Antidata

Carlos C. Rodríguez

March 4, 2006

To Herb.
`http://omega.albany.edu:8008/Herb/`

Introduction

How can I honestly justify to myself and others, the use of one prior instead of another? Where does prior information come from?

These questions are in the mind of anyone confronted with bayesian inference for the first time. The usual: Relax, don't worry, it doesn't really matter much what the prior is... after just a few observations the likelihood dominates over the prior and we all end up agreeing about the posterior...etc, does in fact help to reduce the level of anxiety of a soon to become bayesian acolyte but the problem remains. Since, what if the prior does matter? Surely there must exist situations where it matters what the prior is. Then what?

Herbert Robbins loved to make fun of the non-empirical faithful bayesians by closing his eyes shaking, and saying: *"and they deliver the prior!"*. Almost always he performed this act right after he finished showing how he could estimate the prior from the data and behave asymptotically as if he knew what the prior was, allowing him to shrink the value of the loss in a compound decision problem. He would then get back to those bayesians in his mind, again, screaming loud and clear: *"they can't learn from data!. I can!"* and leaving the room whistling the same military boot-camp tune that he arrived with. I like to think that Herb would have been pleased to know that the conjugate priors, that he often used, were in fact best possible in a precise sense. Thus, not only the parameters of the prior, but the family of priors as well, are chosen by objective data.

I started writing this as class notes for an undergraduate statistics course at SUNY Albany. The reader I initially had in mind was a typical student from my class for whom the course was likely to be the first encounter with bayesian inference. But, right after dotting all the i's for the simplest, most canonical, gaussian examples, in one dimension, I realized to my own surprise, that I had learned a few things that were not in the textbooks. First of all, a pedagogical (if not plain conceptual) point. When writing the conjugate prior for a gaussian mean μ with known variance σ^2 do not write $N(\mu_0, \sigma_0^2)$ but $N(\mu_0, \sigma^2/\nu_0)$. With the second form (which by the way is the form that comes

from the entropic prior) the posterior variance is simply $\sigma^2/(\nu_0+n)$. Second, the virtual data interpretation for the natural conjugate prior for the exponential family, essentially breaks down unless the correct information volume element dV is used. Third, and the most important lesson that I learned from my own notes, the 1-priors even though they are not conjugate they are more ignorant than the 0-priors. The 1-priors, just like the 0-priors can be thought as based on $\alpha > 0$ virtual observations. However, where the 0-priors add the α virtual observations to the actual n sample points, the 1-priors subtract the α from the $n!$. I call this *anti-data* since α of these points annihilate α of the observations leaving us with a total of $n - \alpha$. I find this, as far as I know, new statistics phenomenon very pleasing. True ignorance, that claims only the model and the observed data, has a price. To build the prior we must spend some of the information cash in hand. No free lunches. Thus, the posterior confidence intervals for a gaussian mean with unknown variance could end up a little larger, than the ones from sampling theory.

The Exponential Family

Let vector $x \in \mathcal{X}$ have scalar probability density,

$$p(x|\theta) = \exp\left(\sum_{j=1}^{k} C_j(\theta)T_j(x) - T(x) - \log Z(\theta)\right)$$

defined for all $\theta \in \Theta \subset R^k$. We say that the distribution of x is in the k-parameter exponential family generated by the volume element dV in $\mathcal{X}$ when the above is the density of probability with respect to this dV, i.e.,

$$P[x \in A|\theta] = \int_A p(x|\theta)dV.$$

The data space $\mathcal{X}$ is assumed to be independent of θ. The functions T, C_j, T_j for $j = 1, \ldots k$ and the normalizing constant Z (also known as *the partition function*) are assumed to be (nice) functions of their arguments.

Many well known families of probability distributions are of this *exponential* type. For example, the Bernoulli, binomial, Poisson, beta, gamma, and Normal families are all of this kind. For the case of the two parameter families (beta, gamma and Normal) we can assume the first parameter, or the second parameter or none of them to have a given fix value and the resulting families are also members of the general exponential family. Let's consider the three cases generated by the Normal as typical illustrations.

Gaussian with known variance: $N(\theta, \sigma^2)$

This is a one parameter family with density (w.r.t. dx) given by,

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

by expanding the square and bringing the normalization constant into the exponential we can write,

$$p(x|\theta) = \exp\left(\frac{\theta}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \log Z(\theta)\right)$$

where the partition function is,

$$Z(\theta) = \sqrt{2\pi\sigma^2}\ \exp(\frac{\theta^2}{2\sigma^2}),$$

and, $C_1 = \theta/\sigma^2, T_1 = x, T = x^2/(2\sigma^2)$.

Gaussian with known mean: $N(\mu, \theta)$

The density is,

$$p(x|\theta) = \exp\left(-\frac{(x-\mu)^2}{2\theta} - \log\sqrt{2\pi\theta}\right).$$

In this case, $T = 0, T_1 = (x-\mu)^2, C_1 = -1/(2\theta)$ and, $Z = \sqrt{2\pi\theta}$.

General Gaussian: $N(\mu, \sigma^2), \theta = (\mu, \sigma^2)$

Here,

$$p(x|\theta) = \exp\left(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right)$$

thus, $T = 0, T_1 = x, C_1 = \mu/\sigma^2, T_2 = x^2, C_2 = -1/(2\sigma^2)$ and

$$Z = \sqrt{2\pi\sigma^2}\ \exp(\mu^2/2\sigma^2)$$

Natural Priors for the Exponential Family

When the distribution of a vector x is in the k-parameter exponential family, there is a simple recipe for priors. Make a $(k+1)$-parameter exponential family of priors for θ by defining their scalar probability densities with respect to the information volume element dV in $\{p(\cdot|\theta) : \theta \in \Theta\}$ by,

$$p(\theta|t) = \frac{1}{W(t)}\exp\left(\sum_{j=1}^{k} t_j C_j(\theta) - t_0 \log Z(\theta)\right)$$

where, $t = (t_0, t_1, \ldots, t_k) \in \{t : W(t) < \infty\}$ with,

$$W(t) = \int_\Theta \exp\left(\sum_{j=1}^{k} t_j C_j(\theta) - t_0 \log Z(\theta)\right) dV$$

Prior information is encoded in the values of $t_1, t_2, \ldots, t_k$. The strength of the information supplied with t is measured by t_0. Hence,

$$P[\theta \in A|t] = \int_A p(\theta|t)\, dV$$

where the information volume element dV is given by,

$$dV = \sqrt{\det(g_{ij}(\theta))}\, d\theta$$

and,

$$g_{ij}(\theta) = 4 \int_{\mathcal{X}} \partial_i \sqrt{p(x|\theta)}\ \partial_j \sqrt{p(x|\theta)}\ p(x|\theta)\, dx$$

are the entries of the information matrix.

