Parameter priors for Ising models

research notes

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

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

25 June 2018; updated 26 June 2018

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

'Flat priors do not exist'
(anonymous)

1 A two-unit model with sufficient statistics

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

Assume that 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. Also assume that only the number, the mean, and the second moments of these past observations are relevant to forecast the new one; that is,

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

are sufficient statistics. These assumptions are collectively denoted I.

There is a series of mathematical results, which we call the Koopman-Pitman-Lauritzen theorem, that says that our probabilistic forecasts must

assume this general form, for any N:

$$\mathrm{p}(\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)}|\, I) = \int \left[\prod_{i=1}^{N} g(\boldsymbol{s}^{(i)})\, \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \lambda s_1^{(i)} s_2^{(i)}\bigr)}{Z(\mu_1,\mu_2,\lambda)}\right] \times \mathrm{p}(\mu_1,\mu_2,\lambda|\, I)\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2\,\mathrm{d}\lambda,$$

$$= \int \bigl[\textstyle\prod_{i=1}^{N} g(\boldsymbol{s}^{(i)})\bigr]\, \frac{\exp[N\,(\mu_1 \bar{s}_1 + \mu_2 \bar{s}_2 + \lambda \overline{ss})]}{Z(\mu_1,\mu_2,\lambda)^N} \times \mathrm{p}(\mu_1,\mu_2,\lambda|\, I)\,\mathrm{d}\mu_1\,\mathrm{d}\mu_2\,\mathrm{d}\lambda,$$

$$\text{with}\quad Z(\mu_1,\mu_2,\lambda) \coloneqq 1 + \exp(\mu_1) + \exp(\mu_2) + \exp(\mu_1+\mu_2+\lambda). \quad (2)$$

Denote the three parameters that appear in this formula by $\theta \coloneqq (\mu_1,\mu_2,\lambda) \in \mathbf{R}^3$.

The distribution $g(\boldsymbol{s})$ and the density $\mathrm{p}(\mu_1,\mu_2,\lambda|\, I)$ 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 $\mathrm{p}(\theta|\, I)$ is called *prior parameter density*.

Formula (2) may appear deceivingly specific in its dependence on the parameters θ; let's summarize in words the content of the theorem:

(*a*) our joint probability for the observations $\boldsymbol{s}^{(1)},\dots,\boldsymbol{s}^{(N)}$ is given by a convex combination of joint probabilities;

(*b*) each joint probability in this convex combination factorizes into N independent probabilities for the N observations;

(*c*) each joint probability in the convex combination is identified by a triple of parameters θ; it therefore belongs to a three-dimensional submanifold of joint probabilities. Note that the full manifold of joint probabilities is (2^N-1)-dimensional;

(*d*) the weight assigned to the probability labelled by θ is $\mathrm{p}(\theta|\, I)\,\mathrm{d}\theta$.

Point (*c*) shows that the Pitman-Koopman-Lauritzen theorem greatly reduces our freedom in specifying the joint probability. This is the effect of assuming that the statistics (1) are sufficient; it's a very strong assumption.

Points (*c*) and (*d*) show that the parameters θ are just coordinates of a three-dimensional manifold. There is nothing special about these coordinates, besides the fact that they appear as coefficients of the linear

combination of statistics in the exponential (2). We could choose different coordinates $\boldsymbol{t}$, with one-one coordinate transformations $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{\theta})$, $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{t})$. In these new coordinates the mixed joint probabilities are

$$\frac{\exp\{N\,[\mu_1(\boldsymbol{t})\,\bar{s}_1 + \mu_2(\boldsymbol{t})\,\bar{s}_2 + \lambda(\boldsymbol{t})\,\overline{ss}]\}}{Z[\boldsymbol{\theta}(\boldsymbol{t})]^N}, \tag{3}$$

and the weights of the convex combination are given by $\mathrm{p}(\boldsymbol{t}|\, I)\,\mathrm{d}\boldsymbol{t}$, the densities for $\boldsymbol{\theta}$ and for $\boldsymbol{t}$ being related by a Jacobian determinant:

$$\mathrm{p}(\boldsymbol{\theta}|\, I) = \mathrm{p}[\boldsymbol{t}(\boldsymbol{\theta})|\, I]\,\det\!\left(\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}}\right). \tag{4}$$

Using Bayes's theorem with the probabilities (2) we find our forecast for a new observation $\boldsymbol{s}$ conditional on observations $\bigl(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)}\bigr)$:

$$\mathrm{p}(\boldsymbol{s}|\, \boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)}, I) = \int \frac{\exp\bigl(\mu_1 s_1 + \mu_2 s_2 + \lambda s_1 s_2\bigr)}{Z(\boldsymbol{\theta})}\, \mathrm{p}(\boldsymbol{\theta}|\, \boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} I)\,\mathrm{d}\boldsymbol{\theta} \tag{5a}$$

with

$$\mathrm{p}(\boldsymbol{\theta}|\, \boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} I) \propto \left[\prod_{i=1}^{N} \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \lambda s_1^{(i)} s_2^{(i)}\bigr)}{Z(\boldsymbol{\theta})}\right] \mathrm{p}(\boldsymbol{\theta}|\, I) = \exp\{N\,[\mu_1 a_1 + \mu_2 a_2 + \lambda\,\overline{ss} - \ln Z(\boldsymbol{\theta})]\}\, \mathrm{p}(\boldsymbol{\theta}|\, I). \tag{5b}$$

