# Parameter priors for Ising models

## research notes

P.G.L. Porta Mana
<piero.mana@ntnu.no>

Y. Roudi
<yasser.roudi@ntnu.no>

25 June 2018; updated 23 July 2018

Study of uniform priors in parameter space and in constraint space for Ising models

'Flat priors do not exist'
(anonymous)

## 1 Initial assumptions for models with second-order sufficient statistics

### 1.1 A two-unit model with sufficient statistics

Consider a population consisting in two binary units $\boldsymbol{s} \coloneqq (s_1, s_2)$ with values in $\{0, 1\}$. One observation of this population can give four results: $\boldsymbol{s} \in \{00, 01, 10, 11\}$.

We have $N$ observations $(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)})$ of this or other populations prepared in similar conditions, so that knowledge of these observations is relevant for our forecast of a new observation $\boldsymbol{s}$, again in similar conditions. We assume that only the number, the mean, and the second moments of these past observations are relevant to forecast the new one; that is, we have these sufficient statistics:

$$N, \quad \tfrac{1}{N}\bigl(\boldsymbol{s}^{(1)} + \dots + \boldsymbol{s}^{(N)}\bigr) =: \bar{\boldsymbol{s}}, \quad \tfrac{1}{N}\bigl(s_1^{(1)}s_2^{(1)} + \dots + s_1^{(N)}s_2^{(N)}\bigr) =: \overline{ss} \qquad (1)$$

These assumptions are collectively denoted $I$.

A series of mathematical results, which we call the Koopman-Pitman-Lauritzen theorem, says that our probabilistic forecasts must assume this general form, for any $N$:

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} | \, I) \\
&\quad = \int \left[ \prod_{i=1}^{N} g\bigl(\boldsymbol{s}^{(i)}\bigr) \, \frac{\exp\bigl(\mu_1 \, s_1^{(i)} + \mu_2 \, s_2^{(i)} + \mu_{12} \, s_1^{(i)} s_2^{(i)}\bigr)}{Z(\boldsymbol{\mu})} \right] \mathrm{p}(\boldsymbol{\mu} | \, I) \, \mathrm{d}\boldsymbol{\mu} \\
&\quad = \int \bigl[\textstyle\prod_{i=1}^{N} g\bigl(\boldsymbol{s}^{(i)}\bigr)\bigr] \, \frac{\exp[N \, (\mu_1 \, \bar{s}_1 + \mu_2 \, \bar{s}_2 + \mu_{12} \, \overline{ss})]}{Z(\boldsymbol{\mu})^N} \, \mathrm{p}(\boldsymbol{\mu} | \, I) \, \mathrm{d}\boldsymbol{\mu} \\
&\qquad\qquad \text{with} \quad \boldsymbol{\mu} \coloneqq (\mu_1, \mu_2, \mu_{12}) \in \mathbf{R}^3, \\
&Z(\boldsymbol{\mu}) \coloneqq g(00) + g(10) \, \exp(\mu_1) + g(01) \, \exp(\mu_2) + g(11) \, \exp(\mu_1 + \mu_2 + \mu_{12}).
\end{aligned} \tag{2}$$

The distribution $g(\boldsymbol{s})$ and the density $\mathrm{p}(\boldsymbol{\mu} | \, I) \, \mathrm{d}\boldsymbol{\mu}$ in the formula above are not determined by the theorem: they need to be determined by additional assumptions. The distribution $g$ is often determined by symmetry or combinatorial properties of the problem. From now on we assume it to be unity: $g(\boldsymbol{s}) = 1$. The density $\mathrm{p}(\boldsymbol{\mu} | \, I) \, \mathrm{d}\boldsymbol{\mu}$ is called *prior parameter density*.

The formula above says that our joint probability distribution for the $N$ outcomes is given by a mixture of joint, factorizable distributions from a three-parameter family. This family is a submanifold in the space of all possible joint distributions, which has dimension $4^N - 1$. The parameters $\boldsymbol{\mu}$ are coordinates in this submanifold, each triplet identifying a particular factorizable distribution

$$\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} | \, \boldsymbol{\mu}, I) = \prod_{i=1}^{N} \frac{\exp\bigl(\mu_1 \, s_1^{(i)} + \mu_2 \, s_2^{(i)} + \mu_{12} \, s_1^{(i)} s_2^{(i)}\bigr)}{Z(\boldsymbol{\mu})}. \tag{3}$$

The weights of the mixture are $\mathrm{p}(\boldsymbol{\mu} | \, I) \, \mathrm{d}\boldsymbol{\mu}$.

