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 27 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,

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

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:

$$\begin{aligned}
\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} \,|\, I) &= \int \left[\prod_{i=1}^{N} g(\boldsymbol{s}^{(i)})\, \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \lambda s_1^{(i)} s_2^{(i)}\bigr)}{Z(\mu_1, \mu_2, \lambda)}\right] \times \\
&\qquad\qquad \mathrm{p}(\mu_1, \mu_2, \lambda \,|\, I)\, \mathrm{d}\mu_1\, \mathrm{d}\mu_2\, \mathrm{d}\lambda, \\
&= \int [\textstyle\prod_{i=1}^{N} g(\boldsymbol{s}^{(i)})]\, \dfrac{\exp[N\,(\mu_1 \bar{s}_1 + \mu_2 \bar{s}_2 + \lambda \overline{s s})]}{Z(\mu_1, \mu_2, \lambda)^N} \times \\
&\qquad\qquad \mathrm{p}(\mu_1, \mu_2, \lambda \,|\, I)\, \mathrm{d}\mu_1\, \mathrm{d}\mu_2\, \mathrm{d}\lambda,
\end{aligned}$$

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

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

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 $\mathrm{p}(\boldsymbol{\theta} \,|\, I)$ is called *prior parameter density*.

Formula (2) may appear deceivingly specific in its dependence on the parameters $\boldsymbol{\theta}$; 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:

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

(*b*) each joint probability in this convex combination factorizes into N independent probabilities for the N observations, as is clear from the formula above;

(*c*) each joint probability in the convex combination is identified by a triplet of parameters $\boldsymbol{\theta}$; 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 $\boldsymbol{\theta}$ is $\mathrm{p}(\boldsymbol{\theta} \,|\, I)\, \mathrm{d}\boldsymbol{\theta}$.

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 theorem selects a particular three-dimensional submanifold within each of the $(2^N - 1)$-dimensional manifolds of probability distributions for N observations, for all N. But the theorem doesn't select any particular coordinate system within the submanifold: the parameters $\boldsymbol{\theta}$ are just coordinates, and there is nothing special about them, 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

$$\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} \mid \boldsymbol{t}, I) = \prod_{i=1}^{N} \frac{\exp\bigl[\mu_1(\boldsymbol{t})\, s_1^{(i)} + \mu_2(\boldsymbol{t})\, s_2^{(i)} + \lambda(\boldsymbol{t})\, s_1^{(i)} s_2^{(i)}\bigr]}{Z[\boldsymbol{\theta}(\boldsymbol{t})]}; \tag{4}$$

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

$$\mathrm{p}(\boldsymbol{\theta} \mid I) = \mathrm{p}[\boldsymbol{t}(\boldsymbol{\theta}) \mid I]\, \det\Bigl(\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{\theta}}\Bigr). \tag{5}$$

This coordinate change will be central in the rest of this study.

2 New coordinates and their motivation

Assuming that (1) are sufficient statistics and therefore using formula (2), 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, $\mathrm{p}(\bar{s}, \overline{ss} \mid I, \text{large } N)$.

In this section we show that there is a particular coordinate system $\boldsymbol{t} \coloneqq (m_1, m_2, l)$ 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*:

$$\mathrm{p}[(\bar{s}_1, \bar{s}_2, \overline{ss}) = \boldsymbol{x} \mid I, \text{large } N] \approx \mathrm{p}(\boldsymbol{t} = \boldsymbol{x} \mid I). \tag{6}$$

To see this, consider the parameterized, factorized joint probability $\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} \mid \boldsymbol{\theta}, I)$ of eq. (12). The expectation of the statistics

$(\bar{s}_1, \bar{s}_2, \overline{ss})$ is given by

$$\mathrm{E}[(\bar{s}_1, \bar{s}_2, \overline{ss})|\,\boldsymbol{\theta}, I] = \\ \tfrac{1}{N}\sum_i \mathrm{E}[(s_1^{(i)}, s_2^{(i)}, s_1^{(i)}s_2^{(i)})|\,\boldsymbol{\theta}, I] = \mathrm{E}[(s_1, s_2, s_1 s_2)|\,\boldsymbol{\theta}, I] \tag{7}$$

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 $\boldsymbol{\theta}$. From the properties of the variance we also have

$$\mathrm{V}[(\bar{s}_1, \bar{s}_2, \overline{ss})|\,\boldsymbol{\theta}, I] = \frac{1}{N}\mathrm{V}[(s_1, s_2, s_1 s_2)|\,\boldsymbol{\theta}, I]. \tag{8}$$

