Yet Another Analysis of Dice Problems [1]

Ali Mohammad-Djafari

Laboratoire des Signaux et Systèmes,
Unité mixte de recherche 8506 (CNRS-Supélec-UPS)
Supélec, Plateau de Moulon, 91192 Gif-sur-Yvette, France

Abstract. During the MaxEnt 2002 workshop in Moscow, Idaho, Tony Vignaux asked again a few simple questions about using Maximum Entropy or Bayesian approaches for the famous Dice problems which have been analyzed many times through this workshop and also in other places. Here, there is another analysis of these problems. I hope that, this paper will answer a few questions of Tony and other participants of the workshop on the situations where we can use Maximum Entropy or Bayesian approaches or even the cases where we can actually use both of them.

INTRODUCTION

Dice problems have been analyzed many times (See mainly Ed. Jaynes papers [1, 2, 3, 4] and also [5, 6, 7, 8, 9]), but it seems that still many questions are open. In this note, I will try to answer some of them. Before starting, we need to set up precise notations and describe precisely the context. Let consider an imaginary die with K faces ($K = 6$ is the ordinary die), where on each face there is a number. We note these numbers $\boldsymbol{g} = [g_1, \ldots, g_{K'}]$. K is the number of elementary states and commonly, $K' = K$ and $g_k = k$, but we may also consider the cases where g_k are any other numbers (integer or real) distinct or not. Let represent by X the variable corresponding to face number and by G the variable corresponding to the number written on the faces. So, X may take values $\{1, \ldots, K\}$ and G can take values $\{g_1, \ldots, g_{K'}\}$. Then, we can define $P(X = k)$ and $P(G = g_k)$. If the g_k are distinct numbers, *i.e.*,, $K = K'$, they are equal $P(X = k) = P(G = g_k) = \theta_k$, but note that $\mathrm{E}\{X\} = \sum_k k\theta_k \neq \mathrm{E}\{G\} = \sum_k g_k\theta_k$. If g_k is a monotone function of k, then it is easy to relate $\mathrm{E}\{X\}$ to $\mathrm{E}\{G\}$, but it may not always be the case.

Note also that, in many dice problems, the main hypothesis is that they are fair. Then assigning the probability distributions become a combinatorial computation. For example, suppose we throw two dice and count the sums S of the two faces numbers. We want to assign the probabilities $p_j = P(S = s_j)$. First, we assume $g_k = k$ and note that S can take the values in the set $\Omega = \{2, 3, \ldots, 12\}$ and $|Q| = 11$. We must be careful here because the event $S = s_j$ can occur $q(s_j) = 6 - |7 - s_j|$ times. For example, $S = 2$ occurs one time $E_j = \{(1, 1)\}$, but $S = 5$ occurs 5 time $E_j = \{(1, 4), (2, 3), (3, 2), (4, 1)\}$. Now, using the basic principle of *equal weight* of statistical mechanics or *insufficient reason* of Laplace, we assign $p_j = P(S = s_j) \propto |E_j|$ which gives $p_j = P(S = s_j) = q(s_j)/\sum_{j=1}^{|Q|} q(s_j)$.

---

[1] Draft version of September 18, 2002. All comments and corrections will be welcome.

In a more general case, we may have L dice and may want to define the events such that $E_j = \{(X_1 = x_1, \ldots, X_L = x_L)\}$ or $E_j = \{(X_1 = x_1, \ldots, X_L = x_L) \ : \ \sum_l x_l = s_j\}$ and assign them probabilities. We may also consider the case where we through L dice simultaneously N times which is not the same as throwing N dice simultaneously L times, except the case where the dice are identical. We may also consider the cases where the number of throwing the dice are different, *i.e.,*, the dice l has been thrown N_l times.

In some other analysis, we may not know if the die is loaded or not. This may be one of the question to be answered. To be able to answer to a question, we may need to gather relevant data. These data may be of different form and thus, as we will see in the following, the way to use them to answer to a question may also differ.

Before gathering any data, we may define the *question* to be answered. For example, if we want to know if the die is loaded or not, we may be interested to infer about $\boldsymbol{\theta}$. Also, before gathering any data, we may make hypotheses and we may be able to translate the knowledge contained in these hypotheses by an *a priori* probability law $\pi(\boldsymbol{\theta})$. For example, we may assume that the die is not loaded and assume $\theta_1 = \theta_2 = \ldots = \theta_K = 1/K$ or choose a uniform prior for $\pi(\boldsymbol{\theta})$ over the set $\{\boldsymbol{\theta} \ : \ \theta_k \in [0,1] \& \sum_k \theta_k = 1\}$. Note that, even if they translate the same commonly speaking hypothesis, mathematically speaking, they are not exactly the same. The former says $P(\sum_k \theta_k \neq 1) = 0$ and $P(\theta_k \neq \theta_l, k \neq l) = 0$ and $P(a < \theta_k \leq b) = (b-a), \ \forall 1 > b > a > 0$.

We may also be able to associate a likelihood function $P(D|\boldsymbol{\theta})$ to the data to represent the amount of knowledge about the unknown parameters contained in the data. We will see however that this may not be easy in some cases.

The questions may also be different: We may want to know if the die is loaded or not or we may want to know what is the probability that the next face be the face k, or still, what are the numbers written on the faces of the die.

Let start by a simple and easy problem which, here after, we call Problem 1.

PROBLEM 1

We have observed the complete data $\boldsymbol{x} = [x_1, \ldots, x_N]$ and we know the number of states K (number of faces). The question is to estimate $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$ where $\theta_k = P(X = k)$ is the probability of the event face k up.

Here is a Matlab program which simulates this data generation:

```
K=6;N=100;x=round((K-1)*rand(N,1))+1;
```

and the following is an example (an N sample) of this data set:

$\boldsymbol{x} = [4,2,2,2,1,5,4,5,1,4,3,3,6,6,4,6,6,4,4,1,1,2,1,6,4,2,4,2,3,2,2,6,2,2,1,6,$
$5,5,6,3,5,4,2,2,4,4,4,3,6,6,4,5,2,5,3,5,2,5,1,3,3,4,3,1,3,3,5,3,3,2,5,5,3,4,4,$
$3,3,3,4,1,2,4,4,5,4,5,6,5,6,5,5,5,1,1,4,1,5,2,1,6]$.

Note that, if we re-run again the program, we obtain a different data set. Here is

the results of a second run:

$\boldsymbol{x} = [6,2,4,3,5,5,3,1,5,3,4,5,6,5,2,3,6,6,3,5,1,3,5,1,2,2,2,4,2,2,1,5,3,6,3,3,$
$5,4,2,4,5,1,4,3,5,4,5,3,3,2,2,4,3,4,2,4,3,5,5,4,3,5,5,4,5,4,3,2,3,4,5,3,5,4,3,$
$5,4,3,4,4,5,6,4,5,2,6,2,2,5,5,2,1,5,2,2,4,2,3,1,6]$.

These two data sets can represent two different experiences using the same die.

Let note by n_k the number of times the face k has been shown up $n_k = \#(X = k)$. Then we have $\sum_k n_k = N$. Here is a Matlab program which computes these numbers:

```
nk=zeros(K,1);
for k=1:6
   nk(k)=sum(x==k);
end
```

and here are the results for the two above data sets:

Data set 1: $\boldsymbol{n} = [13,17,17,21,19,13]$ and
Data set 2: $\boldsymbol{n} = [07,19,21,20,25,08]$.

Now, let start by asking about the values of θ_k. If all the θ_k are the same value, we can say that the die is not loaded, but if they are too different from each other, we may say that the die is loaded.

A wise man can say: This is an easy problem. If each trial has been done identically and independently, then it is reasonable to *estimate* each θ_k by $\theta_k = n_k/N$ and no need for more complex mathematics. But if we ask: How *confident* or (how sure) are you about these values? He may say: hum..., let use the probability theory.

