AARHUS
UNIVERSITY

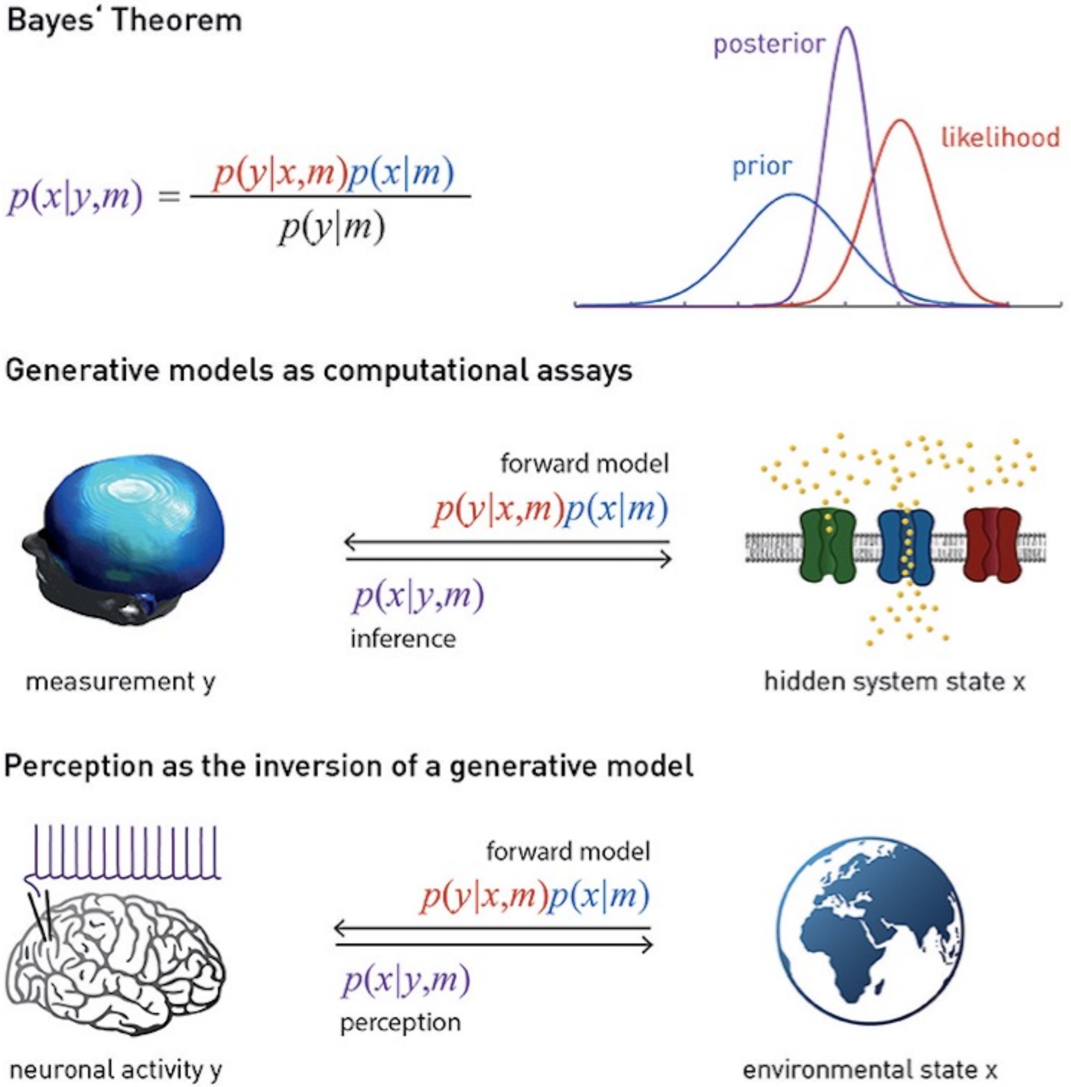
Data Science 5

Chris Mathys

Master's Degree Programme in Cognitive Science

Spring 2023

The mind as a data scientist



Bayes' Theorem

$$p(x|y,m) = \frac{p(y|x,m)p(x|m)}{p(y|m)}$$

