A Brief Introduction to Bayesian Inference

Peter Sykacek[1]

Department of Biotechnology

Bioinfromatics Research Group

BOKU University

peter.sykacek@boku.ac.at

The Next Three Hours

- Why should you bother?
- Introduction to Bayesian data analysis
- Priors, likelihoods and inference
- Bayesian view of the t-test
- Bayesian linear models
- Summary and outlook

Why Bother?

Moore’s Law:

PC 1984 5 MB Hard Drive

PC 2007 2 TB Hard Drive (4*500 GB) $\approx$ 400 Euro

How much paper on one PC in 2007 assuming 10.000 (single byte) characters per page ?

Why Bother?

Moore's Law:

PC 1984 5 MB Hard Drive

PC 2007 2 TB Hard Drive (4*500 GB) $\approx$ 400 Euro

How much paper on one PC in 2007 assuming 10.000 (single byte) characters per page ?

It is actually a stack of paper 20 km high!

2 TB $\approx$ $2 * 10^{12}$ byte

$= 2 * 10^8$ pages, assuming 1000 pages = 10 cm

a stack $2 * 10^5 * 10$ cm = $2 * 10^4$ m = 20 km

What About Data Generation?

Medical monitoring 1:

20 channels EEG+physiological signals 8 hours sleep at 200 Hz and 16 Bit :

$20 * 8 * 3600 * 200 * 2 \approx 230, 410^6$ byte ≈ 250 MB.

A single sleep lab with 8 recording units, operated at nights only, will generate one TB in just over a year.

What About Data Generation?

Medical monitoring 1:

20 channels EEG+physiological signals 8 hours sleep at 200 Hz and 16 Bit :

$20 * 8 * 3600 * 200 * 2 \approx 230, 410^6$ byte ≈ 250 MB.

A single sleep lab with 8 recording units, operated at nights only, will generate one TB in just over a year.

Medical monitoring 2:

An FMRI scanner, 1dm^3 volume, 10s temporal and 1mm^3 spatial resolution, 16 bit.

One scanner generates $10^6 * 360 * 2$ byte ≈ 720 MB per hour which fills 1 TB in about 58 days.

What About Data Generation?

Medical monitoring 1:
20 channels EEG+physiological signals 8 hours sleep at 200 Hz and 16 Bit :
$20 * 8 * 3600 * 200 * 2 \approx 230, 410^6$ byte ≈ 250 MB.
A single sleep lab with 8 recording units, operated at nights only, will generate one TB in just over a year.

Medical monitoring 2:
An FMRI scanner, 1dm^3 volume, 10s temporal and 1mm^3 spatial resolution, 16 bit.
One scanner generates $10^6 * 360 * 2$ byte ≈ 720 MB per hour which fills 1 TB in about 58 days.

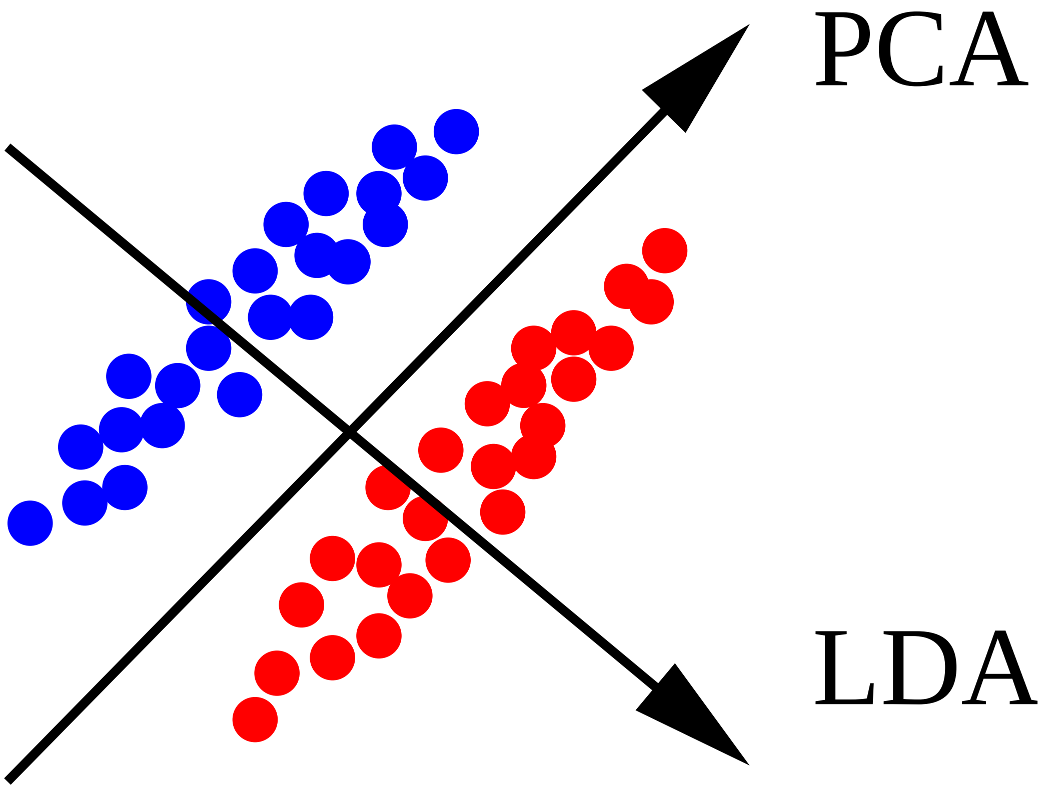
High throughput molecular biology:
A small lab produces up to 12 slides per 24 hours. One slide can contain up to 30.000 probes with ≈ 300 pixels/probe at 16 bit. Since we scan the entire array this is about 240 MB per 24 hours.

Clearly such ammounts can not be analysed manually. Statistics provides means to do that and thus to secure your job!

Why Understand Statistics?

Result = Data + Model!

Linear discriminant and principle component analysis can provide orthogonal projections of the same data.



Two Scenarios in Applied Life Sciences

1. Given measurements $\boldsymbol{x}_n$ and some corresponding dependent information y_n, we might ask: How are they related?

Two Scenarios in Applied Life Sciences

1. Given measurements $\boldsymbol{x}_n$ and some corresponding dependent information y_n, we might ask: How are they related?
2. Given two sets of measurements $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$, we might ask: Which of those are closer related to some corresponding dependent information y_n?

Two Scenarios in Applied Life Sciences

1. Given measurements $\boldsymbol{x}_n$ and some corresponding dependent information y_n, we might ask: How are they related?
2. Given two sets of measurements $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$, we might ask: Which of those are closer related to some corresponding dependent information y_n?

$->$ two instances of “inference” commonly found in applied life sciences.

We do for the moment ignore the problem where we have only some measurements $\boldsymbol{x}_n$ and ask how they are structured.

First Scenario

Suppose a life science experiment provided some noisy data $\mathcal{Z} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$. Note: $\boldsymbol{x}_n$ possibly multivariate i.e. vectors.

Based on $\mathcal{Z}$, we have an inference problem of finding an “optimal” relation between $\boldsymbol{x}$ and y:

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon(\lambda)$$

First Scenario

Suppose a life science experiment provided some noisy data $\mathcal{Z} = \{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N)\}$. Note: $\boldsymbol{x}_n$ possibly multivariate i.e. vectors.

Based on $\mathcal{Z}$, we have an inference problem of finding an “optimal” relation between $\boldsymbol{x}$ and y:

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon(\lambda)$$

Noise requires a deterministic and a random component.

$->$ Inherent uncertainty, y is a random variable!

Inference

Parameter Inference:

Implies knowing $f(\boldsymbol{x};\boldsymbol{\theta})$ and the noise model $\epsilon(\lambda)$ up to unknown parameters ($\boldsymbol{\theta}$ and λ) which we will be inferring from data.

Inference

Parameter Inference:

Implies knowing $f(\boldsymbol{x};\boldsymbol{\theta})$ and the noise model $\epsilon(\lambda)$ up to unknown parameters ($\boldsymbol{\theta}$ and λ) which we will be inferring from data.

Model Inference:

A more realistic assumption is that the model class is unknown and we will be inferring model class and parameters.

Assessing Model Parameters

Idea: subtract the deterministic part from y_n:

$$\epsilon_n = y_n - f(\boldsymbol{x}_n; \boldsymbol{\theta})$$

For convenience introduce $\mathcal{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and $\mathcal{D} = \{y_1, ..., y_N\}$. Assuming that ϵ_n are i.i.d samples, we get the likelihood function:

$$p(\mathcal{D}|\boldsymbol{\theta}, \lambda, \mathcal{X}) = \prod_n p(y_n|\boldsymbol{\theta}, \lambda, \boldsymbol{x}_n)$$

which is a suitable objective function to be maximized for $\boldsymbol{\theta}$ and λ.

A Major Problem

True model - linear regression, Gaussian noise:

$$p(y|\boldsymbol{x}) = f(\boldsymbol{x};\boldsymbol{\theta}) + \epsilon(\lambda)$$

$f(\boldsymbol{x};\boldsymbol{\theta}) = [1, \boldsymbol{x}^T]\boldsymbol{\theta}$ and $\epsilon(\lambda) = \mathcal{N}(\epsilon; 0, \lambda)$, with λ denoting “precision” (i. e. inverse variance).