Let assume to know K and we have given $\boldsymbol{x}$ (and thus we now N) and assume that the die has been thrown always in the same manner and independently. Here then, we can write the complete likelihood function

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_k C_N^{n_k}\, \theta_k^{n_k}\, (1-\theta_k)^{N-n_k}. \tag{1}$$

Note that in the right hand side of this expression, $\boldsymbol{x}$ is present through n_k and we can write $P(\boldsymbol{x}|\boldsymbol{\theta}) = P(\boldsymbol{n}|\boldsymbol{\theta})$.

Then the likelihood $\mathcal{L}(\boldsymbol{\theta}) = P(\boldsymbol{x}|\boldsymbol{\theta})$ and we have

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \sum_k [n_k \ln \theta_k + (N - n_k) \ln(1-\theta_k)] + c \tag{2}$$

where $c = \sum_k \ln C_N^{n_k}$ does not depend on $\boldsymbol{\theta}$.

Knowing that each parameter $\theta_k \in [0,1]$, we can choose a uniform prior $\pi(\theta_k) = 1$ on this interval. However, we know that $\sum_k \theta_k = 1$, then we can define the set $\Theta = \{\boldsymbol{\theta} : \theta_k \in [0,1] \& \sum_k \theta_k = 1\}$ and thus define a uniform prior on this set $\pi(\boldsymbol{\theta}) = 1,\ \forall \boldsymbol{\theta} \in \Theta$ and zero

elsewhere, and thus obtain the *a posteriori* law

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{1}{m(\boldsymbol{x})}\prod_k C_N^{n_k}\,\theta_k^{n_k}\,(1-\theta_k)^{N-n_k}. \tag{3}$$

which is defined on the same set Θ and where $m(\boldsymbol{x})$ is the marginal or evidence function:

$$m(\boldsymbol{x}) = \int_\Theta \mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \int_\Theta \mathrm{d}\boldsymbol{\theta}\prod_k C_N^{n_k}\,\theta_k^{n_k}\,(1-\theta_k)^{N-n_k}. \tag{4}$$

Thus, we have

$$\ln\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \sum_k \left[n_k \ln\theta_k + (N-n_k)\ln(1-\theta_k)\right] - \ln m(\boldsymbol{x}) \tag{5}$$

Now, if we are only interested by the value of $\widehat{\boldsymbol{\theta}}^{MAP}$ which has the highest probability, we can compute it by putting the derivative of $\ln\pi(\boldsymbol{\theta}|\boldsymbol{x})$ with respect to each parameter θ_k to obtain

$$\partial\ln\pi(\boldsymbol{\theta}|\boldsymbol{x})/\partial\theta_k = \frac{n_k}{\theta_k} - \frac{N-n_k}{1-\theta_k} = \frac{n_k - N\theta_k}{\theta_k(1-\theta_k)} = 0 \longrightarrow \widehat{\theta}_k^{MAP} = \frac{n_k}{N}. \tag{6}$$

There is only one possible solution to this equation and there is not any ambiguity. Here is the results for the two above data sets:

Data set 1: $\widehat{\boldsymbol{\theta}}^{MAP} = [0.1300, 0.1700, 0.1700, 0.2100, 0.1900, 0.1300]$ and
Data set 2: $\widehat{\boldsymbol{\theta}}^{MAP} = [0.0700, 0.1900, 0.2100, 0.2000, 0.2500, 0.0800]$.

But, we must be careful here on the interpretations that we can give to these numerical values. We may want to answer the following questions:

- Do these two data sets come from the same die?
- Is this die loaded?
- What is the probability of seeing face k up based on the data set 1 or the data set 2?
- If I throw this die 100 times again, what will be the number of times I will see face k up?

We have still too much to do before being able to give correct answers to these questions.

PROBLEM 2

Assume now that, in place of $\boldsymbol{x}$, we have only access to the data $\boldsymbol{n} = (n_1, \ldots, n_K)$ and know the values of K and N (or if we knew that $\sum_k n_k = N$). It is easy to see that we obtain exactly the same result, because $(\boldsymbol{n}, K, \sum_{k=1}^K n_k = N)$ define perfectly the likelihood and form sufficient statistics about this problem.

Note however that, in both cases, the likelihood $\mathcal{L}(\boldsymbol{\theta})$ is not defined for $\theta_k = 0$ and $\theta_k = 1$ and consequently, the posterior pdf $\pi(\boldsymbol{\theta}|\boldsymbol{n})$ may not be a proper pdf. We are going to analyze properly this point.

First, noting that the likelihood function in previous section $\mathcal{L}(\boldsymbol{\theta}) = \prod_k l(\theta_k)$ and $\pi(\boldsymbol{\theta}) = \prod_k \pi(\theta_k)$, we also have $\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_k \pi(\theta_k|\boldsymbol{x}) = \prod_k \pi(\theta_k|n_k)$. Thus, we can work hereafter only with the functions $l(\theta)$, $\pi(\theta)$, $\pi(\theta|x)$ and $m(x)$ which is given by [10, 11]

$$m(x) = \int_0^1 \pi(\theta)\, C_N^x\, \theta^x\, (1-\theta)^{(N-x)}\, \mathrm{d}\theta. \tag{7}$$

With a uniform prior $\pi(\theta) = 1$ we have

$$m(x) = \int_0^1 C_N^x\, \theta^x\, (1-\theta)^{(N-x)}\, \mathrm{d}\theta = \mathcal{B}(x, N-x) \tag{8}$$

where $\mathcal{B}(\alpha, \beta)$ is the Beta probability density function (pdf)

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{(\beta-1)} \tag{9}$$

which is defined for $\alpha > 0$, $\beta > 0$ and $x \in [0,1]$ and where

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{(\beta-1)}\, \mathrm{d}x \tag{10}$$

and we have:

$$\text{mode}\{x\} = \frac{\alpha-1}{\alpha+\beta-2}, \quad \text{E}\{x\} = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \text{Var}\{x\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Consequently, the posterior law, whose expression is $\pi(\boldsymbol{\theta}|\boldsymbol{n}) = \mathcal{B}(\boldsymbol{n}-1, N-1-\boldsymbol{n})$ or equivalently $\pi(\theta_k|n_k) = \mathcal{B}(n_k - 1, N-1-n_k)$, is only bounded if $N-1 > n_k > 1$. The MAP estimators $\widehat{\theta}_k^{MAP} = \frac{n_k-2}{N-4}$ do not exist if $n_k \leq 1$ or if $n_k \geq N-1$ and if $N \leq 4$. The posterior mean estimators $\widehat{\theta}_k^{PM} = \frac{n_k-1}{N-2}$ exist if $N > 2$ and the posterior variances $\text{Var}\{\theta\}_k = \frac{(n_k-1)(N-n_k-1)}{(N-2)^2(N-1)}$ exist if $N > 2$. Note also that when $n_k = 1$ the correspond mean estimator is $\theta_k = 0$ and when $n_k = N-2$ the correspond mean estimator is $\theta_k = 1$. This shows a kind of negative bias of the estimator toward $\theta_k = 0$ and $\theta_k = 1$ (See Table 1).

One may wants to have a proper posterior law $\pi(\boldsymbol{\theta}|\boldsymbol{n})$ for the whole range of possible values of the parameters $\theta_k \in [0,1]$ and the data $n_k = [0, 1, \ldots, N]$. This can be done via other choices for the prior law. In the two previous cases, we choose a uniform *a priori* for θ_k. Some authors argued that this choice is too biased against extreme values 0 and 1 and proposed to use

$$\pi(\theta_k) = [\theta_k(1-\theta_k)]^{-1} = \theta_k^{-1}(1-\theta_k)^{-1} \tag{11}$$

Note also that, again with this prior, the normalization factor or the evidence function

$m(x)$ is given by

$$m(x) = \int_0^1 [\theta(1-\theta)]^{-1} C_N^x \, \theta^x \, (1-\theta)^{(N-x)} \, \mathrm{d}\theta = \mathcal{B}(x, N-x) \tag{12}$$

which result to $\pi(\boldsymbol{\theta}|\boldsymbol{n}) = \mathcal{B}(\boldsymbol{n}, N-\boldsymbol{n})$ which is bounded if $N > n_k > 1$ (See Table 1).