This means that for a triplet $\boldsymbol{\theta}$, for large N we have a probability distribution for the statistics that is very peaked at particular values $\boldsymbol{t} := (m_1, m_2, l)$ determined by the equations

$$m_1 = \mathrm{E}(s_1|\,\boldsymbol{\theta}, I) \equiv \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial \mu_1}, \qquad m_2 = \mathrm{E}(s_2|\,\boldsymbol{\theta}, I) \equiv \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial \mu_2}, \\ l = \mathrm{E}(s_1 s_2|\,\boldsymbol{\theta}, I) \equiv \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial \lambda}. \tag{9}$$

This system of equations actually puts the parameters $\boldsymbol{\theta} := (\mu_1, \mu_2, \lambda)$ and $\boldsymbol{t} := (m_1, m_2, l)$ into one-one correspondence (Mead et al. 1984). The former belong to $\mathbf{R}^3$; the latter to the bounded domain

$$0 \leqslant m_1, m_2 \leqslant 1, \qquad \max(0, 1 - m_1 - m_2) \leqslant l \leqslant \min(m_1, m_2). \tag{10}$$

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

$$\mathrm{p}(\bar{s}_1, \bar{s}_2, \overline{ss}|\, m_1, m_2, l, I) \approx \delta[(\bar{s}_1, \bar{s}_2, \overline{ss}) - (m_1, m_2, l)]. \tag{11}$$

Taking the convex combination of this expression in $\boldsymbol{t}$ with weights $\mathrm{p}(\boldsymbol{t}|\, I)\,\mathrm{d}\boldsymbol{t}$ we obtain eq. (6).

In the coordinates $\boldsymbol{t}$, the probabilities (2) 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 $\boldsymbol{t} := (m_1, m_2, l)$. Given this knowledge we can combinatorially calculate the probability of observing a finite sequence of

N observations $(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)})$, assuming that all sequences having given statistics are equally likely – this equiprobability corresponds to setting $g(\boldsymbol{s}) = 1$ in eq. (2). An example of this combinatorial calculation is given below.

(2) We then express about the limit statistics with the density $\mathrm{p}(\boldsymbol{t}|\, I)\,\mathrm{d}\boldsymbol{t}$.

(3) The two uncertainties above are finally combined in the usual way.

Let's show, in the simplest case, that the probability conditional on $\boldsymbol{t}$,

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

is indeed given combinatorially assuming equiprobability of all sequences, as claimed above. Suppose that we know the limit statistics are (m_1, m_2, l). Only the outcome $\boldsymbol{s} = 11$ gives a non-vanishing contribution to the second moment l, 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 l. 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 l, the frequency of 10 must be given by $m_1 - l$, 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 $1 - l - (m_1 - 1) - (m_2 - l)$. Summarizing,

$$\begin{aligned}\mathrm{p}(\boldsymbol{s}|\, \boldsymbol{t}, I) &= \begin{cases} 1 + l - m_1 - m_2, & \text{for } \boldsymbol{s} = 00 \\ m_1 - l, & \text{for } \boldsymbol{s} = 10 \\ m_2 - l, & \text{for } \boldsymbol{s} = 01 \\ l, & \text{for } \boldsymbol{s} = 11 \end{cases} \\ &\equiv l^{s_1 s_2}\, (m_1 - l)^{s_1 - s_1 s_2}\, (m_2 - l)^{s_2 - s_1 s_2}\, (1 + l - m_1 - m_2)^{1 - s_1 - s_2 + s_1 s_2}. \end{aligned} \tag{13}$$

This probability distribution is exactly eq. (12), as can be checked by finding $\boldsymbol{\theta}(\boldsymbol{t})$ with the inverse of the coordinate transformations (19),

$$\begin{gathered}\mu_1 = \ln \frac{m_1 - l}{1 + l - m_1 - m_2}, \qquad \mu_2 = \ln \frac{m_2 - l}{1 + l - m_1 - m_2}, \\ \lambda = \ln \frac{(1 + l - m_1 - m_2)\, l}{(m_1 - l)\, (m_2 - l)}, \end{gathered} \tag{14}$$

and substituting it in the right side of eq. (12).