Generative models as computational assays

Perception as the inversion of a generative model

Stephan et al. (2016), Front. Hum. Neurosci., 10:550

The brain as a data scientist

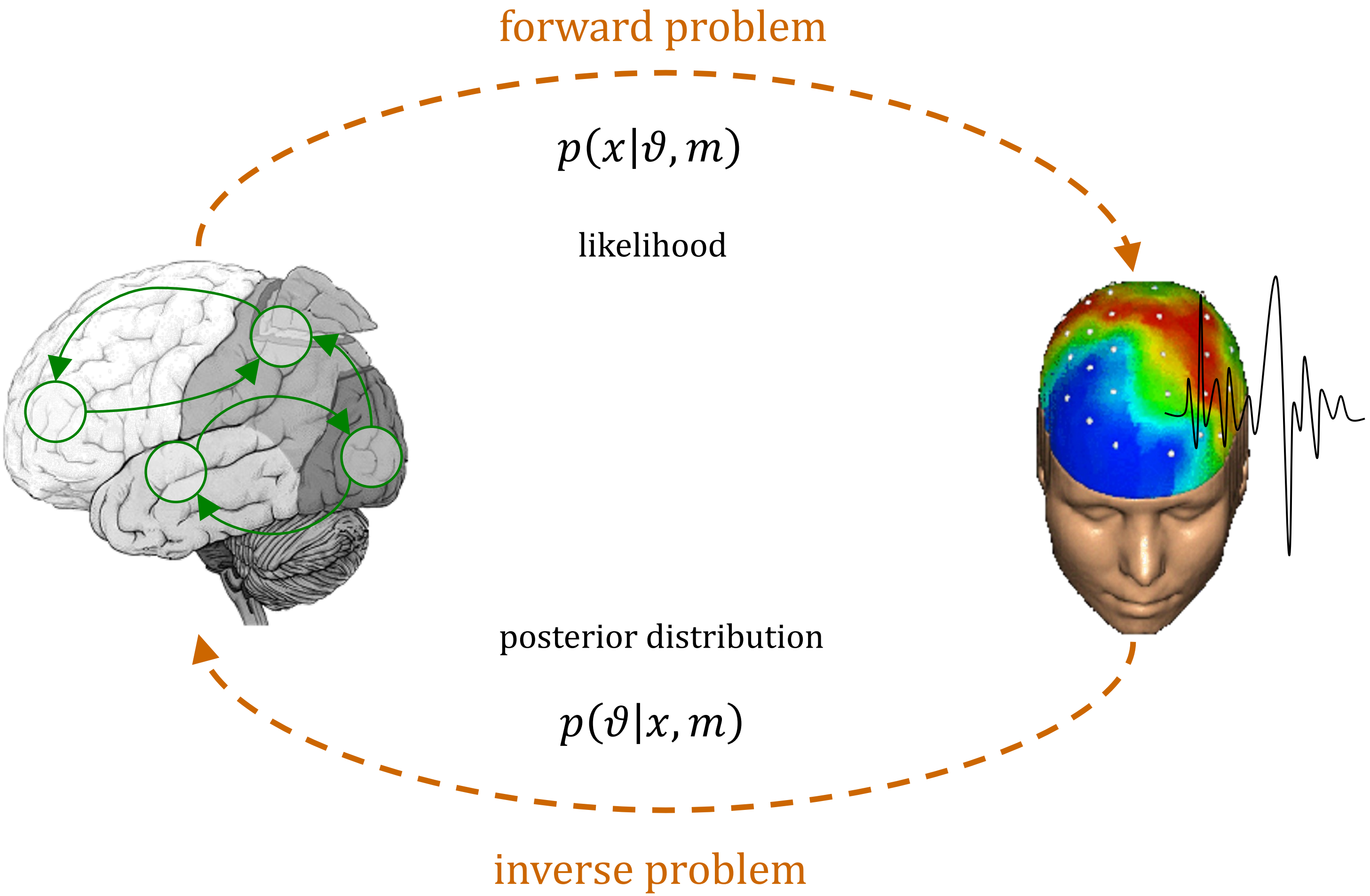
‘Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism.’

