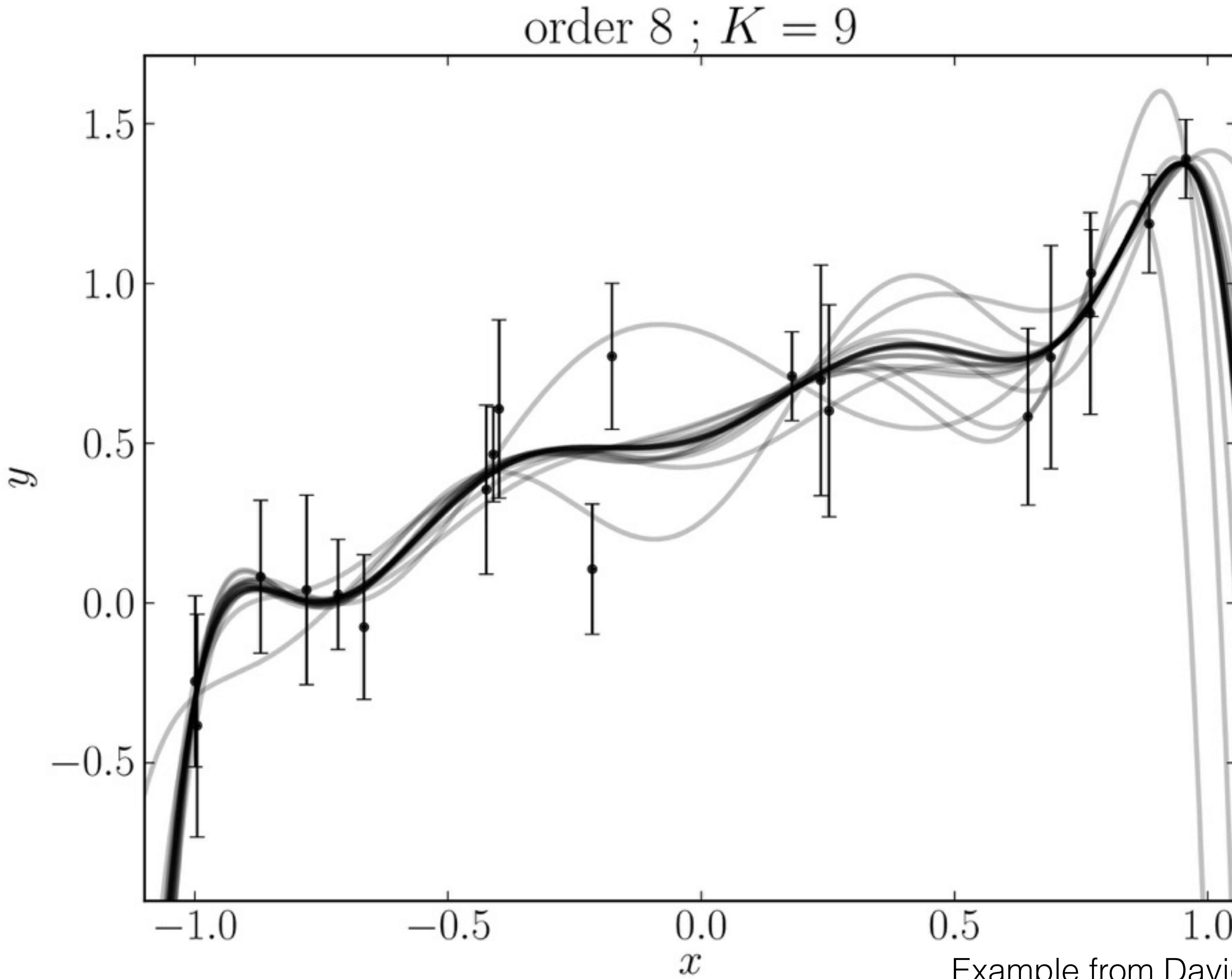
Example from David Hogg



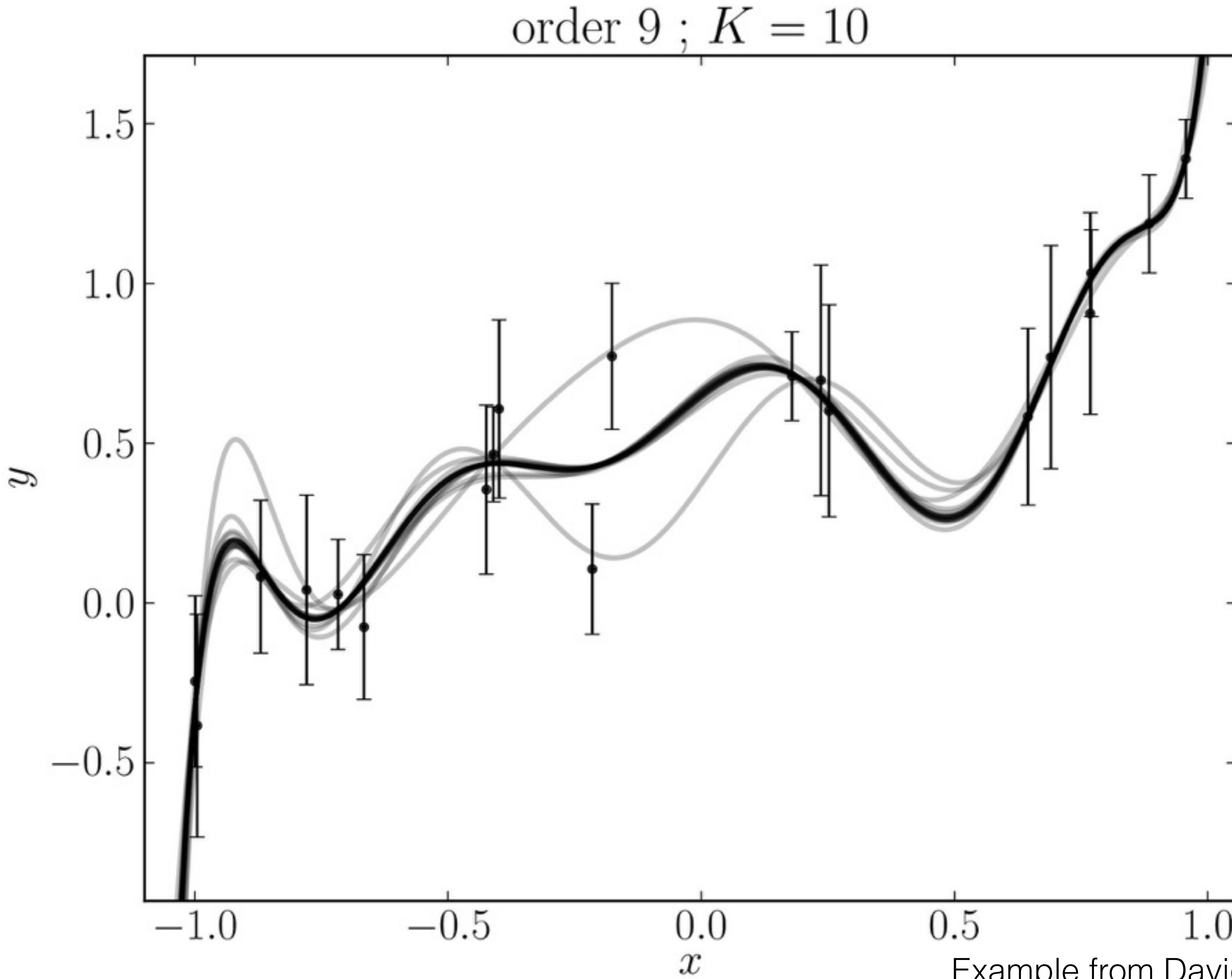
Example from David Hogg



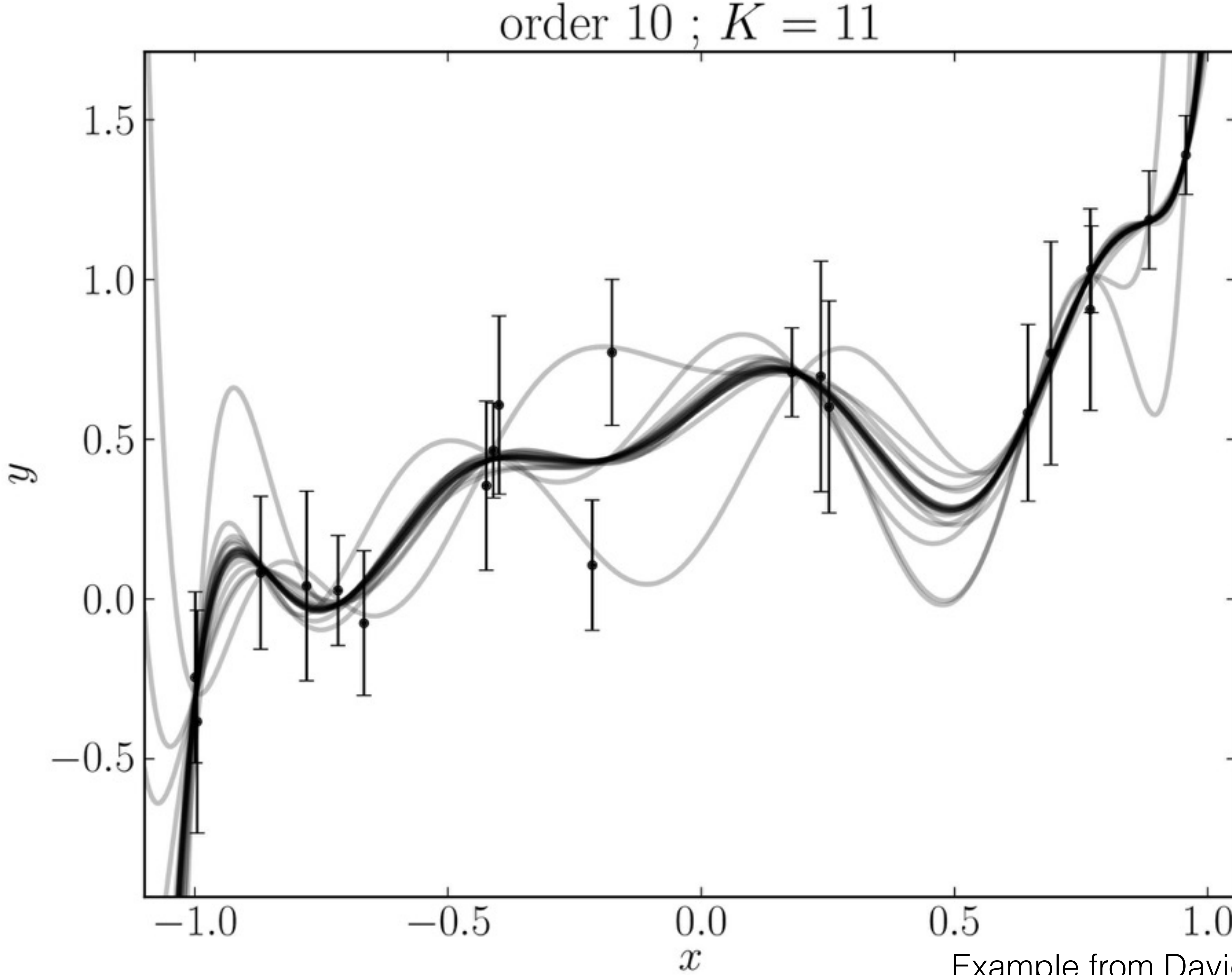
Example from David Hogg