From the integral of the formula (6) and the geometric interpretation above it is clear that the theorem does not select the coordinates $\boldsymbol{\mu}$. We can choose another set of coordinates $\boldsymbol{m} \coloneqq (m_1, m_2, m_{12})$, related to the $\boldsymbol{\mu}$ by a one-one transformation $\boldsymbol{m}(\boldsymbol{\mu})$ with inverse $\boldsymbol{\mu}(\boldsymbol{m})$. The

three-dimensional family is then labelled as

$$\mathrm{p}(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)}|\,\boldsymbol{m},I) = \prod_{i=1}^{N} \frac{\exp\bigl[\mu_1(\boldsymbol{m})\,s_1^{(i)} + \mu_2(\boldsymbol{m})\,s_2^{(i)} + \mu_{12}(\boldsymbol{m})\,s_1^{(i)}s_2^{(i)}\bigr]}{Z[\boldsymbol{\mu}(\boldsymbol{m})]}, \tag{4}$$

the prior parameter density is

$$\mathrm{p}(\boldsymbol{m}|\,I)\,\mathrm{d}\boldsymbol{m} = \det\!\left(\frac{\partial \boldsymbol{m}}{\partial \boldsymbol{\mu}}\right)\mathrm{p}(\boldsymbol{\mu}|\,I)\,\mathrm{d}\boldsymbol{\mu}, \tag{5}$$

and the integral formula becomes, with $g(\boldsymbol{s}) = 1$,

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)}|\,I)\\
&= \int \left[\prod_{i=1}^{N} \frac{\exp\bigl[\mu_1(\boldsymbol{m})\,s_1^{(i)} + \mu_2(\boldsymbol{m})\,s_2^{(i)} + \mu_{12}(\boldsymbol{m})\,s_1^{(i)}s_2^{(i)}\bigr]}{Z[\boldsymbol{\mu}(\boldsymbol{m})]}\right]\mathrm{p}(\boldsymbol{m}|\,I)\,\mathrm{d}\boldsymbol{m}\\
&= \int \frac{\exp\{N\,[\mu_1(\boldsymbol{m})\,\bar{s}_1 + \mu_2(\boldsymbol{m})\,\bar{s}_2 + \mu_{12}(\boldsymbol{m})\,\overline{ss})]\}}{Z[\boldsymbol{\mu}(\boldsymbol{m})]^N}\,\mathrm{p}(\boldsymbol{m}|\,I)\,\mathrm{d}\boldsymbol{m},
\end{aligned} \tag{6}$$

equivalent to (6).

This coordinate change is central to the rest of this study.

### 1.2 New coordinates and their motivation

Assuming that (1) are sufficient statistics and therefore using formula (6), let's ask what's the limit probability of observing particular values of the statistics $\bar{s}$, $\overline{ss}$ for very large $N$; that is, $\mathrm{p}(\bar{s}, \overline{ss}|\,I, \text{large } N)$.

In this section we show that there is a particular coordinate system $\boldsymbol{m} \coloneqq (m_1, m_2, m_{12})$ of the three-dimensional manifold discussed above for which *the prior parameter density coincides, in the large-N limit, with the probability of the observed statistics*:

$$\mathrm{p}[(\bar{s}_1, \bar{s}_2, \overline{ss}) = \boldsymbol{x}|\,I, \text{large } N] \approx \mathrm{p}(\boldsymbol{m} = \boldsymbol{x}|\,I). \tag{7}$$

To see this, consider the parameterized, factorized joint probability $\mathrm{p}(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)}|\,\boldsymbol{\mu},I)$ of eq. (13). The expectation of the statistics $(\bar{s}_1, \bar{s}_2, \overline{ss})$ is given by

$$\begin{aligned}
&\mathrm{E}[(\bar{s}_1, \bar{s}_2, \overline{ss})|\,\boldsymbol{\mu},I] =\\
&\qquad \tfrac{1}{N}\sum_i \mathrm{E}[(s_1^{(i)}, s_2^{(i)}, s_1^{(i)}s_2^{(i)})|\,\boldsymbol{\mu},I] = \mathrm{E}[(s_1, s_2, s_1s_2)|\,\boldsymbol{\mu},I]
\end{aligned} \tag{8}$$

where $(s_1, s_2, s_1 s_2)$ refer to any one of the $N$ observations. The two equalities come from the properties of the expectation and the factorized form of the joint probability conditional on $\boldsymbol{\mu}$. From the properties of the variance we also have

$$\mathrm{V}[(\bar{s}_1, \bar{s}_2, \overline{ss})|\, \boldsymbol{\mu}, I] = \frac{1}{N}\mathrm{V}[(s_1, s_2, s_1 s_2)|\, \boldsymbol{\mu}, I]. \tag{9}$$

This means that for a triplet $\boldsymbol{\mu}$, for large $N$ we have a probability distribution for the statistics that is very peaked at particular values $\boldsymbol{m} \coloneqq (m_1, m_2, m_{12})$ determined by the equations

$$\begin{gathered} m_1 = \mathrm{E}(s_1|\, \boldsymbol{\mu}, I) \equiv \frac{\partial \ln Z(\boldsymbol{\mu})}{\partial \mu_1}, \qquad m_2 = \mathrm{E}(s_2|\, \boldsymbol{\mu}, I) \equiv \frac{\partial \ln Z(\boldsymbol{\mu})}{\partial \mu_2}, \\ m_{12} = \mathrm{E}(s_1 s_2|\, \boldsymbol{\mu}, I) \equiv \frac{\partial \ln Z(\boldsymbol{\mu})}{\partial \mu_{12}}. \end{gathered} \tag{10}$$

This system of equations actually puts the parameters $\boldsymbol{\mu} \coloneqq (\mu_1, \mu_2, \mu_{12})$ and $\boldsymbol{m} \coloneqq (m_1, m_2, m_{12})$ into one-one correspondence (Mead et al. 1984). The former belong to $\mathbf{R}^3$; the latter to the bounded domain

$$0 \leqslant m_1, m_2 \leqslant 1, \qquad \max(0, m_1 + m_2 - 1) \leqslant m_{12} \leqslant \min(m_1, m_2) \tag{11}$$

shown in fig. 1.