What's so Natural About These Priors?

Here is another one of Herb's screams to the rescue: *"Parameters are ghosts. Nobody has ever seen a parameter!"*. How right. I would add, that probability distributions are also ghosts. No one has ever seen a probability distribution either (but that doesn't make me a deFinettian though). In fact all we ever see, as much as we *see* anything, is data. Thus, a natural (if not only) way of providing prior information about the ghostly parameters θ is to write down t_0 typical examples: $x_{-1}, x_{-2}, \ldots, x_{-t_0}$ of data that is expected to resemble, as much as possible, the actual observations. This is sometimes possible to implement even with a set of self-nominated domain experts. It has the added virtue that the more the experts disagree among themselves, the better prior you end up with. With the prior data in hand it is now very reasonable to want the prior probability around θ to be proportional to the likelihood of the prior examples. Just like the rationale for maximum likelihood. In symbols,

$$p(\theta|x_{-1}, x_{-2}, \ldots, x_{-t_0}) \propto p(x_{-1}, x_{-2}, \ldots, x_{-t_0}|\theta).$$

Or if you like, just bayes theorem with uniform (Jeffreys) prior. If the likelihood is assumed to be in the same k-parameter exponential family of the actual data (and if that is not a reasonable assumption for the virtual data in hand, then you need to throw away your model and/or interrogate some of your experts) then,

$$p(\theta|x_{-1}, x_{-2}, \ldots, x_{-t_0}) \propto \exp\left(\sum_{j=1}^{k} C_j(\theta) \sum_{i=1}^{t_0} T_j(x_{-i}) - t_0 \log Z(\theta)\right)$$

which is the natural conjugate prior for the exponential family with prior parameters,

$$t_j = \sum_{i=1}^{t_0} T_j(x_{-i}) \text{ for } j = 1, \ldots, k$$

Hence, there is a simple and useful interpretation for the inferences obtained with the aid of these priors. A bayesian using the natural conjugate prior for the exponential family acts as if s/he has extra t_0 observations with sufficient statistics $t_1, \ldots, t_k$. Thus, reliable prior information should be used with a large value for t_0 and weak prior information should be used with a small value for t_0. The values for t (and therefore the prior examples) need to be restricted to the set that allows the resulting prior to be normalizable, otherwise they can't be used. The smallest possible value for t_0 that still defines a normalizable prior, could be used to define a simple notion of ignorance in this context. We illustrate the general points with the three Normal families considered above.

Posterior Parameters

When the likelihood is a k-parameter exponential family and the natural conjugate prior with prior parameter t is used, the posterior after a sample $x^n = (x_1, \ldots, x_n)$ of n iid observations is collected is given by bayes theorem. Using the iid assumption for the data, collecting the exponents of the exponentials, and dropping overall multiplicative constants independent of θ we obtain,

$$\begin{aligned} p(\theta|x^n, t) &\propto p(x^n|\theta)p(\theta|t) \\ &\propto \exp\left(\sum_{j=1}^{k}(t_j + \sum_{i=1}^{n} T_j(x_i))C_j(\theta) - (t_0 + n)\log Z(\theta)\right) \propto p(\theta|t^{(n)}) \end{aligned}$$

where the $(k+1)$ new parameters $t^{(n)}$ are obtained from the simple formulas,

$$t_0^{(n)} = t_0 + n, \; t_j^{(n)} = t_j + \sum_{i=1}^{n} T_j(x_i) \text{ for } j = 1, \ldots, k$$

Natural Prior for $\mu|\sigma^2$ is $N(\mu_0, \frac{\sigma^2}{\nu_0})$

When the likelihood is Gaussian with a given variance σ^2, the natural prior for the mean $\theta = \mu$ is the two parameter family obtained by replacing the sufficient statistic $T_1 = x \in R$ by the prior parameter $t_1 \in R$, replacing $\log Z$ by $t_0 \log Z$ and dropping multiplicative constants independent of θ in the exponential family likelihood for the $N(\theta, \sigma^2)$. The scalar probability density with respect to $dV = d\mu$ is,

$$p(\mu|t_0, t_1) \propto \exp\left(-t_0\frac{\mu^2}{2\sigma^2} + t_1\frac{\mu}{\sigma^2}\right) \propto N(\mu_0, \frac{\sigma^2}{\nu_0})$$

the middle expression is integrable over $\mu \in R$ (the real line) only when $t_0 > 0$ and $t_1 \in R$. Thus, $\mu_0 = \frac{t_1}{t_0} \in R$ and $\nu_0 = t_0 > 0$. To compute the posterior we just apply the general updating formulas. In this case, the posterior is $N(\mu_n, \sigma^2/\nu_n)$ where,

$$\nu_n = \nu_0 + n, \ \mu_n = \frac{\nu_0 \mu_0 + n\bar{x}_n}{\nu_0 + n}$$

is immediately obtained from the updating formulas for t_0 and t_1. Notice that the posterior parameters ν_n and μ_n are the number of observations and the mean of the observed data $x_1, x_2, \ldots, x_n$ augmented by ν_0 extra observations $x_{n+1}, x_{n+2}, \ldots, x_{n+\nu_0}$ with mean μ_0.