— Helmholtz (1867), Handbuch der physiologischen Optik
[Treatise on Physiological Optics], p. 428

In other words: the mind actively infers the objects of its perception in what Helmholtz calls ‘unconscious inference’ (p. 430).

If this is true, it means that **to understand the mind and the brain, we need to understand the mechanics of inference,** i.e. the mechanics of updating predictions.

Neuroimaging as a data science problem



forward problem

$p(x|\vartheta, m)$

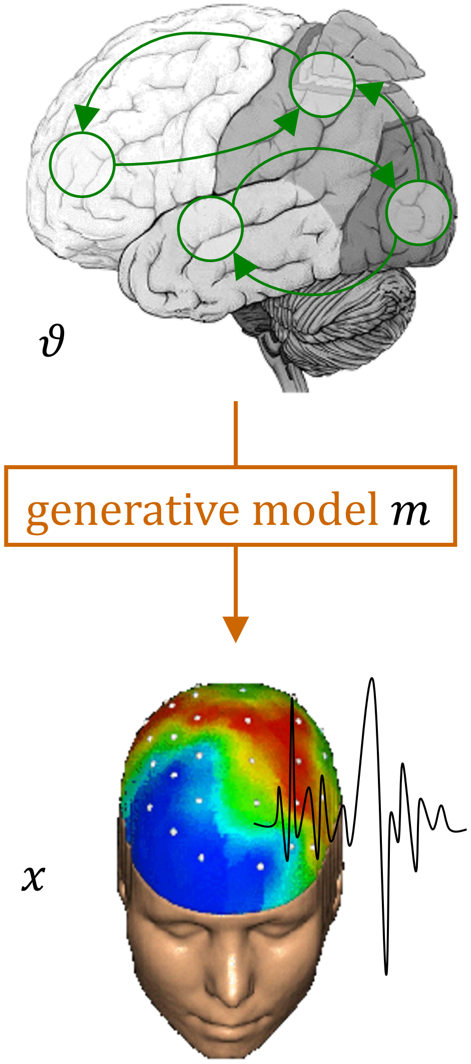
likelihood

posterior distribution

$p(\vartheta|x, m)$

inverse problem

Inference in neuroimaging

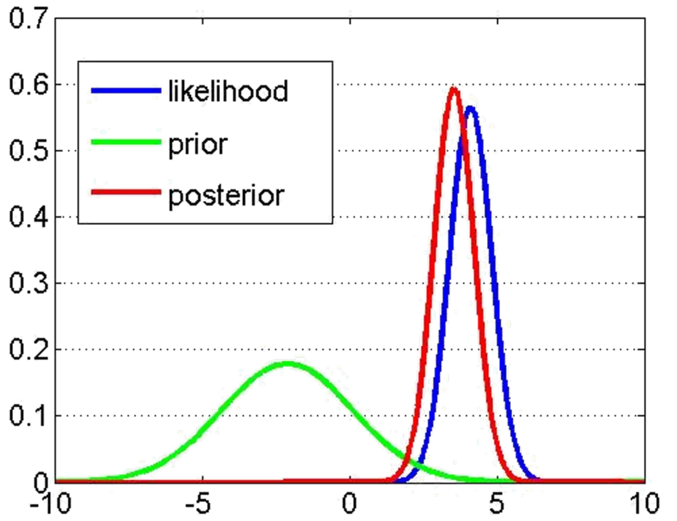


generative model m

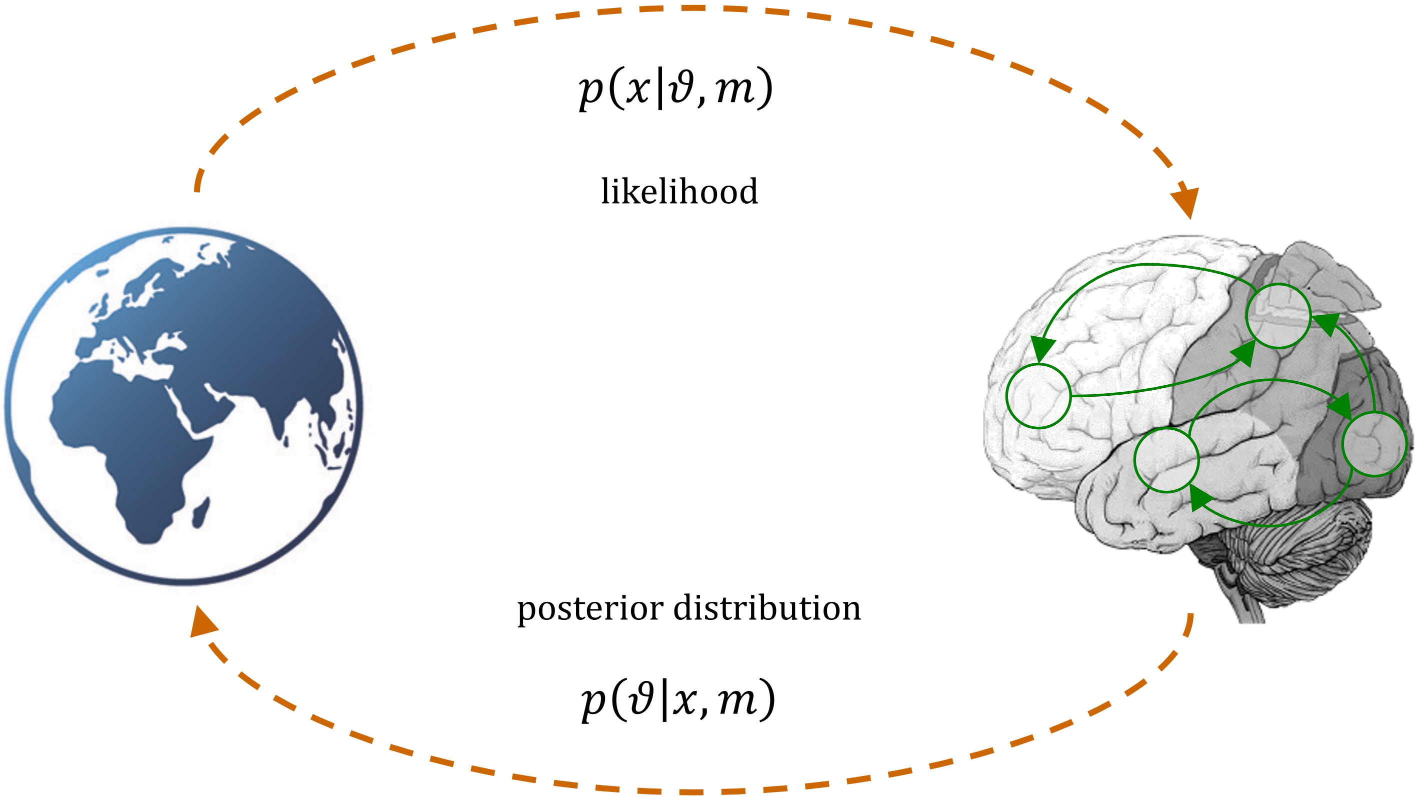
Likelihood: $p(x|\vartheta, m)$

Prior: $p(\vartheta|m)$

Bayes’ theorem: $p(\vartheta|x, m) = \dfrac{p(x|\vartheta, m)p(\vartheta|m)}{p(x|m)}$



And again: the brain as data scientist



forward problem

$p(x|\vartheta, m)$

likelihood

posterior distribution

$p(\vartheta|x, m)$

inverse problem

A very simple example of updating in response to new information

Imagine the following situation:

You're on a boat, you're lost in a storm and trying to get back to shore. A lighthouse has just appeared on the horizon, but you can only see it when you're at the peak of a wave. Your GPS etc., has all been washed overboard, but what you can still do to get an idea of your position is to measure the angle between north and the lighthouse. These are your measurements (in degrees):

76, 73, 75, 72, 77

What number are you going to base your calculation on?