Using these new parameters, the probability for statistics $(\bar{s}, \overline{ss})$ becomes for large $N$

$$\mathrm{p}(\bar{s}_1, \bar{s}_2, \overline{ss}|\, m_1, m_2, m_{12}, I) \approx \delta[(\bar{s}_1, \bar{s}_2, \overline{ss}) - (m_1, m_2, m_{12})]. \tag{12}$$

Taking the convex combination of this expression in $\boldsymbol{m}$ with weights $\mathrm{p}(\boldsymbol{m}|\, I)\, \mathrm{d}\boldsymbol{m}$ we obtain eq. (7).

In the coordinates $\boldsymbol{m}$, the formula (6) given by the theorem can be interpreted in the following way.

1. We first assume to know that the limit statistics in a very large number of observations is $\boldsymbol{m} \coloneqq (m_1, m_2, m_{12})$. Given this knowledge we can combinatorially calculate the probability of observing a finite sequence of $N$ observations, $\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)}|\, \boldsymbol{m}, I)$, assuming that all sequences having given statistics are equally likely – this equiprobability corresponds to setting $g(\boldsymbol{s}) = 1$ in eq. (6). An example of this combinatorial calculation is given below.

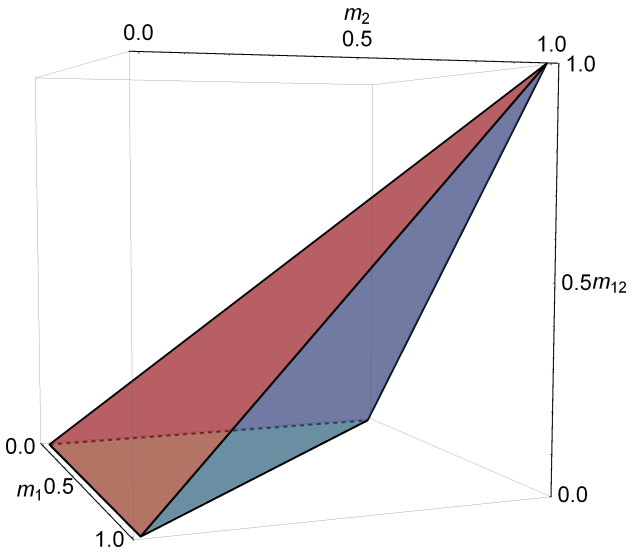


Figure 1

2. We then express our uncertainty about the limit statistics with the probability density $\mathrm{p}(\boldsymbol{m}|\, I)\,\mathrm{d}\boldsymbol{m}$.
3. The two uncertainties above are finally combined in the usual way using the law of total probability.

Let's show, in the simplest case, that the probability conditional on the statistics $\boldsymbol{m}$,

$$\mathrm{p}(\boldsymbol{s}|\, \boldsymbol{m}, I) = \frac{\exp\bigl[\mu_1(\boldsymbol{m})\, s_1 + \mu_2(\boldsymbol{m})\, s_2 + \mu_{12}(\boldsymbol{m})\, s_1 s_2\bigr]}{Z[\boldsymbol{\mu}(\boldsymbol{m})]} \tag{13}$$

is indeed given combinatorially assuming equiprobability of all sequences, as claimed above. Suppose that we know the limit statistics are $(m_1, m_2, m_{12})$. Only the outcome $\boldsymbol{s} = 11$ gives a non-vanishing contribution to the second moment $m_{12}$, eq. (1). This number is therefore equal to the limit relative frequency of 11. Assuming equiprobability we set the probability of this outcome in the next observation equal to this frequency $m_{12}$. Only the outcomes 10 and 11 give non-vanishing contributions to the mean $m_1$; this number is therefore equal to their joint limit relative frequencies. Since the frequency of 11 is given by $m_{12}$, the frequency of 10 must be given by $m_1 - m_{12}$, which is then our probability

for this outcome in the next observation. Analogous reasoning holds for the outcome 01. Finally, the limit relative frequencies of all four outcomes must sum to 1; thus the limit frequency and probability of outcome 00 must be $1 - m_{12} - (m_1 - 1) - (m_2 - m_{12})$. Summarizing,

$$\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{m}, I) = \begin{cases} 1 - m_1 - m_2 + m_{12} & \text{for } \boldsymbol{s} = 00 \\ m_1 - m_{12} & \text{for } \boldsymbol{s} = 10 \\ m_2 - m_{12} & \text{for } \boldsymbol{s} = 01 \\ m_{12} & \text{for } \boldsymbol{s} = 11 \end{cases}$$

$$\equiv m_{12}^{s_1 s_2}\,(m_1 - m_{12})^{s_1 - s_1 s_2}\,(m_2 - m_{12})^{s_2 - s_1 s_2}\,(1 - m_1 - m_2 + m_{12})^{1 - s_1 - s_2 + s_1 s_2}. \tag{14}$$

This probability distribution is exactly eq. (13), as can be checked by finding $\boldsymbol{\mu}(\boldsymbol{m})$ with the inverse of the coordinate transformations (10),

$$\mu_1 = \ln \frac{m_1 - m_{12}}{1 + m_{12} - m_1 - m_2}, \qquad \mu_2 = \ln \frac{m_2 - m_{12}}{1 + m_{12} - m_1 - m_2},$$
$$\mu_{12} = \ln \frac{(1 + m_{12} - m_1 - m_2)\, m_{12}}{(m_1 - m_{12})\,(m_2 - m_{12})}, \tag{15}$$

and substituting it in the right side of eq. (13).