To eliminate this problem, one can choose

$$\pi(\theta_k) = \theta_k^{a-1}(1-\theta_k)^{b-1} \tag{13}$$

which results to

$$m(x) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} C_N^x \, \theta^x \, (1-\theta)^{(N-x)} \, \mathrm{d}\theta = \mathcal{B}(x+a, N+b-x) \tag{14}$$

which result to $\pi(\boldsymbol{\theta}|\boldsymbol{n}) = \mathcal{B}(\boldsymbol{n}+a, N+b-\boldsymbol{n})$ which is bounded if $N-b > n_k > 1-a$. Then, the mean values $\theta_k = (n_k+a)/(N+b+a)$ have the limit value $\theta_k = n_k/N$ when $a = b \mapsto 0$. The following Table summarizes these points.

$\alpha > 0$	$\beta > 0$	$\alpha+\beta$	mode	mean	variance
$\pi(\theta_k) = 1, \quad \pi(\theta_k\|n_k) = \mathcal{B}(n_k-1, N-n_k-1)$					
$n_k - 1$	$N-n_k-1$	$N-2$	$\frac{n_k-2}{N-4}$	$\frac{n_k-1}{N-2}$	$\frac{(n_k-1)(N-n_k-1)}{(N-2)^2(N-1)}$
$n_k > 1$	$n_k < N-1$	$N > 2$	$n_k > 1$, $N > 4$	$n_k > 0$, $N > 2$	$n_k > 0$, $n_k < N$, $N > 2$
$\pi(\theta_k) = \theta_k^{-1}(1-\theta)^{-1}, \quad \pi(\theta_k\|n_k) = \mathcal{B}(n_k+1, N-n_k)$					
$n_k + 1$	$N-n_k$	$N+1$	$\frac{n_k}{N-1}$	$\frac{n_k+1}{N+1}$	$\frac{n_k+1)(N-n_k)}{(N+1)^2(N+2)}$
$n_k > 0$	$n_k < N$	$N > 0$	$n_k \geq 0$, $N > 1$	$n_k \geq 0$, $N \geq 0$	$n_k \geq 0$, $n_k \leq N$, $N \geq 0$
$\pi(\theta_k) = \theta_k^{a-1}(1-\theta)^{b-1}, \quad \pi(\theta_k\|n_k) = \mathcal{B}(n_k+a, N+b-n_k)$					
$n_k + a$	$N-n_k+b$	$N+a+b$	$\frac{n_k+a-1}{N+a+b-2}$	$\frac{n_k+a}{N+a+b}$	$\frac{(n_k+a)(N-n_k+b)}{(N+a+b)^2(N+a+b+1)}$
$n_k \geq 0$	$n_k \leq N$	$N \geq 0$	$n_k \geq 0$, $N > 0$	$n_k \geq 0$, $N \geq 0$	$n_k \geq 0$, $n_k \leq N$, $N \geq 0$

Note also that, when we have the expression of the *a posteriori* law $\pi(\boldsymbol{\theta}|\boldsymbol{n})$, we may define other estimators than the MAP or the posterior mean (PM). We may also answer the questions of type $P(a < \theta_k < b)$.

Note however that, all these computed numbers depend on the data and our prior knowledge we included. For any other data set we obtain other numbers. One may want to study the sensitivity of the solution to a kind of variability of data. This can be done by Monte Carlo simulations or by repeating the experience (but very often this may not be possible).

Also, in general the number of data or, more precisely, the contrast between the number of data and the number of parameters, is a crucial parameter. One may want to know the convergence of the solution to the hypothetical case where the number of data goes to infinity.

Now, let see if we can answer some of the questions at the end of the last section.

- What is the probability of seeing face k up based on the data set 1 or the data set 2? For each data set, we can give the following answers:
  - The most probable values of θ_k are $\theta_k^{MAP} = [xxxxxxxx]$, or
  - The mean values of θ_k are $\theta_k^{MP} = [xxxxxxxx]$, or
  - The variance values of θ_k are $v_k = [xxxxxxxx]$, or still
  - The lower values a_k and upper values b_k for which the probabilities $P(a_k < \theta_k < b_k) = .9$ are $a_k = [xxxxxxxx]$ and $b_k = [xxxxxxxx]$.
- Do these two data sets come from the same die?
  We can try to answer this question by comparing the probability laws $\pi_1(\theta_k|\boldsymbol{x}_1)$, $\pi_2(\theta_k|\boldsymbol{x}_2)$ and $\pi(\theta_k|\boldsymbol{x}_1, \boldsymbol{x}_2)$. But how to do this comparison? We may try to compute the relative entropy

$$KL(\pi_1\pi_2;\pi) = \int \pi_1(\theta_k|\boldsymbol{x}_1)\,\pi_2(\theta_k|\boldsymbol{x}_2)\ln\frac{\pi_1(\theta_k|\boldsymbol{x}_1)\,\pi_2(\theta_k|\boldsymbol{x}_2)}{\pi(\theta_k|\boldsymbol{x}_1,\boldsymbol{x}_2)}\,\mathrm{d}\theta_k \qquad (15)$$

  If this value is near to zero, this means that the two data sets comes from different dice.
- Is this die loaded?
  We can answer this question by computing the probabilities of two hypotheses $H_1 = (\theta_1 = \theta_2 = \ldots = \theta_K)$ and $H_0 = (\theta_k \neq \theta_l)$, *i.e.*, $P(H_1|\boldsymbol{x})$ and $P(H_0|\boldsymbol{x})$.

$$P(H_1|\boldsymbol{x}) = \prod_k \int \mathrm{d}\theta\pi(\theta_k = \theta|\boldsymbol{x}_1) \qquad (16)$$

$$P(H_0|\boldsymbol{x}) = \int \mathrm{d}\theta_1 \ldots \int \mathrm{d}\theta_K \prod_k \pi(\theta_k|\boldsymbol{x}_1) \qquad (17)$$

- If I throw this die $N' = 100$ times again, what will be the number of times I will see the face k up?
  To answer to this question, there are two methods:
  i) Use the data set $\boldsymbol{x} = \{\boldsymbol{n}, N, K\}$ to compute $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ and estimate $\widehat{\boldsymbol{\theta}}$ by one of the previous methods (MAP, PM, ...) and then compute $P(\boldsymbol{n}'|\widehat{\boldsymbol{\theta}}, N' = 100, K)$.

ii) Try to find the expression of $P(\boldsymbol{n}'|\boldsymbol{n}, N, K, N')$ by following

$$\begin{aligned}
P(\boldsymbol{n}|\boldsymbol{\theta}, N, K) &= \prod_k C_N^{n_k}\, \theta_k^{n_k}\, (1-\theta_k)^{N-n_k} \\
P(\boldsymbol{n}'|\boldsymbol{\theta}, N', K) &= \prod_k C_{N'}^{n'_k}\, \theta_k^{n'_k}\, (1-\theta_k)^{N'-n'_k} \\
P(\boldsymbol{n}, \boldsymbol{n}'|\boldsymbol{\theta}, N, K, N') &= \prod_k C_{N+N'}^{n_k+n'_k}\, \theta_k^{n_k+n'_k}\, (1-\theta_k)^{N+N'-n_k-n'_k} \\
P(\boldsymbol{n}'|\boldsymbol{n}, \boldsymbol{\theta}, N, K, N') &= P(\boldsymbol{n}, \boldsymbol{n}'|\boldsymbol{\theta}, N, K, N')/P(\boldsymbol{n}'|\boldsymbol{\theta}, N', K)
\end{aligned}$$

and then integrate out $\boldsymbol{\theta}$ to obtain $P(\boldsymbol{n}'|\boldsymbol{n}, N, K, N')$.