Right. The mean: 74.6. How do you calculate that?

Updating the mean of a series of observations

The usual way to calculate the mean $\bar{u}$ of $u_1, u_2, \ldots, u_n$ is to take

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$$

This requires you to remember all u_i, which can become inefficient. Since the measurements arrive sequentially, we would like to update $\bar{u}$ sequentially as the u_i come in – without having to remember them.

It turns out that this is possible. After some algebra (see next slide), we get

$$\bar{u}_{n+1} = \bar{u}_n + \frac{1}{n+1}(u_{n+1} - \bar{u}_n)$$

Updating the mean of a series of observations

Proof of sequential update equation:

$$\bar{u}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} u_i = \frac{1}{n+1}\left(u_{n+1} + n \cdot \frac{1}{n}\sum_{i=1}^{n} u_i\right) =$$

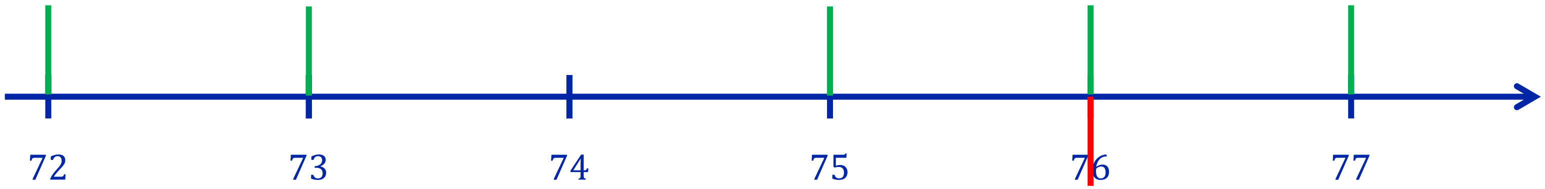
$$= \frac{1}{n+1}(u_{n+1} + n\bar{u}_n) = \frac{1}{n+1}(u_{n+1} - \bar{u}_n + (n+1)\bar{u}_n)$$

$$= \bar{u}_n + \tfrac{1}{n+1}(u_{n+1} - \bar{u}_n)$$

q.e.d.

Updating the mean of a series of observations

The seqential updates in our example now look like this:



$$\bar{u}_1 = 76$$

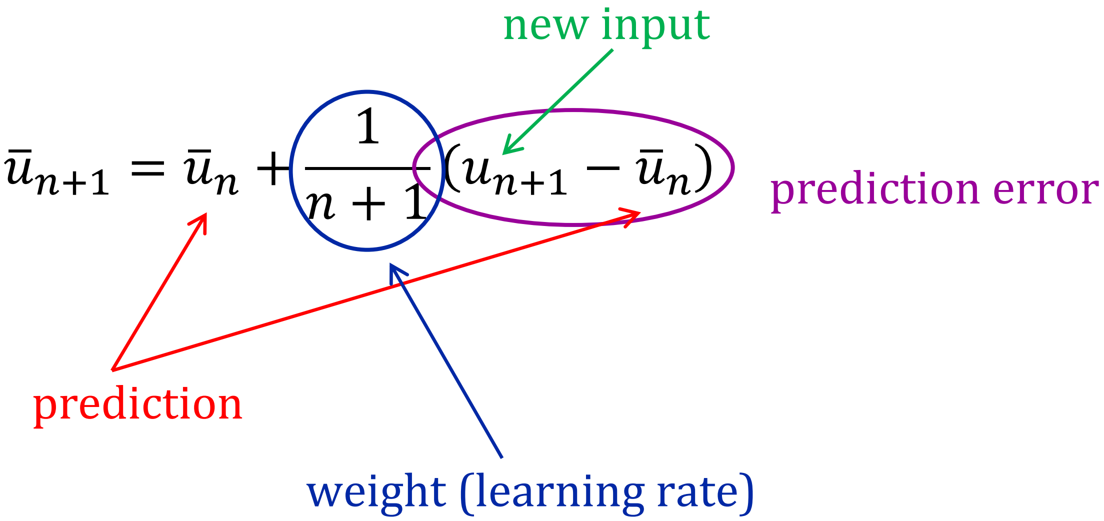
$$\bar{u}_2 = 76 + \frac{1}{2}(73 - 76) = 74.5$$