### 1.3 Scientifically motivated prior parameter densities

The interpretation of the Koopman-Pitman-Lauritzen formula (6) explained in the previous section gives us more intuitive grounds to choose the prior parameter density $\mathrm{p}(\boldsymbol{m}|\,I)\,\mathrm{d}\boldsymbol{m}$: given the interpretation of the observables $\boldsymbol{s}$ in a particular scientific context, which limit statistics $\boldsymbol{m}$ would we expect to observe?

If $\boldsymbol{s}$ represents the binned activity of a neural population in the brain, for example, from our research experience we consider more likely to find low mean values $\bar{s}_1 = m_1$, $\bar{s}_2 = m_2$ than high ones, close to 1. We may also have some vague expectations about the second moments $\overline{s_1 s_2} = m_{12}$. Even vague prior knowledge can be expressed by a probability density with particular features, and this leads to better predictions. Let's examine this possibility more concretely.

Using Bayes's theorem with formula (6) we find our probability for a new outcome $\boldsymbol{s}$ conditional on observations $(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)})$:

$$\mathrm{p}(\boldsymbol{s}\,|\,\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)}, I) = \int \frac{\exp\bigl(\mu_1 s_1 + \mu_2 s_2 + \mu_{12} s_1 s_2\bigr)}{Z(\boldsymbol{\mu})}\, \mathrm{p}(\boldsymbol{\mu}\,|\,\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} I)\, \mathrm{d}\boldsymbol{\mu} \tag{16a}$$

with

$$\mathrm{p}(\boldsymbol{\mu}\,|\,\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} I) \propto \left[\prod_{i=1}^{N} \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \mu_{12} s_1^{(i)} s_2^{(i)}\bigr)}{Z(\boldsymbol{\mu})}\right] \mathrm{p}(\boldsymbol{\mu}\,|\,I) \equiv \exp\{N\,[\mu_1 a_1 + \mu_2 a_2 + \mu_{12}\overline{ss} - \ln Z(\boldsymbol{\mu})]\}\, \mathrm{p}(\boldsymbol{\mu}\,|\,I). \tag{16b}$$

The density $\mathrm{p}(\boldsymbol{\mu}\,|\,\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} I)$ is called *posterior parameter density*.

The last expression shows that the $N$ observations affect our forecast only through the averages $\bar{\boldsymbol{s}}$ and $\overline{ss}$, eq. (1), as we assumed.

The proportionality relation of the last formula reminds us that we must perform an integral over $\boldsymbol{\mu}$ to calculate the posterior parameter density. We must also perform an integral over $\boldsymbol{\mu}$ to calculate the conditional probability for $\boldsymbol{s}$. These integrals are difficult when we consider populations with many units. When the number $N$ of known observations is large, the posterior parameter density is often approximated by a Dirac delta centred on the maximum of the posterior,

$$\boldsymbol{\mu}_{\mathrm{m}} \coloneqq \arg\sup_{\boldsymbol{\mu}} \{N\,[\mu_1 a_1 + \mu_2 a_2 + \mu_{12}\overline{ss} - \ln Z(\boldsymbol{\mu})] + \ln \mathrm{p}(\boldsymbol{\mu}\,|\,I)\}. \tag{17}$$

The probability for $\boldsymbol{s}$ then equals the exponential calculated at $\boldsymbol{\mu}_{\mathrm{m}}$. If the prior parameter density $\mathrm{p}(\boldsymbol{\mu}\,|\,I)$ is constant or very broad, it can be dropped in the calculation of the maximum, as an approximation.

The literature indeed often assumes a prior parameter density $\mathrm{p}(\boldsymbol{\mu}\,|\,I)$ that is constant in $\boldsymbol{\mu}$. This is an 'improper', non-normalizable prior, because $\boldsymbol{\mu} \in \mathbf{R}^3$. So we are properly considering a *sequence* of normalizable priors of increasing width – for example, normal distributions with increasing variance – and the resulting limit if it exists.

How reasonable is prior density constant in $\boldsymbol{m}$? Let's find the equivalent density for the parameters $\boldsymbol{m}$ discussed in the previous section.

