

Prior Probability

April 5, 2006

We have learned that inverse problems are “fixed” by adding external information. A general framework for handling the external information is to use conjunction of states of information. More traditionally, Bayesian calculations are used. In this sense, inverse problems are just an example of general inference and Bayes rule or conjunction of probability density functions gives us the starting point. If we have data \mathbf{d} and we are trying to infer object \mathbf{m} we know that

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}.$$

Alternatively we could write using conjunction of probability density functions

$$\sigma(\mathbf{d}, \mathbf{m}) = \frac{\rho(\mathbf{d}, \mathbf{m})\Theta(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})}.$$

We make no apology that the inference for \mathbf{m} must depend on the prior knowledge of \mathbf{m} . In fact, we celebrate the simplicity of this fact because it exposes a fundamental reality that in our ivory towers we can often pridefully ignore. This fact is “You cannot make inference without assumptions.” We have actually not fully addressed the situation because in addition to the data \mathbf{d} we also have some hypothesis \mathcal{H} which is underlying the thinking process (the model that relates \mathbf{m} to \mathbf{d} is part of this Hypothesis). Thus we have more fully

$$p(\mathbf{m}|\mathbf{d}, \mathcal{H}) = \frac{p(\mathbf{d}|\mathbf{m}, \mathcal{H})p(\mathbf{m}|\mathcal{H})}{p(\mathbf{d}|\mathcal{H})}.$$

Now we are in a position to evaluate more than one Hypothesis given the data:

$$p(\mathcal{H}_i|\mathbf{d}, I) = \frac{p(\mathbf{d}|\mathcal{H}_i, I)p(\mathcal{H}_i|I)}{p(\mathbf{d}|I)}$$

where I represents information that we are not questioning. You can see that probability theory exposes the fact that we cannot reason intelligently without making assumptions. Joseph Smith stated this fundamental idea somewhat differently when he discussed in his “Lectures on Faith” that faith is the first principle of action. One could say, “you cannot act without faith” just as we have seen you cannot reason without making assumptions. As you learn more and more be careful of falling into the trap of believing that somehow you have lost the need to be humble. You must always be willing to question your assumptions. The most reliable assumptions are those that hold up to questioning (in our own minds) over and over again. Probability theory helps us realize that.

1 Choosing priors

Now that we have emphasized the importance of prior probability. The question still nags as to how to go about constructing priors that are “reasonable” in that they properly transmit our assumptions. This is an area that can still use active work, but there are a couple of ideas that have percolated to the top of the knowledge stew. We will discuss maximum entropy priors, and talk a bit about general non-informative priors, the Jeffrey’s prior (again), and transformation groups.

Our discussion will be guided by the common desire to choose a “minimally-informative” prior. By this, we mean that given a set of constraints, we wish to otherwise add no more information. This requires us to come up with some definition of information. In the discrete, finite world, this problem was usefully solved by Shannon. It is not difficult to see that a “non-informative” prior when we have n choices for m is $p(m) = \frac{1}{n}$. This prior expresses our lack of knowledge about m . In the discrete, finite world this makes complete sense. It becomes more of a struggle as we pass to the continuous and/or infinite limit, so we start with finite, discrete spaces and look at generalizing the notion of minimally-informative when we additionally want to constrain the average of some value and would like to choose a “minimally-informative” prior that has this same average.

2 Entropy and Shannon Information

It is a good idea before determining what is non-informative to establish a useful definition of information. Shannon was thinking about problems arising from communicating “messages” from one place to another when he came up with his definition of entropy as a measure of information. In fact what Shannon showed is quite profound and has implications beyond his original goals. Shannon was trying to derive a measure of the “information” content of a finite, discrete probability space. Define the average information content of this probability assignment as

$$H(p) = E[\phi] = \sum_k \phi(k) p(k).$$

Suppose we impose three requirements on our measure of average information.