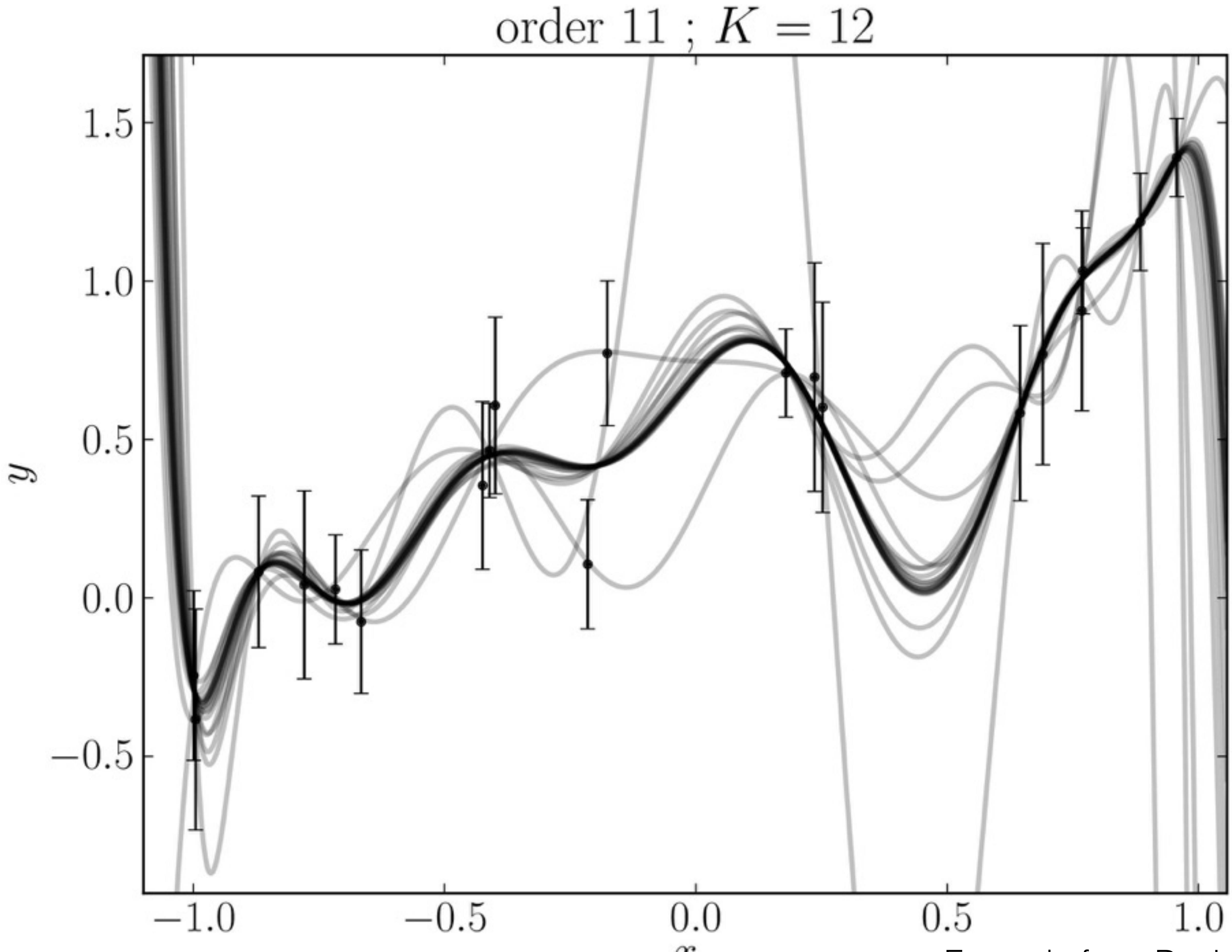
Example from David Hogg



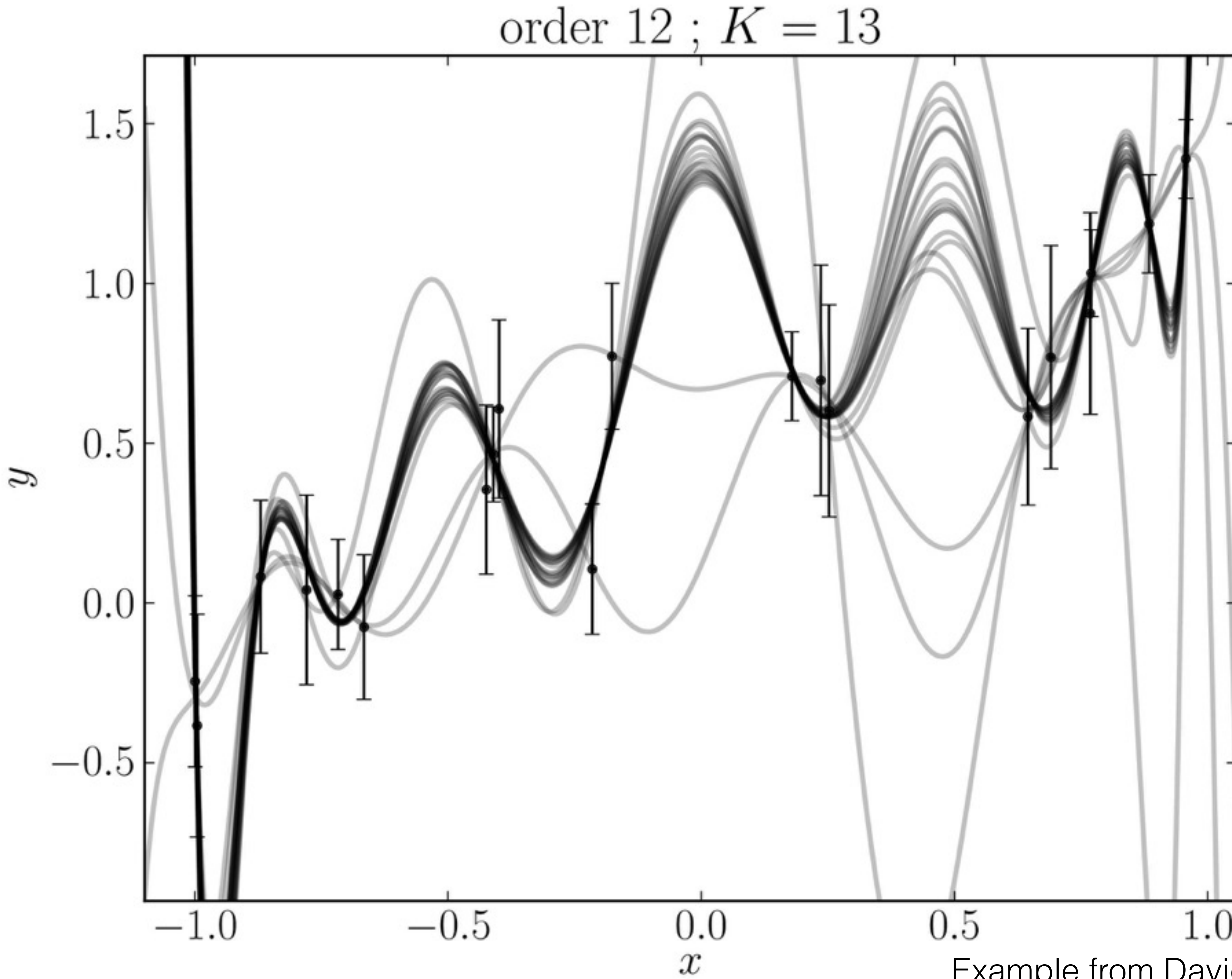
Example from David Hogg



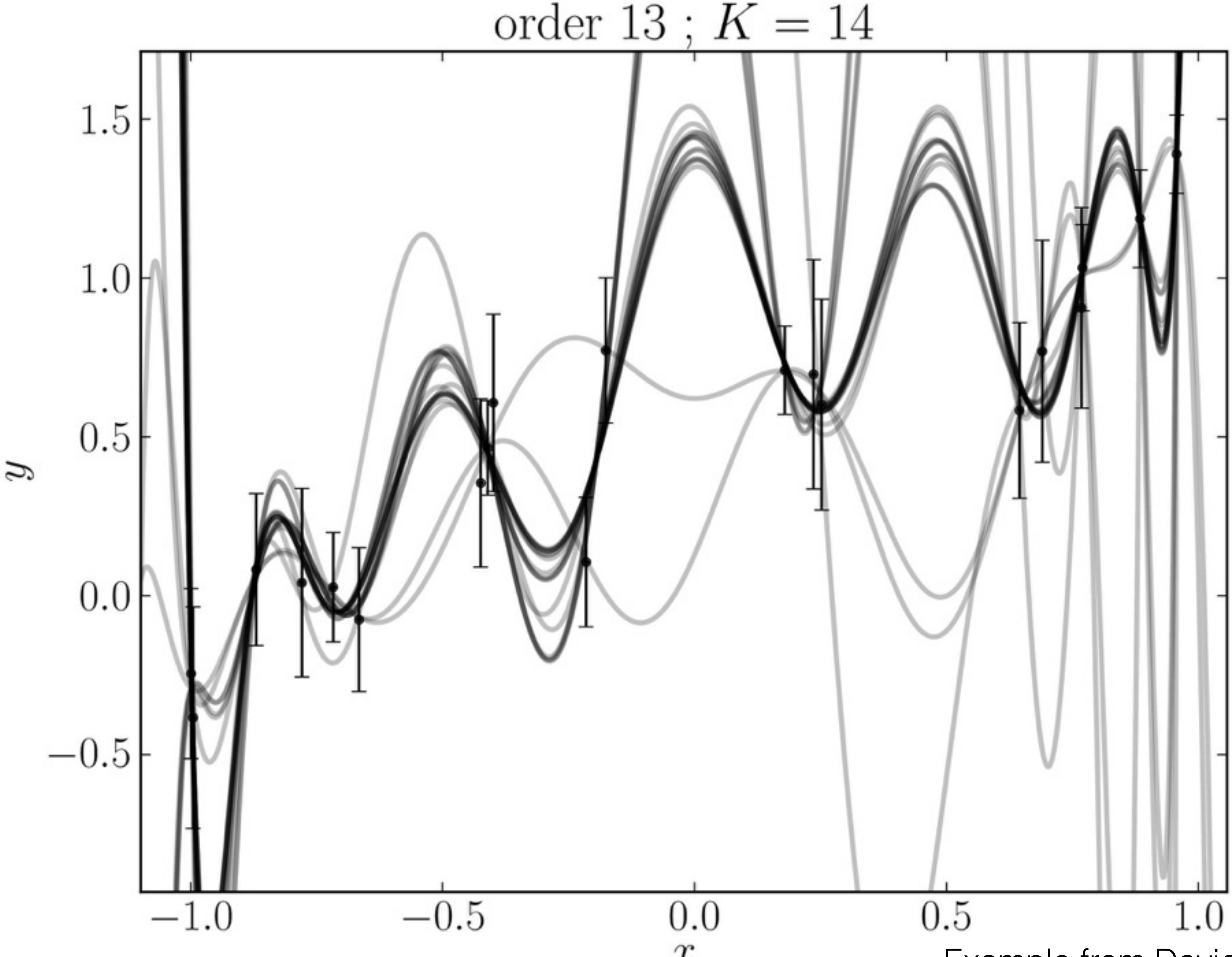
Example from David Hogg



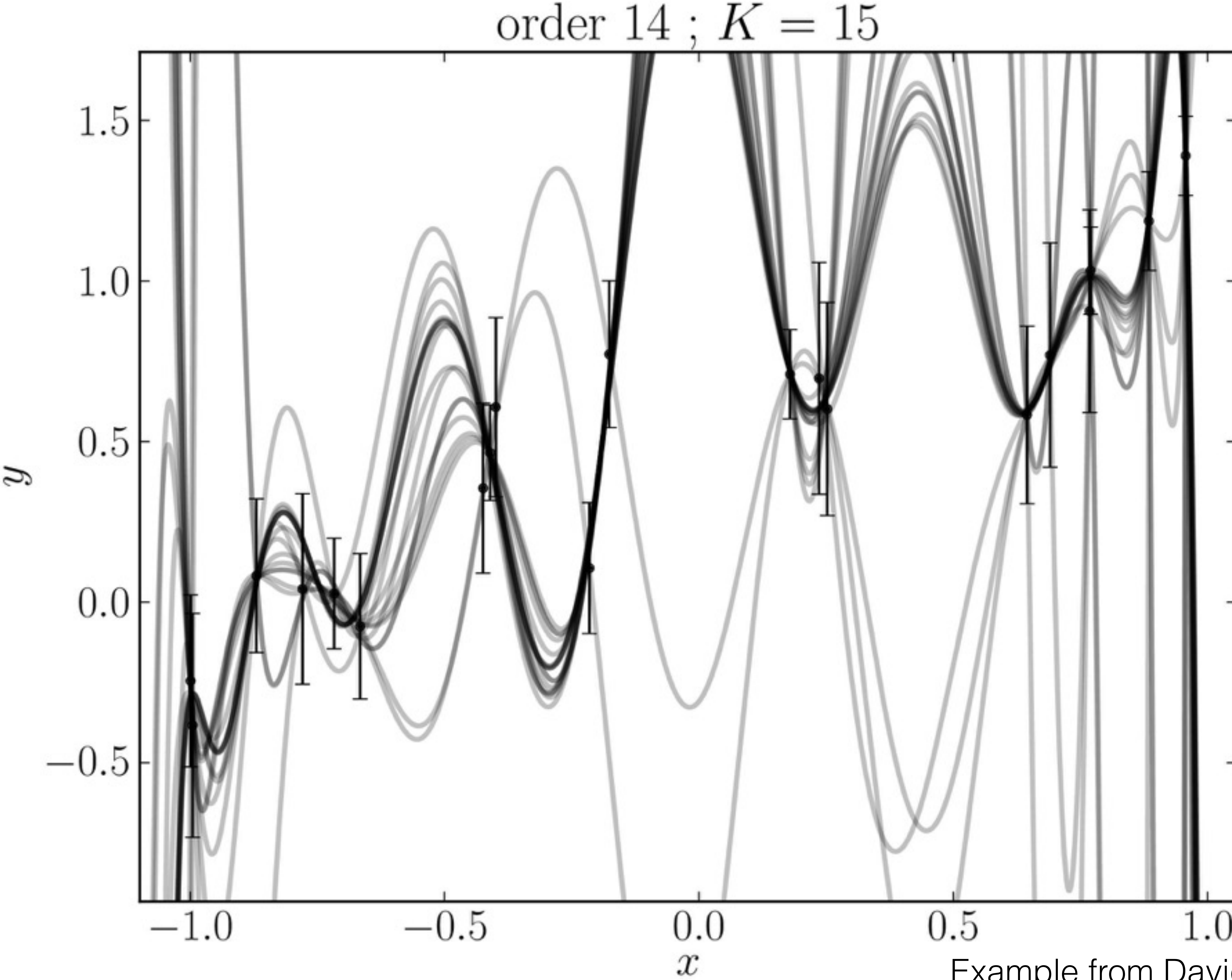
