Lecture 4

Introduction to Modelling and Learning

- We look at two of the simplest forms of learning, proportions and polynomials.
- These also serve to illustrate some basic principles.

Overview

- **Example learning and inference for “Visit to Asia” graph**

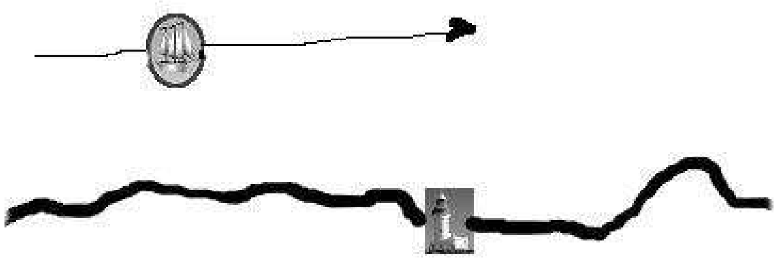
  We’ll put some data through B-Course and look at the results. More data lead to better results.
- Probability prerequisites
- Learning a proportion
- Learning a polynomial

Overview

- Example learning and inference for "Visit to Asia" graph
- **Probability prerequisites**
- Learning a proportion
- Learning a polynomial

The Abstraction of Continuity

- Nothing is continuous in the real world. All data is gathered with A-D converters, all measurements discrete. Physics presents a continuous abstraction of a discrete world.
- You are in a lighthouse at night and expect a ship to travel on a straight shipping lane past you. Typically modelled with a Cauchy.



1. Measurements assumed to infinite precision.
2. Model assumes path is linear and infinite in both directions, i.e., off towards Pluto.
3. Typical prior assumes equally likely to be anywhere on the line, e.g., in the phi-delta-kappa galaxy.

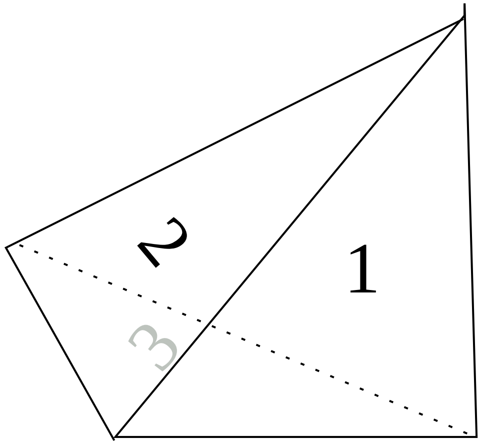
Prior Knowledge

Suppose you are predicting whether someone has a thyroid problem based on factors such as age, sex, pregnancies, blood levels of hormones, etc.

- If you use logistic regression or depth 4 decision tree as your predictive distribution, does it mean the “truth” follows these restrictions?
- In practice you may know many things:
  - Incidence increases with age and certain hormone levels.
  - Incidence quiet rare.
  - Hormone levels modify dramatically during pregnancy.

What are the outcome proportions?

With this irregular 4-sided die, what are the outcome probabilities?



- Each outcome is reasonably likely.
- Wont be uniform.
- Are larger sides more or less likely?

Models for Learning

Probability of sample of size N as a vector of data $\vec{x}$, given model $\mathscr{M}$ and parameters θ (possibly a vector), assuming *independently and identically distributed* (i.i.d.) data:

$$p(\vec{x}|\theta,\mathscr{M}) \;=\; \prod_{i=1,\ldots,N} p(x_i|\theta,\mathscr{M})$$

For **prediction**, one has output variables too, $\vec{y}$.

$$p(\vec{x},\vec{y}|\theta,\mathscr{M}) \;=\; \prod_{i=1,\ldots,N} p(x_i,y_i|\theta,\mathscr{M})$$

$$p(\vec{y}|\vec{x},\theta,\mathscr{M}) \;=\; \prod_{i=1,\ldots,N} p(y_i|x_i,\theta,\mathscr{M}) \qquad \text{conditional model}$$

Models for Learning, cont.

We'll look at methods for the conditional model. Some methods are based on the probability model being "truth" (ha ha!):

maximum likelihood (ML): maximize $\log p(\vec{y}\,|\,\vec{x},\theta,\mathcal{M})$
maximum *a posterior* **(MAP):** maximize $\log p(\theta\,|\,\vec{x},\vec{y},\mathcal{M})$
evidence: maximize for k from multiple models $\mathcal{M}_k$, $\log p(\vec{x}\,|\,\vec{y},\mathcal{M}_k)$

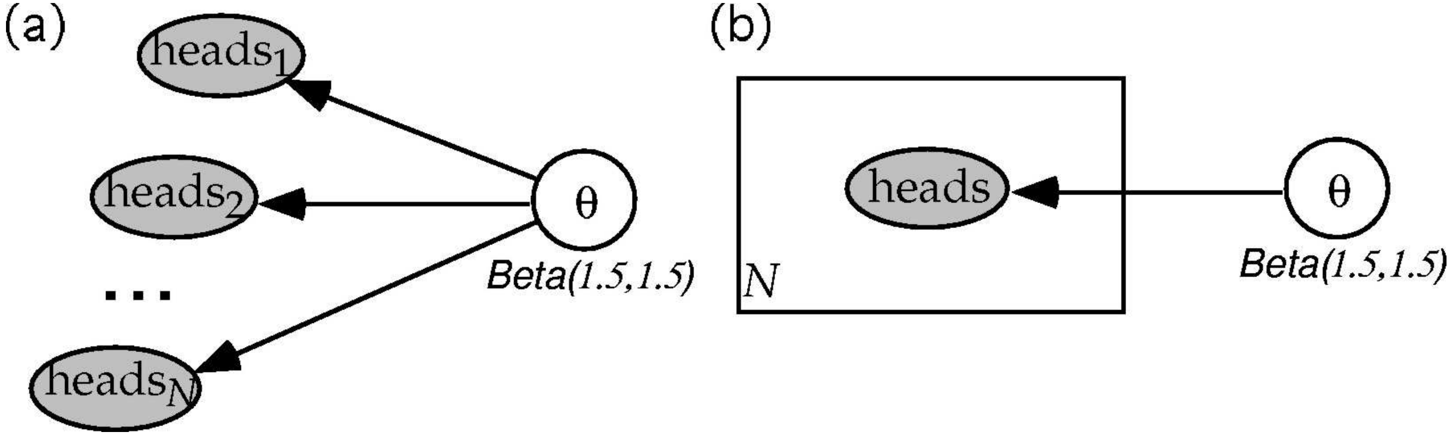
Other methods are based on optimizing a "cost" measure for each prediction, done empirically *without* the probability model: minimum error, minimum squared error, half Brier score, Hellinger distance, . . . , e.g., Let $f_\theta(x_i)$ be the model's prediction for y_i, then optimize for θ

$$\sum_{i=1,\ldots,N} \text{distance}(y_i, f_\theta(x_i))$$

Overview

- Example learning and inference for “Visit to Asia” graph
- Probability prerequisites
- **Learning a proportion**
- Learning a polynomial

Why care about Proportions?