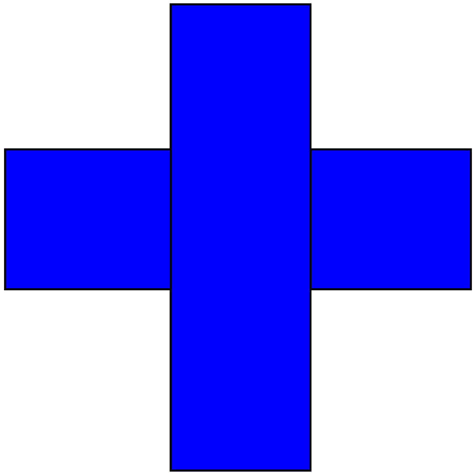
Finite sample size and different model classes: What is the maximum of the likelihood?

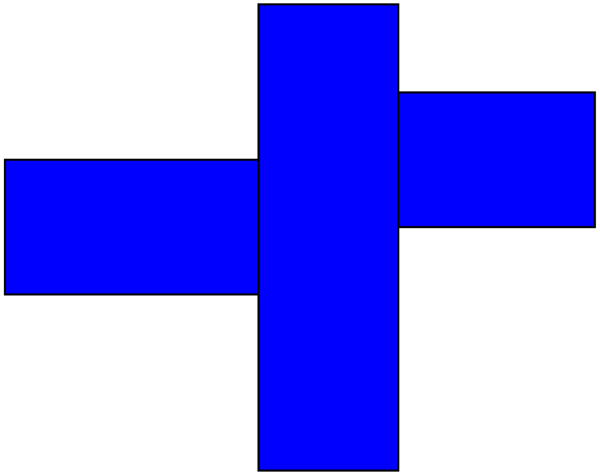
Think “phone book”: Perfect memorizing of all y_n, modelling error 0, $\lambda - > \infty$, $p(\mathcal{D}|\boldsymbol{\theta}, \lambda, \mathcal{X}) - > \infty$.

$- >$ likelihood unsuitable objective for model inference!

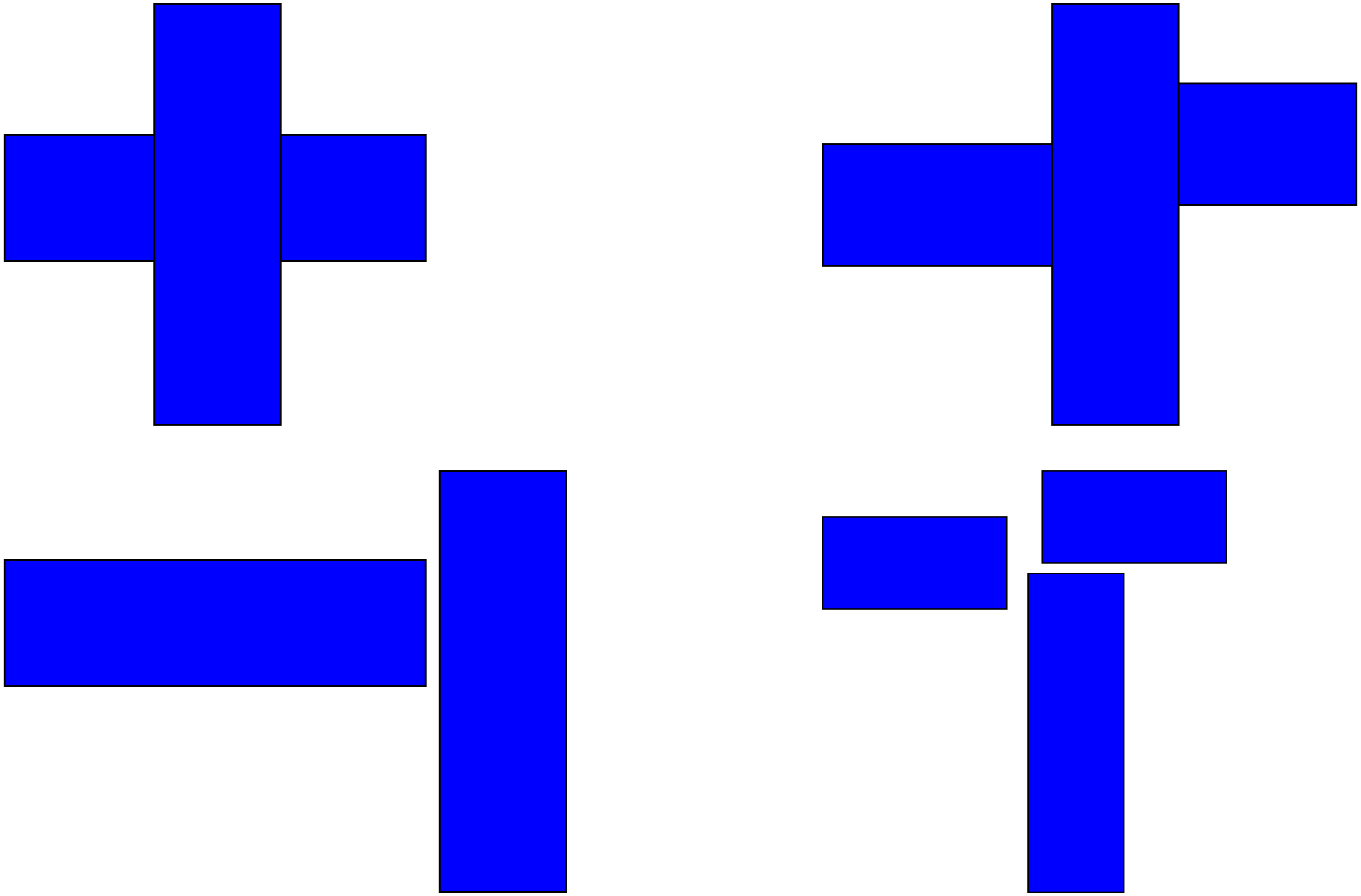
Why is memorizing useless?

Guess the Correct “Model”





Guess the Correct “Model”



Model comparison requires putting external objectives on top of likelihood! (AIC, BIC, etc.)

Occam's Razor

We implicitly apply Occam's Razor

William of Occam (or Ockham) (1288 - 1348)

Entia non sunt multiplicanda sine necessitate: Entities are not to be multiplied without necessity.

Interpretation: One should always opt for an explanation in terms of the fewest possible number of causes, factors, or variables.

Material from `http://en.wikipedia.org/wiki/William_of_Ockham`.

Bayesian Inference

Thomas Bayes (1701 - 1763)

Occam's Razor built in!
Two important consequences for "learning from data". Inference based on a decision theoretic framework

Bayesian Inference

Thomas Bayes (1701 - 1763)

Occam's Razor built in!

Two important consequences for "learning from data". Inference based on a decision theoretic framework

$$p(I|\mathcal{D}) = \frac{p(\mathcal{D}|I)p(I)}{p(\mathcal{D})}$$

1) Revise beliefs by Bayes theorem

Bayesian Inference

Thomas Bayes (1701 - 1763)

Occam's Razor built in!

Two important consequences for "learning from data". Inference based on a decision theoretic framework

$$p(I|\mathcal{D}) = \frac{p(\mathcal{D}|I)p(I)}{p(\mathcal{D})}$$

1) Revise beliefs by Bayes theorem

$\alpha_{opt} = \mathrm{argmax}_{\alpha} < u(\alpha) >$, where $< u(\alpha) >= \int_G u(\alpha, I)p(I|\mathcal{D})dI$.

2) Decisions by maximising expected utility

A Bayesian Dice Model - the Likelihood

Goal: inferring probabilities observing sides of a dice, i.e. $\pi = \{\pi_1, .., \pi_5, 1 - \sum_{k=1}^{5} \pi_k\}$
Data: N observations from rolling the dice.

A Bayesian Dice Model - the Likelihood

Goal: inferring probabilities observing sides of a dice, i.e. $\pi = \{\pi_1, .., \pi_5, 1 - \sum_{k=1}^{5} \pi_k\}$
Data: N observations from rolling the dice.

We need a likelihood function:
Throwing the dice once results in a multinomial one distribution over sides, i.e.
$P(I_n|\pi) = \prod_{k=1}^{6} \pi_k^{\delta(I_n=k)}$, where $I_n \in \{1, .., 6\}$.

Independence assumption $- >$ likelihood:
$p(\mathcal{D}|\pi) = \prod_n P(I_n|\pi)$, where $\mathcal{D} = \{I_1, ..., I_N\}$ denotes the N outcomes.

What is the final expression of the likelihood?

Bayesian Dice Model - the Prior

We typically use a conjugate prior: a convenient choice to remain within a functional family which is a known distribution. The Multinomial suggests a Dirichlet prior over π:

$$p(\pi) = \frac{\Gamma(\sum_{k=1}^{6} \alpha_k)}{\prod_{k=1}^{6} \Gamma(\alpha_k)} \prod_{k=1}^{6} \pi_k^{\alpha_k - 1}$$

$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$ is known as gamma function.

Write the definition of $\Gamma(\alpha)$ down! You will need it later during the lecture!

The α_k are hyper parameters of our model.

What is their logical meaning?

Bayesian Dice Model: the Posterior

Multiplying prior and likelihood and renormalising gives the posterior distribution over π as the result of Bayesian inference of the dice model:

$$p(\pi|\mathcal{D}) = \frac{1}{p(\mathcal{D})} \frac{\Gamma(\sum_{k=1}^{6} \alpha_k)}{\prod_{k=1}^{6} \Gamma(\alpha_k)} \prod_{k=1}^{6} \pi_k^{\alpha_k + n_k - 1}$$

where $p(\mathcal{D}) = \int_{\pi_1,..,\pi_6} p(\pi, \mathcal{D}) d\pi$ denotes the marginal likelihood, which is useful for model selection.

What is the functional form of the marginal likelihood ?

