Refining a Prior Probability Distribution with Testable Constraints

Marshall Bradley

Two examples of the refinement of a prior probability distribution based upon the requirement of minimum information and testable constraints are presented. Information is the probability distance between the updated distribution and the prior. Constraints are additional background information not incorporated into the initial choice of the prior. They are imposed by the technique of Lagrange multipliers. An example of a testable constraint would be that the updated prior have a prescribed mean or mean and variance. Information as defined here is just the negative of entropy so the approaches of minimum information and maximum entropy are equivalent. If the testable constraint takes the form of imposing moments on the prior then a fat tailed prior is turned into a thin tailed update. This may or may not be desirable. Minimizing information is a way to refine a prior. It is not a replacement for Bayes theorem.

Introduction

The following discussion is presented within the context of Bayesian probability theory. In a typical problem of Bayesian inference we seek to determine the posterior probability distribution of some parameter θ in light of some data D that is at our disposal. That is we ask a question about the parameter θ. In order to answer the question we must supply a likelihood function $L(D \mid \theta)$ and a prior probability distribution $\pi(\theta)$ that describes our knowledge of θ prior to the availability of data. The likelihood function is in essence our model of the data given a knowledge of the parameter. Bayes theorem tells us that the posterior probability distribution $P(\theta \mid D)$ is

$$P(\theta \mid D) = \frac{1}{E} \int \pi(\theta) L(D \mid \theta) d\theta$$

where the evidence E is

$$E = \int \pi(\theta) L(D \mid \theta) d\theta.$$

No matter how we choose the prior $\pi(\theta)$ it must be the case that

$$\int \pi(\theta d\theta = 1$$

since $\pi(\theta)$ is a probability density function. The foregoing three equations are the complete calculus of inference in the presence of uncertainty (Skilling, 2010). The difference between the prior and the posterior is the information

$$H(P \mid \pi) = \int P(\theta \mid D) \log\left[\frac{P(\theta \mid D)}{\pi(\theta)}\right] d\theta .$$

Sometimes we begin with a vague idea as to the nature of the prior and would like to constrain it with some testable background information. This can be accomplished by minimizing the information subject to the constraints. As an example we might like to impose first and second moments on the prior. In this case the quantity that we must minimize is

$$\int p(\theta) \log\left[\frac{p(\theta)}{\pi(\theta)}\right] d\theta + \lambda_0 \int p(\theta)\, d\theta + \lambda_1 \int \theta p(\theta)\, d\theta + \lambda_2 \int \theta^2\, p(\theta)\, d\theta$$

where $p(\theta)$ is our updated prior and $(\lambda_0, \lambda_1, \lambda_2)$ are Lagrange multipliers that are determined by minimization with respect to θ subject to the constraints

$$\int p(\theta)\, d\theta = 1, \quad \int \theta p(\theta)\, d\theta = \mu, \quad \int \theta^2\, p(\theta)\, d\theta = \mu^2 + \sigma^2.$$

The discrete version of the minimization problem is

$$\frac{\partial}{\partial p_j}\left(\sum_{i=1}^{n} p_i \, \log(p_i / \pi_i) + \lambda_0 \sum_{i=1}^{n} p_i + \lambda_1 \sum_{i=1}^{n} \theta_i p_i + \lambda_2 \sum_{i=1}^{n} \theta_i^2 \, p_i\right) = 0.$$

This has solution

$$p_j = \pi_j \exp\left[-\left(1 + \lambda_0 + \lambda_1 \theta_j + \lambda_2 \theta_j^2\right)\right].$$

