Priors with Randomly-distributed Hyperparameters
Bill Menke, October 29, 2020

Suppose that one introduces prior information about the value of a model parameter m via a prior p.d.f. $p(m)$. The p.d.f. might have a variance parameter, say x, that controls its properties. For instance, x might be a variance that expresses how confident we are of the prior information that m is close to some prescribed value.

However, suppose that we are uncertain of x, too. We can treat x as a random variable with its own p.d.f. $p(x)$, in which case it is called a "hyperparameter". Then, the overall uncertainty in m is a combination of the uncertainty expressed in the "nominal" prior, which is now understood to be a conditional p.d.f., say $p(m|x)$, and the p.d.f. $p(x)$ of the hyperparameter. According to the usual rules of probability, the joint probability of m and x is:

$$p(m,x) = p(m|x)\, p(x)$$

and the "real" prior is:

$$p(m) = \int p(m,x)\, dx$$

Typically, the "real" prior $p(m)$ will be wider than the "nominal" prior $p(m|x)$, because it combines uncertainty from two sources.

An important question is whether this two-step process has advantages over the alternative of initially specifying a different, wider prior.

Suppose that the "nominal" prior is $p(m;\mu,\sigma^2,\kappa)$ where μ, σ^2 and κ are respectively the mean, variance and kurtosis of the p.d.f. Here, parameters to the right of the semi-colon denote known parameters, not hyperparameters. If the variance were very poorly known, then one might consider it to be a hyperparameter, in which case the "nominal" prior is $p(m|\sigma^2;\mu,\kappa)$ and the p.d.f. for variance $p(\sigma^2)$. The "real" prior is then:

$$p(m;\mu,k) = \int p(m|\sigma^2;\mu,\kappa)\, p(\sigma^2)\, d\sigma^2$$

Presumably, one proceeded in this fashion because one believes that the introduction of the variance hyperparameter has not altered the mean and kurtosis; that is, the mean and kurtosis of $p(m;\mu,\sigma^2,k)$ are the same as the mean and kurtosis of is $p(m;\mu,\kappa)$. This reasoning is correct – although arguably for deceptive reasons – and is the major advantage of using hyperparameters.

A general result can be constructed for the expectations x_n of a set of functions $F^{(n)}(m)$, that is:

$$x_n \equiv E(F^{(n)}) = \int F^{(n)}(m)\, p(m)\, dm$$

Suppose that the p.d.f. $p(m;\mathbf{x})$ is parameterized in terms of $x_i, i = 0 \cdots N$. Let the symbol $\mathbf{x}^{(\sim n)}$ mean all the x_is except x_n. Then, for a single hyperparameter x_n and the kth function:

$$\int F^{(k)}(m)\, p\left(m;\mathbf{x}^{(\sim n)}\right) dm = \int F^{(k)}(m) \left\{ \int p\left(m|x_n;\mathbf{x}^{(\sim n)}\right) p(x_n)\, dx_n \right\} dm =$$

$$\int \left\{ \int F^{(k)}(m)\, p\left(m|x_n; \mathbf{x}^{(\sim n)}\right) dm \right\} p(x_n)\, dx_n = \int x_k\, p(x_n)\, dx_n = \begin{cases} x_k & \text{if } k \neq n \\ E(x_n) & \text{if } k = n \end{cases}$$

Here, we use the fact that $p\left(m|x_n; \mathbf{x}^{(\sim n)}\right)$ is the same function as $p\left(m; x_n, \mathbf{x}^{(\sim n)}\right)$, and we presume that all the integrals exist (which may not always be the case). When $k \neq n$, the "real" prior has the same $E(F^{(k)})$ as the "nominal" prior, but when $k = n$ the "real" prior's is $E\left(F^{(n)}\right) \equiv \int F^{(n)}\, p(x_n)\, dx_n$.