PROBLEM 3

Now, consider the case where, the observer has given to us only a subset $(n_1, \ldots, n_{K'})$ of the whole data $\boldsymbol{n} = (n_1, \ldots, n_K)$ with $K' < K$. (He just has forgotten to count and report the numbers $\{n_k, k = K'+1, \ldots, K\}$, but he is sure that the used die has K faces. In this case we can only obtain an expression for the likelihood function if we know the total number of the observations $N = \sum_{k=1}^{K} n_k \geq N' = \sum_{k=1}^{K'} n_k$ which is

$$P(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{k=1}^{K'} \theta_k^{n_k}\, (1-\theta_k)^{N-n_k}. \tag{18}$$

Note that this likelihood expression does not depend on the parameters $\{\theta_k, k = K'+1, \ldots, K\}$. Thus, the maximum likelihood (ML) estimation approach is unable to propose any values for them, while the Bayesian approach and in particular the MAP estimation can propose a solution which depends on the choice of *a priori*. For example, with a uniform prior, we have:

$$\theta_k = \begin{cases} \frac{n_k}{N} & k = 1, \ldots, K' \\ \frac{(N-N')}{(K-K')N} & k = K'+1, \ldots, K \end{cases} \tag{19}$$

where the first row is common with ML and the second row is due to the uniform prior and the normalization.

It is important to note that, while in the two previous cases, the prior law $\pi(\boldsymbol{\theta})$ has less important role, here the classical ML approach cannot give any answer to the problem and the role of prior information is crucial.

PROBLEM 4

Another interesting case is the one where we do not know the number of states (faces of the die). For example, we have observed the following data:

$\boldsymbol{x} = [4,2,2,2,1,*,4,*,1,4,3,3,*,*,4,*,*,4,4,1,1,2,1,*,4,2,4,2,3,2,2,*,2,2,1,$
$*,*,*,*,3,*,4,2,2,4,4,4,3,*,*,4,*,2,*,3,*,2,*,1,3,3,4,3,1,3,3,*,3,3,2,$
$*,*,3,4,4,3,3,3,4,1,2,4,4,*,4,*,*,*,*,*,*,*,1,1,4,1,*,2,1,*]$

where $*$ may mean *anything else greater than 4* or *do not know*. Note that these two cases are different. In the following, we first consider the first case which is, in fact, very close to the Problem 3 in previous section, because we know exactly the n_k for $k = 1, \ldots, K'$ but we do not know other n_k, $k > K'$ nor the the true value of $K > K'$ itself. However, N is given. We can only give an expression for the likelihood if we fix the value of K. Then, we can consider $K = 5, 6, 7, ...$ and for each case compute the results using (**??**):

For $K = 5$ we obtain: $\boldsymbol{\theta} = [0.1300, 0.1700, 0.1700, 0.2100, 0.03200]$
For $K = 6$ we obtain: $\boldsymbol{\theta} = [0.1300, 0.1700, 0.1700, 0.2100, 0.01600, 0.01600]$
For $K = 7$ we obtain: $\boldsymbol{\theta} = [0.1300, 0.1700, 0.1700, 0.2100, 0.01067, , 0.01067, 0.01067]$
and so on.

A difficult question remains: How to fixe K? We may try to compare $\pi(\boldsymbol{\theta}|\boldsymbol{x}, K)$ for different values of K through their entropies. We may also choose a prior for it and compute $\pi(\boldsymbol{\theta}, K|\boldsymbol{x})$ or still integrate out $\boldsymbol{\theta}$ to obtain $\pi(K|\boldsymbol{x})$ from which we can estimate K.

The case where, the $*$ in the data means *do not know* is more complex. If at least we know K, then it may still be possible to write the expression of the likelihood. Let note the true values of n_k by $N\nu_k$. Then, we know that $N\nu_k =\in [n_k, n_k + n_*],\ k = 1, \ldots, K'$ and $N\nu_k =\in [0, n_*],\ k = K', \ldots, K$ with $n_* = N - \sum_{k=1}^{K'} n_k$. Then, we may write

$$P(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\nu}, K, K') \;=\; \prod_{k=1}^{K'} C_{N-n_*}^{n_k}\, \theta_k^{n_k}\, (1-\theta_k)^{N-n_*-n_k} \prod_{k=1}^{K} C_{n_*}^{\nu_k}\, \theta_k^{N\nu_k}\, (1-\theta_k)^{n_*-N\nu_k}$$

or

$$P(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\nu}, K) \;=\; \prod_{k=1}^{K} C_{N}^{N\nu_k}\, \theta_k^{N\nu_k}\, (1-\theta_k)^{n_*-N\nu_k}$$

We can then try to integrate out $\boldsymbol{\theta}$ from this expression to obtain $P(\boldsymbol{x}|\boldsymbol{\nu}, K)$ or integrate out $\boldsymbol{\nu}$ to obtain $P(\boldsymbol{x}|\boldsymbol{\theta}, K)$. But, what to do if we do not know K? Can we also integrate out K by summing over all values of K?

Another question that may arise in this problem and the previous ones, is to estimate the frequencies $\nu_k = n_k/N$ which is not exactly the same question of estimating θ_k. In the following, we consider this problem.

First consider the case of complete data $\{\boldsymbol{n}, N, K\}$ of problems 1 and 2. We may note that, if we assume that the die is fair, the knowledge of the past experience ($\{\boldsymbol{n}, N, K\}$) does not change anything on the results of the future experience. But, if we do not know if the die is loaded, then from the past experience, we can estimate $\boldsymbol{\theta}$ and use it to compute the probability of observing any event.

The situation becomes more complexe if we do not know K or N or if some data are missing as is the case in problems 3 or 4, or more generally the cases where we cannot write easily the exact expression of the likelihood.

Let consider the incomplete data problem 4 where we know N, n_k and n_*, but we do not know K and assume that the $*$ are *a priori* distributed uniformly between 1 and K (or between K' and K) and compute the numbers $d_k = (n_k + n_*/K)/N$ (or $d_k = (n_k + n_*/(K - K'))/N$). We can then say that these computed d_k are good approximations to the true unobserved ν_k. The question is how to model this approximation. Two models can then be used:

i) Assume d_k as the mean values of the unknown frequencies ν_k

$$d_k = \mathrm{E}\{\nu_k\} = \int \nu_k p(\nu_k)\,\mathrm{d}\nu_k \tag{20}$$

or

ii) Assume each d_k to be the sum of the true ν_k and a random error ϵ_k:

$$d_k = \nu_k + \epsilon_k \tag{21}$$

where ϵ_k is assumed to be centered with unknown pdf. In both cases, we are interested by finding $p(\nu_k|d_k)$ or $p(\boldsymbol{\nu}|\boldsymbol{d})$.

But, before going further, it is important to note that, in the following, we are not going to analyze the original data $\boldsymbol{x}$ but the *pre-processed* data $\boldsymbol{d}$. We changed the problem to a new one: Given $\boldsymbol{d}$ can we assign or compute $p(\boldsymbol{\nu}|\boldsymbol{d})$.
Two approaches can then be used.