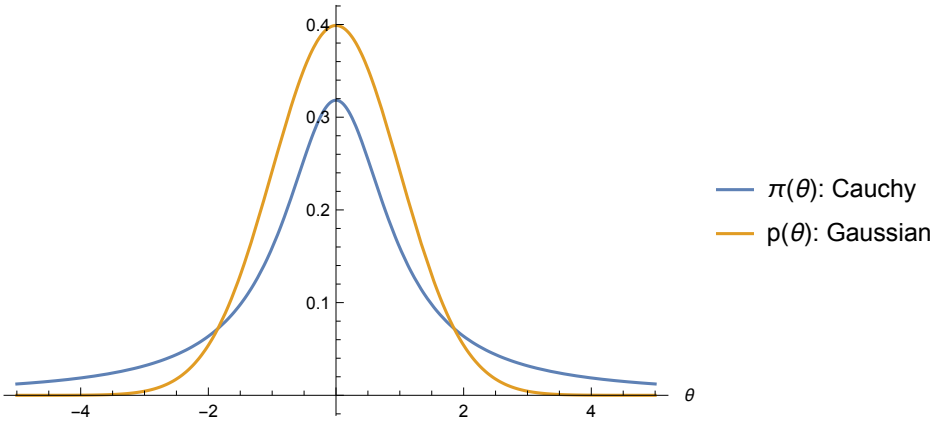
This is a Gaussian in θ. Details will be presented later, but we can already anticipate the final result. If $\pi(\theta)$ is not sharply peak then the Gaussian will dominate and we will find the the updated prior is

$$p(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \mu)^2\right\}, \quad -\infty < \theta < \infty.$$

If our initial choice for $\pi(\theta)$ was a Cauchy probability density function centered on μ with width σ and we impose the constraint of minimum information and the additional requirements that the first and second moments are respectively μ and $\mu^2 + \sigma^2$ then the new prior relative to the old prior looks like this:

In[ ]:=
```
Plot[{PDF[CauchyDistribution[0, 1], θ],
  PDF[NormalDistribution[0, 1], θ]}, {θ, -5, 5}, ··· ]
```

Out[ ]=



In essence the minimum information update has turned a Cauchy distribution into a Gaussian.

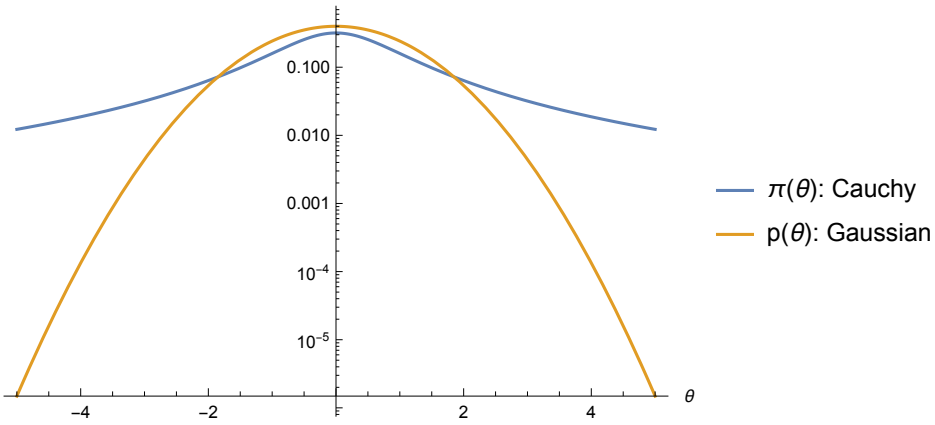
Our definition of information is just the negative of what is commonly called the entropy of a probability distribution. That is minimum information and maximum entropy are the same thing. So imposing

the requirement of maximum entropy takes the fat tailed Cauchy distribution and turns it into a thin tailed Gaussian. This precludes our prior from providing support for very large or small values of θ. This follows from the fact that it is very unlikely to obtain values outside the range $(\mu - 3\,\sigma,\ \mu + 3\,\sigma)$ for a Gaussian.

The difference in the tails between is even more visually dramatic when viewed on a log scale:

In[ ]:=
```
LogPlot[{PDF[CauchyDistribution[0, 1], θ],
  PDF[NormalDistribution[0, 1], θ]}, {θ, -5, 5}, ···]
```

Out[ ]=



An analytic example of the evaluation of the information integral

In the following section we define a prior probability distribution and construct a likelihood function. From these we use Bayes theorem to find the posterior. Just to be clear let's review our notation. The unknown parameter in question is denoted by θ and the data is D. The prior is $\pi(\theta)$. The difference between this function and 3.14159 … will be clear from context. The posterior is $P(\theta \mid D)$ and the information is

$$H(P \mid \pi) = \int P(\theta \mid D) \log\left[\frac{P(\theta \mid D)}{\pi(\theta)}\right] d\theta\,.$$

The information is a measure of the distance between the prior and posterior. As we shall see in our example, if the posterior is sharply peaked in comparison to the prior, then the information will be large.