The general result applies to mean, kurtosis and variance because the they are so closely related to moments; that is, the functions $F^{(k)}(m) = m^k$. However, were the "nominal" prior parameterized by its median, which unlike the mean, cannot be written as the expected value of any function, the medians of the "nominal" and "real" priors would, in general, be different from one another. Another way in which the result is deceptive is that, although the mean and kurtosis of the nominal and real priors are equal, the underlying p.d.f.s are *not* the same. For instance, if the "nominal" prior is Normal, the "real" prior might not be.

Case 1. A Uniformly-distributed prior with a Uniformly-distributed hyperparameter

The nominal prior for model parameter m is uniform between 0 and hyperparameter L, with L being uniform between 1 and 2. As shown in the derivation, the functional form of the "real" prior $p(m)$ is:

$$p(m) = \begin{cases} \ln 2 & \text{if } 0 \leq m \leq 1 \\ \ln 2 - \ln m & \text{if } 1 \leq m \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

It's not uniform, but it's well-defined (Figure 1).

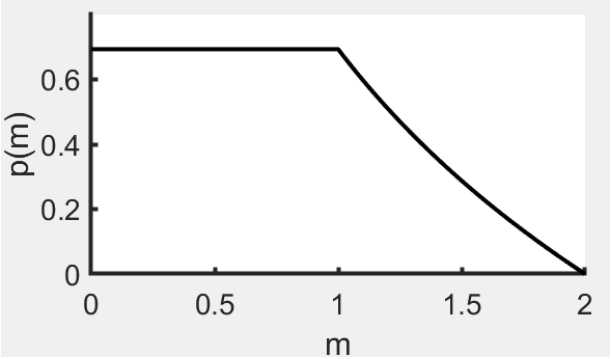

Fig.2. The prior $p(m)$.

The "real" prior $p(m)$ is computed from the "nominal" prior $p(m|L)$ and the p.d.f. $p(L)$ of the hyperparameter:

$$p(m|L) = \begin{cases} L^{-1} \; if \; 0 \leq m \leq L \\ 0 \quad \text{otherwise} \end{cases} \quad \text{and} \quad p(L) = \begin{cases} 1 \quad if \; 1 \leq L \leq 2 \\ 0 \quad \text{otherwise} \end{cases}$$

The joint probability of m and L is computed using the rule $p(m, L) = p(m|L)p(L)$. It has the value of L^{-1} in a trapezoidal area of the (m, L) plane (Figure 2).

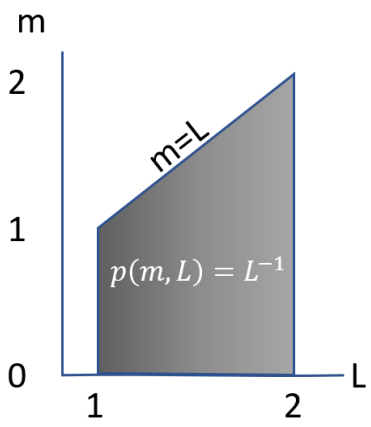

Fig. 1. The joint p.d.f. $p(m, L)$ is non-zero in a trapezoidal region of the (m, L) plane. Inside the region it has the value L^{-1}.

That the total probability P of $p(m, L)$ is unity can be verified by integration over the trapezoid:

$$P = \int_1^2 \left\{ L^{-1} \int_0^L p(m,L)\, dm \right\} dL = \int_1^2 \{L^{-1} L\} dL = 2 - 1 = 1$$

The real prior can be formed by integrating $p(m, L)$ over L to form a univariate p.d.f. $p(m)$:

$$p(m) = \int_1^2 p(m,L)\, dL = \begin{cases} \int_1^2 L^{-1} dL & \text{if } 0 \le m \le 1 \\ \int_m^2 L^{-1} dL & \text{if } 1 \le m \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Using $\int L^{-1} dL = \ln L$ yields:

$$p(m) = \begin{cases} \ln 2 & \text{if } 0 \le m \le 1 \\ \ln 2 - \ln m & \text{if } 1 \le m \le 2 \\ 0 & \text{otherwise} \end{cases}$$

The total probability P is unity:

$$P = \int_0^2 p(m)\, dm = \int_0^1 p(m)\, dm + \int_1^2 p(m)\, dm = \ln 2 + \ln 2 - [m \ln m - m]_1^2$$

$$= 2\ln 2 - 2 \ln 2 + 2 - 0 - 1 = 1$$

Here we have used the rule $\int \ln x \; dx = x \ln x - x + C$.