The Jacobian determinants of the transformations $\boldsymbol{\mu}(\boldsymbol{m})$ and $\boldsymbol{m}(\boldsymbol{\mu})$, from eqs (15) and (10), are

$$\det\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{m}}\right) = \frac{1}{m_{12}\,(m_1 - m_{12})\,(m_2 - m_{12})\,(1 + m_{12} - m_1 - m_2)}, \tag{18a}$$

$$\det\left(\frac{\partial \boldsymbol{m}}{\partial \boldsymbol{\mu}}\right) = \det \frac{\partial^2 \ln Z(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}\, \partial \boldsymbol{\mu}} = \frac{\exp(2\mu_1 + 2\mu_2 + \mu_{12})}{Z(\boldsymbol{\mu})^4}. \tag{18b}$$

These expressions are worthy of notice, because they can be uniquely written as

$$\det\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{m}}\right) = \frac{1}{\mathrm{p}(00|\,\boldsymbol{m}, I)\,\mathrm{p}(10|\,\boldsymbol{m}, I)\,\mathrm{p}(01|\,\boldsymbol{m}, I)\,\mathrm{p}(11|\,\boldsymbol{m}, I)} \equiv \frac{1}{\prod_{\boldsymbol{s}} \mathrm{p}(\boldsymbol{s}|\,\boldsymbol{m}, I)}, \tag{19a}$$

$$\det\left(\frac{\partial \boldsymbol{m}}{\partial \boldsymbol{\mu}}\right) = \mathrm{p}(00|\,\boldsymbol{\mu}, I)\,\mathrm{p}(10|\,\boldsymbol{\mu}, I)\,\mathrm{p}(01|\,\boldsymbol{\mu}, I)\,\mathrm{p}(11|\,\boldsymbol{\mu}, I) \equiv \prod_{\boldsymbol{s}} \mathrm{p}(\boldsymbol{s}|\,\boldsymbol{\mu}, I), \tag{19b}$$

as can be checked from eq. (13) for $N = 1$.

Consider an assumption, denoted $I_{\boldsymbol{\mu}}$, that lead us to assign a constant prior parameter density function $\mathrm{p}(\boldsymbol{\mu}|\,I_{\boldsymbol{\mu}})$. For $\boldsymbol{m}$ such an assumption corresponds to the density

$$\mathrm{p}(\boldsymbol{m}|\,I_{\boldsymbol{\mu}})\,\mathrm{d}\boldsymbol{m} \propto \det\left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{m}}\right)\mathrm{d}\boldsymbol{m} \equiv \frac{\mathrm{d}\boldsymbol{m}}{m_{12}\,(m_1 - m_{12})\,(m_2 - m_{12})\,(1 + m_{12} - m_1 - m_2)}, \tag{20}$$

obtained from eqs (19) and (14). This density, besides being improper, gives very high probability to the extreme values of all three statistics. It doesn't seem much appropriate in the context of brain activity discussed above, for example.

Thus two questions appear: What densities are more appropriate? Do they lead to more difficult computations than those used at present?

A simple parameter density that is normalizable and doesn't give too high probability to extreme values of the statistics is that constant in the $\boldsymbol{m}$ coordinates; denote this assumption by $I_{\boldsymbol{m}}$:

$$\mathrm{p}(\boldsymbol{m}|\,I_{\boldsymbol{m}})\,\mathrm{d}\boldsymbol{m} = 6\,\mathrm{d}\boldsymbol{m}, \tag{21}$$

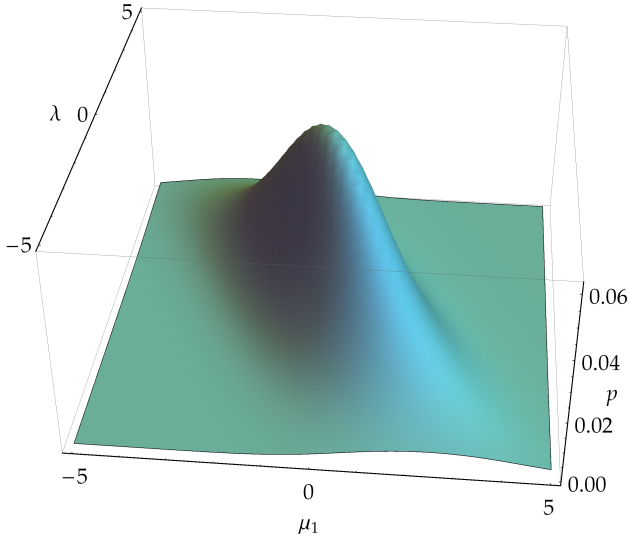


Figure 2

the normalization constant calculated from solid geometry looking at the pyramid of fig. 1, having volume 1/6. In terms of $\boldsymbol{\mu}$ coordinates, using the Jacobian determinant (19), it is

$$\mathrm{p}(\boldsymbol{\mu}|\, I_{\boldsymbol{m}})\,\mathrm{d}\boldsymbol{\mu} = 6\left[\prod_{s}\mathrm{p}(s|\,\boldsymbol{\mu}, I_{\boldsymbol{m}})\right]\mathrm{d}\boldsymbol{\mu} \equiv 6\,\frac{\exp(2\mu_1+2\mu_2+\mu_{12})}{Z(\boldsymbol{\mu})^4}\,\mathrm{d}\boldsymbol{\mu}. \quad (22)$$

Its marginal for $(\mu_1, \mu_{12})$ is shown in fig. 2.