Natural Prior for $\sigma^2|\mu$ is $\chi^{-2}(\nu_0, \sigma_0^2)$

When the likelihood of an observation is $x|\theta \sim N(\mu, \theta)$ we have,

$$p(x|\theta) \propto \exp\left(-\frac{1}{2\theta}(x-\mu)^2 - \frac{1}{2}\log\theta\right)$$

and the natural prior is given with respect to $dV = d\theta/\theta$ as,

$$p(\theta|t_0, t_1) \propto \theta^{-\frac{t_0}{2}} \ \exp(-\frac{t_1}{2\theta}).$$

In order for this last function to be integrable over the region $0 < \theta < \infty$ with respect to $d\theta/\theta$, it is necessary that $t_0 > 0$ and $t_1 > 0$. This can be seen by finding the density of $\xi = t_1/\theta$ to be Gamma$(t_0/2, 1/2)$, i.e. a chi-square with $\nu_0 = t_0$ degrees of freedom. We define an inverse chi-square by,

$$\theta \sim \chi^{-2}(\nu_0, \theta_0) \Longleftrightarrow \frac{\nu_0\theta_0}{\theta} \sim \chi^2_{\nu_0}$$

and we say that θ follows an inverse chi-square with ν_0 degrees of freedom and scale θ_0. Also,

$$\theta \sim \chi^{-2}(\nu_0, \theta_0) \Longleftrightarrow p(\theta)\, d\theta \propto \theta^{-\frac{\nu_0}{2}} \ \exp\left(-\frac{\nu_0\theta_0}{2\theta}\right) \frac{d\theta}{\theta}.$$

The natural family of priors for $\sigma^2|\mu$ is then $\chi^{-2}(\nu_0, \sigma_0^2)$ with $\nu_0\sigma_0^2 = t_1 > 0$ and $\nu_0 = t_0 > 0$.

The posterior parameters ν_n and σ_n^2 are obtained from the prior parameters and the observed data as,

$$\nu_n = \nu_0 + n, \text{ and } \nu_n\sigma_n^2 = \nu_0\sigma_0^2 + \sum_{i=1}^{n}(x_i - \mu)^2.$$

Again, these parameters are the number of observations and the sum of the squared deviations with respect to μ of the observed data $x_1, \ldots, x_n$ augmented by ν_0 extra points $x_{n+1}, \ldots, x_{n+\nu_0}$ with,

$$\sum_{i=1}^{\nu_0}(x_{n+i} - \mu)^2 = \nu_0 \ \sigma_0^2$$

The Natural Prior for $\theta = (\mu, \sigma^2)$

We show here that when both the mean and the variance of a Gaussian are unknown the natural prior for the vector (μ, σ^2) is simply the product of the two distributions obtained above, i.e.,

$$(\mu, \sigma^2) \sim \chi^{-2}(\nu_0, \sigma_0^2)\ N(\mu_0, \sigma^2/\nu_0)$$

which is a three parameter family. To see it, just write the likelihood for one observation disregarding overall proportionality constants independent of $\theta = (\mu, v)$ where we now let $v = \sigma^2$),

$$\begin{aligned} p(x|\theta) &\propto \exp\left(-\frac{1}{2v}(x-\mu)^2 - \frac{1}{2}\log v\right) \\ &\propto \exp\left(-\frac{1}{2v}\mu^2 + \frac{x}{v}\mu - \frac{x^2}{2v} - \frac{1}{2}\log v\right) \end{aligned}$$

and use the recipe for the prior: Replace $T_1 = x \in R$ by a parameter $t_1 \in R$, replace $T_2 = x^2 > 0$ by a parameter $t_2 > 0$ and multiply the rest of the terms in the exponent by the strength parameter t_0. In general the vector of prior parameters t needs to be restricted to the set for which the resulting prior is proper. In this case, the information volume element is,

$$dV \propto \frac{d\mu d\sigma^2}{\sigma^3} \propto \frac{d\mu\, dv}{v^{3/2}}$$

and the scalar probability density with respect to this dV is,

$$\begin{aligned} p(\theta|t)dV &\propto \exp\left(-t_0\left(\frac{\mu^2}{2v} + \frac{1}{2}\log v\right) + \frac{t_1}{v}\mu - \frac{t_2}{2v}\right)\frac{d\mu dv}{v^{3/2}} \\ &\propto \left\{\exp\left(-\frac{t_2}{2v} - \frac{t_0}{2}\log v\right)\frac{dv}{v}\right\}\left\{\exp\left(-\frac{t_0}{2v}\mu^2 + \frac{t_1}{v}\mu\right)\frac{1}{v^{1/2}}\right\}\ d\mu \\ &\propto \chi^{-2}(\nu_0, \sigma_0^2)\ N(\mu_0, \sigma^2/\nu_0)\ d\mu\, dv \end{aligned}$$

where we made the substitutions,

$$\nu_0 = t_0,\ \mu_0 = \frac{t_1}{t_0},\ \sigma_0^2 = \frac{t_2}{t_0}.$$

To obtain the posterior parameters ν_n, μ_n and σ_n^2 we apply the following general recipe. Combine the sufficient statistics for the observed data $x_1, x_2, \ldots, x_n$ which in this case are the observed mean and variance, i.e. $\sum_{i=1}^n x_i = n\bar{x}_n$ and $\sum_{i=1}^n (x_i - \bar{x}_n)^2$, with the sufficient statistics for the ν_0 virtual extra observations, $x_{n+1}, \ldots, x_{n+\nu_0}$, namely $\sum_{i=1}^{\nu_0} x_{n+i} = \nu_0\mu_0$ and $\sum_{i=1}^{\nu_0}(x_{n+i} - \mu_0)^2 = \nu_0\sigma_0^2$. To obtain the updating formulas we simply pool together all the data. The actual observations with the virtual observations. We then have,

$$\nu_n = \nu_0 + n, \;\; \mu_n = \frac{\nu_0\mu_0 + n\bar{x}_n}{\nu_0 + n}$$

as the total number of points and the overall mean. Finally, the new variance is obtained simply from,

$$\nu_n\sigma_n^2 = \sum_{i=1}^{\nu_0+n} (x_i - \mu_n)^2.$$

This, however needs to be simplified to an expression containing only available data since the virtual observations are only given through the value of the sufficient statistics, $\nu_0\mu_0$ and $\nu_0\sigma_0^2$. Towards this end, we split the sum and add and subtract the sample mean $\bar{x}_n$ to the first term and add and subtract the prior mean μ_0 to the second term, obtaining,

$$\begin{aligned}
\nu_n\sigma_n^2 &= \sum_{i=1}^{n}(x_i - \bar{x}_n + \bar{x}_n - \mu_n)^2 + \sum_{i=1}^{\nu_0}(x_{n+i} - \mu_0 + \mu_0 - \mu_n)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu_n)^2 + \nu_0\sigma_0^2 + \nu_0(\mu_0 - \mu_n)^2 \\
&= \nu_0\left[\sigma_0^2 + (\mu_0 - \mu_n)^2\right] + n\left[\hat{\sigma}_n^2 + (\bar{x}_n - \mu_n)^2\right] \\
&= \nu_0\sigma_0^2 + n\hat{\sigma}_n^2 + \frac{n\nu_0}{\nu_0 + n}(\bar{x}_n - \mu_0)^2.
\end{aligned}$$

Any of the last two identities can be computed from the values of the sample and prior means and variances.