Case 2. A Normally-distributed prior with a Normally-distributed hyperparameter

The "nominal" prior for a model parameter m is a Normally-distribution with known mean $\bar{m}$ and hyperparameter variance σ^2. The "certainty" $s = \sigma^{-1}$ is Normally-distributed with zero mean and known reciprocal variance k^2. Thus, s^2 has a high probability of being less than k, while the variance σ^2 has a high probability of being greater than k (but probability falls off to zero as $\sigma \to \infty$). As shown below, the true prior $p(m)$ is Cauchy-distributed with mode $\bar{m}$ and scale parameter k.

$$p(m|s) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-½\sigma^{-2}(m-\bar{m})^2\} = \frac{s}{\sqrt{2\pi}}\exp\{-½s^2(m-\bar{m})^2\} \quad \text{with } -\infty \le m \le \infty$$

$$p(s) = \frac{2k}{\sqrt{2\pi}}\exp\{-½k^2s^2\} \quad \text{with} \quad 0 \le s \le \infty$$

$$p(m,s) = p(m|s)p(s) = \frac{s}{\sqrt{2\pi}}\exp\{-½s^2(m-\bar{m})^2\}\frac{2k}{\sqrt{2\pi}}\exp\{-½k^2s^2\}$$

$$p(m,s) = \frac{k}{\pi}s\exp\{-½s^2[(m-\bar{m})^2 + k^2]\}$$

$$p(m,s) = \frac{k}{\pi}s\exp\{-½Z^2s^2\} \quad \text{with} \quad Z^2 = [(m-\bar{m})^2 + k^2]$$

$$p(m) = \frac{k}{\pi}\int_0^\infty s\exp\{-½Z^2s^2\}\ ds$$

$$x = Zs \quad \text{and} \quad s = Z^{-1}x \quad \text{and} \quad ds = Z^{-1}dx \quad \text{and} \quad x \to 0 \text{ as } s \to 0 \quad \text{and} \quad x \to \infty \text{ as } s \to \infty$$

$$p(m) = \frac{k}{\pi}Z^{-2}\int_0^\infty x\exp\{-½x^2\}\ dx = \frac{k}{\pi}Z^{-2}\int_0^\infty x\exp\{-½x^2\}\ dx$$

Wikipedia says if $\varphi(x) = \frac{1}{\sqrt{2\pi}}\exp\{-½x^2\}$ then $\int x\varphi(x)dx = -\varphi(x) + C$, so:

$$p(m) = \frac{k}{\pi}Z^{-2} = \left(\frac{1}{\pi}\right)\left(\frac{k}{(m-\bar{m})^2 + k^2}\right) = \left(\frac{1}{\pi k}\right)\left(\frac{1}{\left(\frac{m-\bar{m}}{k}\right)^2 + 1}\right)$$

which is a Cauchy distribution with median $\bar{m}$ and scale factor k. The Cauchy distribution is extremely long-tailed.

The distribution $p(\sigma)$ is:

$$s = \sigma^{-1} \quad \text{and} \quad \frac{ds}{d\sigma} = -\sigma^{-2}$$

$$p(\sigma) = p[s(\sigma)]\left|\frac{ds}{d\sigma}\right| = \frac{2k}{\sqrt{2\pi}}\sigma^{-2}\exp\{-½k^2\sigma^{-2}\}$$