Example from David Hogg



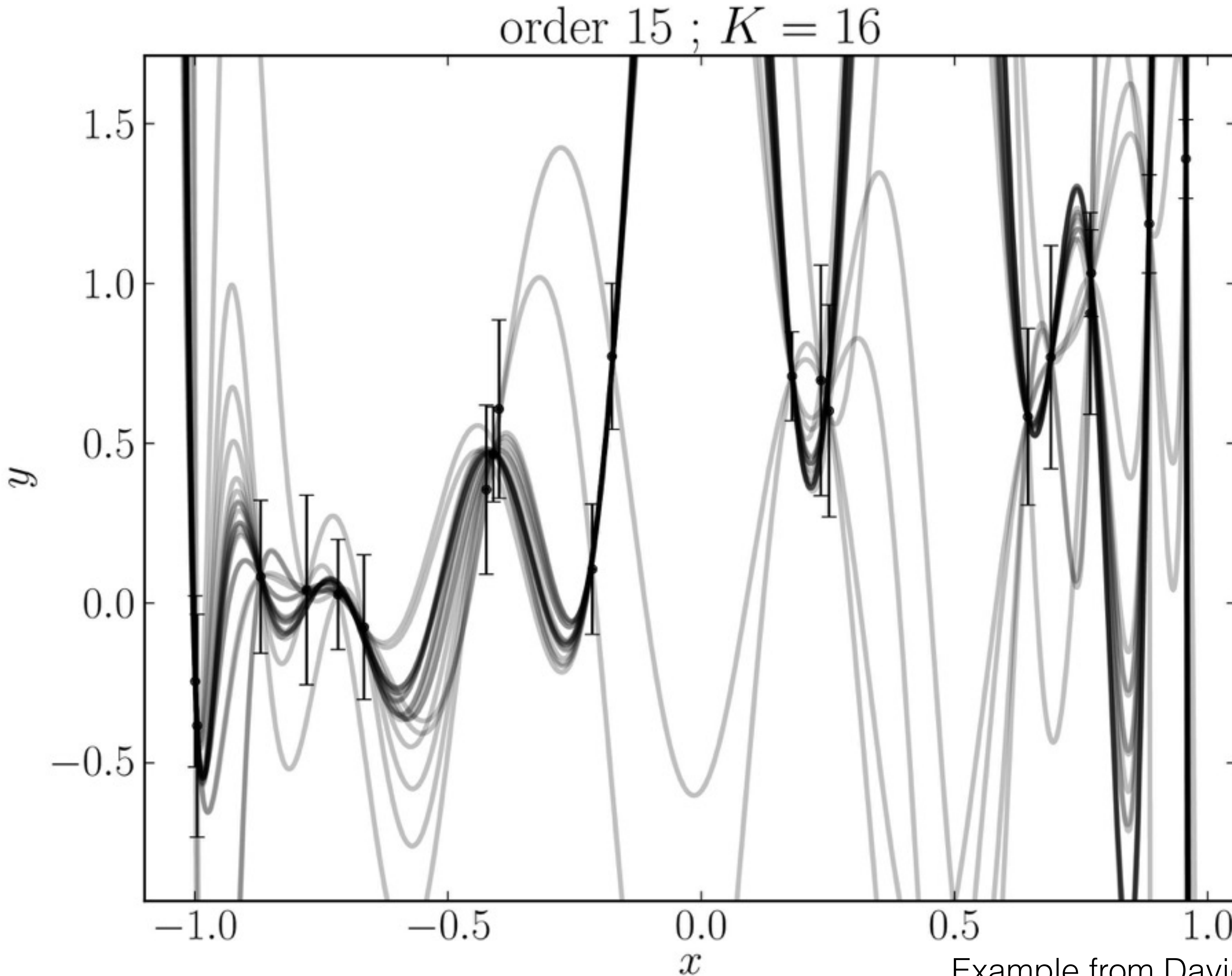
Example from David Hogg



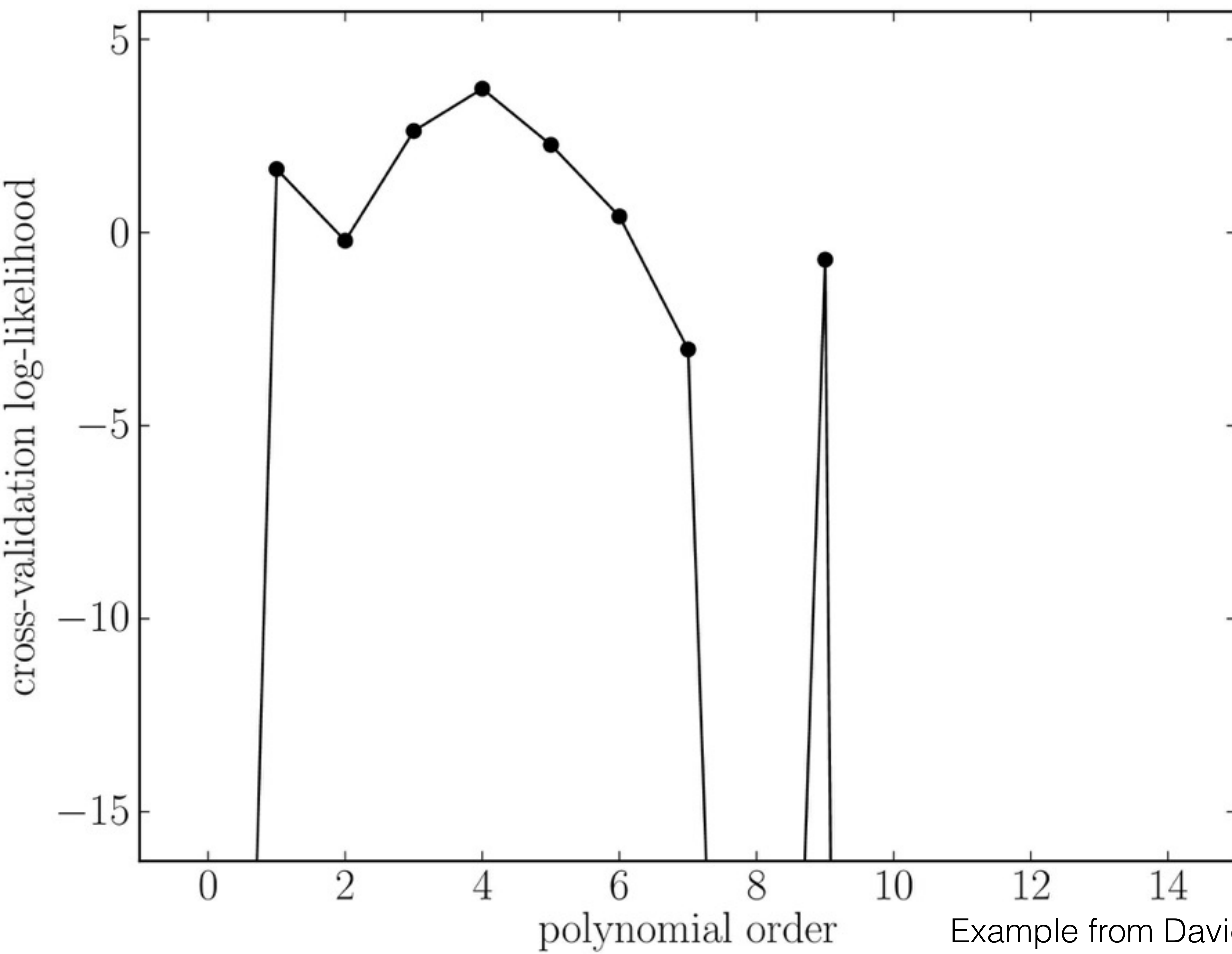
Example from David Hogg



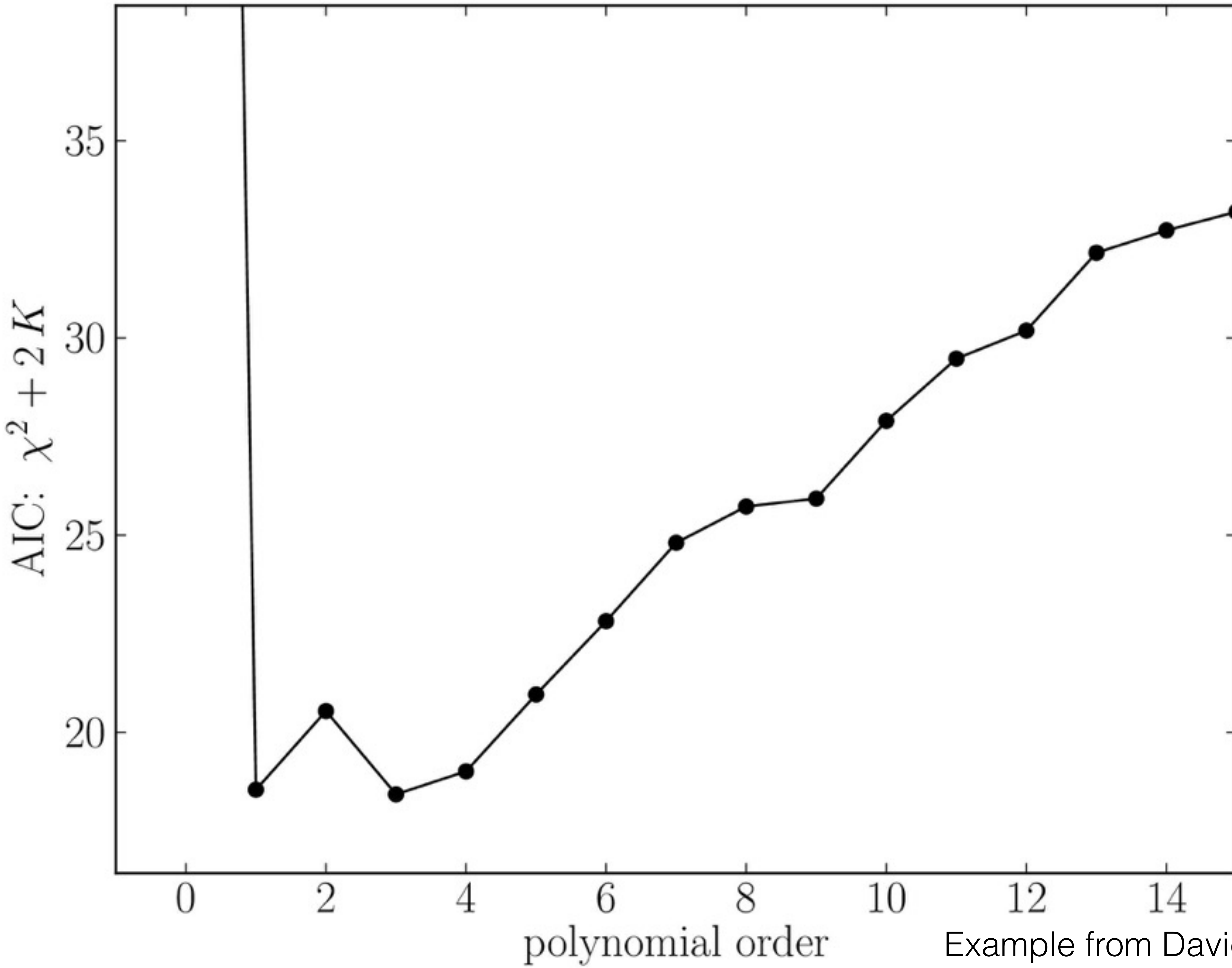
Example from David Hogg