Information Theory or Maximum Entropy approach:
This approach is based on the first equation between d_k and ν_k. It is obvious that, there are infinite number of possible solution to this equation. Let note by $\mathcal{P}$ this ensemble:

$$\mathcal{P} = \{p \,:\, \mathrm{E}\{\nu_k\} = \int \nu_k p(\nu_k)\,\mathrm{d}\nu_k = d_k\} \tag{22}$$

The Maximum Entropy principle chooses the one $p^{ME}(\nu_k)$ with the highest entropy

$$p^{ME}(\nu_k) = \arg\max_{p\in\mathcal{P}} \{H(p)\} \tag{23}$$

where

$$H(p) = -\int p(x)\ln p(x)\,\mathrm{d}x, \tag{24}$$

or, more generally, if we assume to have a reference (prior?) distribution $q(\nu_k)$, the one $p^{MKL}(\nu_k)$ which has minimum Cross Entropy or Kullback-Leibler (KL) divergence [7, 12, 13], of p with respect to to q:

$$p^{MKL}(\nu_k) = \arg\min_{p\in\mathcal{P}} \{KL(p,q)\} \tag{25}$$

where

$$KL(p,q) = \int p(x)\ln(p(x)/q(x))\,\mathrm{d}x. \tag{26}$$

We note that when q is uniform $KL(p,q) = -H(p)$ and thus $p^{MKL}(\nu_k) = p^{ME}(\nu_k)$.

The unique solution, if exists, is given by

$$p^{MKL}(\nu_k) = \frac{1}{Z(\lambda_k)} q(\nu_k)\exp\{-\lambda_k\nu_k\} \tag{27}$$

where

$$Z(\lambda_k) = \int q(\nu_k)\exp\{-\lambda_k\nu_k\}\,\mathrm{d}\nu_k \tag{28}$$

and it can be shown that λ_k is the solution of the equation

$$-\partial \ln Z(\lambda_k)/\partial\lambda_k = d_k \tag{29}$$

which can be computed numerically. It is evident that the expressions of $p^{MKL}(\nu_k)$, $Z(\lambda_k)$, and consequently any numerical values for the estimate

$$\nu_k^{MKL} = \mathrm{E}\{\nu_k\} = \int \nu_k p^{MKL}(\nu_k)\,\mathrm{d}\nu_k \tag{30}$$

depend on the choice of q.

As a matter of algorithmic and computation of $\widehat{\boldsymbol{\lambda}}$ (solution of the equation (29)) and $\widehat{\boldsymbol{\nu}}$ defined in (30), it is interesting to know that they can be computed through:

$$\left\{ \begin{array}{ll} \widehat{\boldsymbol{\lambda}} & = \arg\min_{\boldsymbol{\lambda}} \left\{ D(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^t \boldsymbol{d} \right\} \\ \widehat{\boldsymbol{\nu}} & = \arg\min_{\boldsymbol{\nu}\in\mathcal{C}} \left\{ H(\boldsymbol{\nu}, \boldsymbol{\nu}^{(0)}) \right\} \end{array} \right. \tag{31}$$

where $D(\boldsymbol{\lambda})$ is called *dual criterion* and $H(\boldsymbol{\nu}, \boldsymbol{\nu}^{(0)})$ is called *primal criterion* and where $\nu_k^{(0)} = \mathrm{E}_q\{\nu_k\} = \int \nu_k q(\nu_k)\,\mathrm{d}\nu_k$.

The expressions of dual and primal criteria also depend on the expression of q. For example, when q is uniform on $\mathcal{C}$, p is exponential we have $Z(\boldsymbol{\lambda}) = \prod_k (1/\lambda_k)$, $\ln Z(\boldsymbol{\lambda}) = -\sum_k \ln\lambda_k$, $D(\boldsymbol{\lambda}) = -\sum_k \ln\lambda_k + \sum_k \lambda_k d_k$ and $P(\boldsymbol{\nu}, \boldsymbol{\nu}^{(0)}) = -\sum_k \ln(\nu_k/\nu_k^{(0)}) + \sum_k (\nu_k - \nu_k^{(0)})$. For other choices of q and more details on these relations refer to [14, 15, 16, 17, 18, 19, 20, 21, 22, 23].

Bayesian approach:

The Bayesian approach is based on the second equation, *i.e.*,, $d_k = \nu_k + \epsilon_k$ and we have to find an expression for the likelihood $\mathcal{L}(\boldsymbol{\nu}) = P(\boldsymbol{d}|\boldsymbol{\nu})$ and assign a prior $q(\nu_k)$ or

$q(\boldsymbol{\nu})$. When this done we can give an expression for the posterior $\pi^B(\boldsymbol{\nu}|\boldsymbol{d})$. Note that, in both cases, we have to choose $q(\boldsymbol{\nu})$. The first step which is to find an expression for $\mathcal{L}(\boldsymbol{\nu}) = P(\boldsymbol{d}|\boldsymbol{\nu})$ is not easy. Here are a few approaches:

Assuming $\boldsymbol{\theta} = \boldsymbol{\nu}$:
The first approach consists in assuming $\boldsymbol{\theta} = \boldsymbol{\nu}$. Then, if we are also given N, the problem becomes equivalent to the Problem 2 and we have:

$$P(\boldsymbol{d}|\boldsymbol{\nu}, N) = \prod_k C_N^{Nd_k} \nu_k^{Nd_k} (1-\nu_k)^{N(1-d_k)} \tag{32}$$

Then, again choosing a uniform prior $q(\boldsymbol{\nu}) = \frac{1}{Z_0}\delta(1 - \sum_k \nu_k)$, we obtain

$$\pi(\nu_k|\boldsymbol{d}, N) = \mathcal{B}(Nd_k - 1, N(1-d_k) - 1) \tag{33}$$

and then we have

$$\mathrm{E}\{\nu_k|\boldsymbol{d}, N\} = \frac{Nd_k - 1}{N-2}. \tag{34}$$

We see that $\mathrm{E}\{\nu_k|\boldsymbol{d}, N\} \mapsto d_k$ when N goes to infinity.

But, if we do not know N, we can try to integrate out N.

CAN WE DO IT EASILY ?

Frequentiste point of view:
Here, we assume *a priori* that the die is fair and try to obtain an expression for the likelihood $\mathcal{L}(\boldsymbol{d}|\boldsymbol{\nu}, N)$ using the following arguments:
Given N and K and assuming that each through of the die is independent of all others, we may argue on the number of possible outcomes resulting to a particular data set using the multinomial coefficient

$$W(\boldsymbol{n}, N, K) = \frac{N!}{n_1! \ldots n_k!} = \frac{N!}{\prod_{k=1}^K (n_k!)}. \tag{35}$$

$W(\boldsymbol{n}, N, K)$ is the number of possible outcomes $\boldsymbol{x}$ such that the face k appears n_k times between the total possible outcomes which is K^N. Thus, we may assign

$$P(\boldsymbol{n}|N, K) = W(\boldsymbol{n}, N, K)/(K^N) = \frac{N!}{(K^N)\prod_{k=1}^K (n_k!)}. \tag{36}$$

It is known that, using the Stirling approximation [2] the expression of this probability, when N is large, converges to

$$\lim_{N \mapsto \infty} \ln P(\boldsymbol{n}|N, K) = H(\boldsymbol{\nu}) = -\sum_{k=1}^K \nu_k \ln \nu_k \tag{37}$$

---

[2] Stirling (1692-1770) showed that $x_n = \frac{n! e^n}{n^{n+1/2}}$ converges to $\sqrt{2\pi n}$ when n goes to ∞. This means that, for large n we get the approximation $\ln(n!) = \frac{1}{2}\ln(2\pi n) + n \ln n$. However, even if this is usually called Stirling's formula, in fact, it may have been known earlier to Abraham de Moivre (see `http://www-gap.dcs.st-and.ac.uk/history/Mathematicians/De_Moivre.html`).

where $\nu_k = \lim_{N \mapsto \infty} \frac{n_k}{N}$.