Comparing the prior parameter density above with the general formula (16b) for the posterior density, we see that *the prior density $I_{\boldsymbol{m}}$ is equivalent to the posterior density $I_{\boldsymbol{\mu}}$ conditional on having observed all four possible outcomes once*:

$$\mathrm{p}(\boldsymbol{\mu}|\, I_{\boldsymbol{m}})\,\mathrm{d}\boldsymbol{\mu} = \mathrm{p}(\boldsymbol{\mu}|\,00, 10, 01, 11, I_{\boldsymbol{\mu}})\,\mathrm{d}\boldsymbol{\mu}. \quad (23)$$

We can write this as $I_{\boldsymbol{m}} = I_{\boldsymbol{\mu}} \wedge A$, where $A$ represents the observation of the four outcomes.

This result is computationally important. If we want to make inferences conditional on some data $D$ using a density constant in $\boldsymbol{m}$, we can use the same algorithms and approximations used for the density constant in $\boldsymbol{\mu}$, but augmenting the data $D$ with the 'auxiliary data' $A$.

If it can be proven that the formula for the Jacobian determinant (19) holds for any number of units, then this method requires to add $2^n$ auxiliary data if we consider $n$ units. This should have a big influence on our predictions even when the observed data are numerous.

### 1.4 Explicit formulae for belief update: connection with Dirichlet priors

The three-dimensional family of probabilities $\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{\mu}, I)$ or $\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{m}, I)$ covers the full simplex of probability distributions for the four states of our units. Our model is actually 'non-parametric'. This is clear from formulae (14): the limit statistics $\boldsymbol{m}$ determine the limit relative frequencies of the four states and vice versa.

Consider then a new parameterization $\boldsymbol{q} \coloneqq (q_1, q_2, q_3)$ defined by

$$q_1 = m_1 - m_{12}, \quad q_2 = m_2 - m_{12}, \quad q_3 = m_{12}, \tag{24}$$

and denote

$$q_0 \coloneqq 1 - q_1 - q_2 - q_3. \tag{25}$$

In terms of these parameters our three-dimensional family of distributions is

$$\begin{aligned} \mathrm{p}(\boldsymbol{s}|\,\boldsymbol{q}, I) &= \begin{cases} q_0 & \text{for } \boldsymbol{s} = 00 \\ q_1 & \text{for } \boldsymbol{s} = 10 \\ q_2 & \text{for } \boldsymbol{s} = 01 \\ q_3 & \text{for } \boldsymbol{s} = 11 \end{cases} \\ &\equiv q_{\boldsymbol{s}}, \quad \text{where } \boldsymbol{s} \text{ is interpreted as a binary number.} \end{aligned} \tag{26}$$

The parameters $(q_0, \boldsymbol{q})$ belong to the three-dimensional simplex

$$\Delta \coloneqq \{\boldsymbol{x} \in \mathbf{R}^4 \mid \textstyle\sum_{s=1}^4 x_s = 1, x_s \geqslant 0\}. \tag{27}$$

The probability distribution for a sequence of states becomes

$$\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)}|\, I) = \int \left[\prod_{i=1}^{N} q_{s^{(i)}}\right] \mathrm{p}(\boldsymbol{q}|\, I)\, \mathrm{d}\boldsymbol{q}. \tag{28}$$

From eq. (14), the quantities $(1 - q_1 - q_2 - q_3, q_1, q_2, q_3)$ can be interpreted as the limit relative frequencies of the states $(00, 10, 01, 11)$,

and the parameter density $\mathrm{p}(\boldsymbol{q}|\, I)\,\mathrm{d}\boldsymbol{q}$ as our uncertainty about these frequencies.

The parameter transformation (24) has unit Jacobian, so we have

$$\mathrm{p}(\boldsymbol{m}|\, I)\,\mathrm{d}\boldsymbol{m} = \mathrm{p}(\boldsymbol{q}|\, I)\,\mathrm{d}\boldsymbol{q}. \tag{29}$$

A constant parameter density in $\boldsymbol{m}$ is also constant in $\boldsymbol{q}$, and is a special case of the parameter density function

$$\mathrm{p}(\boldsymbol{q}|\, \Lambda, \boldsymbol{\nu}, I) = \frac{\Gamma(\Lambda)}{\prod_{i=0}^{3}\Gamma(\Lambda\,\nu_i)} \prod_{i=0}^{3} {q_i}^{\Lambda\,\nu_i - 1}, \qquad \text{with} \quad \Lambda > 0,\ \boldsymbol{\nu} \coloneqq (\nu_i) \in \Delta. \tag{30}$$

This density was used by Laplace (1781) and motivated by Johnson (1924; 1932), and its integrals can be analytically calculated with Dirichlet's (1839) formula (32). This density is commonly called 'Dirichlet prior' (see Good 1965 ch. 4; Zabell 1982; Jaynes 1996; Gupta et al. 2001) but we'll call it 'Johnson density' here, since he determined its mathematical form from informational desiderata.