Iterative Inference

Given prior counts $\{\alpha_1, ..\alpha_k\}$ and data sets $\mathcal{D}_1 = \{I_1, ..., I_N\}$ and $\mathcal{D}_2 = \{I_{N+1}, ..., I_{N+M}\}$, using $p(\pi|\mathcal{D}_1)$ as prior for $\mathcal{D}_2$ will result in the same posterior $p(\pi|\mathcal{D}_1, \mathcal{D}_2)$ we get from the original prior and the pooled data $\mathcal{D} = \{I_1, .., I_{N+M}\}$:

$$p(\pi|\mathcal{D}_1) = \frac{\Gamma(\sum_k (\alpha_k + n_k))}{\prod_k \Gamma(\alpha_k + n_k)} \prod_k \pi_k^{\alpha_k + n_k - 1}$$

$$p(\pi|\mathcal{D}_1, \mathcal{D}_2) = \frac{\Gamma(\sum_k (\alpha_k + n_k + m_k))}{\prod_k \Gamma(\alpha_k + n_k + m_k)} \prod_k \pi_k^{\alpha_k + n_k + m_k - 1}$$

Since $n_k + m_k$ is the overall number of observations of side k this is equivalent to $p(\pi|\mathcal{D})$.

Applied Bayesian Decision Theory

Horse betting: bet x; choice α; uncertain outcome of race I. Bookmakers "odds" r_A and r_B (one + odds ratio) imply utility function $u(\alpha, I)$:

$\alpha \backslash I$	"A" wins	"B" wins
bet "A"	xr_A	0
bet "B"	0	xr_B
no bet	x	x

Need probability of $I = [A, B]$ i.e. respective horse wins. From previous observations (races) $\mathcal{D}$: $P(I = A|\mathcal{D}) = 0.7$ and $P(I = B|\mathcal{D}) = 0.3$.

Horse Betting ctd.

Calculate expected utility $u(\alpha) = \sum_I u(\alpha, I) P(I|\mathcal{D})$:

bet "A"	bet "B"	no bet
$0.7xr_A$	$0.3xr_B$	x

Maximise expected utility!

case	I	II	III
r_A	1.4	1.9	1.3
r_B	3.2	2.5	4.5

What are your decisions?

Horse Betting ctd.

Calculate expected utility $u(\alpha) = \sum_I u(\alpha, I)P(I|\mathcal{D})$:

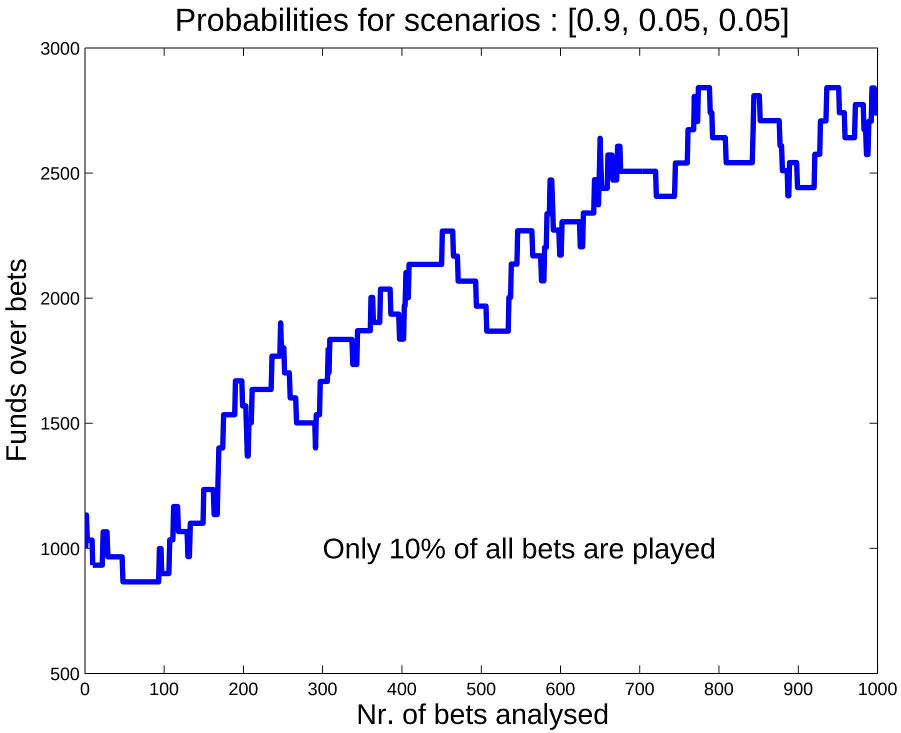
bet “A”	bet “B”	no bet
$0.7xr_A$	$0.3xr_B$	x

Maximise expected utility!

case	I	II	III
r_A	1.4	1.9	1.3
r_B	3.2	2.5	4.5

What are your decisions?

Can we earn money?



Inferring a Univariate Gaussian

Data $\mathcal{D} = \{x_1, .., x_N\}$: drawn from a univariate Gaussian with mean μ and precision λ.
Goal: inferring μ and λ, i.e. apply Bayes theorem:

$$p(\mu, \lambda|\mathcal{D}, g, h, l_0) = \frac{p(\mathcal{D}|\mu, \lambda)p(\mu|l_0)p(\lambda|g, h)}{p(\mathcal{D}|g, h, l_0)}$$

What is the precision?

Inferring a Univariate Gaussian

Data $\mathcal{D} = \{x_1, .., x_N\}$: drawn from a univariate Gaussian with mean μ and precision λ.
Goal: inferring μ and λ, i.e. apply Bayes theorem:

$$p(\mu, \lambda|\mathcal{D}, g, h, l_0) = \frac{p(\mathcal{D}|\mu, \lambda)p(\mu|l_0)p(\lambda|g, h)}{p(\mathcal{D}|g, h, l_0)}$$

What is the precision?
Univariate Gaussian distribution:

$$p(x_n|\mu, \lambda) = (2\pi)^{-\frac{1}{2}} |\lambda|^{\frac{1}{2}} \exp\left(-0.5\lambda(x_n - \mu)^2\right)$$

and Likelihood: $p(\mathcal{D}|\mu, \lambda) = \prod_n p(x_n|\mu, \lambda)$
Functional form of the likelihood?

Priors over μ and λ

Likelihood:

$$p(\mathcal{D}|\mu,\lambda) = (2\pi)^{-\frac{N}{2}}|\lambda|^{\frac{N}{2}} \exp\left(-0.5\lambda(N\mu^2 - 2\mu\sum_n x_n + \sum_n x_n^2)\right)$$

Conjugate prior for μ ?

Priors over μ and λ

Likelihood:

$$p(\mathcal{D}|\mu,\lambda) = (2\pi)^{-\frac{N}{2}}|\lambda|^{\frac{N}{2}}\exp\left(-0.5\lambda(N\mu^2 - 2\mu\sum_n x_n + \sum_n x_n^2)\right)$$

Conjugate prior for μ ?

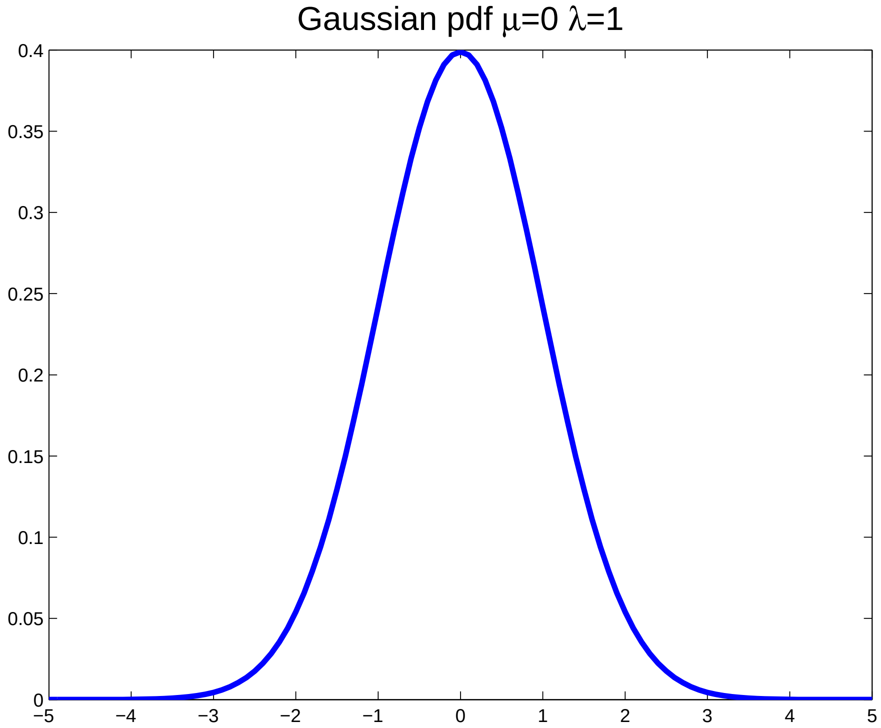
Priors:

$p(\mu|l_0) = (2\pi)^{-0.5}|l_0|^{0.5}\exp(-0.5 l_0\mu^2)$, zero mean Gaussian with precision $l_0 = \gamma\lambda$ “g-prior”

$p(\lambda|g,h) = \frac{h^g}{\Gamma(g)}|\lambda|^{(g-1)}\exp(-h\lambda)$, Gamma distribution with shape g and inverse scale h.