Suppose we draw data from a Gaussian distribution with unknown mean θ and known deviation σ. If x is a data value then the sampling distribution is

$$h(x \mid \theta) = \frac{1}{(2\,\pi)^{1/2}\,\sigma} \exp\left\{-\frac{(x-\theta)^2}{2\,\sigma^2}\right\}, \quad -\infty < x < \infty.$$

On initial hypothesis we assume that the unknown parameter θ is uniformly distributed on the interval $(\theta_a,\ \theta_b)$. The explicit form of our prior probability distribution is

$$\pi(\theta) = \frac{1}{\theta_b - \theta_a}\ ,\ \theta_a < \theta < \theta_b$$

and zero otherwise.

The likelihood of obtaining the independent data sample D consisting of the n values $x_1\, x_2 \ldots x_n$ is

$$L(D \mid \theta) = \prod_{i=1}^{n} h(x_i \mid \theta) = \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}.$$

If we define the statistics $\bar{x}$ and s of the data $x_1\, x_2 \ldots x_n$ via

$$n\,\bar{x} = \sum_{i=1}^{n} x_i \ \text{ and } \ (n-1)\,s^2 = \sum_{i=1}^{n}(x - \bar{x})^2,$$

then the likelihood of the data can be written

$$L(\bar{x}, s, n \mid \theta) = \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n(\theta - \bar{x})^2\right]\right\}.$$

Bayes theorem tells us that the posterior distribution is

$$P(\theta \mid D) = \frac{1}{E}\ \pi(\theta)\ L(\bar{x}, s, n \mid \theta) = \frac{1}{\mu_b - \mu_a}\ \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\,(n-1)\,s^2\right\} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\}$$

where the evidence E is

$$E = \int_{\mu_a}^{\mu_b} f(\theta)\ L(\bar{x}, s, n \mid \theta)\, d\theta = \frac{1}{\mu_b - \mu_a}\ \frac{1}{(2\pi)^{n/2}\,\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\,(n-1)\,s^2\right\} \int_{\theta_a}^{\theta_b} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\} d\theta.$$

Dividing by the evidence yields an expression for the posterior in which all factors not dependent upon the unknown mean θ cancel out. Namely

$$P(\theta \mid D) = \frac{1}{E}\ f(\theta)\ L(\bar{x}, s, n \mid \theta) = \frac{1}{\int_{\theta_a}^{\theta_b} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\} d\theta} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\},\ \theta_a < \theta < \theta_b.$$

Now if our prior distribution on θ is such that $\theta_a < \bar{x} - 2\,\sigma \big/ \sqrt{n} < \theta < \bar{x} - 2\,\sigma \big/ \sqrt{n} < \theta_b$ then the limits in the normalization integral can be replaced by $-\infty$ and ∞. In this case the posterior is Gaussian distributed with mean $\bar{x}$ and standard deviation $\sigma \big/ \sqrt{n}$. That is

$$P(\theta \mid D) = \frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\},\ -\infty < \theta < \infty.$$

$$H[P(\theta \mid D) \mid \pi(\theta)] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\} \log\left[\frac{\frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\}}{\frac{1}{\theta_b - \theta_a}}\right] d\theta$$

This can be expanded to yield

$$H = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\}\left[\log\left(\frac{\theta_b - \theta_a}{\sqrt{2\pi}\,\sigma/\sqrt{n}}\right) - \frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right] d\theta.$$

This is equal to

$$H = \log\left(\frac{\theta_b - \theta_a}{\sqrt{2\pi}\,\sigma/\sqrt{n}}\right) - \frac{1}{\sqrt{2\pi}\,\sigma/\sqrt{n}}\ \frac{n}{2\sigma^2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\,n(\theta - \bar{x})^2\right\} (\theta - \bar{x})^2\, d\theta$$

The last term in the information integral is tedious so we evaluate it using Mathematica. Surprisingly it simplifies to 1/2:

In[•]:=
```
Clear[n, σ, xbar];
n/(2 σ^2) 1/(Sqrt[2 π] σ/Sqrt[n])
Integrate[Exp[-n/(2 σ^2) (θ - xbar)^2] (θ - xbar)^2, {θ, -∞, ∞}, Assumptions → {n > 0, σ > 0}]
```