- Many statistic problems have learning proportions as an inner loop,

  e.g., decision trees, hidden Markov models, Bayesian networks with simple probability tables, variable length n-grams, i.e., basic core of best performing general compression algorithms, . . .

- Reflects most of the major techniques used.

Likelihood for Proportions

Have a variable x_i taking on K outcomes, $1,\ldots,K$. Model is discrete distribution with outcome probabilities $\theta_1,\ldots,\theta_K$, which sum to 1.

$$p(\vec{x}\,|\,\theta,\mathcal{M}) \;=\; \prod_{i=1,\ldots,N} p(x_i\,|\,\theta,\mathcal{M})$$

Summarize the N data by $n_1,\ldots,n_K$ the count for each outcome.

$$\log p(\vec{x}\,|\,\theta,\mathcal{M}) \;=\; \sum_{k=1,\ldots,K} n_k \log\theta_k$$

Sometimes exclude ordering information, getting a multinomial, which adds the term $\log C^N_{n_1,\ldots,n_K}$ for

$$C^N_{n_1,\ldots,n_K} \;=\; \prod_{k=1,\ldots,K} n_k! \Big/ N!$$

Maximum Likelihood Parameters

To maximize, add Lagrange multiplier term:

$$\sum_{k=1,\ldots,K} n_k \log\theta_k + \lambda\left(1 - \sum_{k=1,\ldots,K}\theta_k\right)$$

and optimize for $\theta_1,\ldots,\theta_K$ setting λ to make the constraint $1 = \sum_{k=1,\ldots,K}\theta_k$ hold.

Differentiation w.r.t. θ_k and setting to zero yields $n_k/\theta_k = \lambda$. Thus $\lambda = N$ and

$$\hat{\theta}_k = \frac{n_k}{N}$$

Maximum *A Posterior* Params

The Bayesian method requires a prior. It is only simple when the prior is the same functional form as the likelihood. This form is a called a Beta(α_1, α_2) distribution for $K = 2$ and a Dirichlet$(\alpha_1, \ldots, \alpha_K)$ distribution for $K > 2$.

$$\log p(\theta \,|\, \mathscr{M}) \;=\; \sum_{k=1,\ldots,K} (\alpha_k - 1) \log \theta_k + \text{ constant}$$

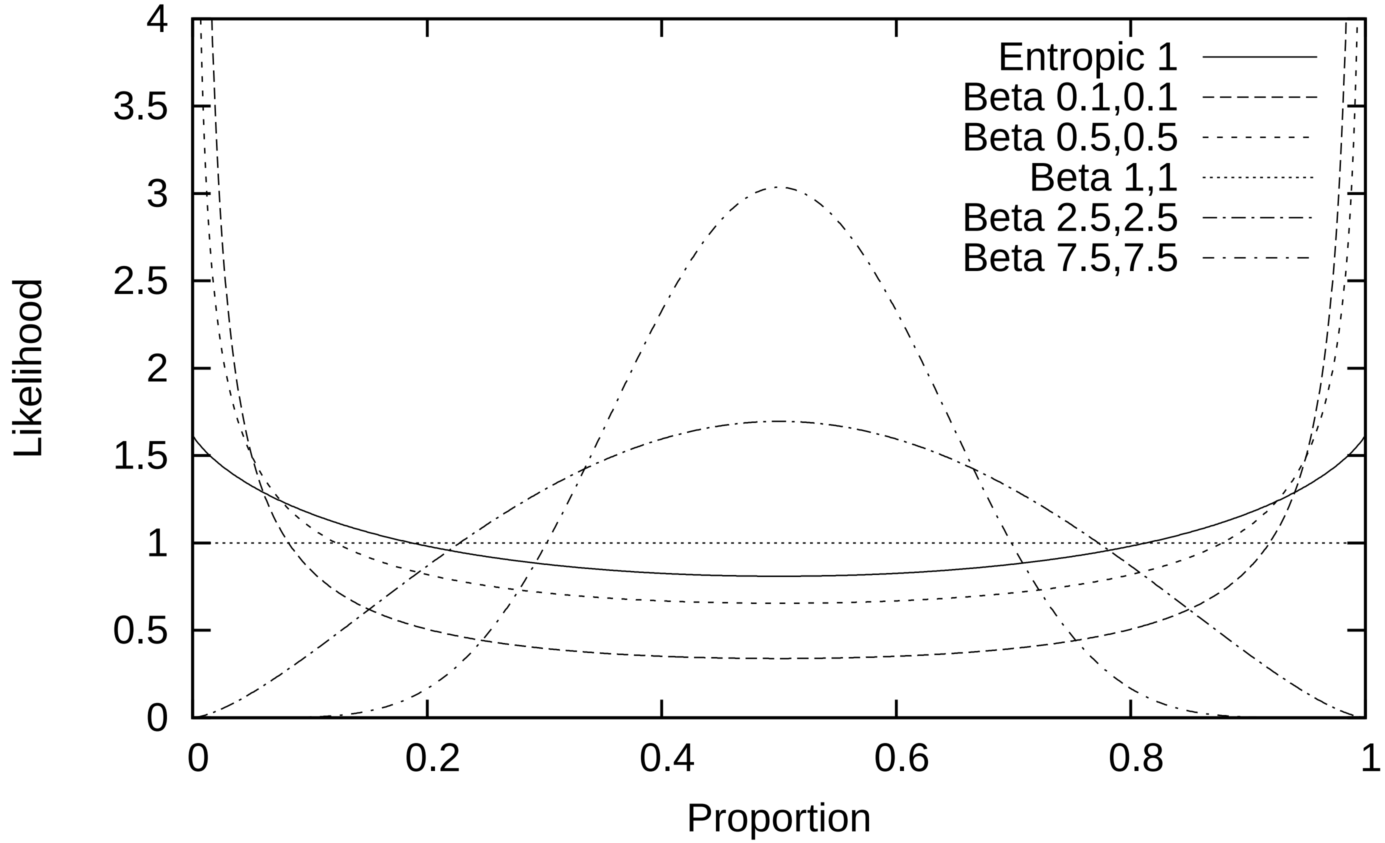
Similarly, the maximum *a posterior* parameters become (for $\alpha_0 = \sum_{k=1,\ldots,K} \alpha_k$):