1. We want the functional $H(p)$ to take its largest value for $p(k) = \frac{1}{n}$ where n is the number of elementary events having non-zero probability. In other-words we want the largest value of $H(p)$ to correspond to the case when we are the most un-informed about p . (So, entropy is “lack of information”).
2. Let $p(k, m)$ be the joint probability for two independent probability schemes with $p(k, m) = p_A(k) p_B(m)$ then we would like $H(p) = H(p_A) + H(p_B)$ so that our uncertainty measure adds the uncertainty from two independent schemes.
3. Suppose we have a probability scheme defined by $p(k) = P\{A_k\}_{k=1}^n$ and we define a new scheme $p'(k) = P\{A_k\}_{k=1}^{n+1}$ with $p'(k) = p(k)$ for $1 \leq k \leq n$ and $p'(n+1) = 0$. Clearly we have not altered the problem in any real way and so we require that $H(p) = H(p')$.

Shannon proved that the unique function which satisfies these requirements is

$$H(p) = - \sum_{k=1}^n p(k) \log_b p(k)$$

where $b > 0$ is any constant. In other words, the information measure is

$$\phi(k) = - \log_b p(k) = \log_b \left(\frac{1}{p(k)} \right).$$

The choice of logarithm is often described as choosing the units of the information measure. In communication theory often $b = 2$ and the units of information are “bits.” In inference problems we generally choose $b = e$ and call the units of information “nats.” Shannon called $H(p)$ the entropy of the probability scheme. If $p(k)$ represents a degree of belief on some parameter, then $H(p)$ represents how much information we know about that parameter. High $H(p)$ means we know very little; low $H(p)$ means we know quite a bit.

With this definition of entropy as average information we can see how it plays a role in choosing discrete priors. The most non-informative prior is the one which maximizes the entropy. With no other constraint we know that $p(k) = \frac{1}{n}$ gives us the maximum entropy (by design) and is the most non-informative prior. A more interesting use of maximum entropy priors occurs when we have some constraints to apply. For example suppose we have a set of r constraints with $r < n$ that can be written as

$$E[h_i(k)] = \mu_i \quad i = 1 \dots r.$$

What is the maximum entropy probability mass function that is still consistent with these constraints? We, of course, also have the constraint $\sum_k p_k = 1$. Using Lagrange multipliers we see that the Lagrangian for this problem is

$$J(\mathbf{p}) = -\sum_k p_k \log p_k + (\lambda_0 - 1) \left(1 - \sum_k p_k\right) + \sum_i \lambda_i \left[\mu_i - \sum_k h_i(k) p(k)\right].$$

To find the $p_k \equiv p(k)$ that maximizes this functional we take the derivative and set it equal to 0:

$$\begin{aligned} \frac{\partial J}{\partial p_n} &= -1 - \log p_n - (\lambda_0 - 1) - \sum_i \lambda_i h_i(n) = 0 \\ p_n &= \exp \left[-\lambda_0 - \sum_i \lambda_i h_i(n) \right]. \end{aligned}$$

To find the factors λ_i we take the derivative of the augmented function with respect to λ_i (apply the constraints) giving:

$$\mu_i = \sum_k h_i(k) p(k).$$

First,

$$\sum_k p_k = e^{-\lambda_0} \sum_k \exp \left[-\sum_i \lambda_i h_i(k) \right] = 1.$$

Define the function

$$Z(\lambda_1, \dots, \lambda_m) = Z(\boldsymbol{\lambda}) = \sum_k \exp \left(-\sum_{i=1}^r \lambda_i h_i(k) \right).$$

Then $\sum_k p_k = 1$ means that

$$\lambda_0 = \log Z(\boldsymbol{\lambda}).$$

The other constraints are

$$\begin{aligned} \mu_i &= \exp(-\lambda_0) \sum_{k=1}^n h_i(k) \exp \left[-\sum_{i=1}^r \lambda_i h_i(k) \right] \\ &= -\frac{\partial \lambda_0}{\partial \lambda_i} = -\frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \lambda_i}. \end{aligned}$$