Direct Computation of the Posterior

It is a bit harder to check that the above formulas actually come from a direct application of bayes rule. For completeness we show here all the steps. We have data $x^n = (x_1, \ldots, x_n)$ iid $N(\mu, v)$, the parameter is $\theta = (\mu, v)$ and the prior is the natural prior obtained above, i.e., the joint prior distribution of (μ, v) is given as the marginal distribution $v \sim \chi^{-2}(\nu_0, \sigma_0^2)$ multiplied by the conditional distribution $\mu|v \sim N(\mu_0, v/\nu_0)$ with both densities given with respect to the standard Lebesgue measure. By bayes theorem and the underlying assumptions we have that the posterior density, but now with respect to the usual $dV = d\mu dv$, is given by,

$$\begin{aligned}
p(\theta|x^n) &\propto p(x^n|\theta)\; p(\theta) \\
&\propto v^{-n/2} \exp\left(\frac{-1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2\right)
\end{aligned}$$

$$v^{-\nu_0/2-1}\exp\left(\frac{-\nu_0\sigma_0^2}{2v}\right)$$
$$v^{-1/2}\exp\left(\frac{-\nu_0}{2v}(\mu-\mu_0)^2\right)$$

adding and subtracting the sample mean inside the summation, expanding the squares, simplifying, collecting terms, and letting $\hat{\sigma}_n^2$ to be the sample variance, i.e., $n\hat{\sigma}_n^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2$ we get,

$$\begin{aligned} p(\theta|x^n) \quad &\propto \quad v^{\frac{-(\nu_0+n)}{2}-1}\exp\left(\frac{-1}{2v}(\nu_0\sigma_0^2 + n\hat{\sigma}_n^2)\right) \\ &\qquad v^{-1/2}\exp\left(\frac{-1}{2v}[(\nu_0+n)\mu^2 - 2(\nu_0\mu_0 + n\bar{x}_n)\mu]\right) \\ &\qquad \exp\left(\frac{-1}{2v}[\nu_0\mu_0^2 + n(\bar{x}_n)^2]\right) \end{aligned}$$

now, the middle line clearly shows that $\mu|v \sim N(\mu_n, v/\nu_n)$ but we need to complete the square explicitly to be able to identify the marginal distribution of v. We must be careful, not to add extra multiplicative terms containing the variance v and so what needs to be added to complete the square also needs to be killed explicitly. That's the reason for the second exponential in the second line below.

$$\begin{aligned} p(\theta|x^n) \quad &\propto \quad v^{\frac{-(\nu_0+n)}{2}-1}\exp\left(\frac{-1}{2v}(\nu_0\sigma_0^2 + n\hat{\sigma}_n^2)\right) \\ &\qquad \exp\left(\frac{-(\nu_0+n)}{2v}(\mu-\mu_n)^2\right)\ \exp\left(\frac{+(\nu_0+n)}{2v}\mu_n^2\right) \\ &\qquad \exp\left(\frac{-1}{2v}[\nu_0\mu_0^2 + n(\bar{x}_n)^2]\right) \end{aligned}$$

finally collect all the terms and simplify,

$$\begin{aligned} p(\theta|x^n) \quad &\propto \quad v^{-1/2}\exp\left(\frac{-(\nu_0+n)}{2v}(\mu-\mu_n)^2\right) \\ &\qquad v^{\frac{-(\nu_0+n)}{2}-1}\exp\left(\frac{-1}{2v}(\nu_0\sigma_0^2 + n\hat{\sigma}_n^2 + A)\right) \end{aligned}$$

where,

$$\begin{aligned} (\nu_0+n)A \quad &= \quad (\nu_0+n)(\nu_0\mu_0^2 + n(\bar{x}_n)^2) - (\nu_0\mu_0 + n\bar{x}_n)^2 \\ &= \quad \nu_0 n(\mu_0^2 + (\bar{x}_n)^2 - 2\mu_0\bar{x}_n) \\ &= \quad \nu_0 n(\bar{x}_n - \mu_0)^2. \end{aligned}$$

Which shows that the marginal distribution of v is indeed $\chi^{-2}(\nu_n, \sigma_n^2)$ with the updating formulas for ν_n and σ_n^2 as previously given.

The Natural Priors are Entropic

The natural conjugate priors for the exponential family are a special case of a larger and more general class of priors known as entropic priors. Entropic priors maximize ignorance and therefore the natural conjugate priors for the exponential family inherit that optimality property. Thus, the priors introduced above are not only convenient, they are also best in a precise objective sense.

Entropy and Entropic Priors

Here I try to minimize technicalities and full generality in favor of easy access to the main ideas. Let's start with *entropy*. It is a number associated to two probability distributions for the same data space $\mathcal{X}$. It measures their intrinsic dissimilarity (or separation) as probability distributions. For distributions P, Q with densities p and q, with respect to a dV in $\mathcal{X}$, define

$$I(P:Q) = E_p\left[\log \frac{p(x)}{q(x)}\right] = \int_{\mathcal{X}} p(x)\ \log \frac{p(x)}{q(x)}\ dV$$

as their entropy. The I is for Information and the : is for ratio. This notation makes $I(P:Q) \neq I(Q:P)$ explicit. The operator E_p denotes expectation assuming $x \sim p$. Even though the definition of $I(P:Q)$ uses the densities p, q and a volume element dV, the actual value is independent of all of that!. The number $I(P:Q)$ depends only on the probability distributions P, Q not on the choice of dV that it is needed for defining densities p, q. There is a continuum of measures of separation $I_\delta(P:Q)$ for $0 \leq \delta \leq 1$ that fills up the gap between $I_0 = I(Q:P)$ and $I_1 = I(P:Q)$. In fact, $I(P:Q)$ is the mean amount of information for discrimination of P from Q when sampling from P; $I(Q:P)$ same but when sampling from Q instead; $I_\delta(Q:P)$ sort of when sampling from a mixture in between P and Q. What's important is that these I_δ are essentially the only quantities able to measure the intrinsic separation between probability distributions. In here we'll be considering only the two extremes I_0 and I_1.