✠ continue

Denote by $K(\bar{\boldsymbol{s}}, \overline{ss}; N)$ the number of possible sequences of N observations $(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)})$ that have statistics $(\bar{\boldsymbol{s}}, \overline{ss})$. Note that $\sum_{\bar{\boldsymbol{s}}, \overline{ss}} K(\bar{\boldsymbol{s}}, \overline{ss}; N) = 4^N$, the total number of possible sequences of N observations. For example, if $N = 4$, there are 12 possible sequences that have statistics $\bar{s}_1 = 2/4$, $\bar{s}_2 = 1/4$, $\overline{ss} = 1/4$: the sequence $(00, 00, 10, 11)$ and its $4!/2$ distinct permutations; therefore $K(2/4, 1/4, 1/4; 4) = 12$.

, noting that it can be written

$$\mathrm{p}(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)} | \, \boldsymbol{\theta}, I) = \frac{\exp[N\,(\mu_1 \bar{s}_1 + \mu_2 \bar{s}_2 + \lambda \overline{ss})]}{Z(\boldsymbol{\theta})^N}. \tag{15}$$

This probability has the same value for all sequences $(\boldsymbol{s}^{(1)}, \dots, \boldsymbol{s}^{(N)})$ having the same statistics $(\bar{\boldsymbol{s}}, \overline{ss})$. Therefore, the probability of observing a particular statistics, given $\boldsymbol{\theta}$, is

$$\mathrm{p}(\bar{\boldsymbol{s}}, \overline{ss} | \, \boldsymbol{\theta}, I) = K(\bar{\boldsymbol{s}}, \overline{ss}; N)\, \frac{\exp[N\,(\mu_1 \bar{s}_1 + \mu_2 \bar{s}_2 + \lambda \overline{ss})]}{Z(\boldsymbol{\theta})^N}, \tag{16}$$

and considering the convex combination we have

$$\mathrm{p}(\bar{\boldsymbol{s}}, \overline{ss} | \, I) = K(\bar{\boldsymbol{s}}, \overline{ss}; N) \int \frac{\exp[N\,(\mu_1 \bar{s}_1 + \mu_2 \bar{s}_2 + \lambda \overline{ss})]}{Z(\boldsymbol{\theta})^N}\, \mathrm{p}(\boldsymbol{\theta} | \, I)\, \mathrm{d}\boldsymbol{\theta}. \tag{17}$$

When N is large, the exponential term with the largest argument dominates as $\boldsymbol{\theta}$ varies. The parameter values $\boldsymbol{\theta}_{\mathrm{m}}$ of this term are therefore

$$\boldsymbol{\theta}_{\mathrm{m}} := \arg\sup_{\boldsymbol{\theta}} [\mu_1 \bar{s}_1 + \mu_2 \bar{s}_2 + \lambda \overline{ss} - \ln Z(\boldsymbol{\theta})], \tag{18}$$

and they can be shown to be unique (Mead et al. 1984); in fact they are implicitly given by the equations (Mead et al. 1984; Porta Mana 2017)

$$\begin{gathered} m_1 = \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial \mu_1} \equiv \mathrm{E}(s_1 | \, \boldsymbol{\theta}, I), \qquad m_2 = \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial \mu_2} \equiv \mathrm{E}(s_2 | \, \boldsymbol{\theta}, I), \\ l = \frac{\partial \ln Z(\boldsymbol{\theta})}{\partial \lambda} \equiv \mathrm{E}(s_1 s_2 | \, \boldsymbol{\theta}, I). \end{gathered} \tag{19}$$

$$\text{for} \qquad (m_1, m_2, l) = (\bar{s}_1, \bar{s}_2, \overline{ss}), \quad \boldsymbol{\theta} = \boldsymbol{\theta}_{\mathrm{m}}. \tag{20}$$

The system of equations (19) gives a one-one correspondence between the parameters $\boldsymbol{t} := (m_1, m_2, l)$ and $\boldsymbol{\theta} := (\mu_1, \mu_2, \lambda)$. The latter belong to $\mathbf{R}^3$; the former to the bounded domain

$$0 \leqslant m_1, m_2 \leqslant 1, \qquad \max(0, 1 - m_1 - m_2) \leqslant l \leqslant \min(m_1, m_2). \tag{21}$$