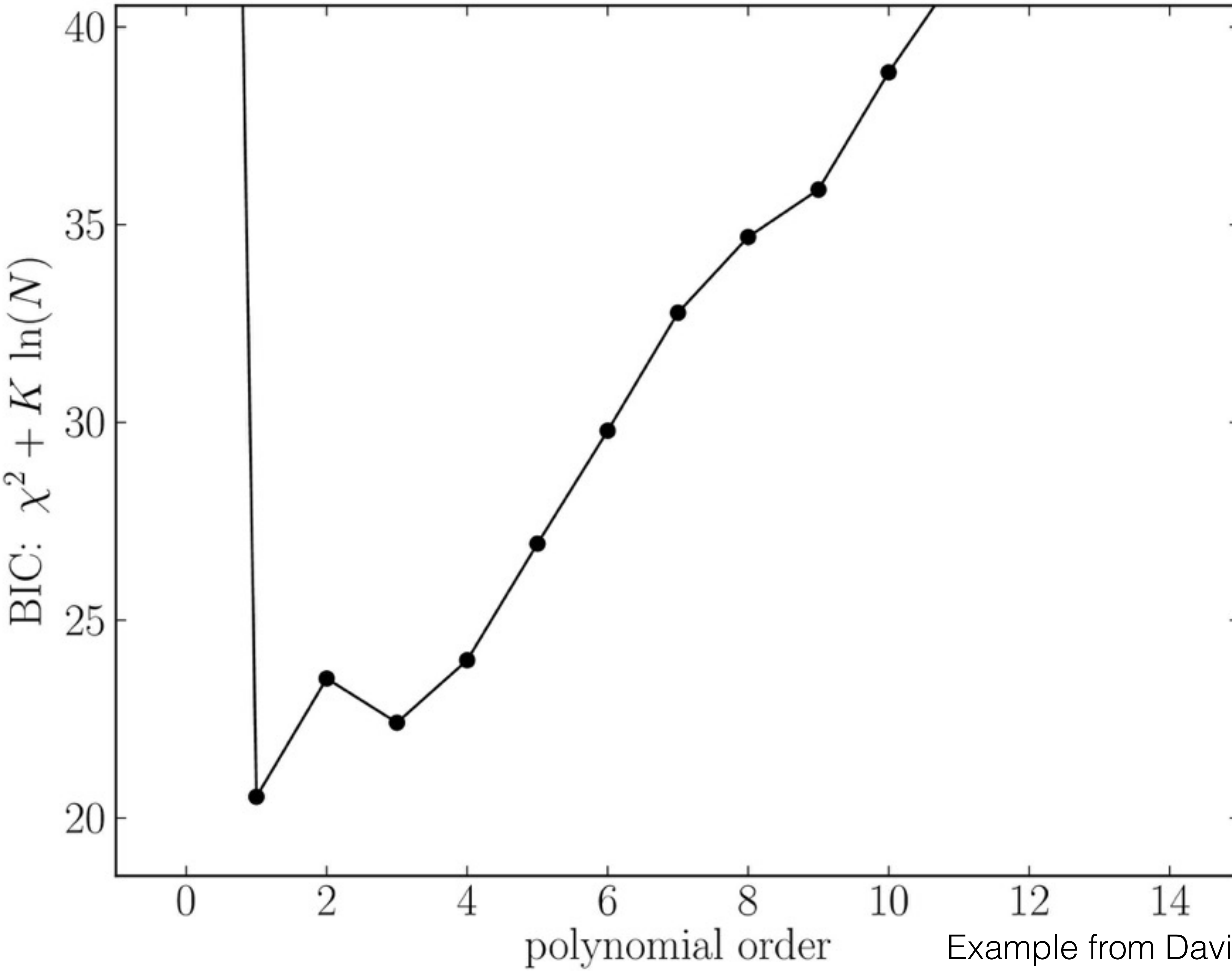
Example from David Hogg



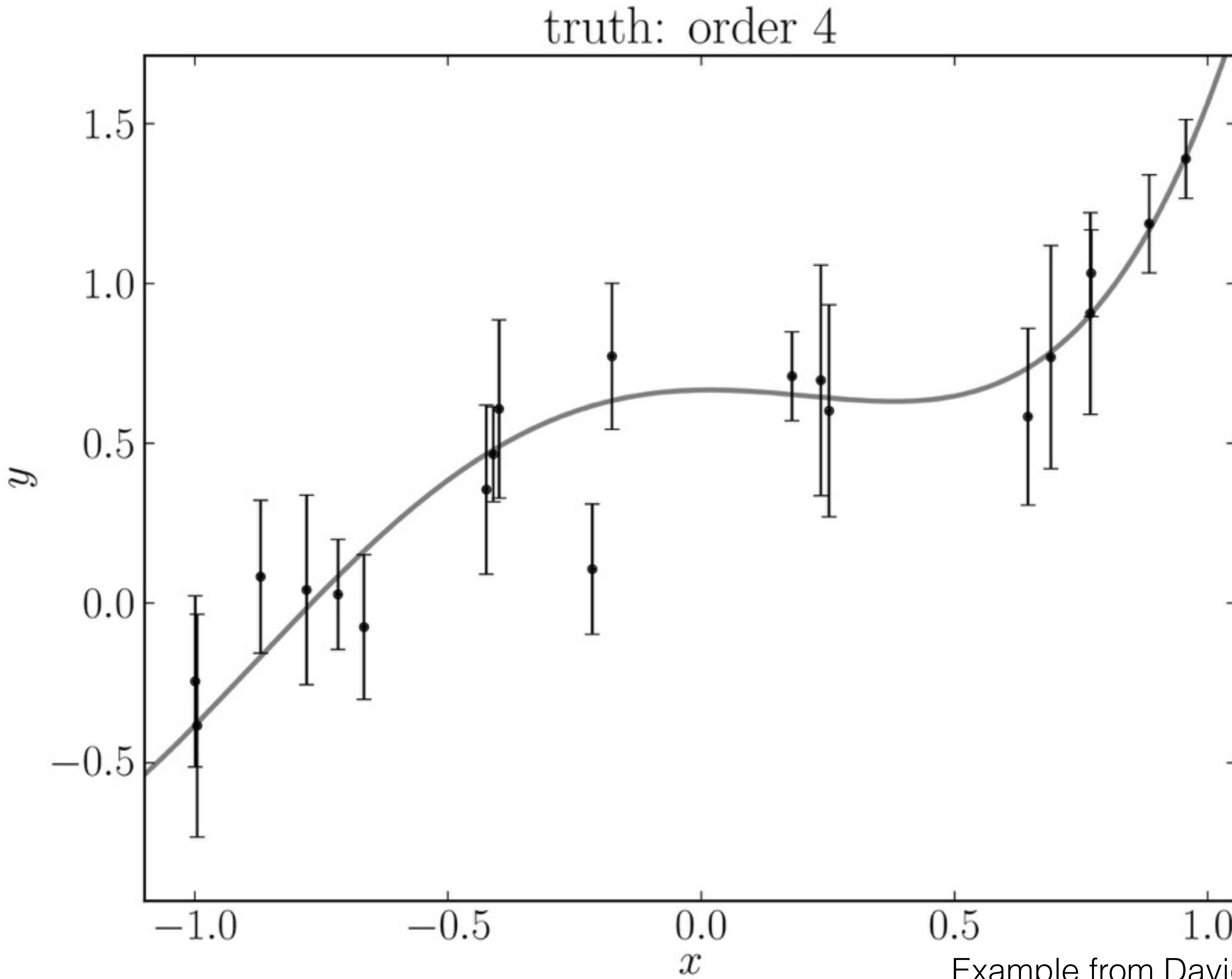
Example from David Hogg



Example from David Hogg



Example from David Hogg



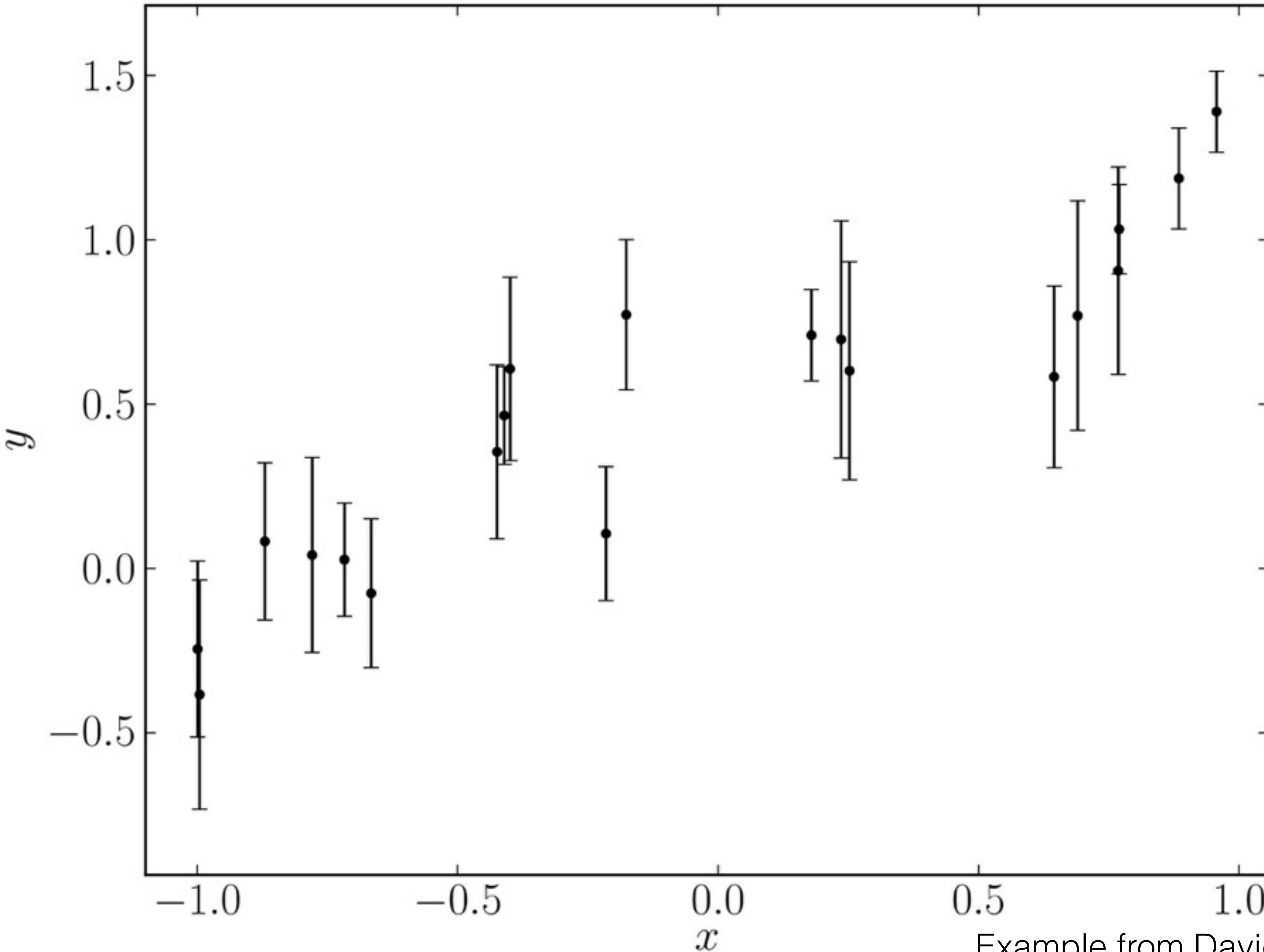