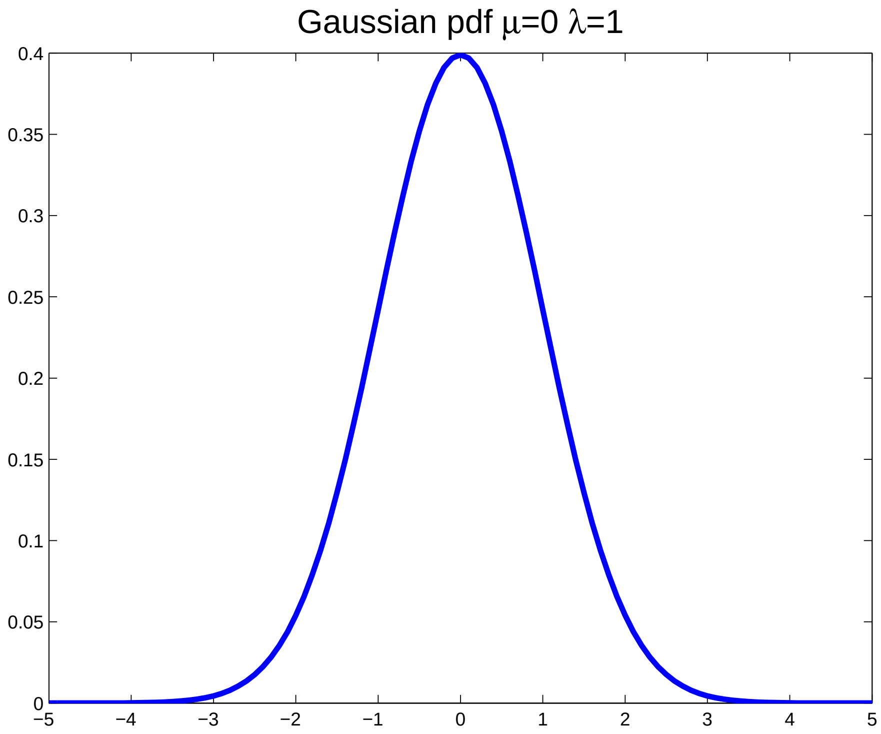
Priors ctd.

Gaussian defined for $x \in \Re$

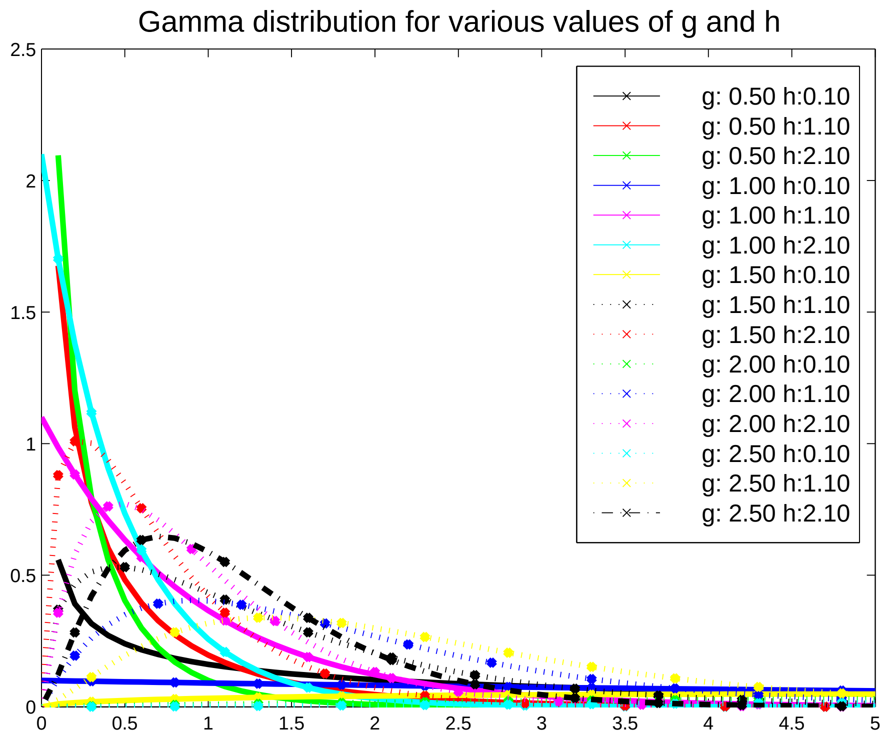


Priors ctd.

Gaussian defined for $x \in \Re$



Gamma defined for $x \in \Re | x > 0$



Prior Times Likelihood

$$p(\mathcal{D},\mu,\lambda|g,h,\gamma)=p(\mathcal{D}|\mu,\lambda)p(\mu|\lambda\gamma)p(\lambda|g,h)$$

$$=(2\pi)^{-\frac{N+1}{2}}\frac{h^g}{\Gamma(g)}|\gamma|^{\frac{1}{2}}|\lambda|^{(\frac{N+1}{2}+g-1)}$$

$$\times\exp\Big(-\lambda\Big(h+0.5\Big((\gamma+N)\mu^2-2\mu\sum_n x_n+\sum_n x_n^2\Big)\Big)\Big)$$

$$=(2\pi)^{-\frac{N+1}{2}}\frac{h^g}{\Gamma(g)}|\gamma|^{\frac{1}{2}}|\lambda|^{(\frac{N+1}{2}+g-1)}$$

$$\times\exp\Big(-\lambda\Big(h+0.5\Big(\Big(\sum_n x_n^2-\frac{(\sum_n x_n)^2}{\gamma+N}\Big)\Big)\Big)\Big)$$

$$\times\exp\Big(-\lambda 0.5(\gamma+N)\Big(\mu-\frac{\sum_n x_n}{\gamma+N}\Big)^2\Big)$$

For normalisation, integrate over λ and μ.

Integrating out λ

We need to solve:

$$\int_{\lambda=0}^{\infty} |\lambda|^{(\frac{N+1}{2}+g-1)} \exp(-\lambda\beta_0)d\lambda$$

Any ideas?

Integrating out λ

We need to solve:

$$\int_{\lambda=0}^{\infty} |\lambda|^{(\frac{N+1}{2}+g-1)} \exp(-\lambda\beta_0) d\lambda$$

Any ideas?

Setting $x = \lambda\beta_0$, and $d\lambda = \frac{dx}{\beta_0}$ we convert to a Gamma type integral $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$ and get:

$$\begin{aligned} p(\mathcal{D}, \mu | g, h, \gamma) = {} & (2\pi)^{-\frac{N+1}{2}} \frac{h^g}{\Gamma(g)} |\gamma|^{\frac{1}{2}} \Gamma\left(\frac{N+1}{2} + g\right) \\ & \times \left(h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma + N}\Big)\Big) + 0.5(\gamma + N)\Big(\mu - \frac{\sum_n x_n}{\gamma + N}\Big)^2\right)^{-\left(\frac{N+1}{2}+g\right)} \end{aligned}$$

Further Analysis of $p(\mathcal{D}, \mu | g, h, \gamma)$

$$p(\mathcal{D}, \mu | g, h, \gamma) = (2\pi)^{-\frac{N+1}{2}} \frac{h^g}{\Gamma(g)} |\gamma|^{\frac{1}{2}} \Gamma\left(\frac{N+1}{2} + g\right)$$

$$\times \left(h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma + N}\Big)\Big)\right)^{-\left(\frac{N+1}{2}+g\right)}$$

$$\times \left(1 + \frac{0.5(\gamma+N)\Big(\mu - \frac{\sum_n x_n}{\gamma+N}\Big)^2}{h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma+N}\Big)\Big)}\right)^{-\left(\frac{N+1}{2}+g\right)}$$

Compare with student-t distribution:

$$p(\mu|\theta, \kappa, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} |\kappa|^{0.5} (\nu\pi)^{-0.5} \left(1 + \frac{(\mu - \theta)^2 \kappa}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$->$ last factor proportional to student-t distrbution over μ

Analysis of $p(\mathcal{D}, \mu | g, h, \gamma)$ ctd.

Comparing coefficients:

$$\theta = \frac{\sum_n x_n}{N+\gamma} \quad , \; \nu = N + 2g$$

$$\kappa = \frac{(N+2g)(N+\gamma)}{2h + \sum_n x_n^2 - \left(\sum_n x_n\right)^2/(N+\gamma)}$$

$$p(\mathcal{D}, \mu|g,h,\gamma) = (2\pi)^{-\frac{N+1}{2}} \frac{h^g}{\Gamma(g)} |\gamma|^{\frac{1}{2}} \Gamma\left(\frac{N+1}{2} + g\right)$$

$$\times \left(h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma + N}\Big)\Big)\right)^{-\left(\frac{N+1}{2}+g\right)}$$

$$\times \frac{\Gamma\left(\frac{N+2g}{2}\right)}{\Gamma\left(\frac{N+2g+1}{2}\right)} \left|\frac{(N+2g)(N+\gamma)}{2h + \sum_n x_n^2 - \left(\sum_n x_n\right)^2/(N+\gamma)}\right|^{-0.5} ((N+2g)\pi)^{0.5}$$

$$\times \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} |\kappa|^{0.5} (\nu\pi)^{-0.5} \left(1 + \frac{(\mu-\theta)^2\kappa}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Any ideas how to get the marginal likelihood $p(\mathcal{D}|g,h,\gamma)$?