$$\bar{u}_3 = 74.5 + \frac{1}{3}(75 - 74.5) = 74.\overline{6}$$

$$\bar{u}_4 = 74.\overline{6} + \frac{1}{4}(72 - 74.\overline{6}) = 74$$

$$\bar{u}_5 = 74 + \frac{1}{5}(77 - 74) = 74.6$$

What are the building blocks of the updates we've just seen?



$$\bar{u}_{n+1} = \bar{u}_n + \frac{1}{n+1}(u_{n+1} - \bar{u}_n)$$

new input

prediction error

prediction

weight (learning rate)

Is this a general pattern?

More specifically, does it generalize to Bayesian inference?

Indeed, it turns out that in many cases, Bayesian inference can be based on parameters that are updated using **precision-weighted prediction errors.**

Updates in a simple Gaussian model

Think boat, lighthouse, etc., again, but now we're doing Bayesian inference.

Before we make the next observation, our belief about the true value of the state x can be described by a Gaussian prior:

$$p(x) \sim \mathcal{N}(\mu_x, \pi_x^{-1})$$

The likelihood of an observation u is also Gaussian, with precision π_ε :

$$p(u|x) \sim \mathcal{N}(x, \pi_\varepsilon^{-1})$$

Bayes' rule now tells us that the posterior is Gaussian again:

$$p(x|u) = \frac{p(u|x)p(x)}{\int p(u|x')p(x')\mathrm{d}x'} \sim \mathcal{N}\left(\mu_{x|u}, \pi_{x|u}^{-1}\right)$$

Updates in a simple Gaussian model

Here’s how the updates to the sufficent statistics μ and π describing our belief look like:

$$\pi_{x|u} = \pi_x + \pi_\varepsilon$$

prediction error

$$\mu_{x|u} = \mu_x + \frac{\pi_\varepsilon}{\pi_{x|u}}(u - \mu_x)$$

prediction

$$\text{weight (learning rate)} = \frac{\text{how much we're learning here}}{\text{how much we already know}}$$

The mean is updated by an uncertainty-weighted (more specifically: precision-weighted) prediction error.

The size of the update is proportional to the likelihood precision and inversely proportional to the posterior precision.

This pattern is not specific to the univariate Gaussian case, but generalizes to Bayesian updates for all exponential families of likelihood distributions with conjugate priors (i.e., to all formal descriptions of inference you are ever likely to need).

Reduction to mean updating

Reminder (Gaussian update):

$$\mu_{x|u} = \mu_x + \frac{\pi_\varepsilon}{\pi_{x|u}}(u - \mu_x) = \mu_x + \frac{\pi_\varepsilon}{\pi_x + \pi_\varepsilon}(u - \mu_x)$$

Reducing by $\pi_\varepsilon$the fraction of precisions that make the learning rate, we get

$$\mu_{x|u} = \mu_x + \frac{1}{\frac{\pi_x}{\pi_\varepsilon} + 1}(u - \mu_x)$$

As we shall see, this is the equation for updating an arithmetic mean, but with the number of observations n replaced by $\frac{\pi_x}{\pi_\varepsilon}$.

This shows that Bayesian inference on the mean of a Gaussian distribution entails nothing more than updating the arithmetic mean of observations with $\frac{\pi_x}{\pi_\varepsilon} =: \nu$ as a proxy for the number of prior observations, i.e. for the **weight of the prior relative to the observation**.

Generalization to all exponential families of distributions

Many of the most widely used probability distributions are families of exponential distributions.