$$\hat{\theta}_k = \frac{n_k + \alpha_k - 1}{N + \alpha_0 - K}$$

The mean parameters:

$$\text{Expec}_{\theta \,|\, \vec{x}, \mathscr{M}}(\theta_k) \;=\; \frac{n_k + \alpha_k}{N + \alpha_0}$$

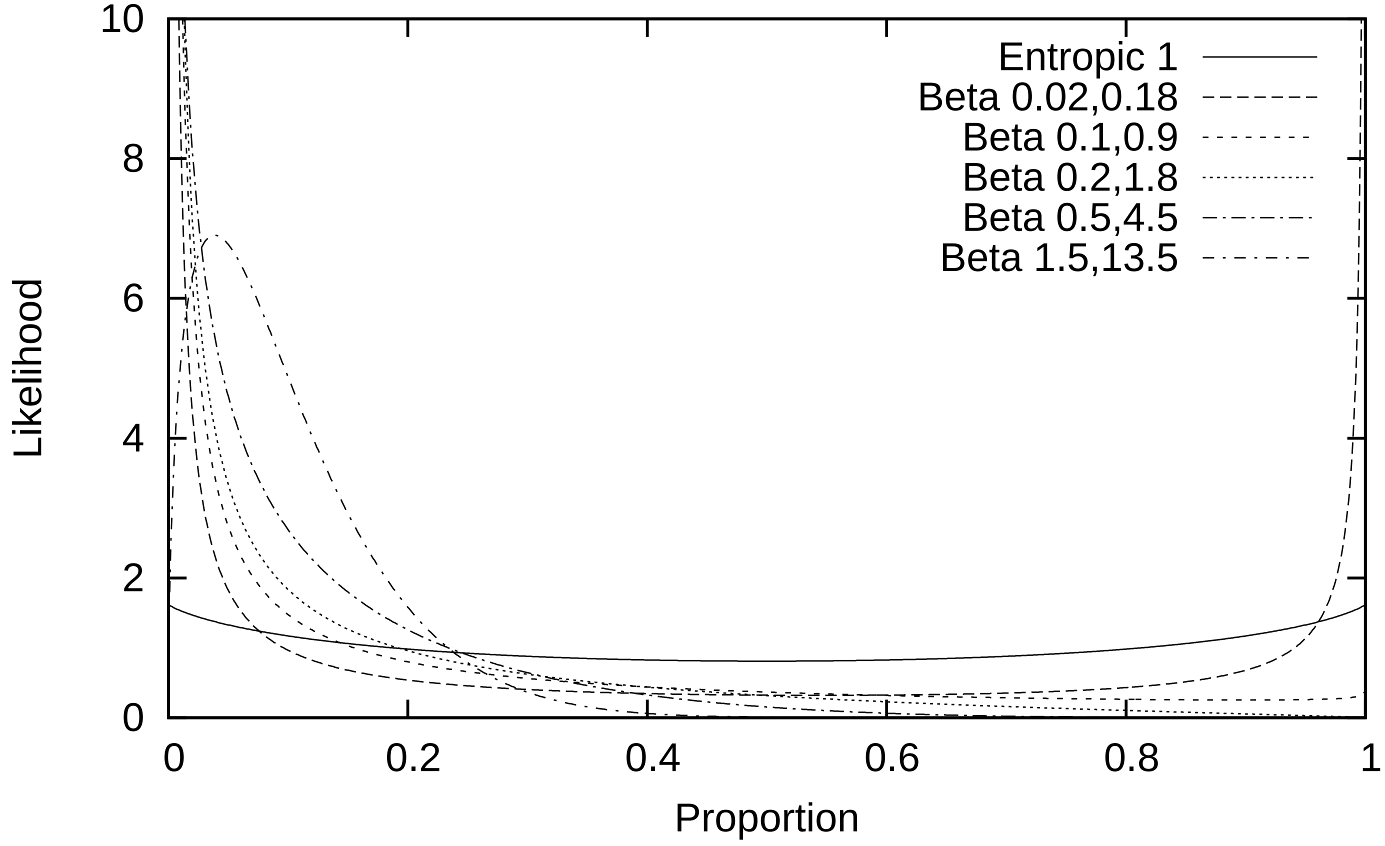
Prior Probabilities for Binary Proportions