Marginal Likelihood and Posterior

$$p(\mathcal{D}|g,h,\gamma) = (2\pi)^{-\frac{N+1}{2}} \frac{h^g}{\Gamma(g)} |\gamma|^{\frac{1}{2}} \left(h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma + N}\Big)\Big)\right)^{-\left(\frac{N+1}{2}+g\right)}$$

$$\times\Gamma\left(\frac{N+2g}{2}\right)\left|\frac{(N+2g)(N+\gamma)}{2h+\sum_n x_n^2 - \left(\sum_n x_n\right)^2/(N+\gamma)}\right|^{-0.5} ((N+2g)\pi)^{0.5}$$

$$p(\mu,\lambda|\mathcal{D},g,h,\gamma) = \left(h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma + N}\Big)\Big)\right)^{\left(\frac{N+1}{2}+g\right)}$$

$$\times\frac{1}{\Gamma\left(\frac{N+2g}{2}\right)\sqrt{((N+2g)\pi)}}\left|\frac{(N+2g)(N+\gamma)}{2h+\sum_n x_n^2 - \left(\sum_n x_n\right)^2/(N+\gamma)}\right|^{0.5}$$

$$\times|\lambda|^{(\frac{N+1}{2}+g-1)} \exp\Big(-\lambda\Big(h+0.5\Big((\gamma+N)\mu^2 - 2\mu\sum_n x_n + \sum_n x_n^2\Big)\Big)\Big)$$

A MatLab Implementation

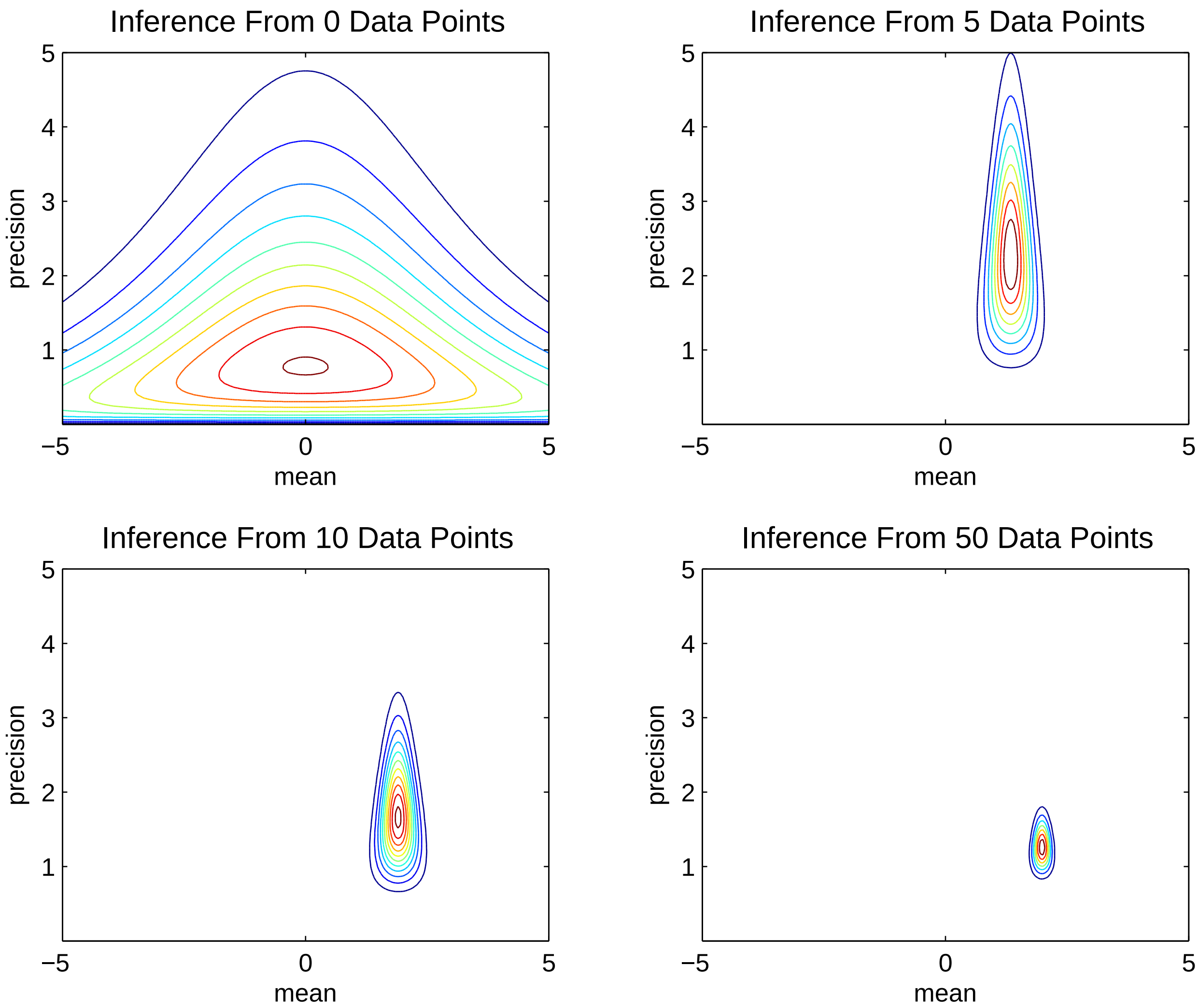
Note the implementation on the log scale!

```
function [mrgllh]=prcmn_gauss_mrglh(data, g, h, gam)
% function [mrgllh]=prcmn_gauss_mrglh(data, g, h, gam)
% calculates the log marginal likelihood of inferring a
% univariate Gaussian under a g-prior like scenario.
%
% (C) P. Sykacek 2007 <peter@sykacek.net>

data=data(:);
ndat=length(data);
sum_x_sqr=sum(data.^2);
sqr_sum_x=sum(data).^2;
mrgllh=-(ndat+1)/2 * log(2*pi) + g*log(h) - gammaln(g) + 0.5*log(gam);
mrgllh=mrgllh-((ndat+1)/2+g)*log(h+0.5*(sum_x_sqr-sqr_sum_x/(ndat+gam)));
mrgllh=mrgllh+gammaln(ndat/2+g)-0.5*(log(ndat+2*g)+log(ndat+gam)-...
    log(2*h+sum_x_sqr-sqr_sum_x/(ndat+gam)));
mrgllh=mrgllh+0.5*(log(ndat+2*g)+log(pi));
```

Posterior Dependency on Data Size

Prior settings: $g = 1.2$, $h = 0.9$ and $\gamma = 0.1$



Bayesian Model Selection I

All aspects of Bayesian inference:

Parameter inference:

$$p(\boldsymbol{\theta}|\mathcal{D}, I) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)}{p(\mathcal{D}|I)}$$

Note: $p(\mathcal{D}|I) = \int_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}, I)p(\boldsymbol{\theta}|I)d\boldsymbol{\theta}$

Novel part: By including an indicator I, we made the model class explicit.

Bayesian Model Selection II

Reasoning about different model classes I:

$$P(I|\mathcal{D}) = \frac{P(I)p(\mathcal{D}|I)}{p(\mathcal{D})}$$

Note: $p(\mathcal{D}|I)$ is just the normalisation constant from parameter inference.

The above denominator is the normalisation constant $p(\mathcal{D}) = \sum_I P(I)p(\mathcal{D}|I)$.

Renormalising the maginal likelihood of model class I multiplied by its prior probability gives thus the posterior probability of model class I under the data $\mathcal{D}$.

Bayesian Model Selection III

If we have K models, we may chose $P(I) = \frac{1}{K}$ to reflect “ignorance”.

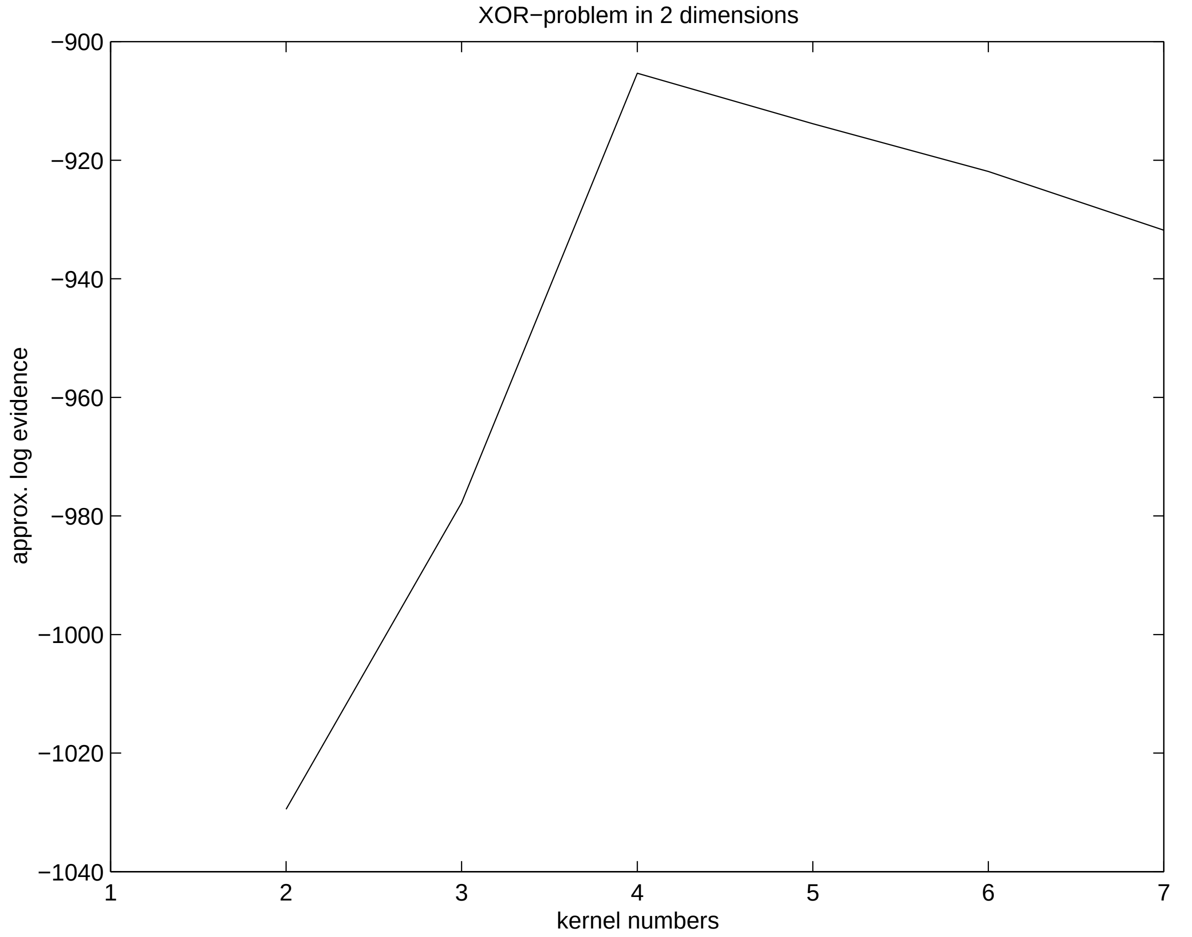
Model selection will choose model I with the largest posterior probability.