It has extreme values $p(\sigma \to 0) = 0$ and $p(\sigma \to \infty) = 0$ and has a peak at $\sigma = \sqrt{½}k$ (Figure 3):

$$p(\sigma) \propto k^2\sigma^{-2}\exp(-½k^2\sigma^{-2}) = x^{-2}\exp(-½x^{-2}) \quad \text{with} \quad k^2\sigma^{-2} = x^{-2} \quad \text{or} \quad kx = \sigma$$

$$\frac{dp}{dx} = 0 = -2x^{-3}\exp(-½x^{-2}) + (-½)(-2x^{-3})x^{-2}\exp(-½x^{-2})$$

$$0 = -2 + x^{-2} \quad \text{or} \quad x^{-2} = 2 \quad \text{or} \quad x = \sqrt{½} \quad \text{so} \quad \sigma = \sqrt{½}k$$

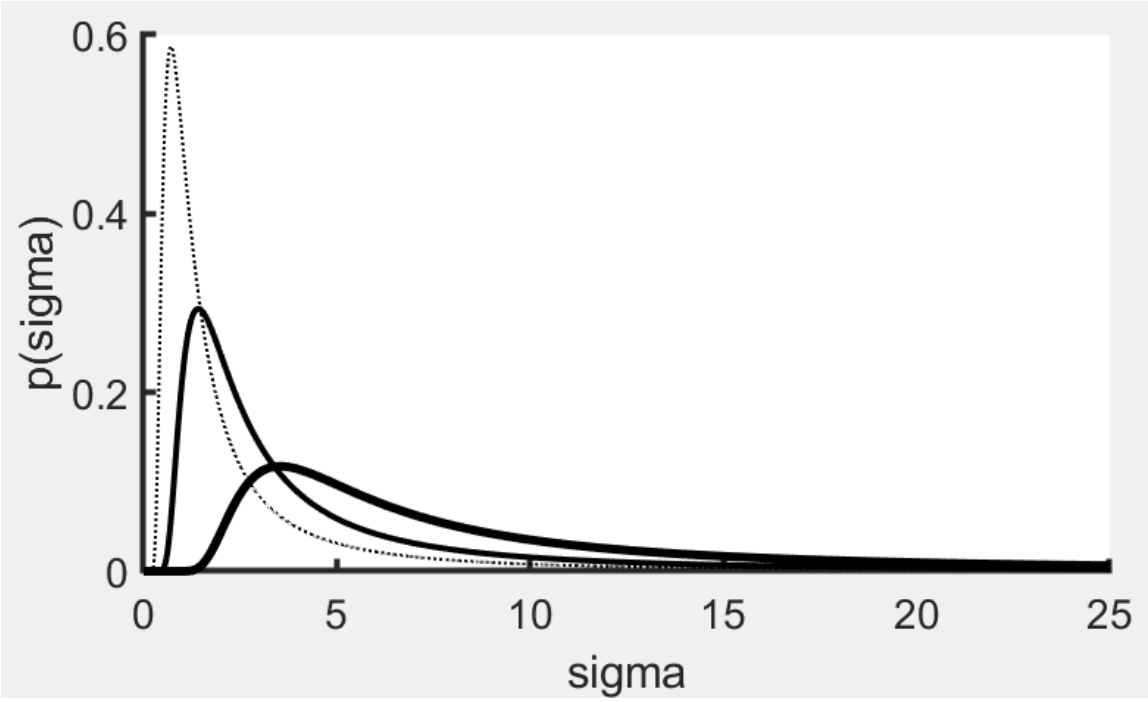

Fig. 3. The probability density function $p(\sigma)$ for $k = 1$ (dotted curve), $k = 2$ (solid curve) and $k = 5$ (bold curve). The peak is at $\sigma = \sqrt{½}k$.

The p.d.f. $p(\sigma)$ is very long-tailed and has no expected value:

$$E(\sigma) = \int_0^\infty \sigma\, p(\sigma)\, d\sigma = \frac{2k}{\sqrt{2\pi}} \int_0^\infty \sigma^{-1} \exp\{-½k^2\sigma^{-2}\}\, d\sigma$$

$$x = k\sigma^{-1} \quad \text{and} \quad \sigma = kx^{-1} \quad \text{and} \quad d\sigma = -kx^{-2}dx \quad \text{and} \quad \sigma \to 0, x \to \infty \quad \text{and} \quad \sigma \to \infty, x \to 0$$

$$\bar{\sigma} = \frac{2k^3}{\sqrt{2\pi}} \int_\infty^0 (k^{-1}x) \exp\{-½k^2x^2\} (-kx^{-2})\, dx = \frac{2k}{\sqrt{2\pi}} \int_0^\infty x^{-1} \exp\{-½k^2x^2\}\, dx$$

For $x \to 0$, the exponential in the integrand $\to 1$, while $x^{-1} \to \infty$, so the integrand is not integrable and the expected value does not exist.