Example from David Hogg

Cross-validation: advantages

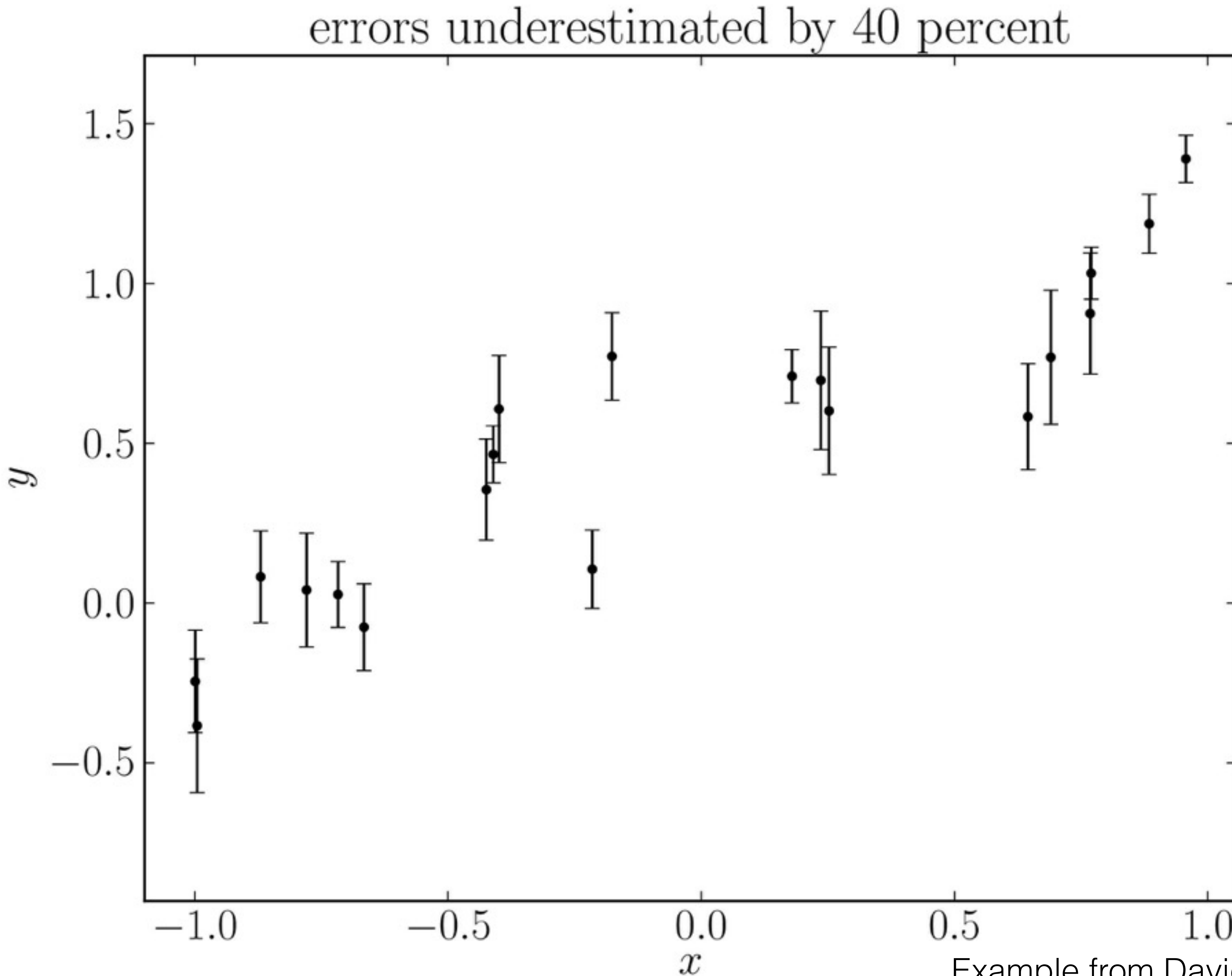
- Makes sense and does not make strong assumptions (just that all data points are typical)