This explanation and this approximation have also been used to justify the choice of an expression for entropy $H(\boldsymbol{\nu}) = -\sum_k \nu_k \ln \nu_k$ and a prior law for nb which is $\pi(\boldsymbol{n}) \propto \exp\{\alpha H(\nu)\}$, so that, given a set of constraints on ν_k, finding the most probable (sampling argument or maximum likelihood approach) value of $\boldsymbol{n}$ subject to those constraints become equivalent to maximizing $H(\boldsymbol{\nu})$ subject to those constraints:

$$\widehat{\boldsymbol{n}}_k = \arg\max_{\boldsymbol{n}} \{\ln \pi(\boldsymbol{n})\} = \arg\max_{\boldsymbol{n}} \{H(\boldsymbol{n})\} \tag{38}$$

But, we do not know either N or K. We may however try to use these expressions to find approximations to the likelihood function we need. First, we may assign

$$P(\boldsymbol{d}|\boldsymbol{\nu}, N, K) = P(N\boldsymbol{d}|N, K)(1 - P(N\boldsymbol{\nu}|N, K)) \tag{39}$$

and replacing for $P(N\boldsymbol{d}|N, K)$ and $P(N\boldsymbol{\nu}|N, K)$ and using again the Stirling formula we may find an expression which may be independent of N.

I COULD NOT GO FARTHER !

Integration of nuisance parameter $\boldsymbol{\theta}$:
Again here, we start by assuming N known. Then, we know the expressions of $P(\boldsymbol{d}|\boldsymbol{\theta}, N)$ and $P(\boldsymbol{\nu}|\boldsymbol{\theta}, N)$:

$$P(\boldsymbol{d}|\boldsymbol{\theta}, N) = \prod_k C_N^{Nd_k} \theta_k^{Nd_k} (1 - \theta_k)^{N(1-d_k)}. \tag{40}$$

$$P(\boldsymbol{\nu}|\boldsymbol{\theta}, N) = \prod_k C_N^{N\nu_k} \theta_k^{N\nu_k} (1 - \theta_k)^{N(1-\nu_k)}. \tag{41}$$

Then we can write

$$\begin{aligned} P(\boldsymbol{d}|\boldsymbol{\nu}, \boldsymbol{\theta}, N) &= (1 - P(\boldsymbol{\nu}|\boldsymbol{\theta}, N))\ P(\boldsymbol{d}|\boldsymbol{\theta}, N) \\ &= \left(1 - \prod_k C_N^{N\nu_k} \theta_k^{N\nu_k} (1 - \theta_k)^{N(1-\nu_k)}\right) \\ &\quad \times \prod_k C_N^{Nd_k} \theta_k^{Nd_k} (1 - \theta_k)^{N(1-d_k)}. \end{aligned} \tag{42}$$

Then, we have to integrate out $\boldsymbol{\theta}$ to obtain the likelihood $\mathcal{L}(\boldsymbol{\nu}) = P(\boldsymbol{d}|\boldsymbol{\nu}, N)$.

CAN WE OBTAIN SIMPLE EXPRESSIONS OR CAN WE INTEGRATE OUT N ?

Ad hoc empirical approach:
Another approach is to assign the two pdfs $p(\epsilon) = p(d_k - \nu_k)$ and the prior $q(\nu_k)$ from which we can compute

$$\pi^B(\nu_k|\,\mathrm{d}_k) = p(d_k - \nu_k)\, q(\nu_k)\ /\ m(d_k). \tag{43}$$

Here too, the expression of the posterior pdf $\pi(\nu_k|\,\mathrm{d}_k)$ and thus any inference about ν_k depends on the choice of $p(\epsilon)$ and $q(\nu_k)$.

A question may arise here:
Can we first fix $q(\nu_k)$ and compute $p^{MKL}(\nu_k)$ and use it again as a prior in this Bayesian approach?
The answer is "No", because $p^{MKL}(\nu_k)$ is in fact $p^{MKL}(\nu_k|d_k)$ and doing so, we have used two times the same data d_k.

Another question is how to compare and how to use $p^{MKL}(\nu_k|d_k)$ and $\pi^B(\nu_k|\,\mathrm{d}_k)$?
My answer is that π^B contains more information than that of p^{MKL}, because to obtain π^B, we combined information about both ϵ_k through $p(\epsilon)$ and ν_k through $q(\nu_k)$ while to obtain p^{MKL} we used only $q(\nu_k)$. Indeed, it seems that the only consistent point estimator of ν_k from p^{MKL} is its posterior mean, while, there is not any such restriction on π^B.

(NEEDS MORE CLARIFICATION)

PROBLEM 5

An important case is the one where we have only given the mean value of the faces numbers $\sum_k k\,\theta_k = d_0$ or the more general case of the mean value of the numbers written on the faces $\sum_k g_k\,\theta_k = d$ without any other knowledge and, in particular, without knowing N. We need however to know K.

Remember also that $\mathrm{E}\{X\} = \sum_k k\,\theta_k$ and $\mathrm{E}\{G\} = \sum_k g_k\,\theta_k$ are not the same. They become equivalent if $g_k = k$.

Thus, we consider the case:

$$\sum_k g_k\,\theta_k = d \tag{44}$$

and we assume to know the number of states K. The objective is to find θ_k.

MaxEnt solution:
The classical answer to this problem is MaxEnt which can be described as follows:
It is obvious that, there are infinite number of possible solutions to the equation (44). The Maximum Entropy principle chooses the one with the highest entropy

$$H(\boldsymbol{\theta}) = -\sum_k \theta_k \ln\theta_k \tag{45}$$

The solution has the form

$$\theta_k(\lambda) = \frac{1}{Z(\lambda)}\exp\{-\lambda\,g_k\} = \exp\{-(\ln Z(\lambda) + \lambda\,g_k)\} \tag{46}$$

where

$$Z(\lambda) = \sum_k \exp\{-\lambda\,g_k\} \tag{47}$$

and λ is the solution of the following equation

$$-\partial \ln Z(\lambda)/\partial\lambda = d. \tag{48}$$

which can be computed numerically.

It is also easy to show that the maximum value of the entropy is

$$H_{max}(\boldsymbol{\theta}) = -\sum_k \theta_k \ln \theta_k = \ln Z(\lambda) + \lambda d = \max_{\lambda} \ln \theta_k(\lambda) \tag{49}$$

which can also be written

$$\max_{\lambda} \boldsymbol{\theta}(\lambda) = \exp\{H_{max}(\boldsymbol{\theta})\}. \tag{50}$$

Bayesian solution:
If we knew N, we could write the expression of the likelihood $P(D = d|\boldsymbol{\theta}, N)$ with $d = \sum_k g_k n_k$ and $\sum_k n_k = N$:

$$P(D = d|\boldsymbol{\theta}, N) = \prod_{n_k=0}^{N} P(\boldsymbol{n}|\boldsymbol{\theta}) \delta(N - \sum_k n_k) \delta(d - \sum_k g_k n_k) \tag{51}$$

We can also try to integrate out N:

$$P(D = d|\boldsymbol{\theta}) = \sum_{N=0}^{\infty} \prod_{n_k=0}^{N} P(\boldsymbol{n}|\boldsymbol{\theta}) \delta(N - \sum_k n_k) \delta(d - \sum_k g_k n_k) \tag{52}$$

These computations seem to me intractable. In the following, I propose another approach:

The main idea here is that, we may account for uncertainty of this data (in particular, because we do not know the value of N) by assuming

$$p(d|\boldsymbol{\theta}) = \mathcal{N}\left(d - \sum_k g_k \theta_k, \sigma^2\right) \tag{53}$$

and by arguing on the additivity and positivity of $\boldsymbol{\theta}$ we choose

$$\pi(\boldsymbol{\theta}) = \exp\{-H(\boldsymbol{\theta})\} \tag{54}$$

Then, the posterior is

$$\pi(\boldsymbol{\theta}|d) = \exp\left\{-\frac{1}{2\sigma^2}(d - \sum_k g_k \theta_k)^2 - H(\boldsymbol{\theta})\right\} \tag{55}$$

and the MAP solution is

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{(d - \sum_k g_k \theta_k)^2 - \alpha H(\boldsymbol{\theta})\right\} \tag{56}$$

with $\alpha = 2\sigma^2$.