Entropic Priors for the Exponential Family

When the distributions P, Q are on the same exponential family we label them by their parameters η and θ and compute,

$$I(\eta:\theta) = \sum_{j=1}^{k} (C_j(\eta) - C_j(\theta))\ \tau_j - \log \frac{Z(\eta)}{Z(\theta)}$$

straight from the above definition. Where,

$$\tau_j = E_\eta\left[T_j(x)\right] = E\left[T_j(x)|\eta\right] = \tau_j(\eta)$$

is the expected value of the j-th sufficient statistic when $x \sim p(x|\eta)$.

The 0-entropic family of prior distributions for θ with parameters $\alpha > 0$ and $\theta_0 \in \Theta$ is given by the following scalar densities with respect to the information volume in Θ,

$$\begin{aligned} p(\theta|\alpha,\theta_0) &\propto e^{-\alpha I_0(\theta:\theta_0)} \propto e^{-\alpha I(\theta_0:\theta)} \\ &\propto \exp\left(\sum_{j=1}^{k} C_j(\theta)\ \alpha\tau_j - \alpha \log Z(\theta)\right) \end{aligned}$$

where now $\tau_j = \tau_j(\theta_0)$. This is exactly the density of the natural conjugate prior for the exponential family with prior parameter $t = (\alpha, \alpha\tau) = \alpha(1, \tau_1, \tau_2, \ldots, \tau_k)$.

Example1: $x|\theta \sim N(\theta, \sigma^2)$

In this case,

$$\begin{aligned} I(\theta_0 : \theta) &= \left(\frac{\theta_0}{\sigma^2} - \frac{\theta}{\sigma^2}\right)\ \theta_0 - \log \frac{e^{\theta_0^2/2\sigma^2}}{e^{\theta^2/2\sigma^2}} \\ &= \frac{\theta_0^2}{2\sigma^2} - \frac{\theta_0\theta}{\sigma^2} - \frac{\theta_0^2}{2\sigma^2} + \frac{\theta^2}{2\sigma^2} \\ &= \frac{\theta_0^2}{2\sigma^2} + \frac{\theta^2}{2\sigma^2} - \frac{2\theta_0\theta}{2\sigma^2}. \end{aligned}$$

Thus,

$$I(\theta_0 : \theta) = I(\theta : \theta_0) = \frac{(\theta - \theta_0)^2}{2\sigma^2}$$

and the 0-entropic prior coincides with the 1-entropic prior. The density w.r.t. $d\theta$ is,

$$p(\theta|\alpha,\theta_0) \propto \exp\left(\frac{-\alpha(\theta-\theta_0)^2}{2\sigma^2}\right) \propto N(\theta_0, \frac{\sigma^2}{\alpha})$$

Example2a: 0-prior when $x|\theta \sim N(\mu, \theta)$

For this case we have, $C_1 = -1/2\theta, T_1 = (x-\mu)^2$ and $Z = \sqrt{2\pi\theta}$. Hence,

$$\begin{aligned} I(\theta_0 : \theta) &= \left(\frac{-1}{2\theta_0} - \frac{-1}{2\theta}\right)\ \theta_0 - \frac{1}{2}\log\frac{\theta_0}{\theta} \\ &= \frac{\theta_0}{2\theta} - \frac{1}{2} - \frac{1}{2}\log\frac{\theta_0}{\theta} \end{aligned}$$

and the element of probability for the 0-entropic prior is computed with $dV = d\theta/\theta$ as,

$$\begin{aligned}\exp\left(-\alpha I(\theta_0:\theta)\right)\ \frac{d\theta}{\theta} &\propto \theta^{-\alpha/2}\exp\left(\frac{-\alpha\theta_0}{2\theta}\right)\frac{d\theta}{\theta}\\ &\propto \chi^{-2}(\alpha,\theta_0)\ d\theta\end{aligned}$$

and as expected, coincides with the previously obtained natural conjugate prior for this case. Recall that the conjugate posterior for this case is, $\chi^{-2}(n+\alpha,\hat{\sigma}^2_{n+\alpha})$ where we have written σ^2_n with a hat and with index $n+\alpha$ to make explicit the fact that it is the variance associated to the sample extended by the α virtual points.

Example2b: 1-prior when $x|\theta \sim N(\mu,\theta)$

By interchanging θ with θ_0 in the previous formula for the entropy we obtain the element of probability for the 1-entropic prior,

$$\begin{aligned}\exp\left(-\alpha I(\theta:\theta_0)\right)\ \frac{d\theta}{\theta} &\propto \theta^{\alpha/2}\exp\left(\frac{-\alpha\theta}{2\theta_0}\right)\frac{d\theta}{\theta}\\ &\propto \chi^{2}(\alpha,\theta_0)\ d\theta\end{aligned}$$

where we say that $\theta \sim \chi^2(\alpha,\theta_0)$ when the density (w.r.t. the usual $d\theta$) of $\alpha\theta/\theta_0$ is χ^2_α, i.e. a chi-square with α degrees of freedom.

By bayes theorem, the 1-posterior (i.e. the posterior when the 1-prior is used) w.r.t. the usual $d\theta$ has the form,

$$p(\theta|x^n,\alpha,\theta_0) \propto \theta^{\frac{-(n-\alpha)}{2}-1}\ \exp\left(\frac{-n\hat{\sigma}^2_n}{2\theta}-\frac{\alpha}{2\theta_0}\theta\right)$$

where $n\hat{\sigma}^2_n = \sum_{i=1}^n (x_i-\mu)^2$.

The family of GIGs

The above distribution is known as Generalized Inverse Gaussian (or GIG). A GIG(a,b,c) has density proportional to $\theta^{a-1}\exp(-b/\theta - c\theta)$ defined for $\theta > 0$ and it is normalizable whenever $a \in R$, $b > 0$ and $c > 0$. When, either $b = 0$ or $c = 0$, but not both zero, the GIG becomes a gamma or an inverse-gamma. The normalization constant involves the BesselK function and it is expensive to compute. It is easy to see that the GIGs have the following property,

$$\theta \sim GIG(a,b,c) \iff 1/\theta \sim GIG(-a,c,b).$$

When $a > 0$, and $c > 0$ the best second order Gamma(α,β) approximation to a GIG(a,b,c) is obtained by matching the quadratic Taylor polynomials for the log likelihoods expanded about the mode of the GIG. The values of α and β are given by the simple formulas,

$$\begin{aligned} \alpha &= a + \frac{2b}{m} \\ \beta &= \frac{\alpha - 1}{m} \end{aligned}$$

where m is the mode of the GIG(a, b, c) located at,

$$m = \frac{1}{2c}\left(a - 1 + \sqrt{(a-1)^2 + 4bc}\right).$$

Furthermore, when $2b/m << a$ and $bc/(a-1)^2 << 1$ the formulas simplify to $\alpha = a$ and $\beta = c$ (just expand the square root to first order).