- Easy to implement, but can be expensive if each fit is expensive (can use leave-N-out instead)

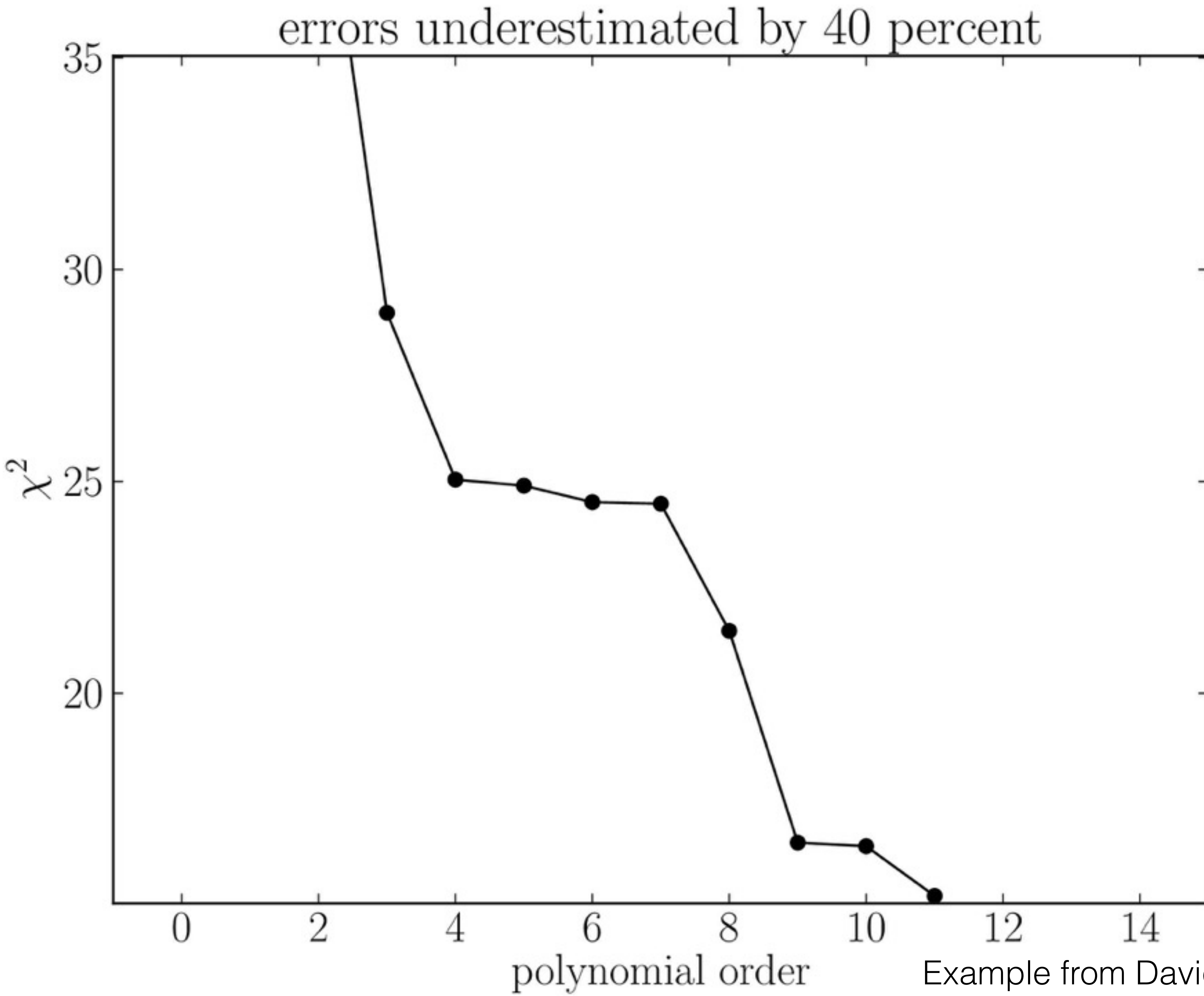
- Robust against certain types of under/overestimates of the uncertainties (similar to bootstrap)