Out[•]=
$$\frac{1}{2}$$

So the information is

$$H = \log\left(\frac{\theta_b - \theta_a}{\sqrt{2\pi}\,\sigma/\sqrt{n}}\right) - \frac{1}{2}.$$

To begin with let us define our prior and posterior:

In[•]:=
```
Clear[θ, θa, θb, xbar, n, σ];
prior[θ_, θa_, θb_] := 1/(θb - θa)
posterior[θ_, xbar_, σ_, n_] := Sqrt[n]/Sqrt[2 π] Exp[-n/(2 σ^2) (θ - xbar)^2]
```

A numerical computation of the information *H* is:

In[•]:=
```
With[{n = 200, xbar = 0.0, σ = 1.0, θa = -5.0, θb = 5.0},
 NIntegrate[posterior[θ, xbar, σ, n] Log[posterior[θ, xbar, σ, n]/prior[θ, θa, θb]], {θ, θa, θb}]]
```

Out[•]=
```
3.53281
```

This agrees with the theoretical computation:

In[•]:=
```
Log[(θb - θa)/(Sqrt[2 π] σ/Sqrt[n])] - 1/2 /. {θa → -5.0, θb → 5.0, σ → 1, n → 200}
```

Out[•]=
```
3.53281
```

Constrained mean and standard deviation

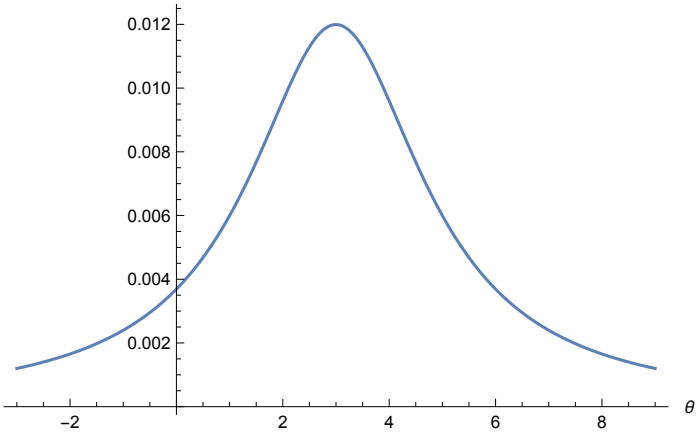
The prior probability distribution $\pi(\theta)$ for the unknown parameter θ is assumed to be Cauchy with center c and width w. Specifically $\pi(\theta)$ is

$$\pi(\theta) = \frac{w/3.14159\ldots}{(\theta - c)^2 + w^2}$$

Our prior looks like this:

In[ ]:=
```
center = 3.0; width = 2.0;
θmin = center - 3 width; θmax = center + 3 width; Nθ = 201;
θ = Table[θ, {θ, θmin, θmax, (θmax - θmin)/(Nθ - 1)}];
prior = Map[1/(π width (1 + (-center+#)^2/width^2)) &, θ];
prior = prior / Total[prior];
ListPlot[prior, Joined → True, ...]
```

Out[ ]=



We minimize the information $H(p \mid \pi) = \int p(\theta) \log(p(\theta)/\pi(\theta))\, d\theta$ subject to normalization and constraints.

The variational equation is

$$\delta\left(\int p(\theta) \log[p(\theta)/\pi(\theta)]\, d\theta + \lambda_0 \int p(\theta)\, d\theta + \lambda_1 \int \theta p(\theta)\, d\theta + \lambda_2 \int \theta^2\, p(\theta)\, d\theta\right).$$

The discrete version of this equation is

$$\frac{\partial}{\partial p_j}\left(\sum_{i=1}^n p_i \log(p_i/\pi_i) + \lambda_0 \sum_{i=1}^n p_i + \lambda_1 \sum_{i=1}^n \theta_i\, p_i + \lambda_2 \sum_{i=1}^n \theta_i^2\, p_i\right) = 0.$$

This has solution

$$p_j = \pi_j \exp\left[-\left(1 + \lambda_0 + \lambda_1\, \theta_j + \lambda_2\, \theta_j^2\right)\right].$$