The density $\mathrm{p}(\boldsymbol{\theta}|\, \boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} I)$ 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 $\boldsymbol{\theta}$ to calculate the posterior parameter density. We must also perform an integral over $\boldsymbol{\theta}$ 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,

$$\boldsymbol{\theta}_{\mathrm{m}} \coloneqq \arg\sup_{\boldsymbol{\theta}} \{N\,[\mu_1 a_1 + \mu_2 a_2 + \lambda\,\overline{ss} - \ln Z(\boldsymbol{\theta})] + \ln \mathrm{p}(\boldsymbol{\theta}|\, I)\}. \tag{6}$$

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

2 Other prior parameter densities

The literature often assumes a prior parameter density $\mathrm{p}(\theta|\, I)$ that is constant in θ. This is an 'improper', non-normalizable prior, because $\theta \in \mathbf{R}^3$. 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.

As noted before, the parameters θ are just coordinates in a manifold of distributions for $\boldsymbol{s}$. A constant density in these coordinates corresponds to an non-constant density in other coordinates. But are these coordinates 'special' in any way, to consider a constant density in them? Are there other coordinates in which it makes more sense to consider a constant density? How does a different choice affect our inference about $\boldsymbol{s}$?

To consider other coordinates it is useful to give the predictive formula (2) a particular interpretation.

First of all it must be noted that each choice of parameters θ gives a distribution

$$\mathrm{p}(\boldsymbol{s}|\, \theta, I) = \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \lambda s_1^{(i)} s_2^{(i)}\bigr)}{Z(\theta)} \tag{7}$$

with different first and second moments

$$\begin{aligned} \mathrm{E}(\boldsymbol{s}|\, \theta, I) &\coloneqq \sum_{\boldsymbol{s}} \boldsymbol{s}\, \mathrm{p}(\boldsymbol{s}|\, \theta, I), \\ \mathrm{E}(s_1 s_2|\, \theta, I) &\coloneqq \sum_{\boldsymbol{s}} s_1 s_2\, \mathrm{p}(\boldsymbol{s}|\, \theta, I). \end{aligned} \tag{8}$$

In other words the set $\theta \in \mathbf{R}^3$ is in one-one correspondence with the possible values of the three expectations above. We can therefore introduce coordinates $\boldsymbol{t} \coloneqq (m_1, m_2, l)$ that identify the distributions

above via the equations

$$\begin{aligned} m_1 &= \mathrm{E}(s_1|\,\boldsymbol{\theta}, I) \equiv \frac{\exp(\mu_1) + \exp(\mu_1+\mu_2+\lambda)}{Z(\mu_1,\mu_2,\lambda)},\\ m_2 &= \mathrm{E}(s_2|\,\boldsymbol{\theta}, I) \equiv \frac{\exp(\mu_2) + \exp(\mu_1+\mu_2+\lambda)}{Z(\mu_1,\mu_2,\lambda)},\\ l &= \mathrm{E}(s_1 s_2|\,\boldsymbol{\theta}, I) \equiv \frac{\exp(\mu_1+\mu_2+\lambda)}{Z(\mu_1,\mu_2,\lambda)}, \end{aligned} \tag{9}$$

which are coordinate transformations with the inverse

$$\begin{gathered} \mu_1 = \ln\frac{m_1-l}{1+l-m_1-m_2}, \qquad \mu_2 = \ln\frac{m_2-l}{1+l-m_1-m_2},\\ \lambda = \ln\frac{(1+l-m_1-m_2)\,l}{(m_1-l)\,(m_2-l)}. \end{gathered} \tag{10}$$

In terms of the coordinates $\boldsymbol{t}$ the family of probability distributions for $\boldsymbol{s}$ has the form

$$\begin{aligned} &\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{t}, I) = (1+l-m_1-m_2)\times\\ &\quad\left(\frac{m_1-l}{1+l-m_1-m_2}\right)^{s_1}\left(\frac{m_2-l}{1+l-m_1-m_2}\right)^{s_2}\left[\frac{(1+l-m_1-m_2)\,l}{(m_1-l)\,(m_2-l)}\right]^{s_1 s_2}\\ &\quad\equiv l^{s_1 s_2}\,(m_1-l)^{s_1-s_1 s_2}\,(m_2-l)^{s_2-s_1 s_2}\,(1+l-m_1-m_2)^{1-s_1-s_2+s_1 s_2}. \end{aligned} \tag{11}$$