Now, if we choose $H(\boldsymbol{\theta}) = \sum_k \theta_k \ln \theta_k$ the numerical results obtained by this approach and the the MaxEnt solution become almost identical. However, if we can fix the value of α, we have access to the $\pi(\boldsymbol{\theta}|d)$ which contains more information than only one point estimator.

Combined data fusion solution:
Assume now that, not only we have the data $\boldsymbol{x}$ or $\boldsymbol{n}$, but also d from previous section. How to combine them. Here is my solution.

Follow the Bayesian approach of the sections 1 or 2 to write down the expression of the *a posteriori* law

$$\ln \pi(\boldsymbol{\theta}|\boldsymbol{n}) = \sum_k [n_k \ln \theta_k + (N - n_k) \ln(1 - \theta_k)] + \ln \pi(\boldsymbol{\theta}) + c \tag{57}$$

and use the expression of $\pi(\boldsymbol{\theta}|d)$ in equation (55) as the prior $\pi(\boldsymbol{\theta})$ here.

PROBLEM 6

Assume now that, our observer has repeated the experience L times, and before each experience, he has changed the numbers written on each face. For example, the first time, he has written $g_k = k$ and for the second experience $g_k = k^2$. This is also equivalent to the experiment of using L similar dice with different colors and different labeling on each faces simultaneously. Then, he computed the numbers n_{kl}.

But, assume now that, finally, he gives us only the mean values $\bar{n}_l = (1/N) \sum_k n_{kl}$ or $d_l = (1/N) \sum_k g_{kl}$. The problem is similar to the previous case, but here we have L data:

$$\sum_k g_{kl}\, \theta_k = d_l, \quad l = 1, \ldots, L \tag{58}$$

which can be written $\boldsymbol{G\theta} = \boldsymbol{d}$ where $\boldsymbol{G}$ is the matrix with elements g_{kl}. Thus, we have a linear system of equations with K unknowns and L data. Note that here we know exactly the values g_{kl}.

If the experimenter has made good choices for g_{kl} and if $L = K$, then we may only try to solve that system of equations and obtain an exact solution to the problem. But, what if $L < K$ or if the experimenter has not made a good choice for g_{kl}, for example, if he has naively wrote $g_{kl} = kl$. In both cases, the system of equation has an infinite number of solutions.

MaxEnt solution:
The MaxEnt approach is again straightforward and the solution has the form

$$\theta_k = \frac{1}{Z(\boldsymbol{\lambda})} \exp\left\{ -\sum_l \lambda_l\, g_{kl} \right\} = \exp\left\{ -(\ln Z(\boldsymbol{\lambda}) + \sum_l \lambda_l\, g_{kl}) \right\} \tag{59}$$

where

$$Z(\boldsymbol{\lambda}) = \sum_k \exp\left\{ -\sum_l \lambda_l\, g_{kl} \right\} \tag{60}$$

and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_L]$ is the solution of the following equation

$$-\partial \ln Z(\boldsymbol{\lambda})/\partial \lambda_l = d_l \tag{61}$$

which can be computed numerically. It is also easy to show that the maximum value of the entropy is

$$H_{max}(\boldsymbol{\theta}) = -\sum_k \theta_k \ln \theta_k = \ln Z(\lambda) + \boldsymbol{\lambda}^t \boldsymbol{d} = \max_{\lambda} \ln \boldsymbol{\theta}(\boldsymbol{\lambda}) \tag{62}$$

which can also be written

$$\max_{\boldsymbol{\lambda}} \boldsymbol{\theta}(\boldsymbol{\lambda}) = \exp\{H_{max}(\boldsymbol{\theta})\}. \tag{63}$$

Bayesian solution:
Following the steps of the section 5, we have

$$p(\boldsymbol{d}|\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{d} - \boldsymbol{G\theta}, \sigma^2\right) \tag{64}$$

and by arguing on the additivity and positivity of $\boldsymbol{\theta}$ we choose

$$\pi(\boldsymbol{\theta}) = \exp\{-H(\boldsymbol{\theta})\} \tag{65}$$

Then, the posterior is

$$\pi(\boldsymbol{\theta}|d) = \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{d} - \boldsymbol{G\theta}\|^2 - H(\boldsymbol{\theta})\right\} \tag{66}$$

and the MAP solution is

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{\|\boldsymbol{d} - \boldsymbol{G\theta}\|^2 - \alpha H(\boldsymbol{\theta})\right\} \tag{67}$$

with $\alpha = 2\sigma^2$.

Combined data fusion solution:
Assume now that, not only we have the data $\boldsymbol{x}$ or $\boldsymbol{n}$, but also $\boldsymbol{d}$ from previous section. How to combine them. Here again we can follow the Bayesian approach of the sections 1 or 2 to write down the expression of the *a posteriori* law

$$\ln \pi(\boldsymbol{\theta}|\boldsymbol{n}) = \sum_k [n_k \ln \theta_k + (N - n_k) \ln(1 - \theta_k)] + \ln \pi(\boldsymbol{\theta}) + c \tag{68}$$

and use the expression of $\pi(\boldsymbol{\theta}|\boldsymbol{d})$ in equation (66) as the prior $\pi(\boldsymbol{\theta})$ here.

PROBLEM 7

Consider the same previous experiment, but this time, the experimenter is sure that all dice were absolutely identical and unloaded, but he has forgotten to note the numbers he has written on the dice faces. However, he has also noted the mean values $(1/L)\sum_l g_{kl} = d_k$. Can we be of any help for him to find them?

Thus, this time, $\theta_k = 1/K, k = 1, \ldots, K$ and we have

$$\sum_k g_{kl}\,\theta_k = (1/K)\sum_k g_{kl} = d_l, \quad l = 1, \ldots, L \tag{69}$$

and also $(1/L)\sum_l g_{kl} = d_k, \quad k = 1, \ldots, K$.

The problem becomes an interesting one, we want to compute the elements of a matrix from its row and column sums. This mathematical problem arise in many other applications such as computed tomography where we want to recover the pixel values of an image from its horizontal and vertical projections.

Except the case of $K = L = 2$, we have always less data than unknowns and the problem has an infinite number of solutions. Even in the case $K = L = 2$ where the number of unknowns and data are equal, the problem is still under-determined and has infinite number of solutions. We need to question our experimenter to see if he can remember of any other information about those numbers (prior information or constraints?) which can be helpful to give reasonable answers about this question.

To go further in details of this problem, let change slightly the notations. We want to estimate the elements g_{kl} of a $(K \times L)$ matrix $\boldsymbol{G}$ from its row sums $r_k = \sum_l g_{kl}$ and its column size $c_l = \sum_k g_{kl}$. We may also note $\boldsymbol{r} = [r_1, \ldots, r_K]$, $\boldsymbol{c} = [c_1, \ldots, c_L]$, $\boldsymbol{d} = [\boldsymbol{r}; \boldsymbol{c}]$ and $\boldsymbol{g}$ a vector containing all the elements of the matrix $\boldsymbol{G}$ concatenated column by column. Then, it is easy to see that we can also write $\boldsymbol{c} = \boldsymbol{A}_1\boldsymbol{g}$, $\boldsymbol{r} = \boldsymbol{A}_2\boldsymbol{g}$ and thus $\boldsymbol{d} = \boldsymbol{A}\boldsymbol{g}$ where $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ and $\boldsymbol{A}$ are, respectively, a $(K \times KL)$, a $(L \times KL)$ and a $((K+L) \times KL)$ matrices with $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_2 \\ \boldsymbol{A}_2 \end{bmatrix}$ and whose elements are composed of zeros and ones.