The Johnson density is a conjugate prior (DeGroot 2004 ch. 9; Diaconis et al. 1979): it updates to a density of the same mathematical form. It leads to probabilities with closed-form formulae summarized in appendix A. In particular, the posterior probability distribution for a state is

$$\mathrm{P}(\boldsymbol{s}^{(N+1)}|\, \boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)}, \Lambda, \boldsymbol{\nu}, I) = \frac{\Lambda\,\nu_{\boldsymbol{s}^{(N+1)}} + N\, f_{\boldsymbol{s}^{(N+1)}}}{\Lambda + N}, \qquad \text{with } f_{\boldsymbol{s}^{(N+1)}} \coloneqq \text{rel. frequency of state } \boldsymbol{s}^{(N+1)} \text{ in the } N \text{ observations}. \tag{31}$$

The density $\mathrm{p}(\boldsymbol{m}|\, I_{\boldsymbol{m}})\,\mathrm{d}\boldsymbol{m} = 6\,\mathrm{d}\boldsymbol{m}$ is a special case of the Johnson density with $\Lambda = 4$, $\nu_i = 1/4$. The improper density $\mathrm{p}(\boldsymbol{\mu}|\, I_{\boldsymbol{\mu}})\,\mathrm{d}\boldsymbol{\mu} \propto \mathrm{d}\boldsymbol{\mu}$, which by formula (19) is like $\mathrm{p}(\boldsymbol{m}|\, I_{\boldsymbol{m}})$ 'minus four observations', corresponds to the improper Johnson density with $\Lambda \to 0$, $\nu_i = 1/4$. The formula for the posterior probabilities (35) can be used in both cases. From that formula with $\Lambda = 4$ it is possible to calculate that the number of observations after which the posterior probability for a state will be within 10% of its relative frequency. For a very high relative frequency this number is $N \approx 30$; for a very low relative frequency $f$ this number is $N \approx 10/f$. With $\Lambda \to 0$ instead the predictive probability is always equal to the relative frequency – also meaning that it is zero for states that haven't yet been observed.

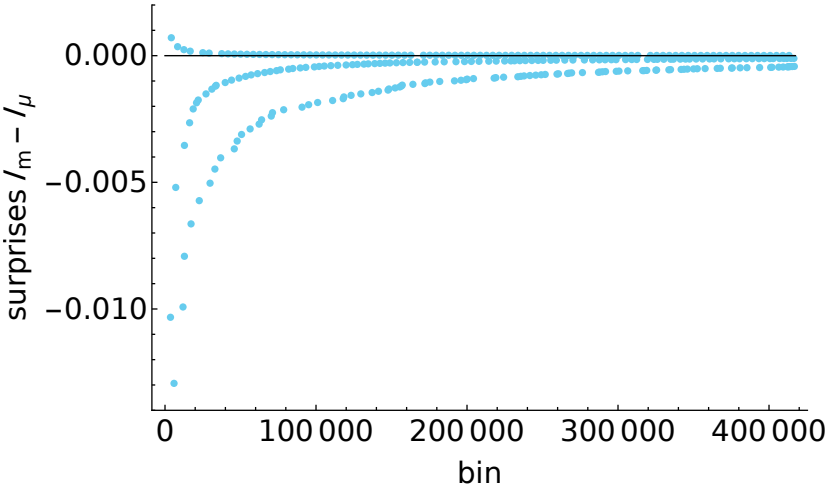


Figure 3

### 1.5 Initial information

If we calculate the

## 2 [Luca:] Does it make sense to test against computer-generated distributions?

> But we must note with sadness that, in much of the current Bayesian literature, very little of the orthodox baggage has been cast off. For example, it is rather typical to see a Bayesian article start with such phrases as: 'Let $X$ be a random variable with density function $p(x|\theta)$, where the value of the parameter $\theta$ is unknown. Suppose this parametric family contains the true distribution of $X$. . . .' Or, one describes a uniform prior $p(\theta|I)$ by saying: '$\theta$ is supposed uniformly distributed'. The analytical solutions thus obtained will doubtless be a valid Bayesian result; but one is still clinging to the orthodox fiction of 'random variables' and 'true distributions'. $\theta$ is simply an unknown constant; it is not 'distributed' at all. What is 'distributed' is our state of knowledge about $\theta$: again there is that persistent

> mind projection fallacy that contaminates all of probability theory, leading inexperienced readers far astray as to what we are doing. Equally bad, those who commit this fallacy seem unaware that this is restricting the application to a small fraction of the real situations where the solution might be useful. In the vast majority of real applications there are no 'random variables' (What defines 'randomness'?) and no 'true distribution' (What defines it? What test could we apply to decide whether some proposed distribution is or is not the 'true' one?); yet probability theory as logic applies to all of them.
>
> Jaynes (2003 § 17.12)

The outcomes of the measurements we make on the system, modelled as a population of binary units, are produced by utterly complex mechanisms. They aren't produced by a computer's pseudo-random-number generator. So it doesn't make sense to check which prior parameter density converges faster to the parameters of some computer algorithm, because algorithm, parameters, distribution do not even exist in the real situation.

There's a circularity in using computer-generated outcomes to test which statistical model or prior density is 'best'. This test would make sense if the computer algorithm mimicked the outcomes of the real experiment well. But if we had such an algorithm we wouldn't need any statistical model to start with.