For equal priors, we select the model with the largest marginal likelihood. Unlike maximising the likelihood this quantity does not necessary lead to the most complex model winning!

If several model classes are equally probable, we should use $P(I|\mathcal{D})$ for model averaging.

Typical Behaviour

Plot of (approximate) log marginal likelihood in a binary regression problem (XOR-structure).



The Bayesian Version of a Paired T-Test

The classical paired t-test infers, whether some data are unlikely under the null hypothesis of being a zero mean Gaussian with unknown variance.

The Bayesian alternative is inferring the posterior probabilities, whether a zero mean Gaussian ($I = 0$), or a generic Gaussian ($I = 1$) are more probable under the dataset.

We choose uninformative priors $P(I = 0) = P(I = 1) = 0.5$ and need in addition the marginal likelihoods. As we know the marginal likelihood of the generic Gaussian already, we need only consider the zero mean Gaussian model.

Zero Mean Gaussian Model

Likelihood:

$$p(\mathcal{D}|\lambda) = (2\pi)^{-\frac{n}{2}} |\lambda|^{\frac{N}{2}} \exp\left(-0.5\lambda \sum_n x_n^2\right)$$

and Gamma prior over λ:

$$p(\lambda|g, h) = \frac{h^g}{\Gamma(g)} \lambda^{g-1} \exp(-\lambda h)$$

Derive the marginal likelihood!

Marginal Likelihoods and $P(I|\mathcal{D})$

Zero mean Gaussian:

$$p(\mathcal{D}|g,h,I=0) = \frac{h^g}{\Gamma(g)}(2\pi)^{-\frac{n}{2}}\left(h + 0.5\sum_n x_n^2\right)^{-\left(\frac{N}{2}+g\right)}\Gamma\left(\frac{N}{2}+g\right)$$

Full Gaussian (from previous calculations):

$$p(\mathcal{D}|g,h,\gamma,I=1) = (2\pi)^{-\frac{N+1}{2}}\frac{h^g}{\Gamma(g)}|\gamma|^{\frac{1}{2}}\left(h + 0.5\Big(\Big(\sum_n x_n^2 - \frac{(\sum_n x_n)^2}{\gamma+N}\Big)\Big)\right)^{-\left(\frac{N+1}{2}+g\right)}$$

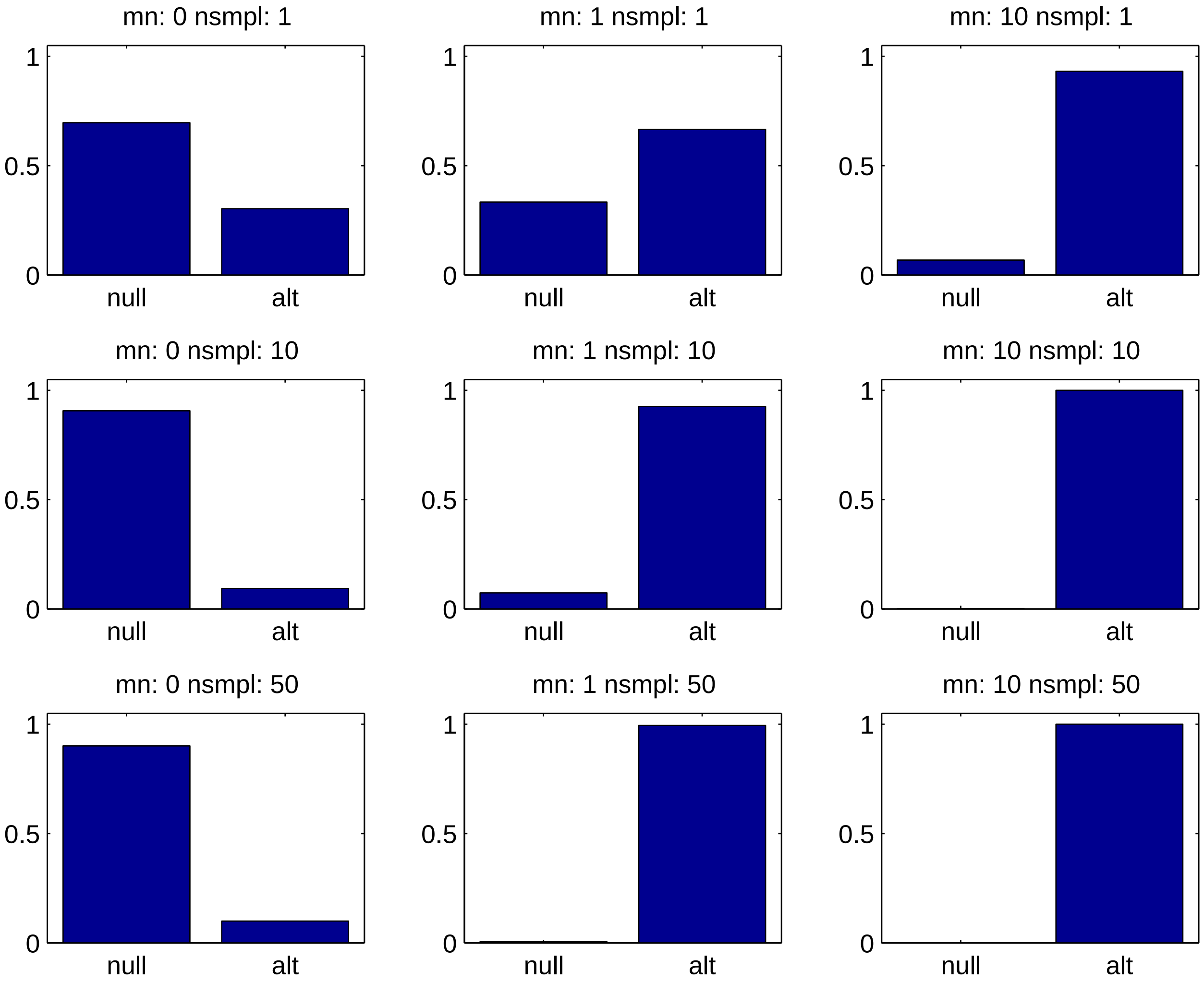
$$\times\Gamma\left(\frac{N+2g}{2}\right)\left|\frac{(N+2g)(N+\gamma)}{2h+\sum_n x_n^2 - \left(\sum_n x_n\right)^2/(N+\gamma)}\right|^{-0.5}((N+2g)\pi)^{0.5}$$

$P(I|\mathcal{D})$ from log marginal likelihoods, where $\log(p(\mathcal{D},I)) = \log(p(\mathcal{D}|I)) + \log(p(I))$:

$$P(I=i|\mathcal{D}) = \frac{1}{1+\sum_{j\neq i}\exp\left(\log(p(\mathcal{D},I=j)) - \log(p(\mathcal{D},I=j))\right)}$$

“Bayesian T-Test” Applied

Priors: $g = 1.2$, $h = 0.9$, $\gamma = 0.1$ and $P(I) = 0.5$



Bayesian Linear Regression

$y_n = \boldsymbol{x}_n^T \boldsymbol{\theta} + \epsilon_n$, where, $\epsilon_n \sim \mathcal{N}(0, \lambda)$, i.i.d. zero mean Gaussian with unknown precision λ:

$$p(y_n|\boldsymbol{\theta}, \lambda, \boldsymbol{x}_n) = (2\pi)^{-\frac{1}{2}} |\lambda|^{\frac{1}{2}} \exp(-0.5\lambda(y_n - \boldsymbol{x}_n^T \boldsymbol{\theta})^2)$$

Gamma prior over λ:

$$p(\lambda|g, h) = \frac{h^g}{\Gamma(g)} \lambda^{g-1} \exp(-\lambda h)$$

g-prior over d-dim vector of reg. coeffs. $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\gamma) = (2\pi)^{-\frac{d}{2}} |\lambda|^{\frac{d}{2}} |\gamma|^{\frac{d}{2}} \exp(-0.5\lambda \boldsymbol{\theta}^T \gamma I \boldsymbol{\theta})$$

Likelihood and Joint Distribution

We use matrix notation:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \dots \\ \dots \\ \boldsymbol{x}_N^T \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ \dots \\ y_N \end{bmatrix}$$

$$p(\mathcal{D}|\boldsymbol{\theta}, \lambda, \boldsymbol{X}) = (2 * pi)^{-\frac{N}{2}} |\lambda|^{\frac{N}{2}} \exp(-0.5\lambda(\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y}))$$

Priors times Likelihood:

$$p(\mathcal{D}, \boldsymbol{\theta}, \lambda | g, h, \gamma, \boldsymbol{X}) = (2\pi)^{-\frac{N+d}{2}} |\gamma|^{\frac{d}{2}} \frac{d^g}{\Gamma(g)} |\lambda|^{\frac{N+d}{2}+g+1}$$

$$\exp(-\lambda(h + 0.5(\boldsymbol{\theta}^T \gamma \boldsymbol{I} \boldsymbol{\theta} + (\boldsymbol{X\theta} - \boldsymbol{y})^T(\boldsymbol{X\theta} - \boldsymbol{y}))))$$

Note again the Gamma type integral ...