Beta prior:
$\propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$

Entropic prior:
$\propto p^{\beta p}(1-p)^{\beta(1-p)}$
$= \ e^{-\beta I(p)}$

Prior Probabilities for Binary Proportions



Beta prior:
$\propto p^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$

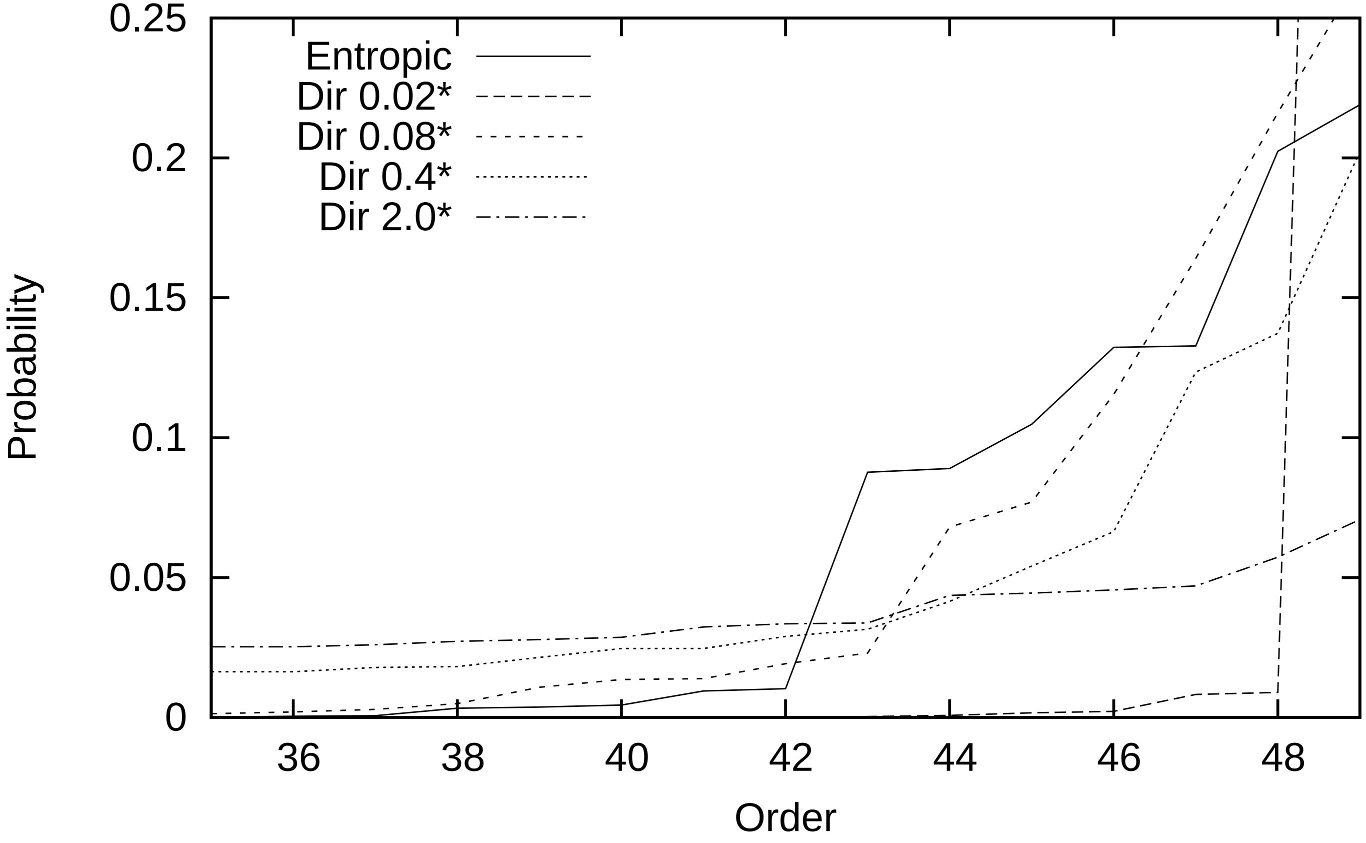
Entropic prior:
$\propto p^{\beta p}(1-p)^{\beta(1-p)}$
$= \ e^{-\beta I(p)}$

Prior Probabilities for 50-way Proportions

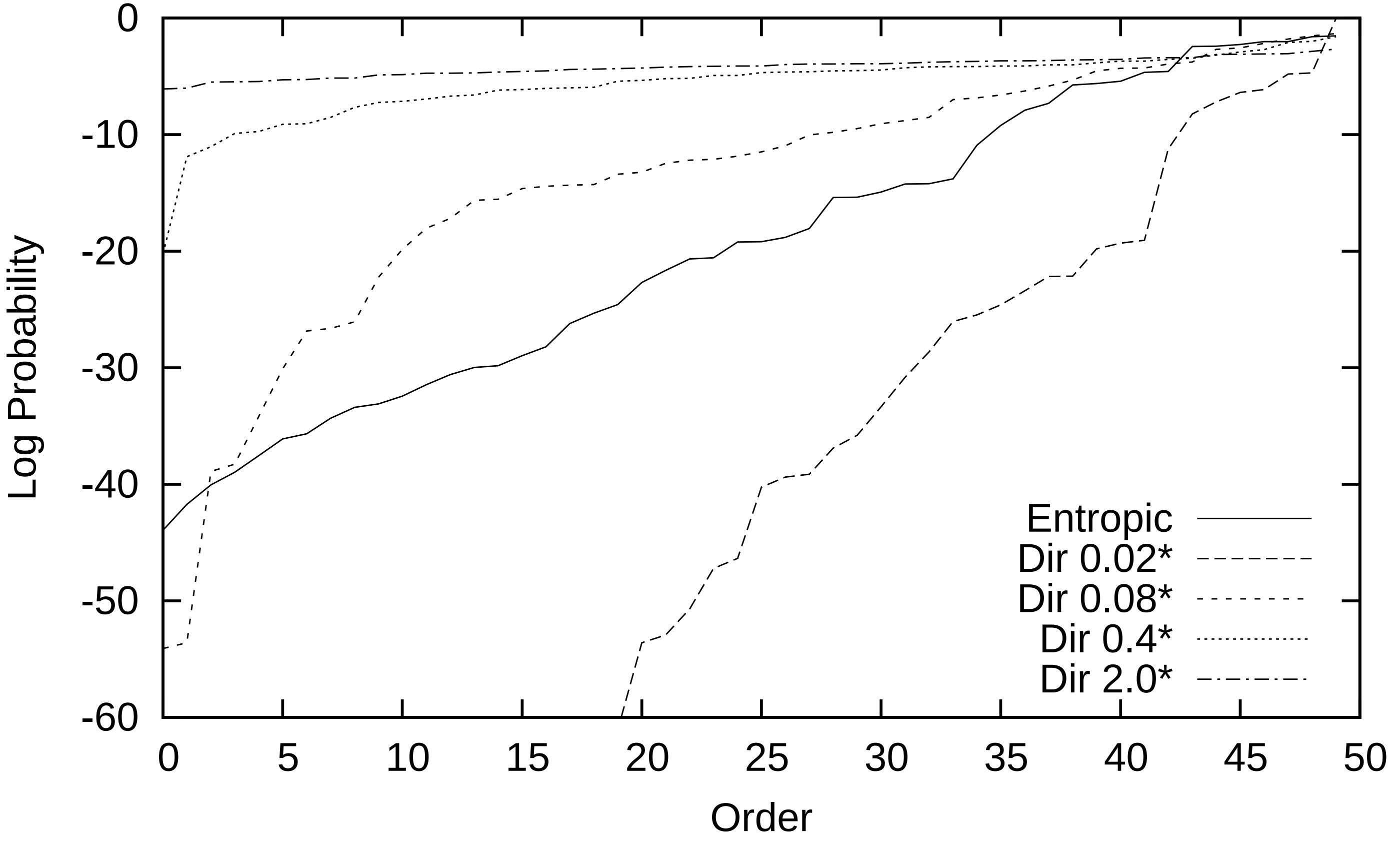
Beta prior:
$\propto \prod_{i=1,\ldots,50} p_i^{\alpha_i - 1}$