If we constrain the first and second moments of the updated prior distribution $p(\theta)$ then we have the normalization and constraints

$$1 = \exp[-(1 + \lambda_0)] \sum_{j=1}^n \pi_j \exp\left[-\left(\lambda_1\, \theta_j + \lambda_2\, \theta_j^2\right)\right]$$

$$\mu = \exp[-(1 + \lambda_0)] \sum_{j=1}^n \pi_j\, \theta_j \exp\left[-\left(\lambda_1\, \theta_j + \lambda_2\, \theta_j^2\right)\right]$$

$$\mu^2 + \sigma^2 = \exp[-(1 + \lambda_0)] \sum_{j=1}^n \pi_j\, \theta_j^2 \exp\left[-\left(\lambda_1\, \theta_j + \lambda_2\, \theta_j^2\right)\right]$$

The unknown Lagrange multipliers λ_1 and λ_2 can be found by solving the pair of nonlinear equations

$$\mu = \frac{\sum_{j=1}^n \pi_j\, \theta_j \exp\left[-\left(\lambda_1\, \theta_j + \lambda_2\, \theta_j^2\right)\right]}{\sum_{j=1}^n \pi_j \exp\left[-\left(\lambda_1\, \theta_j + \lambda_2\, \theta_j^2\right)\right]}$$

$$\mu^2 + \sigma^2 = \frac{\sum_{j=1}^{n} \pi_j \, \theta_j^2 \exp\left[-\left(\lambda_1 \, \theta_j + \lambda_2 \, \theta_j^2\right)\right]}{\sum_{j=1}^{n} \pi_j \exp\left[-\left(\lambda_1 \, \theta_j + \lambda_2 \, \theta_j^2\right)\right]}$$

In order to compute a numerical solution we define:

```
In[•]:= norm[λ1_, λ2_] := Sum[prior[[j]] Exp[-(λ1 * θ[[j]] + λ2 * θ[[j]]^2)], {j, 1, Length[prior]}]
moment1[λ1_, λ2_] :=
 Sum[prior[[j]] × θ[[j]] Exp[-(λ1 * θ[[j]] + λ2 * θ[[j]]^2)], {j, 1, Length[prior]}]
moment2[λ1_, λ2_] :=
 Sum[prior[[j]] θ[[j]]^2 Exp[-(λ1 * θ[[j]] + λ2 * θ[[j]]^2)], {j, 1, Length[prior]}]
```

Now we can compare the prior to the updated prior. We begin by finding the Lagrange multipliers:

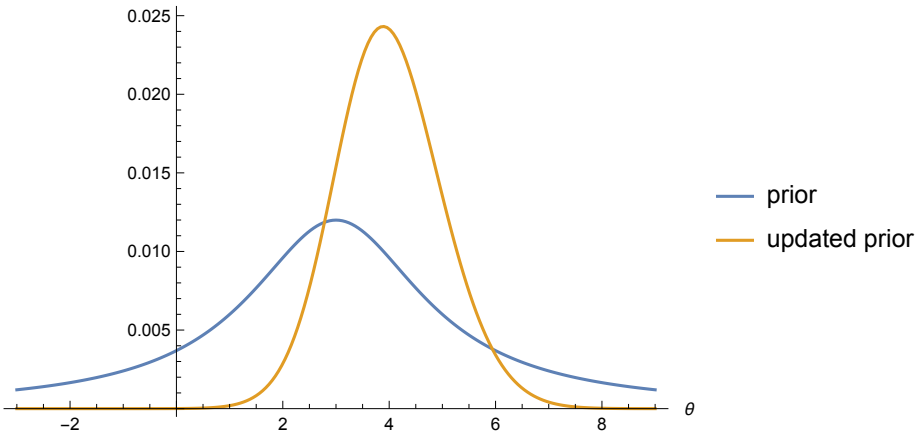
```
In[•]:= μ = 4; σ = 1;
Clear[λ1, λ2];
sol = FindRoot[{μ == moment1[λ1, λ2] / norm[λ1, λ2], μ^2 + σ^2 == moment2[λ1, λ2] / norm[λ1, λ2]}, {{λ1, 0}, {λ2, 0}}];
{λ1, λ2} = {λ1, λ2} /. sol
```

```
Out[•]=
{-3.41415, 0.391454}
```

Once they are found the comparison between prior and updated prior can be made :

```
In[•]:= priorUpdate =
  Table[prior[[j]] Exp[-(λ1 * θ[[j]] + λ2 * θ[[j]]^2)], {j, 1, Length[prior]}] /. sol;
priorUpdate = priorUpdate / Total[priorUpdate];
ListPlot[{prior, priorUpdate}, ...]
```

```
Out[•]=
```



The updated prior $p(\theta)$ looks to be very Gaussian .