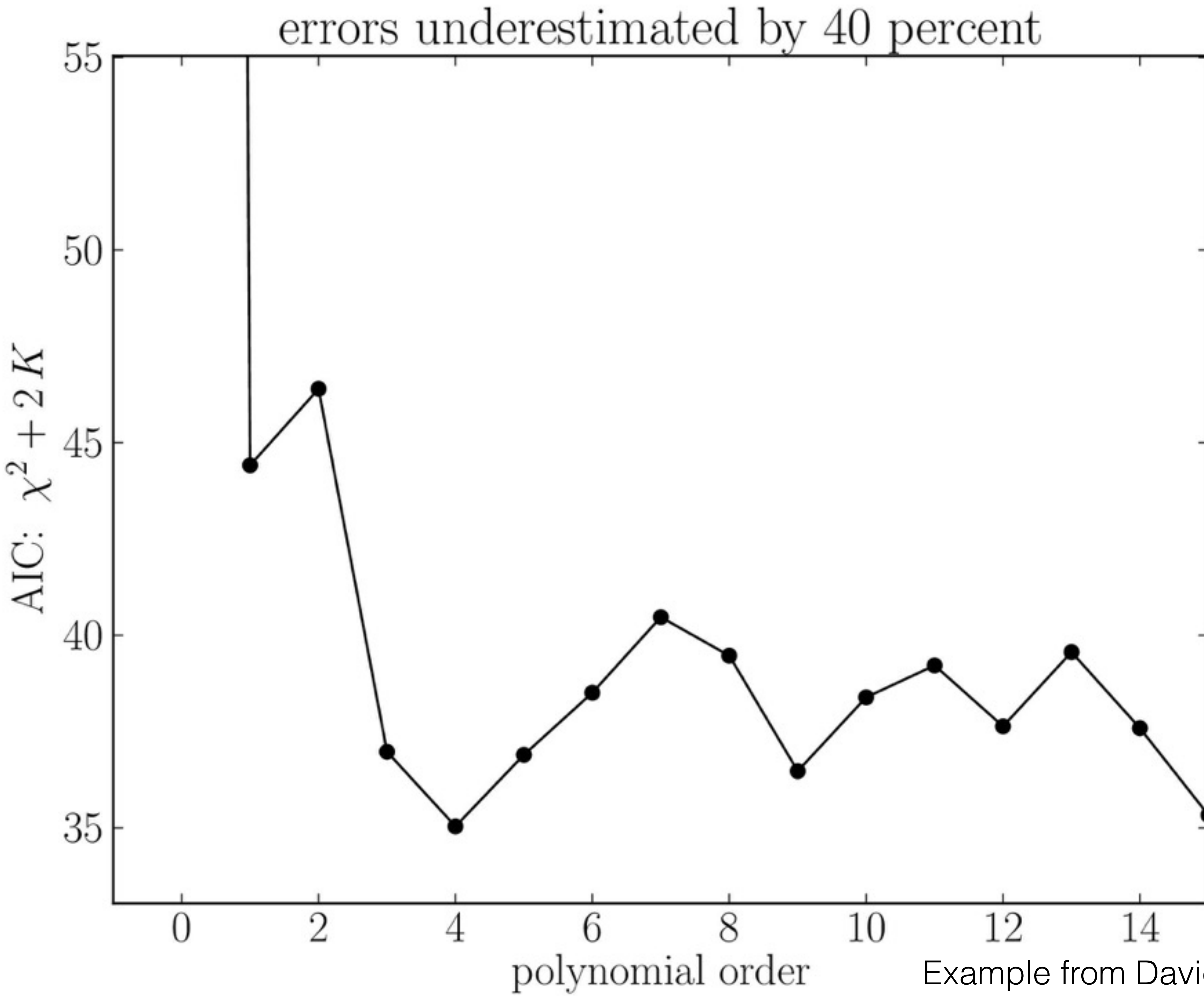
Example from David Hogg



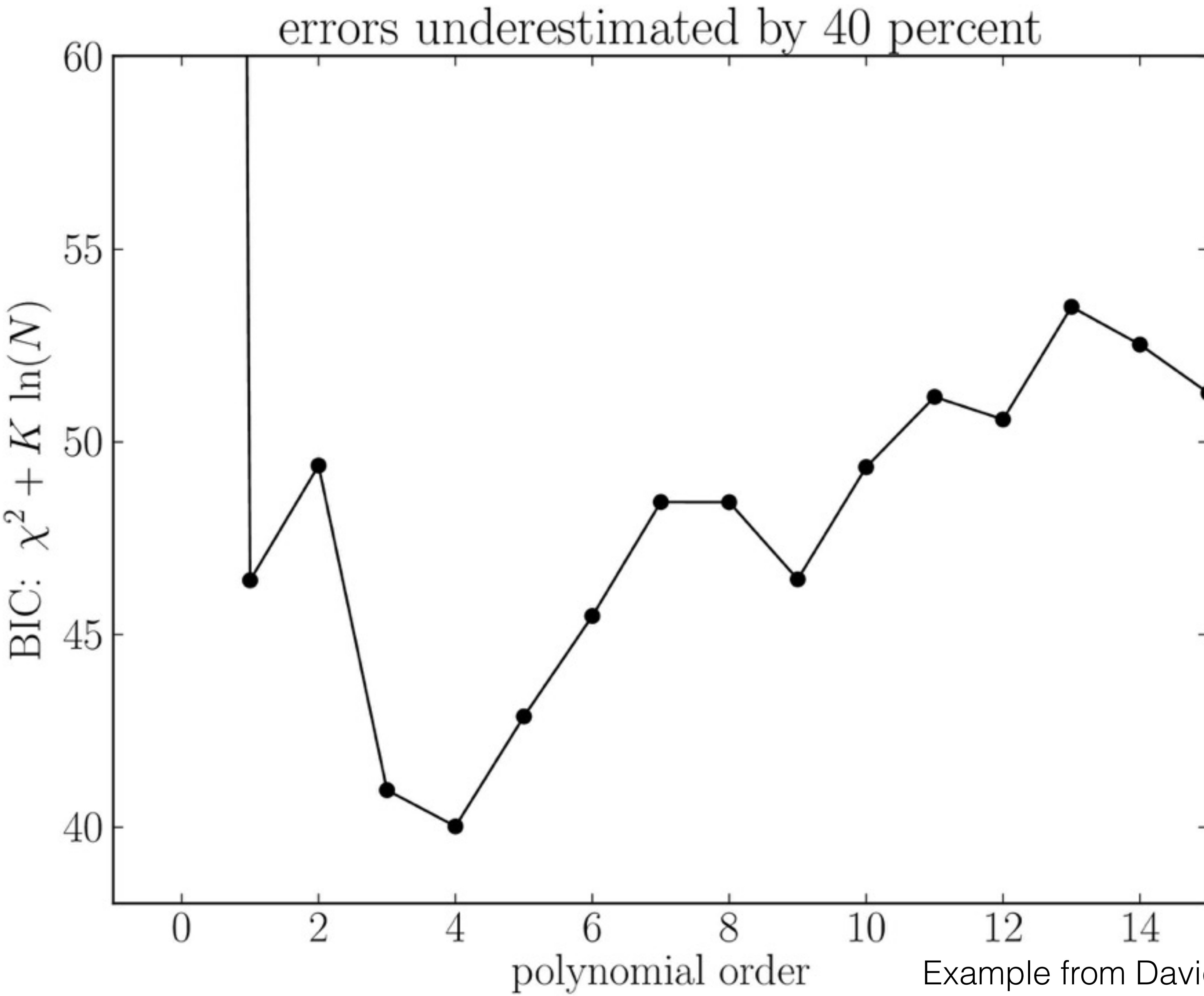
Example from David Hogg



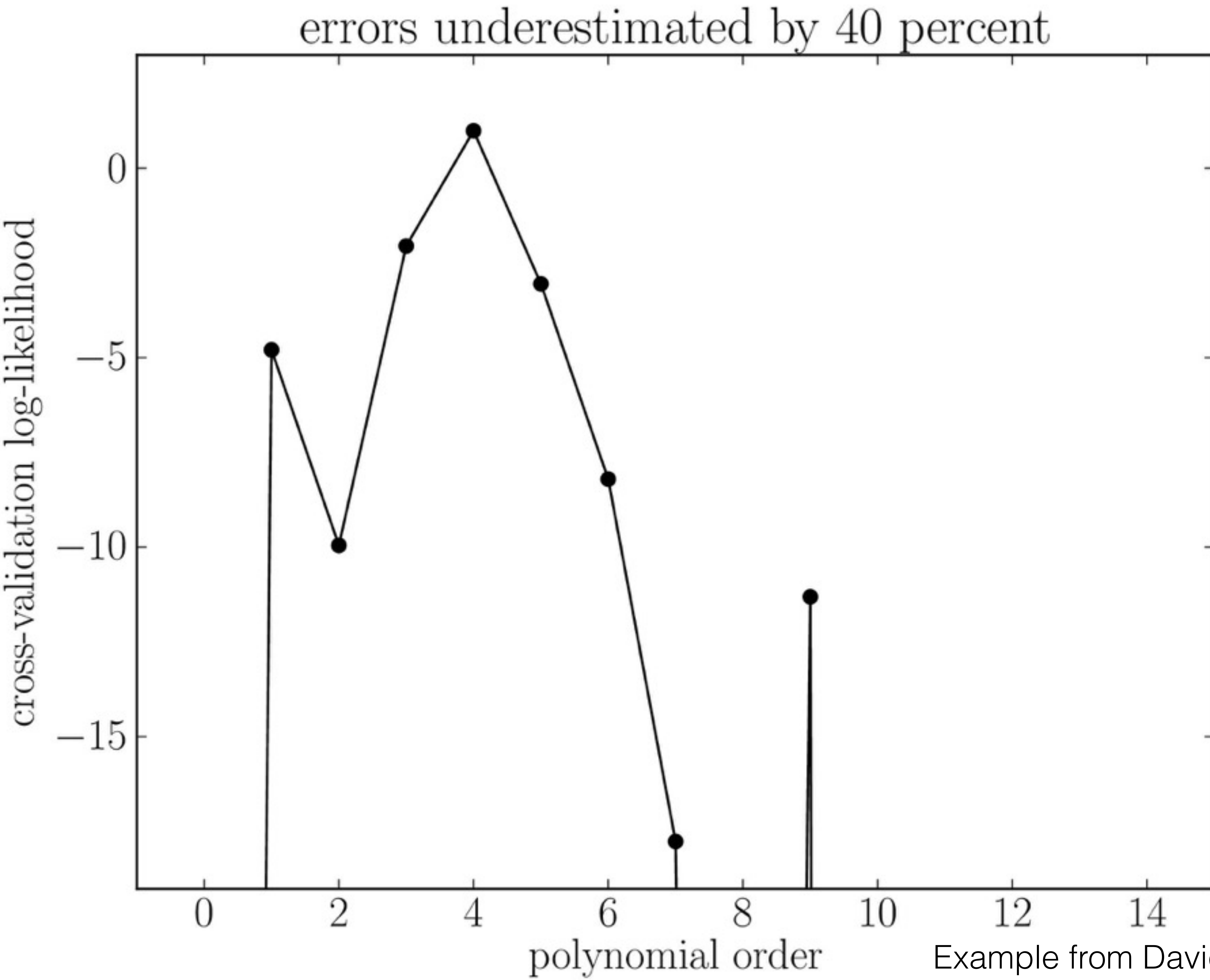
Example from David Hogg



Example from David Hogg

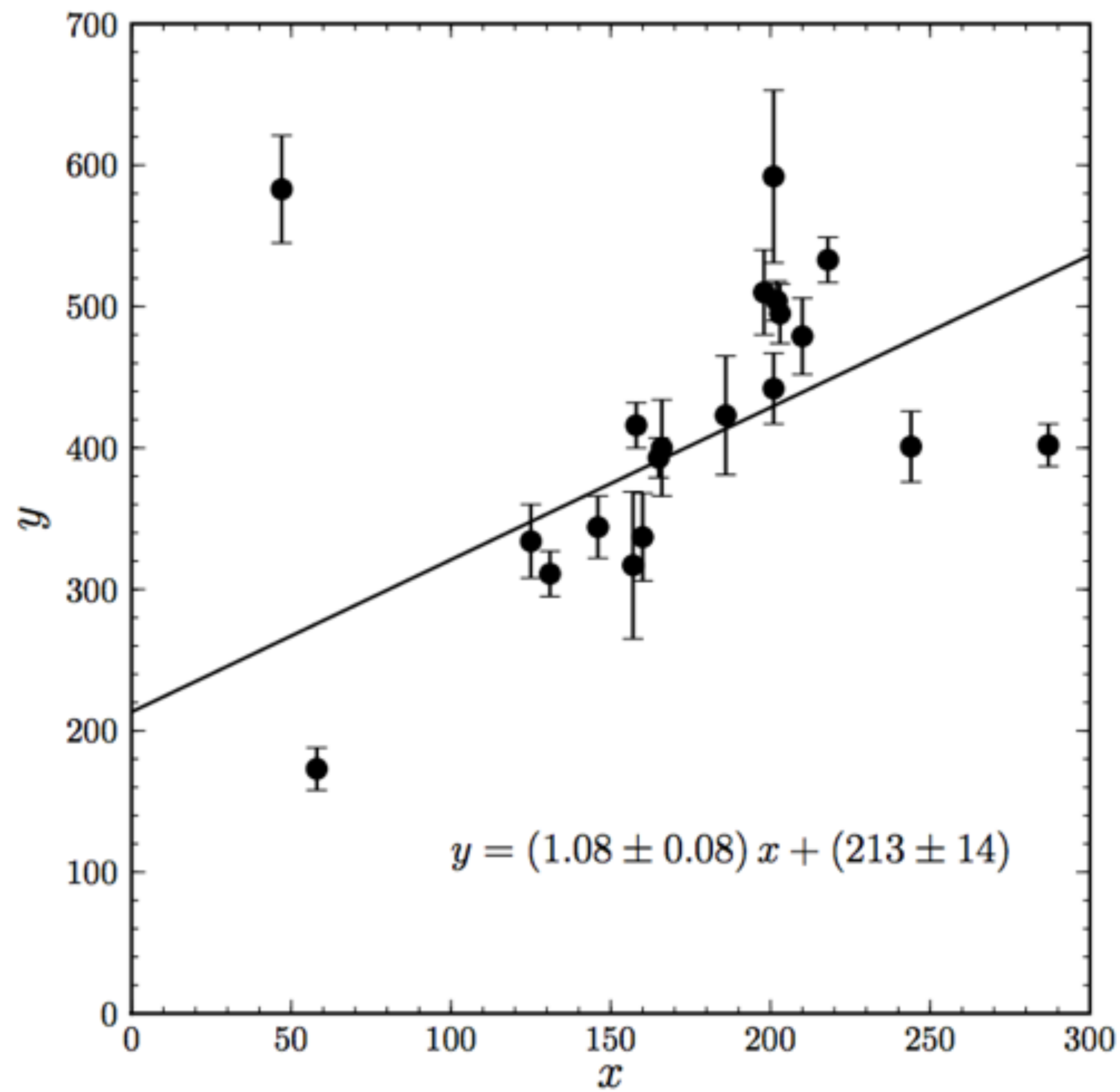


Example from David Hogg



Example from David Hogg

Outliers, robustics statist



Hogg, Bovy, & Lang (2010)

Handling outliers

- Important! Outliers can wreak havoc to model fitting, model validation, model selection

- Beware of simple cuts. Obvious outliers can be removed (e.g., cosmic rays), but cutting out data points induces selection function near actual data

- Typically best to model outliers w/ simple models

- Or can soften the likelihood for small values (large values of chi^2)

Softened chi^2

- Standard chi^2 strongly punishes large deviations from the mode, can make this smaller
- X = Σ_i |chi$_i$|
- $\chi^2_Q = \sum_i \frac{Q^2[y_i - f(x_i)]^2}{Q^2\sigma_i^2 + [y_i - f(x_i)]^2}$
- Any heavier-tailed distribution (e.g., student t)
- Changes the likelihood of *every point*, so reduces sensitivity to all points
- Make most sense if you think you do not understand your uncertainties very well

Modeling outliers

- Can use mixture model: some probability that the data are *inliers* and some probability that the data are *outliers*
- Make sense if you think there are two classes of data
- We discussed this in the last class

Mixture model for outliers

In this case, the likelihood is

$$
\begin{aligned}
\mathscr{L} &\equiv p(\{y_i\}_{i=1}^N | m, b, \{q_i\}_{i=1}^N, Y_\mathrm{b}, V_\mathrm{b}, I) \\
\mathscr{L} &= \prod_{i=1}^N \left[p_\mathrm{fg}(\{y_i\}_{i=1}^N | m, b, I))\right]^{q_i} \left[p_\mathrm{bg}(\{y_i\}_{i=1}^N | Y_\mathrm{b}, V_\mathrm{b}, I)\right]^{[1-q_i]} \\
\mathscr{L} &= \prod_{i=1}^N \left[\frac{1}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(-\frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2}\right)\right]^{q_i} \\
&\quad \times \left[\frac{1}{\sqrt{2\,\pi\,[V_\mathrm{b} + \sigma_{yi}^2]}} \exp\left(-\frac{[y_i - Y_\mathrm{b}]^2}{2\,[V_\mathrm{b} + \sigma_{yi}^2]}\right)\right]^{[1-q_i]} \quad ,
\end{aligned}
$$

Posterior requires prior on q_i, introduces new parameter P_b

$$
\begin{aligned}
p(m, b, \{q_i\}_{i=1}^N, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b} | I) &= p(\{q_i\}_{i=1}^N | P_\mathrm{b}, I)\, p(m, b, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b} | I) \\
p(\{q_i\}_{i=1}^N | P_\mathrm{b}, I) &= \prod_{i=1}^N [1 - P_\mathrm{b}]^{q_i}\, P_\mathrm{b}^{[1-q_i]} \quad ,
\end{aligned}
$$

Hogg, Bovy, & Lang (2010)

Mixture model for outliers

- Can be marginalized over individual q_i individually

$$
\begin{aligned}
\mathscr{L} &\equiv p(\{y_i\}_{i=1}^N | m, b, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b}, I) \\