These equations must be solved for the λ_i to obtain

$$p_k = \frac{1}{Z(\boldsymbol{\lambda})} \exp \left[-\sum_{i=1}^m \lambda_i h_i(k) \right]$$

as the maximum entropy distribution. Notice that at this maximum entropy distribution,

$$\begin{aligned} H_{\max} \equiv S(\boldsymbol{\mu}) &= \lambda_0 + \sum_i \lambda_i \mu_i \\ &= \log Z(\boldsymbol{\lambda}) - \sum_i \lambda_i \frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \lambda_i}. \end{aligned}$$

The relationship between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ is fixed by

$$\mu_i = -\frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \lambda_i}.$$

In general, this can be difficult to solve, however we note that

$$\begin{aligned}\frac{\partial S}{\partial \mu_k} &= \sum_i \frac{\partial \log Z}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \mu_k} + \sum_i \frac{\partial \lambda_i}{\mu_k} \mu_i + \lambda_k \\ &= \sum_i (\mu_i - \mu_i) \frac{\partial \lambda_i}{\partial \mu_k} + \lambda_k\end{aligned}$$

so that

$$\lambda_k = \frac{\partial S(\boldsymbol{\mu})}{\partial \mu_k}.$$

which is another way to express the relationship between λ_k and μ_k . This formula is fundamental in many thermodynamic relationships. This fact led Jaynes to describe how thermodynamic entropy is equivalent to the entropy of inference we are discussing. Thus, when someone tells you that in a closed system entropy can never decrease, this is essentially equivalent to the obvious “with no new data, the amount of information you have about a system can never increase.” The fact that probability theory is deeply connected to physics in ways that are not immediately obvious even to practitioners in the field is an important point. As scientists we cannot escape the fact that we can never “know” the laws of physics and therefore should be casting all of our rules in the language of probability theory. Unfortunately this has not been done directly and so we have the current situation in which shadows of probability theory are used but it is unclear how to interpret what is done. Fortunately, clarity exists now (for some) in thermodynamics. We can only hope that some-day clarity also emerges from the tangled interleaving of quantized energy and probability theory that is modern physics.

2.1 Constrain the mean

Suppose we only constrain the mean of a random variable with N possible outcomes (labeled $1 \dots N$) Suppose the mean is μ . Then the minimum information pmf with this mean is

$$p_k = \frac{1}{Z(\lambda)} \exp[-\lambda k],$$

where

$$\begin{aligned}Z(\lambda) &= \sum_{k=1}^N \exp(-\lambda k) = \frac{e^{-\lambda(N+1)} - e^{-\lambda}}{e^{-\lambda} - 1} \\ &= \begin{cases} \frac{1 - e^{-\lambda N}}{e^\lambda - 1} & \lambda \neq 0 \\ N & \lambda = 0 \end{cases}\end{aligned}$$

To find λ and so we must solve

$$\mu = -\frac{1}{Z(\lambda)} \frac{\partial Z(\lambda)}{\partial \lambda}$$

for λ . Proceeding we get

$$\begin{aligned}\mu &= -\frac{e^\lambda - 1}{1 - e^{-\lambda N}} \left[\frac{(e^\lambda - 1) N e^{-\lambda N} - (1 - e^{-\lambda N}) e^\lambda}{(e^\lambda - 1)^2} \right] \\ &= -\frac{N e^{-\lambda N}}{1 - e^{-\lambda N}} + \frac{e^\lambda}{e^\lambda - 1} \\ &= \frac{N}{1 - \eta^N} + \frac{\eta}{\eta - 1}\end{aligned}$$