For example, the Gaussian distribution is an exponential family of distributions (and so are the beta, gamma, binomial, Bernoulli, multinomial, categorical, Dirichlet, Wishart, Gaussian-gamma, log-Gaussian, multivariate Gaussian, Poisson, and exponential distributions, among others). This means it can be written the following way:

$$p(\boldsymbol{x}|\boldsymbol{\vartheta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta})) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma}\right)$$

with

$$\boldsymbol{x} = x, \qquad \boldsymbol{\vartheta} = (\mu, \sigma)^{\mathrm{T}}, \qquad h(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}}, \qquad \boldsymbol{\eta}(\boldsymbol{\vartheta}) = \left(\frac{\mu}{\sigma}, -\frac{1}{2\sigma}\right)^{\mathrm{T}}, \qquad \boldsymbol{T}(\boldsymbol{x}) = (x, x^2)^{\mathrm{T}}, \qquad A(\boldsymbol{\vartheta}) = \frac{\mu^2}{\sigma} + \frac{\ln\sigma}{2}$$

This allows us to look at Bayesian belief updating in a very general way for all exponential families of distributions.

Generalization to all exponential families of distributions

Our likelihood is an exponential family in its general form:

$$p(\boldsymbol{x}|\boldsymbol{\vartheta}) = h(\boldsymbol{x}) \exp(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta}))$$

The vector $\boldsymbol{T}(\boldsymbol{x})$ (a function of the observation $\boldsymbol{x}$) is called the sufficient statistic.

For the prior, we may assume that we have made ν observations with sufficient statistic $\boldsymbol{\xi}$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{\xi},\nu) = z(\boldsymbol{\xi},\nu) \exp\big(\nu(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi} - A(\boldsymbol{\vartheta}))\big) \quad \text{(where } z(\boldsymbol{\xi},\nu) \text{ is a normlization constant}\big)$$

It then turns out that the posterior has the same form, but with an updated $\boldsymbol{\xi}$ and ν replaced with $\nu + 1$:

$$p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu) = z(\boldsymbol{\xi}',\nu + 1) \exp\big((\nu + 1)(\boldsymbol{\eta}(\boldsymbol{\vartheta}) \cdot \boldsymbol{\xi}' - A(\boldsymbol{\vartheta}))\big)$$

$$\boldsymbol{\xi}' = \boldsymbol{\xi} + \frac{1}{\nu + 1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})$$

Proof of the update equation

$$\overbrace{p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu)}^{\text{posterior}} \propto \overbrace{p(\boldsymbol{x}|\boldsymbol{\vartheta})}^{\text{likelihood}}\ \overbrace{p(\boldsymbol{\vartheta}|\boldsymbol{\xi},\nu)}^{\text{prior}}$$

$$= h(\boldsymbol{x})\exp(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot\boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\vartheta}))z(\boldsymbol{\xi},\nu)\exp\big(\nu(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot\boldsymbol{\xi} - A(\boldsymbol{\vartheta}))\big)$$

$$\propto \exp\big(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi}) - (\nu+1)A(\boldsymbol{\vartheta})\big)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot\frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi}) - A(\boldsymbol{\vartheta})\right)\right)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot\left(\boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) + \nu\boldsymbol{\xi} - (\nu+1)\boldsymbol{\xi})\right) - A(\boldsymbol{\vartheta})\right)\right)$$

$$= \exp\left((\nu+1)\left(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot\underbrace{\left(\boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})\right)}_{=:\boldsymbol{\xi}'} - A(\boldsymbol{\vartheta})\right)\right)$$

$$\Longrightarrow \quad p(\boldsymbol{\vartheta}|\boldsymbol{x},\boldsymbol{\xi},\nu) = z(\boldsymbol{\xi}',\nu')\exp\Big(\nu'\big(\boldsymbol{\eta}(\boldsymbol{\vartheta})\cdot\boldsymbol{\xi}' - A(\boldsymbol{\vartheta})\big)\Big)$$

$$\text{with } \nu' := \nu + 1, \qquad \boldsymbol{\xi}' := \boldsymbol{\xi} + \frac{1}{\nu+1}(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\xi})$$

q.e.d.