The Approximate 1-posterior for $N(\mu, \theta)$

Using this (just obtained above) gamma approximation for the GIG, valid for n large (often $n > 3$ is enough. See the figures),

$$p(\theta | x^n, \alpha, \theta_0) \propto \chi^{-2}\left(n - \alpha, \hat{\sigma}^2_{n-\alpha}\right)$$

where we have used the definition,

$$\hat{\sigma}^2_{n-\alpha} = \frac{n}{n-\alpha}\hat{\sigma}^2_n.$$

This shows that the 1-posterior is just like the 0-posterior but with $n - \alpha$ points instead of $n + \alpha$ points. I find this remarkable. It is the first indication that the 1-prior is indeed less informative than the conjugate 0-prior. This also shows that, in this case, the number of virtual data points acts as a negative number!. Here is a possible rationalization for this paradoxical result. The 1-prior takes into account the inherent higher uncertainty in the prior data relative to the actual data. In fact, as it is usually the case, if the value of θ_0 is estimated from the observed data, it is only natural that the number of degrees of freedom should be reduced due to the double use of the data. This is not unlike the reduction of degrees of freedom for the likelihood ratio statistic when free parameters need to be estimated from the data.

In figure1 the actual 1-posteriors for $1/\theta$ are shown for sample sizes of $1, 2, 5$ and 10 points. Notice that the curves very quickly become indistinguishable from a gamma and eventually a gaussian distribution. The Laplace-gamma and naive gamma approximations are shown in figure2 for the case $n = 1$ and in figure3 for the case $n = 3$. With only one data point the approximations are not good but with just $n = 3$ both approximations become very similar to the actual posterior.

Example3: $x|\theta \sim N(\mu, v)$ with $\theta = (\mu, v)$

Now,

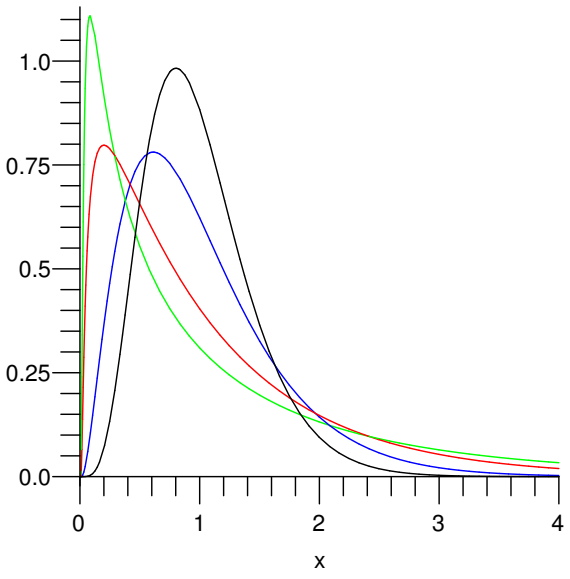

Figure 1: The posterior GIGs for $1/\theta$ when $\alpha = 0.1, \theta_0 = \hat{\sigma}_n^2 = 1$ and $n = 1, 2, 5, 10$ =(green,red,blue,black)

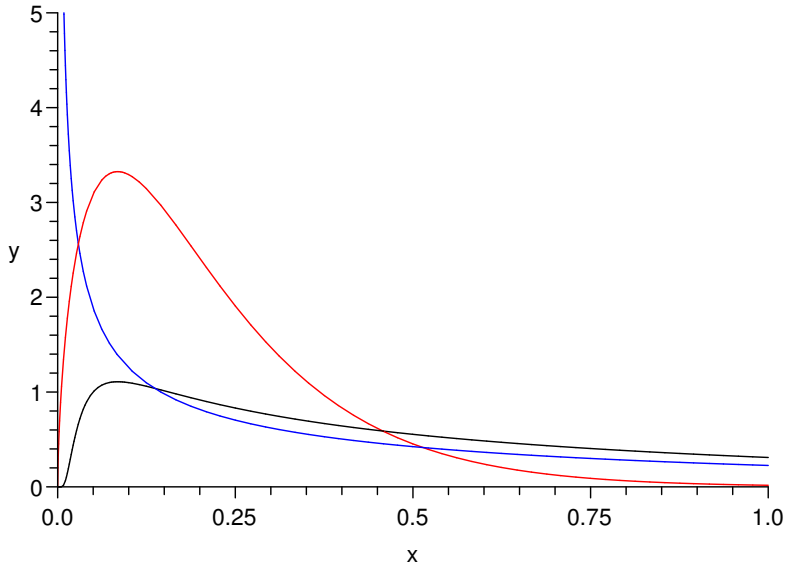

Figure 2: When n=1, GIG(0.45,0.05,0.5) (black); Laplace-Gamma(1.6,7.5) (red); Gamma(0.45,0.5) (blue)

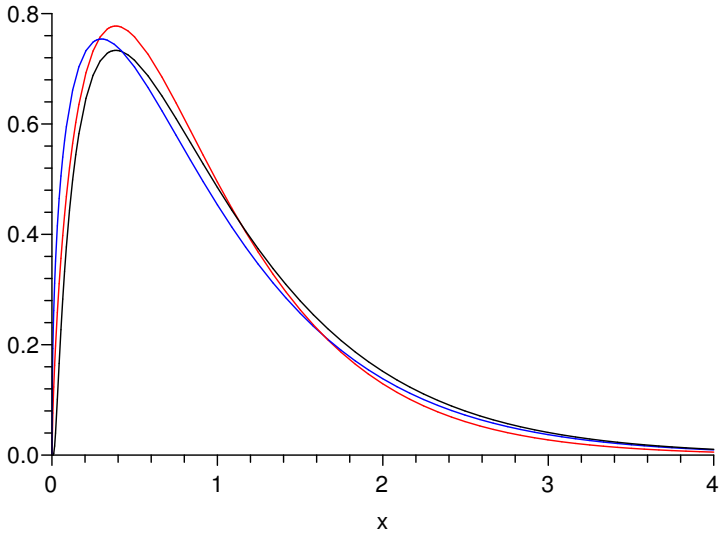

Figure 3: GIG when n=3 (black); Laplace-Gamma approximation (red); Gamma approximation (blue)