where $\eta = e^\lambda$. Thus, η can be found by finding the root of

$$(1 - \eta^N)(\eta - 1)\mu = N(\eta - 1) + \eta(1 - \eta^N)$$

or

$$(\mu - 1)\eta^{N+1} - \mu\eta^N + (N + 1 - \mu)\eta - (N - \mu) = 0.$$

For example, suppose $N = 6$ (i.e. a die) and $\mu = 2$ (loaded die), then

$$\eta^7 - 2\eta^6 + 5\eta - 4 = 0.$$

The only real root $\neq 1$ is $\eta = 1.8768$ which leads to

$$\lambda = 0.629571$$

thus,

$$p_k = \frac{1}{1.1144} \exp[-0.629571k]$$

or explicitly: $p_1 = 0.5938$, $p_2 = 0.3164$, $p_3 = 0.1686$, $p_4 = 0.0898$, $p_5 = 0.0478$, $p_6 = 0.0255$. If instead we had fixed $\mu = 3.5$ then $\lambda = 0$ is the solution and we get the minimum information uniform prior.

3 Information with continuous variables

For a continuous random variable, some care must be taken to define information. We can't just take the limit of the discrete definition of entropy as the discrete-values get closer together because it diverges. Ultimately, we can only define relative information between two probability distributions:

$$I(p_1, p_2) = \int p_1(\mathbf{x}) \log \frac{p_1(\mathbf{x})}{p_2(\mathbf{x})} d\mathbf{x}.$$

The problem fundamentally is that for the discrete case the second axiom of our information measure asserted which distribution would have maximum entropy (minimum information). In the continuous case, we don't know what to pick and so our goal of an absolute information scale must be lessened. If we use the homogeneous distribution for the parameter space, then we recover an absolute scale:

$$I(p) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\mu(\mathbf{x})} d\mathbf{x}.$$

In other-words, we push off our question of prior probability to the choice of the homogeneous distribution.

We will summarize methods for choosing $\mu(\mathbf{x})$ later. The two most-justified choices at this point are $\mu(x) = \text{constant}$ and $\mu(x) = \frac{\text{constant}}{x}$, but we can discern methods for choosing other values for $\mu(\mathbf{x})$ as well.

Once we have defined a prior with minimum information, $I(\mu) = 0$, then we may play the same game as before to find a maximum-entropy distribution satisfying certain constraints. Suppose again we have the constraints

$$E[h_i(\mathbf{x})] = w_i$$

along with $\int p(\mathbf{x}) d\mathbf{x} = 1$. We want to find the minimum information prior. Thus, we set up the augmented Lagrangian:

$$J(p) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{\mu(\mathbf{x})} d\mathbf{x} + (\lambda_0 - 1) \left(\int p(\mathbf{x}) d\mathbf{x} - 1 \right) + \sum_i \lambda_i \left[\int h_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - m_i \right].$$

To find the pdf that minimizes this functional, differentiate with respect to each component of p . To do this we note from the section on derivatives that

$$\begin{aligned} \frac{\partial [\int g(p(\mathbf{x})) d\mathbf{x}]}{\partial p}(\mathbf{y}) &= g'(p(\mathbf{y})) \\ \frac{\partial \int g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}}{\partial p}(\mathbf{y}) &= g(\mathbf{y}) \\ \frac{\partial (J_1 + J_2)}{\partial p} &= \frac{\partial J_1}{\partial p} + \frac{\partial J_2}{\partial p}. \end{aligned}$$

Thus, differentiation produces:

$$0 = 1 + \log \frac{p(\mathbf{y})}{\mu(\mathbf{y})} + \lambda_0 - 1 + \sum_i \lambda_i h_i(\mathbf{y})$$

so that

$$p(\mathbf{y}) = \mu(\mathbf{y}) e^{-\lambda_0} \exp \left[- \sum_i \lambda_i h_i(\mathbf{y}) \right].$$

We can again define the partition function:

$$Z(\boldsymbol{\lambda}) = \int \mu(\mathbf{y}) \exp \left[- \sum_i \lambda_i h_i(\mathbf{y}) \right] d\mathbf{y}$$

so that

$$\lambda_0 = \log Z(\boldsymbol{\lambda})$$

and

$$p(\mathbf{y}) = \frac{\mu(\mathbf{y})}{Z(\boldsymbol{\lambda})} \exp \left[- \sum_i \lambda_i h_i(\mathbf{y}) \right].$$