3 Inference from observed data

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

$$\mathrm{p}(s \mid s^{(1)}, \dots, s^{(N)}, I) = \int \frac{\exp\bigl(\mu_1 s_1 + \mu_2 s_2 + \lambda s_1 s_2\bigr)}{Z(\theta)}\, \mathrm{p}(\theta \mid s^{(1)}, \dots, s^{(N)} I)\, \mathrm{d}\theta \tag{22a}$$

with

$$\mathrm{p}(\theta \mid s^{(1)}, \dots, s^{(N)} I) \propto \left[\prod_{i=1}^{N} \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \lambda s_1^{(i)} s_2^{(i)}\bigr)}{Z(\theta)}\right] \mathrm{p}(\theta \mid I) = \exp\{N\,[\mu_1 a_1 + \mu_2 a_2 + \lambda \overline{ss} - \ln Z(\theta)]\}\, \mathrm{p}(\theta \mid I). \tag{22b}$$

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

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

The proportionality relation of the last formula reminds us that we must perform an integral over θ to calculate the posterior parameter density. We must also perform an integral over θ to calculate the conditional probability for 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,

$$\theta_{\mathrm{m}} := \arg\sup_{\theta} \{N\,[\mu_1 a_1 + \mu_2 a_2 + \lambda \overline{ss} - \ln Z(\theta)] + \ln \mathrm{p}(\theta \mid I)\}. \tag{23}$$

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

4 Other prior parameter densities

The literature often assumes a prior parameter density $\mathrm{p}(\theta \mid I)$ 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 $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}$ gives a distribution

$$\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{\theta}, I) = \frac{\exp\bigl(\mu_1 s_1^{(i)} + \mu_2 s_2^{(i)} + \lambda s_1^{(i)} s_2^{(i)}\bigr)}{Z(\boldsymbol{\theta})} \tag{24}$$

with different first and second moments

$$\begin{aligned} \mathrm{E}(\boldsymbol{s}|\,\boldsymbol{\theta}, I) &\coloneqq \sum_{\boldsymbol{s}} \boldsymbol{s}\,\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{\theta}, I), \\ \mathrm{E}(s_1 s_2|\,\boldsymbol{\theta}, I) &\coloneqq \sum_{\boldsymbol{s}} s_1 s_2\,\mathrm{p}(\boldsymbol{s}|\,\boldsymbol{\theta}, I). \end{aligned} \tag{25}$$

In other words the set $\boldsymbol{\theta} \in \mathbf{R}^3$ 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

$$\begin{aligned} m_1 &= \mathrm{E}(s_1|\,\boldsymbol{\theta}, I) \equiv \frac{\exp(\mu_1) + \exp(\mu_1 + \mu_2 + \lambda)}{Z(\mu_1, \mu_2, \lambda)}, \\ m_2 &= \mathrm{E}(s_2|\,\boldsymbol{\theta}, I) \equiv \frac{\exp(\mu_2) + \exp(\mu_1 + \mu_2 + \lambda)}{Z(\mu_1, \mu_2, \lambda)}, \\ l &= \mathrm{E}(s_1 s_2|\,\boldsymbol{\theta}, I) \equiv \frac{\exp(\mu_1 + \mu_2 + \lambda)}{Z(\mu_1, \mu_2, \lambda)}, \end{aligned} \tag{26}$$

which are coordinate transformations with the inverse

$$\begin{gathered} \mu_1 = \ln \frac{m_1 - l}{1 + l - m_1 - m_2}, \qquad \mu_2 = \ln \frac{m_2 - l}{1 + l - m_1 - m_2}, \\ \lambda = \ln \frac{(1 + l - m_1 - m_2)\, l}{(m_1 - l)\,(m_2 - l)}. \end{gathered} \tag{27}$$

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

$$
\begin{aligned}
\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{t}, I) = {} & (1 + l - m_1 - m_2) \times \\
& \left(\frac{m_1 - l}{1 + l - m_1 - m_2}\right)^{s_1} \left(\frac{m_2 - l}{1 + l - m_1 - m_2}\right)^{s_2} \left[\frac{(1 + l - m_1 - m_2)\, l}{(m_1 - l)\,(m_2 - l)}\right]^{s_1 s_2} \\
& \equiv l^{s_1 s_2}\,(m_1 - l)^{s_1 - s_1 s_2}\,(m_2 - l)^{s_2 - s_1 s_2}\,(1 + l - m_1 - m_2)^{1 - s_1 - s_2 + s_1 s_2}.
\end{aligned}
\tag{28}
$$

Bibliography

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

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.

Porta Mana, P. G. L. (2017): *Geometry of maximum-entropy proofs: stationary points, convexity, Legendre transforms, exponential families*. doi:10.17605/osf.io/vsq5n, arXiv:1707.00624.