What follows is a rewritten version of the first section

Invariance of Expectations of the Prior During Hyperparameter Transformations

Bill Menke, October 31, 2020

Suppose that one introduces prior information about the value of a model parameter m via a prior p.d.f. $p_P(m; x)$. Here, x is a known parameter (like mean or variance) that controls the properties of the prior.

Suppose that we are uncertain of x, too. We can treat x as a random variable with its own p.d.f. $p_x(x)$, in which case x is a hyperparameter. Then, the overall uncertainty in m is a combination of the uncertainty

expressed in a "nominal" prior, which is now understood to be the conditional p.d.f. $p_N(m|x)$, and the $p_x(x)$ of the hyperparameter. The nominal prior $p_N(m|x)$ is the same function as $p_P(m;x)$, but now we view x as a random variable. According to the usual rules of probability, the joint p.d.f. of m and x is:

$$p_J(m,x) = p_N(m|x)\, p_x(x)$$

and the "final" prior is:

$$p_F(m) = \int p_J(m,x)\, dx$$

Suppose that the prior p_P has several parameters, $x_i, i = 1 \cdots N$, only one, say x_n, of which is transformed into a hyperparameter. We denote this situation as $p_N\left(m|x_n;\mathbf{x}^{(\sim n)}\right)$. Here, the symbol $\mathbf{x}^{(\sim n)}$ mean all the x_is except x_n. Now consider the expected value of a function $f^{(n)}(m)$ with respect to a p.d.f. $p(m)$:

$$E\left[f^{(n)},p\right] = \int f^{(n)}(m)\, p(m)\, dm$$

The following general result holds between $x_k^F \equiv E\left[f^{(k)},p_F\right]$ and $x_k \equiv E\left[f^{(k)},p_P\right]$:

$$x_k^F = E\left[f^{(k)},p_F\right] = \int f^{(k)}(m)\, p_F\left(m;\mathbf{x}^{(\sim n)}\right) dm =$$

$$\int f^{(k)}(m)\left\{\int p_N\left(m|x_n;\mathbf{x}^{(\sim n)}\right) p_{x_n}(x_n)\, dx_n\right\} dm =$$

$$\int \left\{\int f^{(k)}(m)\, p_P\left(m|x_n;\mathbf{x}^{(\sim n)}\right) dm\right\} p_{x_n}(x_n)\, dx_n =$$

$$\int \left\{\int f^{(k)}(m)\, p_P\left(m;x_n,\mathbf{x}^{(\sim n)}\right) dm\right\} p_{x_n}(x_n)\, dx_n =$$

$$\int E\left[f^{(k)},p_P\right] p_{x_n}(x_n)\, dx_n = \int x_k\, p_{x_n}(x_n)\, dx_n = \begin{cases} x_k & \text{if } k \neq n \\ E\left[x_n,p_{x_n}\right] & \text{if } k = n \end{cases}$$

Here, we use the fact that $p_N\left(m|x_n;\mathbf{x}^{(\sim n)}\right)$ is literally the same function as $p_P\left(m;x_n,\mathbf{x}^{(\sim n)}\right)$. Also, we presume that all the integrals exist (which may not always be the case). When $k \neq n$, $x_k^F = x_k$, and when $k = n$, $x_n^F = E\left[x_n,p_{x_n}\right] = \int x_n\, p(x_n)\, dx_n$. Thus, as long as the "nominal" prior is parameterized in terms of its expectations with respect to known functions, the parameters obey a kind of invariance. Making one parameter into a hyperparameter does not change the values of the others in the "final" prior, and it changes the hyperparameter itself in a very simple way. While useful, the result is a bit deceptive, because the "form" of p_F may well be different than p_P. For example, when p_P is Normal, p_F will *not* be Normal in general.