$$p(x|\theta) \propto \exp\left(\frac{\mu}{v}x - \frac{x^2}{2v} - \frac{1}{2}\log\ ve^{\mu^2/v}\right)$$

from where we compute,

$$\begin{aligned} I(\theta_0:\theta) &= \left(\frac{\mu_0}{v_0} - \frac{\mu}{v}\right)\ \mu_0 + \left(\frac{-1}{2v_0} - \frac{-1}{2v}\right)(v_0+\mu_0^2) - \frac{1}{2}\log\frac{v_0 e^{\mu_0^2/v_0}}{ve^{\mu^2/v}} \\ &= \frac{\mu^2}{2v} + \frac{\mu_0^2}{2v} - \frac{\mu_0\mu}{v} + \frac{v_0}{2v} - \frac{1}{2} - \frac{1}{2}\log\frac{v_0}{v} \\ &= \frac{(\mu-\mu_0)^2}{2v} + \left\{\frac{v_0}{2v} - \frac{1}{2} - \frac{1}{2}\log\frac{v_0}{v}\right\} \end{aligned}$$

and the 0-prior probability element coincides with what was obtained for the natural conjugate prior,

$$\exp\left(-\alpha I(\theta_0:\theta)\right)\frac{d\mu dv}{v^{3/2}} \propto \chi^{-2}(\alpha, v_0)\ N(\mu_0, \frac{v}{\alpha})\ d\mu dv.$$

On the other hand, the 1-prior, which is non-conjugate in this case, can be easily computed by interchanging θ_0 with θ in the expression for the entropy. We have,

$$\exp\left(-\alpha I(\theta:\theta_0)\right)\frac{d\mu dv}{v^{3/2}} \propto \chi^{2}(\alpha, v_0)\ N(\mu_0, \frac{v}{\alpha})\ d\mu dv.$$

The 1-posterior can be computed from bayes theorem, along the lines of the explicit calculation for the 0-posterior. We obtain,

$$\begin{aligned} p(\mu, v|x^n)\frac{d\mu dv}{v^{3/2}} &\propto \left\{v^{-1/2}\exp\left(\frac{-(\alpha+n)(\mu-\mu_n)^2}{2v}\right)\ d\mu\right\} \\ &\quad v^{-(n-\alpha)/2-1}\exp\left(\frac{-n}{2v}[\hat{\sigma}_n^2 + \frac{\alpha}{\alpha+n}(\bar{x}_n-\mu_0)^2] - \frac{\alpha v}{2v_0}\right)\ dv \\ &\propto \left\{N\left(\mu_n, \frac{v}{\alpha+n}\right)\ \chi^{-2}\left(n-\alpha, \hat{\sigma}_{n-\alpha}^2\right)\right\}\ d\mu dv \end{aligned}$$

where the last expression is valid for n large and α small as before.

The Marginal 1-Posterior for μ

From the last approximation we can integrate over the variance v to obtain the marginal posterior distribution for μ,

$$p(\mu|x^n) = \int_0^\infty p(\mu, v|x^n)\ \frac{dv}{v^{3/2}}$$

$$\propto \int_0^\infty v^{-(n-\alpha+1)/2-1} \exp\left(\frac{-1}{2v}\left[(n-\alpha)\hat{\sigma}^2_{n-\alpha} + (n+\alpha)(\mu-\mu_n)^2\right]\right) dv$$

$$\propto \left[1 + \frac{n+\alpha}{n-\alpha}\left(\frac{\mu-\mu_n}{\hat{\sigma}_{n-\alpha}}\right)^2\right]^{-\left(\frac{n-\alpha+1}{2}\right)}$$

from where we obtain that for n large and α small,

$$\left.\frac{\mu-\mu_n}{\hat{\sigma}_{n-\alpha}/\sqrt{n+\alpha}}\right| x^n \sim t_{n-\alpha}$$

i.e., a student-t distribution with $n-\alpha$ degrees of freedom. Thus, the α virtual prior observations supporting the 1-prior act as anti-data, annihilating an equal number of actual observations and reducing the degrees of freedom to $n-\alpha$. The 1-posterior confidence intervals for μ are larger than the corresponding 0-posterior intervals.

The Actions for Ignorance

Alas the 0-entropic, the 1-entropic prior is in general non conjugate and the inference usually needs to be done by Monte-Carlo. Nevertheless, the 1-priors optimize a notion of ignorance that is remarkably simple (see below).

The 1-prior is the one (and only one) π that makes it most difficult to discriminate the joint distribution of (x^α, θ) (i.e., $\pi(\theta)p(x_1|\theta)p(x_2|\theta)\ldots p(x_\alpha|\theta) \equiv p^\alpha\pi$) from the factorized model $p(x^\alpha|\theta_0)\omega(\theta) \equiv p_0^\alpha\omega$ with $\omega(\theta)$ the uniform (normalized volume) on Θ. In other words, the 1-prior is,

$$\pi^* = \arg\min_\pi I(p^\alpha\pi : p_0^\alpha\omega)$$

The expression for this entropy simplifies to a quantity with an easy interpretation. Compute as follows:

$$\begin{aligned}
I(p^\alpha\pi : p_0^\alpha\omega) &= \int p^\alpha(x|\theta)\pi(\theta)\log\frac{p^\alpha(x|\theta)\pi(\theta)}{p_0^\alpha(x)\omega(\theta)}dx^\alpha d\theta \\
&= \int p^\alpha(x|\theta)\pi(\theta)\left[\log\frac{p^\alpha(x|\theta)}{p_0^\alpha(x)} + \log\frac{\pi(\theta)}{\omega(\theta)}\right]dx^\alpha d\theta \\
&= \int \pi(\theta)\alpha I(\theta:\theta_0)d\theta + \int \pi(\theta)\log\frac{\pi(\theta)}{\omega(\theta)}d\theta \\
&= \alpha\int \pi(\theta)I(\theta:\theta_0)d\theta + \int \pi(\theta)\log\frac{\pi(\theta)}{\omega(\theta)}d\theta.
\end{aligned}$$