Entropic prior:
$\propto \prod_{i=1,\ldots,50} p_i^{\beta p_i}$
$= \; e^{-\beta I(p)}$

Sample from the distribution and place 50 probabilities in increasing order.



Prior Probabilities for 50-way Proportions



Beta prior:
$\propto \prod_{i=1,\ldots,50} p_i^{\alpha_i - 1}$

Entropic prior:
$\propto \prod_{i=1,\ldots,50} p_i^{\beta p_i}$
$= \ e^{-\beta I(p)}$

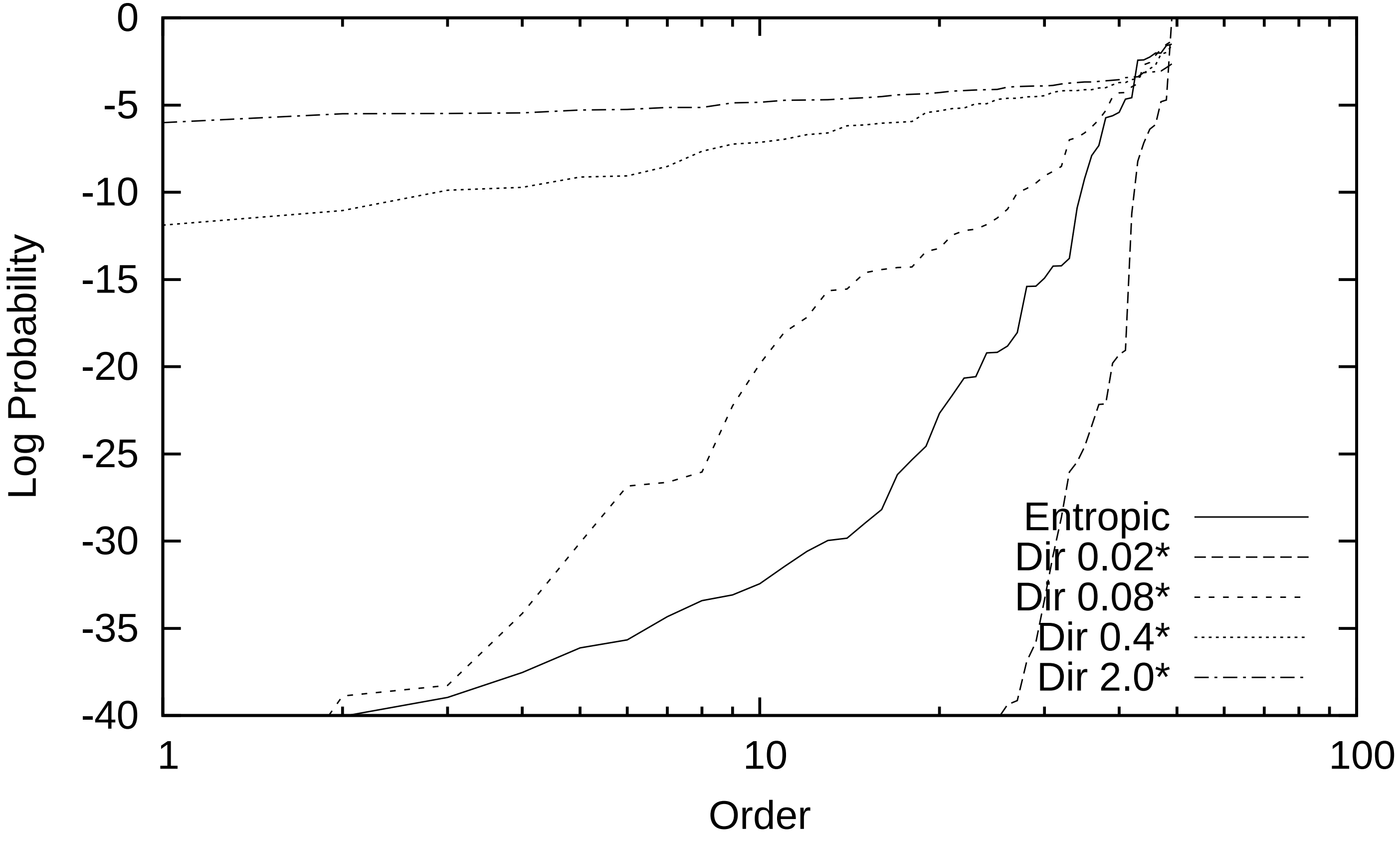
Sample from the distribution and place 50 probabilities in increasing order.

Prior Probabilities for 50-way Proportions

Beta prior:
$\propto \prod_{i=1,\ldots,50} p_i^{\alpha_i - 1}$

Entropic prior:
$\propto \prod_{i=1,\ldots,50} p_i^{\beta p_i}$
$= \; e^{-\beta I(p)}$

Sample from the distribution and place 50 probabilities in increasing order. Plot order to logarithmic scale (i.e., Zipf's law).



Interacting with the Models

Playing with the BetaCoinExperiment demo shows the following:

- Results sensitive for $N < 20$.
- Posterior curve starts quite broad and narrows slowly as N increases, in fact standard deviation is order $1/\sqrt{N}$.
- Most concavity disappears by $N = 1, 2$.