What follows is a rewritten version of the third section

A Normally-distributed Prior with Normally-distributed Certainty is Cauchy-Distributed

Bill Menke, November 1, 2020

The "nominal" prior for a model parameter m is a Normally-distribution with known mean $\overline{m}$ and hyperparameter variance σ^2:

$$p(m|s) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-½\sigma^{-2}(m-\overline{m})^2\} = \frac{s}{\sqrt{2\pi}}\exp\{-½s^2(m-\overline{m})^2\} \quad \text{with } -\infty \le m \le \infty$$

The "certainty" $s = \sigma^{-1}$ is Normally-distributed with zero mean and known variance k^{-2}:

$$p(s) = \frac{2k}{\sqrt{2\pi}}\exp\{-½k^2s^2\} \quad \text{with} \quad 0 \le s \le \infty$$

This p.d.f. says that the certainty is low; that is, $s < k^{-1}$ 68% of the time. When transformed to $p(\sigma)$ it produces the unimodal p.d.f.:

$$p(\sigma) = p[s(\sigma)]\left|\frac{ds}{d\sigma}\right| = \frac{2k}{\sqrt{2\pi}}\sigma^{-2}\exp\{-½k^2\sigma^{-2}\}$$

This p.d.f. has limits $p(\sigma \to 0) = 0$ and $p(\sigma \to \infty) = 0$, has a single peak at $\sigma = \sqrt{½}k$, and is very long-tailed (it has no mean) (Figure 1). Thus, it represents the notion that σ is never zero, that most of te probability is near $\sqrt{½}k$, and that very large values of σ are common. The "final" prior is:

$$p_F(m) = \int_0^\infty p(m|s)\,p(s)\,ds = \left(\frac{1}{\pi k}\right)\left(\frac{1}{\left(\frac{m-\overline{m}}{k}\right)^2 + 1}\right)$$

Thus, $p_F(m)$ is a Cauchy distribution with median $\overline{m}$ and scale factor k. The Cauchy distribution is extremely long-tailed; that is ms far from the median are very common.

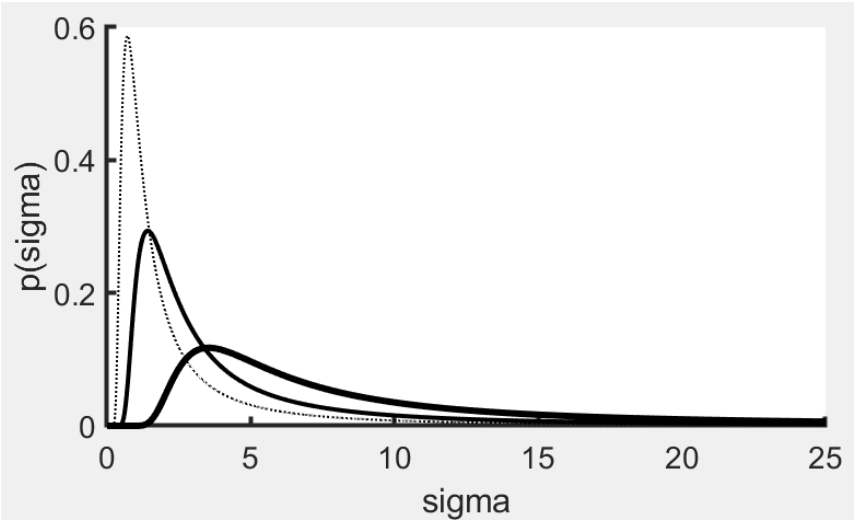

Fig. 1. The probability density function $p(\sigma)$ for $k = 1$ (dotted curve), $k = 2$ (solid curve) and $k = 5$ (bold curve). The peaks are at $\sigma = \sqrt{½}k$.