Thus, the 1-entropic prior $\pi^*(\theta)$ is the result of a compromise between mass concentration about θ_0 so that $I(\theta:\theta_0)$ remains small and uniform spread allover the model so that $I(\pi:\omega)$, which is the second term of the sum above, is also small. To obtain the solution π^* that minimizes this last expression for

the entropy, among all proper priors with $\int \pi = 1$ is an easy exercise in the calculus of variations. However, for our simple case that involves no derivatives of π, the general Euler-Lagrange equations are an overkill. The necessary and sufficient conditions for the optimal can be readily justified if we replace the infinite dimensional vector π by a large but finite dimensional approximation $(\pi_1, \pi_2, \ldots, \pi_N)$ of piecewise constant values on a discretization of Θ into N little volumes. Thus, we only need to add a Lagrange multiplier for the normalization constraint and set all the partial derivatives equal to zero and obtain the usual Euler-Lagrange equations in the limit as $N \to \infty$. Therefore, to minimize $\int \mathcal{L} d\theta$ with,

$$\mathcal{L} = \alpha \pi I + \pi \log \frac{\pi}{\omega} + \lambda \pi$$

we need,

$$\frac{\partial \mathcal{L}}{\partial \pi} = \alpha I + \log \frac{\pi}{\omega} + 1 + \lambda = 0$$

from which we obtain the 1-prior,

$$\frac{\pi^*}{\omega} = C \; e^{-\alpha I}$$

with C chosen to satisfy the constraint $\int \pi^* = 1$.

The split of the original form of the entropy into the two terms $\alpha < I(. : \theta_0) > + I(\pi : \omega)$ suggests several generalizations. First of all, the parameter $\alpha > 0$ does not need to be an integer anymore. Secondly, the two I's can be replaced by I_δ with two different values for δ obtaining the general three-parameter class of invariant actions for ignorance. All the ignorance priors obtained as the minimizers of these actions share a common geometric interpretation illustrated in figure [4]. In particular, the 0-priors minimize,

$$\alpha \int \pi(\theta) I(\theta_0 : \theta) d\theta + \int \pi(\theta) \log \frac{\pi(\theta)}{\omega(\theta)} d\theta$$

with solution identical in form to the 1-priors but with $I(\theta_0 : \theta)$ in the exponent instead of $I(\theta : \theta_0)$. Thus, the natural conjugate priors for the exponential family are most ignorant in this precise objective sense: the 0-priors are the only proper minimizers of the action above.

Virtual Data and Anti-Data

The 0-priors and the 1-priors are in general quite different. However, we expect the posterior distributions computed from these priors to get closer to one another as more data becomes available. In this section we show that the concept of "*anti-data*" associated to 1-priors, discovered for the special case of the estimation of the mean and variance of a gaussian distribution, holds in general in the exponential family where the log likelihood for n observations is,

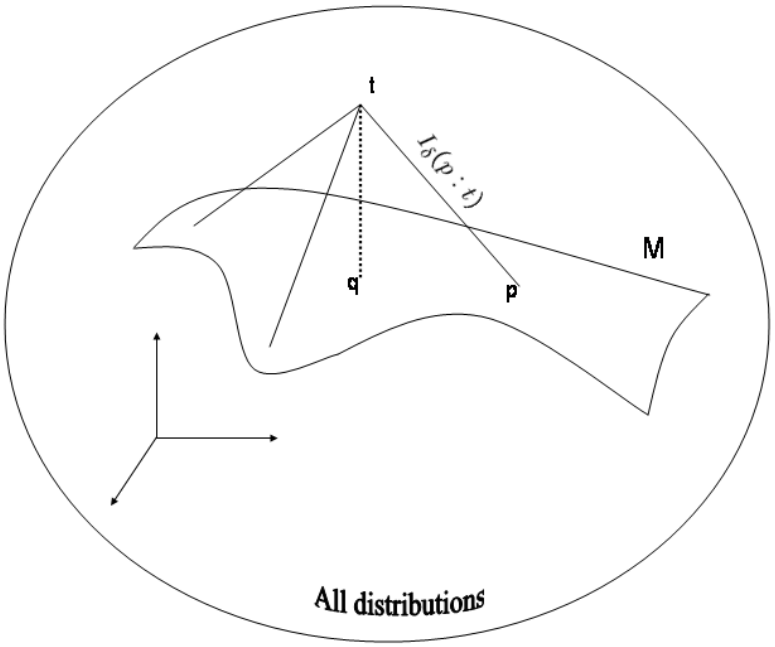

Figure 4: The model $M = \{p\}$, the true distribution t, the projection of the true onto the model is q. Priors are random choices of $p \in M$

$$\log p(x^n|\theta) = \exp(\sum_{j=1}^{k} C_j(\theta) \sum_{i=1}^{n} T_j(x_i)) - n \log Z(\theta) + (\text{other})$$

where the "(other)" terms do not involve θ. For the exponential family the 0-prior π_0, and 1-prior π_1, are such that,

$$\log \pi_0(\theta|\alpha, \theta_0) = \alpha \sum_{j=1}^{k} \tau_j(\theta_0) C_j(\theta) - \alpha \log Z(\theta) + (\text{other}),$$

$$\log \pi_1(\theta|\alpha, \theta_0) = -\alpha \sum_{j=1}^{k} [C_j(\theta) - C_j(\theta_0)] \tau_j(\theta) + \alpha \log Z(\theta) + (\text{other}).$$

Thus, for the 0-posterior

$$\log \pi_0(\theta|x^n, \alpha, \theta_0) = \sum_{j=1}^{k} [\alpha \tau_j(\theta_0) + \sum_{i=1}^{n} T_j(x_i)] C_j(\theta) - (n + \alpha) \log Z(\theta) + (\text{other})$$

and for the 1-posterior,

$$\log \pi_1(\theta|x^n, \alpha, \theta_0) = \sum_{j=1}^{k} (-\alpha [C_j(\theta) - C_j(\theta_0)] \tau_j(\theta) + C_j(\theta) \sum_{i=1}^{n} T_j(x_i)) - (n - \alpha) \log Z(\theta) + (\text{other}).$$

We notice that,

$$\log \pi_1(\theta|x^n, \alpha, \theta_0) = \log \pi_0(\theta|x^n, -\alpha, \theta_0) - \alpha \sum_{j=1}^{k} [C_j(\theta) - C_j(\theta_0)][\tau_j(\theta) - \tau_j(\theta_0)] + (\text{other})$$

which shows that in the limit of weak prior information (i.e., as $\alpha \to 0$) this 1-posterior approaches the 0-posterior but with $-\alpha$ instead of α.