The values of λ are determined by satisfying the constraints which can be written as

$$w_k = - \frac{\partial \log Z(\boldsymbol{\lambda})}{\partial \lambda_k} = - \frac{1}{Z(\boldsymbol{\lambda})} \frac{\partial Z(\boldsymbol{\lambda})}{\partial \lambda_k}.$$

We can also define

$$\begin{aligned} S(\mathbf{w}) &= -I_{\max} = - \int p_{\max}(\mathbf{y}) \log \frac{p_{\max}(\mathbf{y})}{\mu(\mathbf{y})} d\mathbf{y} \\ &= \frac{-1}{Z(\boldsymbol{\lambda})} \int \mu(\mathbf{y}) \exp \left[- \sum_i \lambda_i h_i(\mathbf{y}) \right] \\ &\quad \times \left[- \log Z(\boldsymbol{\lambda}) - \sum_i \lambda_i h_i(\mathbf{y}) \right] \\ &= \lambda_0 + \sum_i \lambda_i w_i \end{aligned}$$

so that again

$$\lambda_i = \frac{\partial S}{\partial w_i}.$$

3.1 Constrain mean

Suppose we've only constrained the mean so we have the constraints

$$E[\mathbf{x}] = \mathbf{m}.$$

Thus, $h_i(\mathbf{x}) = x^i$ for $i = 1 \dots N$. The desired pdf is

$$p(\mathbf{y}) = k \mu(\mathbf{y}) \exp \left[- \sum_i \lambda_i y^i \right].$$

If $\mu(\mathbf{y}) = \text{constant}$ then,

$$p(\mathbf{y}) = k \exp \left[- \boldsymbol{\lambda}^T \mathbf{y} \right]$$

is the appropriate pdf. We see that this solution presents an un-normalizable pdf unless we also constrain the interval.

3.2 Constrain mean and variance

Suppose we have the constraints

$$\begin{aligned} E[\mathbf{x}] &= \mathbf{m} \\ E[\mathbf{x}\mathbf{x}^T] &= \mathbf{R}. \end{aligned}$$

Then using the same technique we see that

$$p(\mathbf{y}) = k\mu(\mathbf{y}) \exp \left[-\sum_i \lambda_i y^i - \sum_i \lambda'_{ij} y^i y^j \right].$$

When $\mu(\mathbf{y}) = \text{constant}$, we have a quadratic exponential which is known to be equivalent to a multivariate Gaussian. Thus, the values of λ_i and λ'_{ij} are such that

$$p(\mathbf{y}) = \frac{1}{\sqrt{2\pi^N |\mathbf{C}|}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}) \right]$$

where

$$\mathbf{C} = \mathbf{R} - \mathbf{m}\mathbf{m}^T.$$

This is the classic result that in a space where $\mu(\mathbf{y}) = \text{constant}$ is appropriate (linear vector space with Euclidean distance metric), the minimum information prior if only the mean and covariance matrices are constrained is a Gaussian.

4 Choosing $\mu(\mathbf{x})$

The probability density function $\mu(\mathbf{x})$ basically defines our state of minimal information. It is the pdf which (by definition) gives us minimal information: $I(\mu(\mathbf{x})) = 0$. Choosing it is equivalent to choosing exactly what we mean by the limiting process that brought us to the continuum. The continuum is ultimately a figment of our imagination. It is a construct defined only by a limit. Our ability to create it is a testament to our imaginative power. It can be a very useful construct (*e.g.* I much prefer working with the Gaussian pdf than the binomial pmf with a very-large value of N). The utility diminishes and can lead to confusion if we don't remain constantly vigilant that the continuum only exists as a limiting process. Many mathematical "paradoxes" associated with probability theory can be ultimately traced to an over-looking of the limiting process (i.e. the limit is taken too early and further "operations" are performed without fully understanding the consequences on the limiting process). Our over-indulgence to the continuum may be at the heart of the difficulties surrounding choosing an appropriate $\mu(\mathbf{x})$.