A good statistical text shows the posterior standard deviation for $K = 2$ is:

$$\sqrt{\frac{(n_1+\alpha_1)(n_2+\alpha_2)}{(N+\alpha_0)^2(N+\alpha_0+1)}} \approx \sqrt{\frac{\hat{\theta}_1\hat{\theta}_2}{N}}$$

History

- Great volumes have been written about supposed "objective" or "reference" priors for this task.
- Maximum Likelihood corresponds to $\alpha_k = 1$. So-called *Laplace Correction* corresponds to using the means for this value.
- So-called *Jeffreys' Method* corresponds to using $\alpha_k = 0.5$. This is an approximate minimax method, as used by the minimum description length and theory communities.
- So-called *Zipf's Law* has that a plot of the log probabilities against log rank (rank equals order in a sort) should be approximately linear. Used for word probabilities.
- No methods work uniformly well.
- In practice people use biased parameters and set α_0 using variance arguments, cross validation, or such.

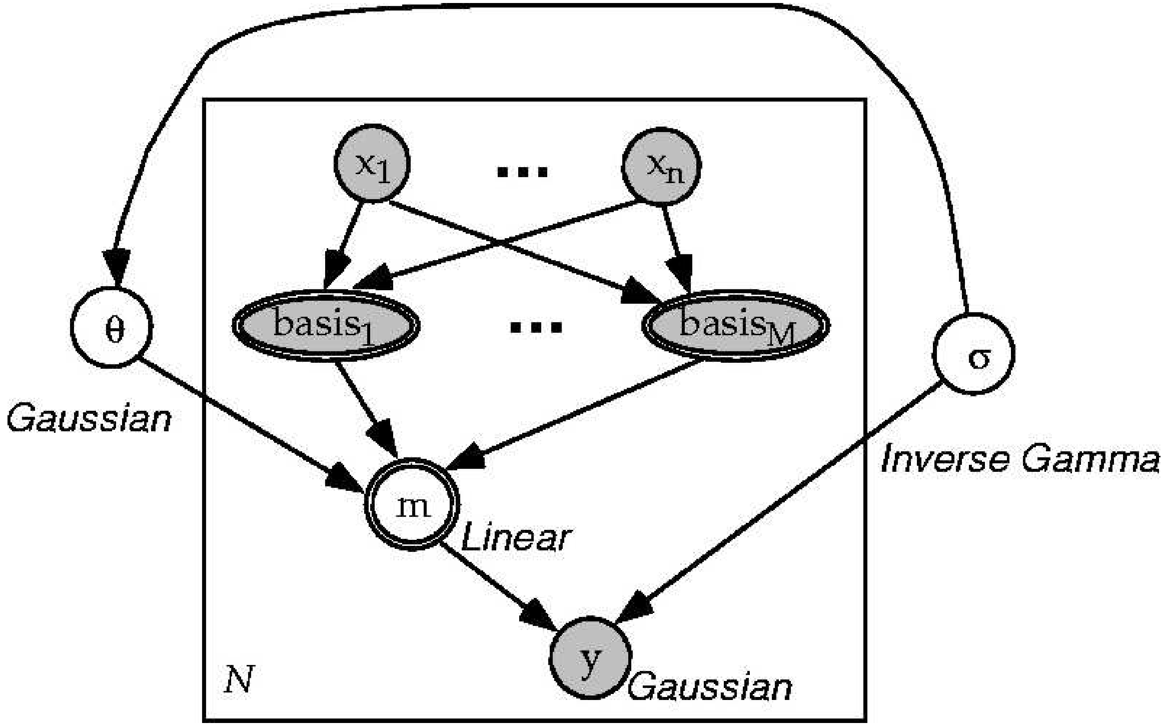
Overview

- Example learning and inference for “Visit to Asia” graph
- Probability prerequisites
- Learning a proportion
- **Learning a polynomial**

Linear Regression

Our interest in the problem is primarily to look at overfitting, and to consider effects of the priors and posteriors. The detail of the math. is not critical.

Linear Regression



$$p(\sigma)\,p(\theta|\sigma)\frac{1}{\left(\sqrt{2\pi}\sigma\right)^N}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i-\sum_{j=1}^{M}\theta_j basis_j(x_{.,i})\right)^2}$$

Linear Regression, Maximum Likelihood

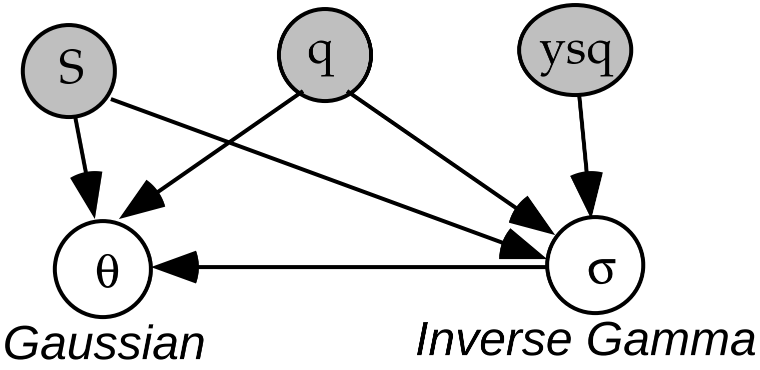
$$
\begin{aligned}
& \log p(\vec{y} \,|\, \vec{x}, \theta, \sigma, \mathscr{M}) \\
= \;& N \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{M} \theta_j basis_j(x_{\cdot,i}) \right)^2 + \text{ constant} \\
= \;& N \log \sigma - \frac{1}{2\sigma^2} \left(\sum_{i=1}^{N} y_i^2 - \sum_{j,k=1}^{M} SS_{j,k} \frac{\theta_j \theta_k}{2\sigma^2} + \sum_{j=1}^{M} m_j \frac{\theta_j}{2\sigma^2} \right) + \ldots
\end{aligned}
$$

$$
\text{where } SS_{j,k} \;=\; \sum_{i=1}^{N} basis_j(\vec{x}_i)\, basis_k(\vec{x}_i)
$$