\mathscr{L} &\equiv \prod_{i=1}^N \left[(1 - P_\mathrm{b})\, p_\mathrm{fg}(\{y_i\}_{i=1}^N | m, b, I)) + P_\mathrm{b}\, p_\mathrm{bg}(\{y_i\}_{i=1}^N | Y_\mathrm{b}, V_\mathrm{b}, I) \right] \\
\mathscr{L} &\propto \prod_{i=1}^N \left[\frac{1 - P_\mathrm{b}}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(-\frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2} \right) \right. \\
&\qquad \left. + \frac{P_\mathrm{b}}{\sqrt{2\,\pi\,[V_\mathrm{b} + \sigma_{yi}^2]}} \exp\left(-\frac{[y_i - Y_\mathrm{b}]^2}{2\,[V_\mathrm{b} + \sigma_{yi}^2]} \right) \right] \quad , \qquad (17)
\end{aligned}
$$

Hogg, Bovy, & Lang (2010)

Mixture model for outliers

- Can be marginalized over individual q_i individually

$$\mathscr{L} \equiv p(\{y_i\}_{i=1}^{N} | m, b, P_\mathrm{b}, Y_\mathrm{b}, V_\mathrm{b}, I)$$

$$\mathscr{L} \equiv \prod_{i=1}^{N} \left[(1 - P_\mathrm{b})\, p_\mathrm{fg}(\{y_i\}_{i=1}^{N} | m, b, I)) + P_\mathrm{b}\, p_\mathrm{bg}(\{y_i\}_{i=1}^{N} | Y_\mathrm{b}, V_\mathrm{b}, I) \right]$$

$$\mathscr{L} \propto \prod_{i=1}^{N} \left[\boxed{\frac{1 - P_\mathrm{b}}{\sqrt{2\,\pi\,\sigma_{yi}^2}} \exp\left(- \frac{[y_i - m\,x_i - b]^2}{2\,\sigma_{yi}^2} \right)} \right. \quad \textbf{Inliers}$$

$$\left. + \boxed{\frac{P_\mathrm{b}}{\sqrt{2\,\pi\,[V_\mathrm{b} + \sigma_{yi}^2]}} \exp\left(- \frac{[y_i - Y_\mathrm{b}]^2}{2\,[V_\mathrm{b} + \sigma_{yi}^2]} \right)} \right] \quad \textbf{Outliers} \;, \qquad (17)$$

Hogg, Bovy, & Lang (2010)

Mixture model for outliers

- Works well in general:
  - Simple model, e.g., Gaussian with some free mean and variance
  - Halo when looking at disk kinematics
  - …
- More expensive than softening the objective function, because additional parameters, but retain more information

Robust statistics

- Most frequentist and Bayesian methods have a lot of assumptions about the distribution of the data

- Robust statistic is robust to deviations in the assumptions, e.g., deviations from Gaussian uncertainties, mis-estimated uncertainties, outliers

- Typically better to properly model one's uncertainties and outliers, but robust statistics useful in general data handling

Median

- Mean is a bad estimator for the central location of the data if the uncertainties are non-Gaussian and have long-tail (~outliers)

- Single bad measurement will throw off the mean

- Median = 50% quantile, does not care if you shift a point much further away —> robust against outliers

- Median is minimum of X = Σ_i $|chi_i|$ = Σ_i $|y_i\text{-median}|$ objective function

Estimates of spread

- Similar to mean, standard deviation is not robust against outliers

- Interquartile range [25%,75%] is robust, similar to the median (or any other range of quantiles near the center)

- Median absolute deviation:
  $MAD = \text{median}(|X_i - \text{median}(X_i)|)$

- For Gaussian: std. dev. = 1.4826 MAD, generally useful to transform any robust range to its Gaussian equivalent