With that in mind, let's look at three ways to come up with $\mu(\mathbf{x})$ for a particular circumstance.

4.1 Homogeneous distribution

Tarantola advocates the assignment $\mu(\mathbf{x})$ as the distribution which assigns equal probability to equal volumes. This implies that there is some defined metric on the space so that volume can be computed. Thus $\mu(\mathbf{x})$ is such that

$$\int_A \mu(\mathbf{x}) d\mathbf{x} = k \int_A dV(\mathbf{x}) = k \int_A v(\mathbf{x}) d\mathbf{x}$$

so that

$$\mu(\mathbf{x}) = kv(\mathbf{x})$$

where if possible k is chosen to cause $\int \mu(\mathbf{x}) = 1$. In certain circumstances improper pdfs are allowable (only to the degree they represent a limiting operation). These pdfs cannot admit normalization and so k is left arbitrary. In these cases, it should not matter what k is. In a Riemann space with a metric tensor defined it can be shown that

$$v(\mathbf{x}) = \sqrt{\det \mathbf{g}(\mathbf{x})}$$

where $g_{ij}(\mathbf{x})$ is the metric associated with the space.

4.2 Limiting process

Another technique for finding $\mu(\mathbf{x})$ is to define the discrete version of the problem and take a limit. This method can work for finite regions quite well. For example, suppose we have the region $a \leq x < b$ and we wish to define a non-informative prior. If we divide this region up into N pieces and define $p_i = \text{Prob}\{i \in [a + \frac{i}{N}(b-a), a + \frac{i+1}{N}(b-a))\}$ for $i = 0 \dots N-1$. A minimally-informative distribution would assign $p_i = \frac{1}{N}$. Then we could define $\mu(\mathbf{x})$ as the limiting distribution:

$$\mu(x) = \lim_{N \rightarrow \infty} \frac{p_i}{\text{width}_i} = \frac{1}{N} \frac{1}{\frac{b-a}{N}} = \frac{1}{b-a}$$

so that the limiting distribution is uniform over the range given. In this process we are implicitly defining a (Euclidean) metric by our choice of a constant interval width (we made the widths identical over the entire range). We could have come up with a different value for $\mu(x)$ had our interval widths varied. So, in fact, the limiting process (while useful) still forces us to think about what the distance metric is (although it lets us do it for a finite-sized problem).

4.3 Transformation groups

Finally, a powerful but under-explored method for finding $\mu(\mathbf{x})$ is to find the pdf that is unchanged under intrinsic transformations that should not change the state of information. In other-words, we think about what kinds of transformations would not change the state of information on the pdf. These transformations imply a change to the underlying pdf. Because the state of information should not change we end up with a functional equation whose solution is an appropriate prior.

The most famous example is the Jeffrey's prior named after Sir Harold Jeffrey's who first postulated this improper prior in the 1930's. The Jeffrey's prior can be defined by requiring that the minimum-information pdf $\mu(x)$ be unchanged when we take powers of the (assumed positive) parameter x . Thus, x and $y = x^m$ should both have the same pdf:

$$\mu(x) = f(x).$$

But,

$$f(y) = \sum_i^m \frac{\mu(y^{1/m})}{my^{\frac{m-1}{m}}},$$

thus,

$$\mu(x) = \frac{1}{x} x^{\frac{1}{m}} \mu(x^{1/m})$$

A solution is

$$\mu(x) = \frac{k}{x}$$

where k is any constant and it is assumed that $x > 0$ because $\mu(x) = 0$ for $x < 0$.

4.4 Conclusion

It appears that minimally-informative priors on infinite domains must be improper (non-normalizable). This is only a conjecture at this point. Choosing an appropriate "minimally-informative" prior seems to come down to choosing a suitable metric on the space. Sometimes this is easiest to work out using transformation groups. Other-times it may be easiest to work out from the metric tensor or from consideration of the variables of interest.