Check that the mean is 4:

```
In[•]:= priorUpdate.θ
```

```
Out[•]=
4.
```

And that the standard deviation is 1:

In[ ]:=
```
(priorUpdate.(θ - priorUpdate.θ)^2)^1/2
```

Out[ ]=
```
1.
```

Constrained mean for positive valued parameter

Minimum information (same as max entropy) refines prior knowledge and provides for an updated prior. The information is

$$H(p \mid \pi) = \sum_{i=1}^{M} p_i \log(p_i / \pi_i)$$

where p is the update and π is prior. Imposing a constraint on the mean beyond prior knowledge leads to minimizing

$$S = \sum_{i=1}^{M} p_i \log(p_i / \pi_i) + \lambda_0\left(\sum_{i=1}^{M} p_i - 1\right) + \lambda_1\left(\sum_{i=1}^{M} x_i p_i - \mu\right)$$

The unknowns are the p_j. The solution is

$$p_j = \pi_j \exp(-\lambda_0) \exp\left(-\lambda_1 \theta_j\right)$$

This is clearly an exponential. To find the Lagrange multiplier λ_1 we must find the root to

$$\mu = \frac{\sum_{i=1}^{M} \pi_j \theta_i \exp(-\lambda_1 \theta_i)}{\sum_{i=1}^{M} \pi_j \exp(-\lambda_1 \theta_i)}$$

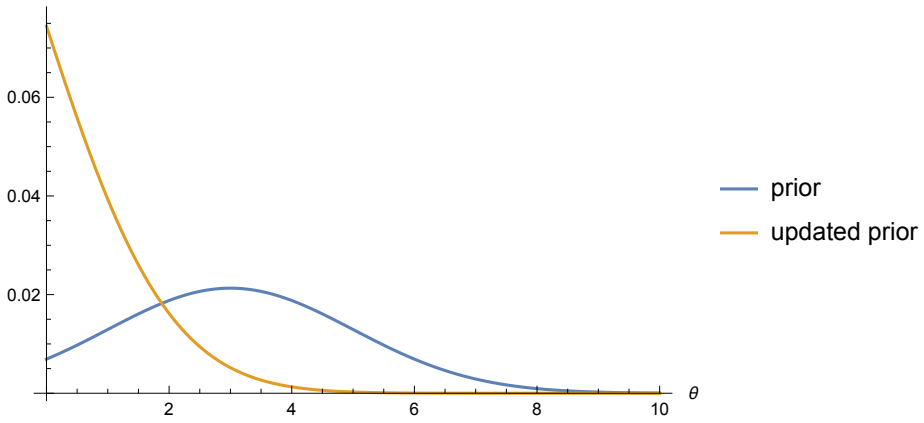
where μ is the imposed mean.

The Lagrange multiplier λ_0 is found by requiring the posterior to sum to zero.

We begin with a Gaussian prior but an exponential update emerges :

In[ ]:=
```
m0 = 3.0; s0 = 2; θmin = 0.0; θmax = 10.0; dθ = 0.1;
θ = Table[θ, {θ, θmin, θmax, dθ}]; M = Length[θ];
prior = Map[Exp[-(# - m0)^2 / (2 s0^2)] &, θ];
prior = prior / Total[prior];
mtrue = 1.0; Clear[λ1];
sol = FindRoot[mtrue == Sum[prior[[j]] θ[[j]] Exp[-λ1 * θ[[j]]], {j, 1, M}] / Sum[prior[[j]] Exp[-λ1 * θ[[j]]], {j, 1, M}], {λ1, 1}];
λ1 = λ1 /. sol;
λ0 = Log[Sum[prior[[j]] Exp[-λ1 * θ[[j]]], {j, 1, M}] ];
posterior = Table[prior[[j]] Exp[-λ0] Exp[-λ1 * θ[[j]]], {j, 1, M}];
ListPlot[{prior, posterior}, ...]
```

Out[ ]=



The updated prior is exponential like.

References

Skilling, John, "Foundations and algorithms", published in *Bayesian Methods in Cosmology*, Cambridge University Press, 2010.