Marginal Likelihood

Similar to the Gaussian before, we integrate out λ, recover a multivariate Student-t distribution and renormalise to find the marginal likelihood:

$$
\begin{aligned}
p(\mathcal{D}|g,h,\gamma,\boldsymbol{X}) = {} & (2\pi)^{-\frac{N+d}{2}} |\gamma|^{\frac{d}{2}} \frac{h^g}{\Gamma(g)} \Gamma\left(\frac{N+2g}{2}\right) \\
& \times \left(h + 0.5\left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\left(\gamma I + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right)\right)^{-\left(\frac{N+d}{2}+g\right)} \\
& \times \left|\frac{(N+2g)(\gamma I + \boldsymbol{X}^T\boldsymbol{X})}{4h + 2\left(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\left(\gamma I + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right)}\right|^{-\frac{1}{2}}
\end{aligned}
$$

Note: matrix equations! Last term is a determinant.

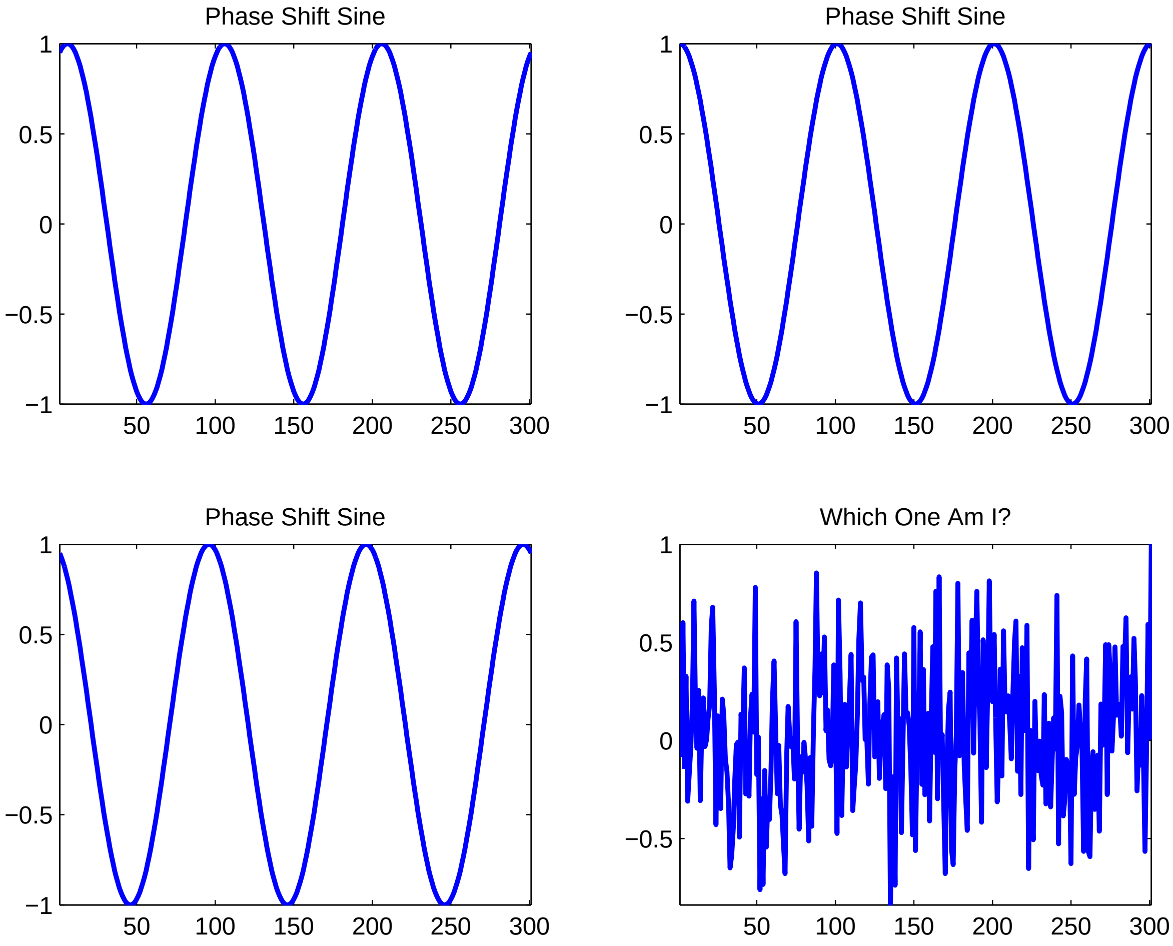
An Implementation

Again on the log scale!

```
function [mrgllh]=bayeslinreg_mrgllh(X, y, g, h, gam)
% calculates the log marginal likelihood for linear regression
% under a g-prior like scenario.
% X:      regressors
% y:      response variables
% g,h:    gamma prior over precision of Gaussian noise residuals
% gam:    g.prior factor for Gaussian prior over regression coefficients
% mrgllh: log marginal likelihood of the model
% (C) P. Sykacek 2007 <peter@sykacek.net>
ndat=size(X,1);
np=size(X,2);
gam_XtX=eye(np)*gam+X’*X;
ytXinvXy=y’*X*pinv(gam_XtX)*X’*y;
mrgllh=-0.5*(ndat+np)*log(2*pi)+0.5*np*log(gam)+g*log(h)-gammaln(g);
mrgllh=mrgllh+gammaln(0.5*(ndat+2*g));
mrgllh=mrgllh-(0.5*(ndat+np)+g)*log(h+0.5*(y’*y-ytXinvXy));
mrgllh=mrgllh-0.5*log(det((ndat+2*g)/(4*h+2*(y’*y-ytXinvXy))*gam_XtX));
```

Can Help Here?

Detect phase of a noisy sine wave.



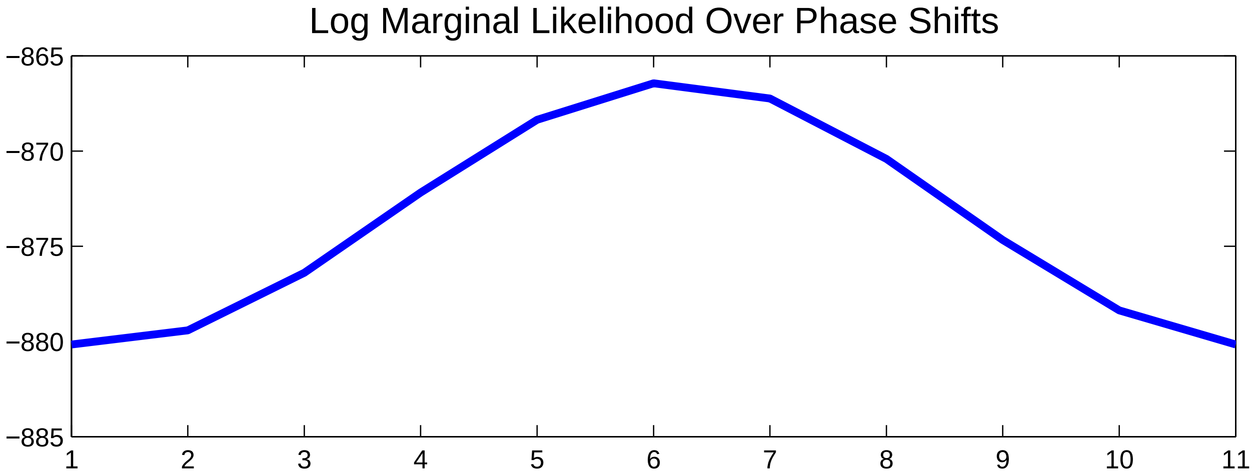
Noisy sine wave as response $\boldsymbol{y}$.

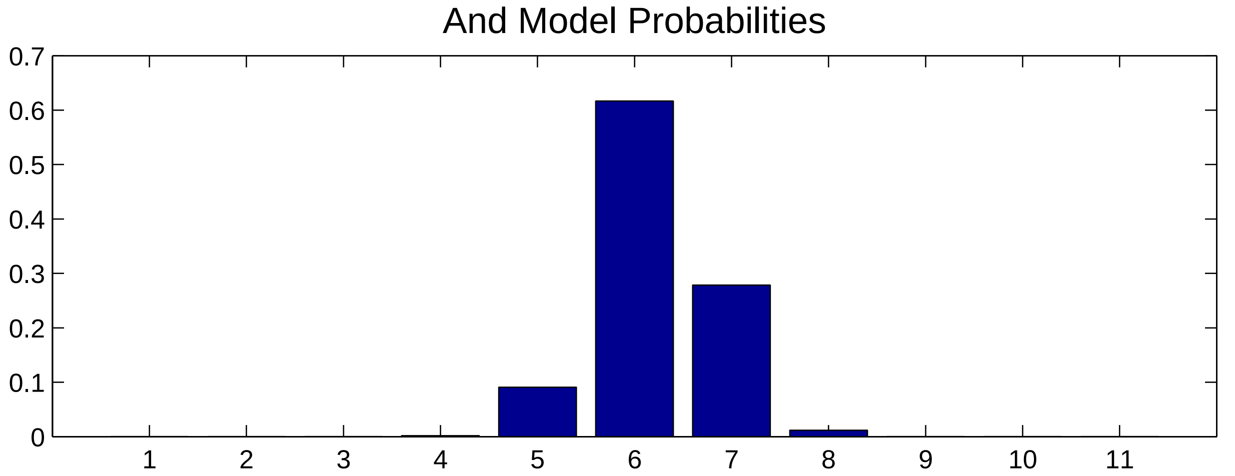
Try all phase shifted sines as regressors $\boldsymbol{x}$.