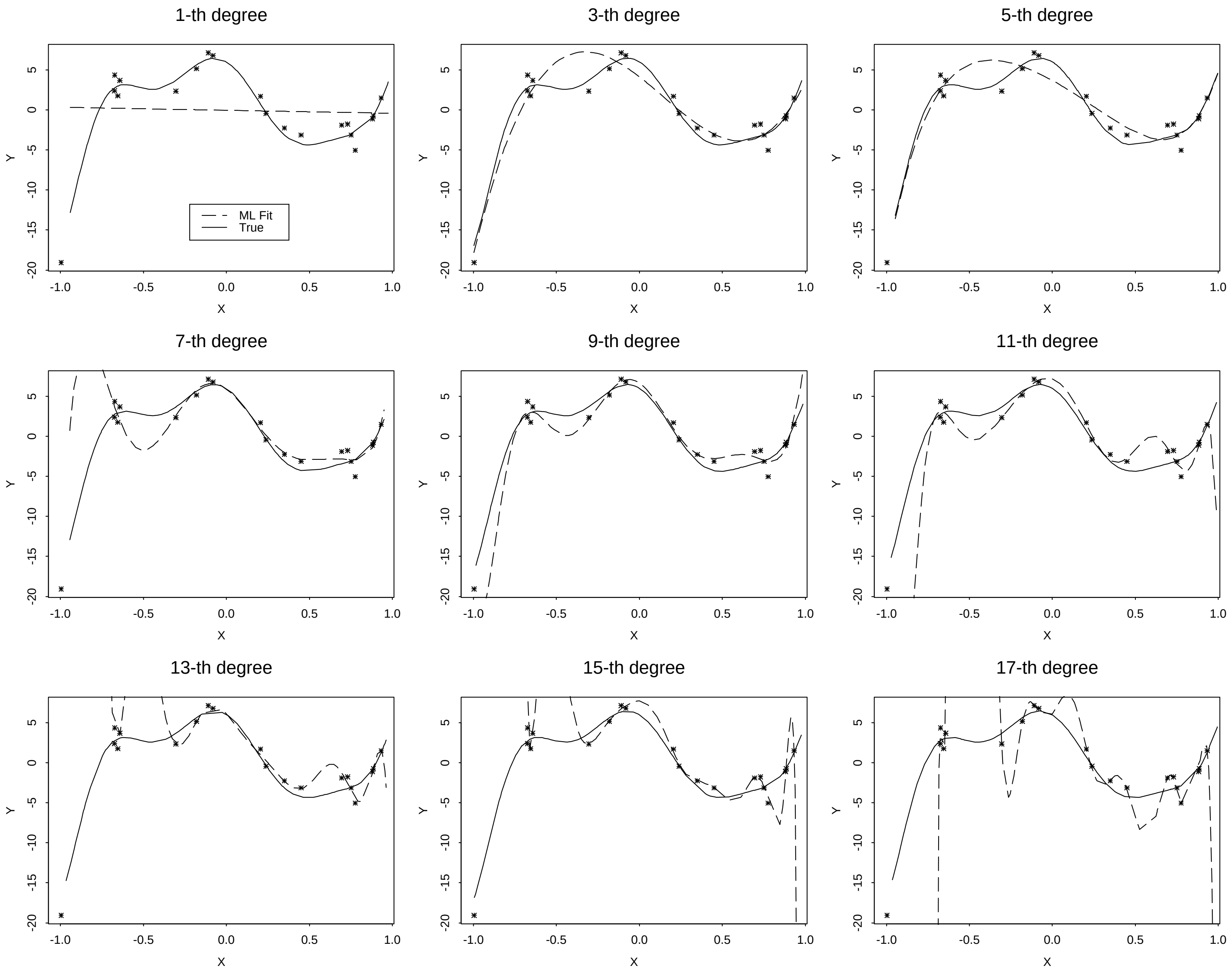
$$
\text{and } m_j \;=\; \sum_{i=1}^{N} y_i basis_j(\vec{x}_i)
$$

Linear Regression, ML cont

- Once again, we see the data is summarized by statistics, in this case: $\vec{SS}, \vec{m}, y^2$.
- This property happens for many common distributions taught in statistics, the Exponential Family, includes Poisson, Gamma, Inverse Gamma, multinomial, ...
- The problem now looks like a Gaussian on θ and an inverse Gamma on σ.



Linear Regression, ML cont



Linear Regression, ML cont

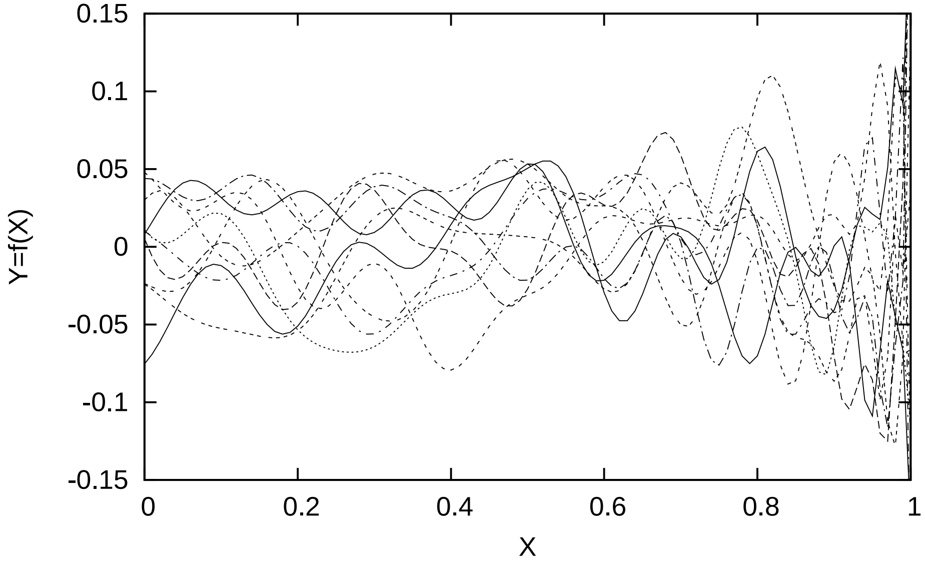
- This phenomena is called *overfitting*.
- As a rough rule of thumb, models require a sample size of at least $10 * K$ where K is the dimension of the parameter set. Otherwise, things become difficult.
- As before, the easiest Bayesian approach is to make the prior look the same as the likelihood, and we just hope that this makes some sense.