The use of computer algorithms has a different purpose. A statistical model or a prior density can be difficult to grasp intuitively. It may have properties that don't reflect the information or assumptions we had in mind; in which case it doesn't correctly represent our beliefs. Testing it against controlled computer-generated outcomes may bring forth such undesired properties. This is the 'device of imaginary results' suggested by Good (1950 § 4.3; 1967; index in Good 1983 p. 280; Jaynes 2003 ch. 5).

✠ To be continued

## Appendix

### A Formulae for the Johnson density

Dirichlet's formula

$$\int_{\Delta} \prod_{i=1}^{N} {q_i}^{x_i-1} \,\mathrm{d}\boldsymbol{q} = \frac{\prod_i \Gamma(x_i)}{\Gamma\bigl(\sum_i x_i\bigr)} \tag{32}$$

allows to calculate all integrals involving the Johnson density.

The probability for the sequence of outcomes is

$$\mathrm{p}(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)}|\,\Lambda,\boldsymbol{\nu},I) = \frac{\Gamma(\Lambda)}{\Gamma(\Lambda+N)}\prod_{s=0}^{3}\frac{\Gamma(\Lambda\nu_s+Nf_s)}{\Gamma(\Lambda\nu_s)}, \qquad \text{with} \quad f_s \coloneqq \begin{cases}\text{rel. frequency of } s\\ \text{in the } N \text{ observations}\end{cases} \tag{33}$$

The posterior density function

$$\mathrm{p}(\boldsymbol{q}|\,\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)},\Lambda,\boldsymbol{\nu},I) = \frac{\Gamma(\Lambda')}{\prod_s\Gamma(\Lambda'\nu'_s)}\prod_{s=0}^{3}q_s{}^{\Lambda'\nu'_s-1}$$

$$\text{with} \quad \Lambda'=\Lambda+N, \quad \boldsymbol{\nu}'=\frac{\Lambda\boldsymbol{\nu}+N\boldsymbol{f}}{\Lambda+N}, \quad f_s \coloneqq \begin{cases}\text{rel. frequency of } s\\ \text{in the } N \text{ observations}\end{cases} \tag{34}$$

The posterior probability distribution for a state:

$$\mathrm{P}(\boldsymbol{s}^{(N+1)}|\,\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)},\Lambda,\boldsymbol{\nu},I) = \frac{\Lambda\nu_{\boldsymbol{s}^{(N+1)}}+N\,f_{\boldsymbol{s}^{(N+1)}}}{\Lambda+N}. \tag{35}$$

## Bibliography


('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

DeGroot, M. H. (2004): *optimal statistical decisions*, reprint. (Wiley, New York).

Diaconis, P., Ylvisaker, D. (1979): *Conjugate priors for exponential families*. Ann. Stat. **7**[2], 269–281.

Dirichlet (Lejeune-Dirichlet), P. G. (1839): *Sur une nouvelle méthode pour la détermination des intégrales multiples*. Comptes rendus hebdomadaires des séances de l'Académie des sciences **8**, 156–160. https://gallica.bnf.fr/ark:/12148/bpt6k2967c/f160.item, http://sites.mathdoc.fr/JMPA/feuilleter.php?id=JMPA_1839_1_4.

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

— (1965): *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. (MIT Press, Cambridge, USA).

— (1967): *A Bayesian significance test for multinomial distributions*. J. Roy. Stat. Soc. B **29**[3], 399–418. See also discussion and reply in Lindley, Plackett, Smith, Mitchell, Kerridge, Armitage, Kingman, Barnard, et al. (1967).

— (1983): *Good Thinking: The Foundations of Probability and Its Applications*. (University of Minnesota Press, Minneapolis, USA).

Gupta, R. D., Richards, D. S. P. (2001): *The history of the Dirichlet and Liouville distributions*. Int. Stat. Rev. **69**[3], 433–446.

Jaynes, E. T. (1996): *Monkeys, kangaroos, and N*. http://bayes.wustl.edu/etj/node1.html. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html, http://omega.albany.edu:8008/JaynesBook.html.

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03john.

— (1932): *Probability: the deductive and inductive problems*. Mind **41**[164], 409–423. With some notes and an appendix by R. B. Braithwaite.

Laplace (Marquis de), P. S. (1781): *Mémoire sur les probabilités*. Histoire de l'Académie Royale des Sciences **année 1778**, 227–332. First read 1778. https://archive.org/details/histoiredelacad78hist, https://gallica.bnf.fr/ark:/12148/bpt6k3577w/f333.item.

Lindley, D. V., Plackett, R. L., Smith, C. A. B., Mitchell, A. F. S., Kerridge, D., Armitage, P., Kingman, J. F. C., Barnard, G. A., et al. (1967): *Discussion on Dr Good's paper*. J. Roy. Stat. Soc. B **29**[3], 418–431. See Good (1967).

Mead, L. R., Papanicolaou, N. (1984): *Maximum entropy in the problem of moments*. J. Math. Phys. **25**[8], 2404–2417. http://bayes.wustl.edu/Manual/MeadPapanicolaou.pdf.

Zabell, S. L. (1982): *W. E. Johnson's "sufficientness" postulate*. Ann. Stat. **10**[4], 1090–1099. Repr. in Zabell (2005 pp. 84–95).

— (2005): *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. (Cambridge University Press, Cambridge).