And compare marginal likelihoods.

Result Using Equal Priors

We find the largest evidence for the middle position:





Predictive Distribution

Bayesian predictions require predicting with uncertainty. This is obtained by providing a predictive distribution:

$$p(\eta|\boldsymbol{\xi}, \mathcal{D}, g, h, \gamma) = \int_{\lambda} \int_{\boldsymbol{\theta}} p(\eta|\boldsymbol{\xi}, \lambda, \boldsymbol{\theta}) p(\lambda, \boldsymbol{\theta}|\mathcal{D}, g, h, \gamma) d\lambda d\boldsymbol{\theta}$$

η: unknown prediction of regressor $\boldsymbol{\xi}$. Here:

$$p(\eta|\boldsymbol{\xi}, \lambda, \boldsymbol{\theta}) = (2\pi)^{-\frac{1}{2}} |\lambda|^{\frac{1}{2}} \exp(-0.5\lambda(\eta - \boldsymbol{\xi}^T\boldsymbol{\theta})^2)$$

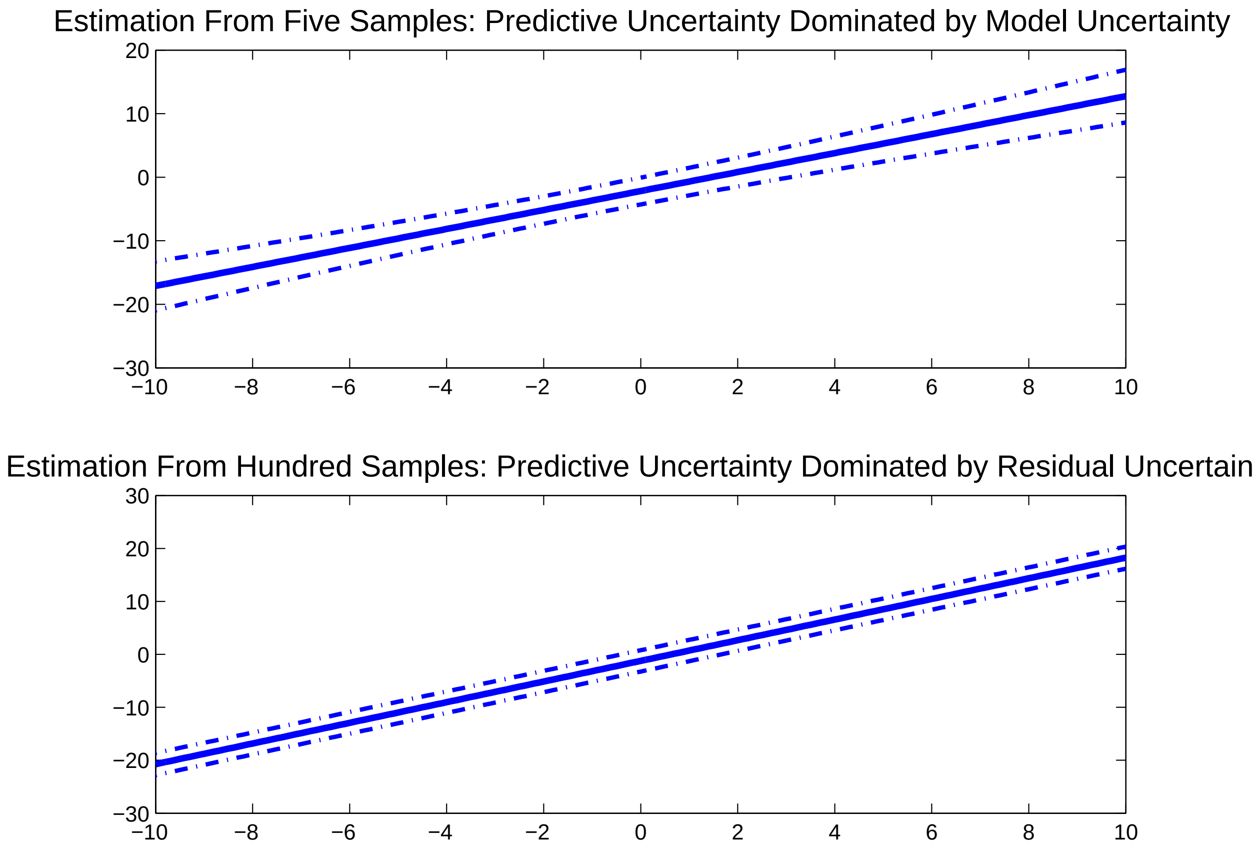
Solution: multiply $p(\eta|\boldsymbol{\xi}, \lambda, \boldsymbol{\theta})$ with the likelihood function and priors and integrate out $\boldsymbol{\theta}$ and λ. Your guess for the functional form of the predictive distribution?

You are Right! It’s a Student-t!

with

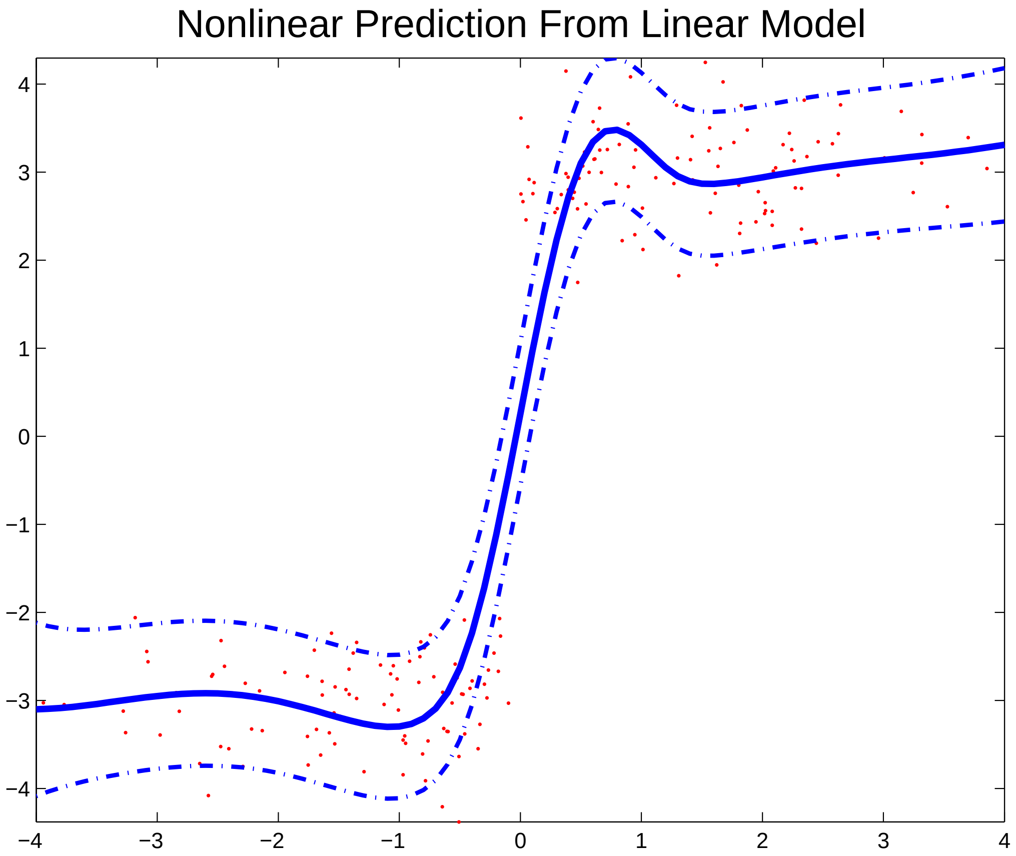
$$
\begin{aligned}
\nu &= N + 2g \\
\hat{\eta} &= \frac{\boldsymbol{\xi}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{X}^T \boldsymbol{y}}{1 - \boldsymbol{\xi}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\xi}} \\
\lambda_t &= \frac{\nu(1 - \boldsymbol{\xi}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\xi})}{2h + \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{\Lambda}^{-1} \boldsymbol{X}^T \boldsymbol{y} - (1 - \boldsymbol{\xi}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\xi}) \hat{\eta}^2} \\
\boldsymbol{\Lambda} &= \gamma I + \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{\xi} \boldsymbol{\xi}^T
\end{aligned}
$$

Linear Predictive Distributons



Nonlinear Predictions

As long as we have a linear in the parameters model!



Trick: project x into a nonlinear space and perform linear regression.

Summary

Model inference is based on Bayes theorem:

$$P(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}$$

and marginalisation:

$$P(I|\mathcal{D}) = \frac{\int_\theta p(\mathcal{D},\theta|I)d\theta P(I)}{\sum_I \int_\theta p(\mathcal{D},\theta|I)p(I)d\theta}$$

Inference results are either decisions after maximising expected utilities or posteriors summarising all uncertainty. An important advantage of Bayesian statistics is to provide a consistent framework for all inference tasks.

Outlook

This lecture captured only very simple models that gave rise to analytically tractable calculations.

For models which include nonlinearities the integrals can not be solved analytically and explicit (exact) solutions do not exist.

If you are interested in advanced Bayesian methods that allow solving more complex problems you are warmly invited to attend 793.492 “Bayesian Data Analysis in the Life Scienes”. This new 3.0 hrs VU starts in winter term 2007/2008. It will cover advanced aspects and include practical analysis sessions (2*8 hrs theory and 3 days blocked in the PC lab).