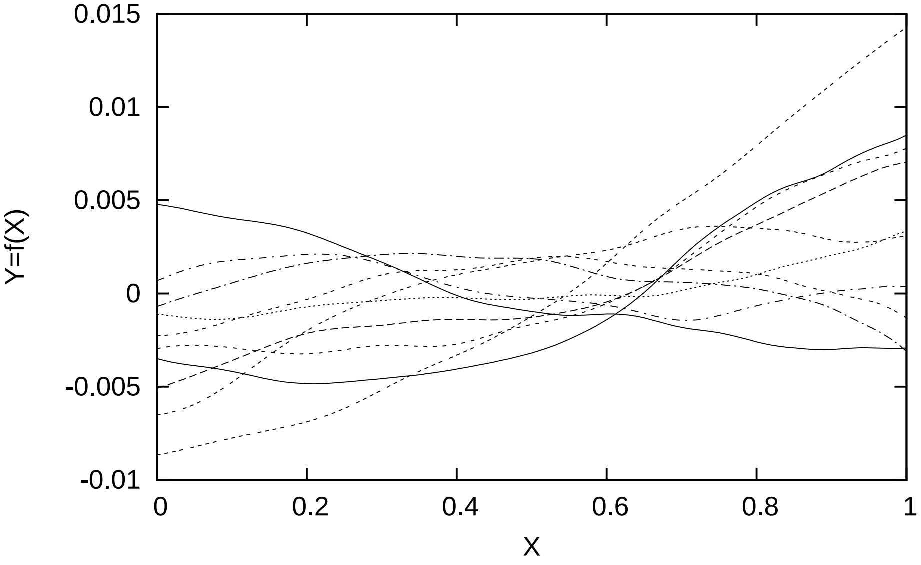
$$p(\theta,\sigma) \;\propto\; \frac{1}{\sigma^a} e^{-\left(\theta^{\dagger}C\theta+b\right)/\sigma^2}$$

In this case its not too bad.

Linear Regression, Prior on θ

With clever choice of the prior covariance matrix on θ (C in previous slide), our prior can generate lines like on the right. The left is uniform in θ for 50 degree Legendre polynomials. Is the right better? That is subjective!



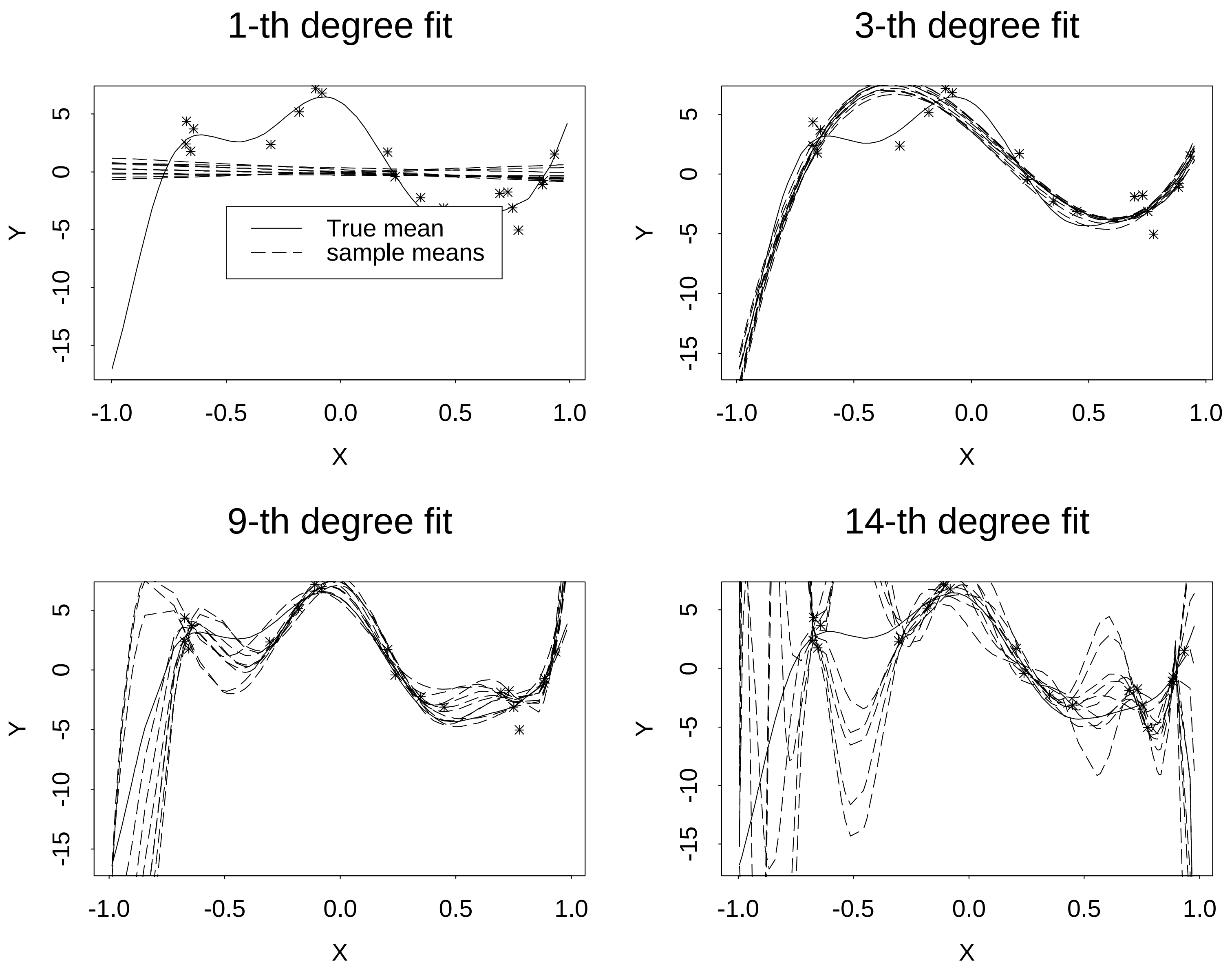


Linear Regression, sampling

To get a better idea of whats happening, we really need to measure the uncertainty in our predictions somehow. One way is to estimate the posterior standard deviation (so-called ”error bars”) and display it. Another is to sample from the posterior and look at the range of curves displayed. We do that next.

- You’ll see the range is tight when there are inadequate parameters for the model.
- And loose in those places where data is sparse.
- Of course, which looks best also depends on resultant error you expected in measurements.

Linear Regression, sampling



1-th degree fit

3-th degree fit

9-th degree fit

14-th degree fit

Next Week

- Review the material in B-Course's library.
- Play with some data on B-Course.