Now, we consider two sets of answers of our experimenter: those who put deterministic constraints on g_{kl} and those who put probabilistic constraints.

Deterministic constraints:

- $g_{kl} = g_k$. Then, we have $r_k = Lg_k$ and we have a unique solution $g_k = r_k/L$ subject to the condition that $\sum_k g_k = \frac{K}{L}\sum_k r_k = c_l, \quad l = 1, \ldots, L$.
- $g_{kl} = g_l$. Then, we have $c_l = Kg_l$ and we have a unique solution $g_l = c_l/K$ subject to the condition that $\sum_l g_l = \frac{L}{K}\sum_l c_l = r_k, \quad k = 1, \ldots, K$.
- $g_{kl} = g_{1_k}\, g_{2_l}$. Then, we have $r_k = g_{1_k}\sum_l g_{2_l}$ and $c_l = g_{2_l}\sum_k g_{1_k}$ and we have $g_{1_k} \propto r_k$ and $g_{2_l} \propto c_l$. There still remains two unknowns $\sum_l g_{2_l}$ and $\sum_k g_{1_k}$. However, if g_{1_k} and g_{2_l} are normalized, then we have a unique solution.
- g_{kl} are normalized as they represent a probability distribution: $\sum_k g_{kl} = \sum_l g_{kl} = \sum_k \sum_l g_{kl} = 1$. This information is not enough to find a unique solution. That becomes true if g_{kl} is separable as in previous case.

- g_{kl} are normalized as they represent a probability distribution: $\sum_k g_{kl} = \sum_l g_{kl} = \sum_k \sum_l g_{kl} = 1$ and and are distributed as uniformly as possible over the grid $\{(k,l), k = 1,\ldots,K, l = 1,\ldots,L\}$.
  This information may be enough to find a solution if it exists, by maximizing $H(\boldsymbol{g}) = -\sum_j g_j \ln g_j$ subject to the data constraint $\boldsymbol{A}\boldsymbol{g} = \boldsymbol{d}$ and the normalization constraint $\sum_j g_j = 1$. Then the solution is given by $\boldsymbol{g} = \frac{1}{Z(\boldsymbol{\lambda})} \exp\{\boldsymbol{A}^t\boldsymbol{\lambda}\}$ where $\boldsymbol{\lambda}$ is the solution of $-\partial \ln Z(\boldsymbol{\lambda})/\partial\lambda_j = d_j$ which can also be computed by $\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left\{ D(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^t\boldsymbol{d} \right\}$. Unfortunately, there is not an explicit expression for this solution, but it is by construction positive ($g_j \propto \exp\{[\boldsymbol{A}^t\boldsymbol{\lambda}]_j\}$) and satisfies the data and normalization constraints for any correct data sets. Note also that this solution is not a linear function of the data. There is only one question remaining: Is there any other criteria $H(\boldsymbol{g})$ which can give these satisfactions?
  To give a partial answer to this question, we may say that any convex criterion can be used to find a unique solution. For example, $H(\boldsymbol{g}) = \sum_j g_j^2 = \|\boldsymbol{g}\|^2$ which gives the minimum norm (generalized inverse) solution $\boldsymbol{g} = \boldsymbol{A}^+\boldsymbol{d}$ which becomes $\boldsymbol{g} = \boldsymbol{A}^t(\boldsymbol{A}\boldsymbol{A}^t)^{-1}\boldsymbol{d}$ if $\boldsymbol{A}\boldsymbol{A}^t$ was invertible. Note that this solution is a linear function of the data, but, this criterion does not guaranty the positivity of the solution. Another example is $H(\boldsymbol{g}) = \sum_j \ln g_j$ which gives the solution of the form $g_j = \frac{1}{[\boldsymbol{A}^t\boldsymbol{\lambda}]_j}$ but, this criterion does not guaranty neither the positivity nor the boundedness of the solution. One can find other convex criteria (see next section).

Probabililistic constraints:

- We know that $\boldsymbol{g} \in \mathcal{C}$ and that we generated $\boldsymbol{g}$ according to a reference measure $q(\boldsymbol{g})$ over $\mathcal{C}$ such that $\mathrm{E}_q\{\boldsymbol{g}\} = \boldsymbol{g}_0$. Now, again, we can use the ME tool and search for $p(\boldsymbol{g})$ such that $\boldsymbol{A}\mathrm{E}_p\{\boldsymbol{g}\} = \boldsymbol{d}$ and minimizes $KL(p,q)$. We know that the solution is $p(\boldsymbol{g}) = \frac{1}{Z(\boldsymbol{\lambda})} q(\boldsymbol{g}) \exp\left\{\boldsymbol{\lambda}^t\boldsymbol{A}^t\boldsymbol{g}\right\}$ where $\boldsymbol{\lambda}$ is the solution of $-\partial \ln Z(\boldsymbol{\lambda})/\partial\lambda_j = d_j$ which can also be computed by $\widehat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \left\{ D(\boldsymbol{\lambda}) = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda}^t\boldsymbol{d} \right\}$ and finally, the solution $\widehat{\boldsymbol{g}} = \mathrm{E}_p\{\boldsymbol{g}\}$ can be computed by $\widehat{\boldsymbol{g}} = \arg\min_{\boldsymbol{A}\boldsymbol{g}=\boldsymbol{d}} \{H(\boldsymbol{g}, \boldsymbol{g}_0\}$. However, as we discussed it before, the expression of H depends on the choice $q(\boldsymbol{g})$:
  For $\mathcal{C}$ a closed set of real numbers and $q(\boldsymbol{g})$ Gaussian, we have $H(\boldsymbol{g}, \boldsymbol{g}^{(0)}) = \|\boldsymbol{g} - \boldsymbol{g}_0\|^2$ and
  for for $\mathcal{C}$ a closed set of real numbers and $q(\boldsymbol{g})$ a Lebesgue measure on $\mathcal{C}$, we have $H(\boldsymbol{g}, \boldsymbol{g}^{(0)}) = Burg(\boldsymbol{g}, \boldsymbol{g}_0) = -\sum_j \ln(g_j/g_{0_j}) + (g_j - g_{0_j})$ and, finally,
  for $\mathcal{C}$ a closed set of integer numbers and $q(\boldsymbol{g})$ Poissonian, we have $H(\boldsymbol{g}, \boldsymbol{g}^{(0)}) = KL(\boldsymbol{g}, \boldsymbol{g}_0) = \sum_j g_j \ln(g_j/g_{0_j}) + (g_j - g_{0_j})$.
  This discussion shows a relation between the classical ME approach of the last section and the ME in the mean as is presented here. Even if here, we have a tool to derive the expression of the needed convex criterion, still an arbitrary remains on the choice of $\mathcal{C}$ and the reference measure $q(\boldsymbol{g})$.
- Each element g_{kl} has been generated independently using a Gaussian random number generator: $g_{kl} \sim \mathcal{N}(k, \lambda)$.
- Each element g_{kl} has been generated independently using a Gaussian random number generator: $g_{kl} \sim \mathcal{N}(l, \lambda)$.

- Two sets of numbers g_{1_k} and g_{2_l} have been generated using a Gaussian random number generator $g_{1_k} \sim \mathcal{N}(k, \lambda_1)$ and $g_{2_l} \sim \mathcal{N}(l, \lambda_2)$, then normalized and pointwisely multiplied: $g_{kl} = g_{1_k} g_{2_l}$.
- Each element g_{kl} has been generated independently using a random number generator. We consider three interesting cases: $g_{kl} \sim \mathcal{N}(\lambda, 1), \quad g_{kl} \sim \mathcal{N}(\lambda, \lambda)$ and $g_{kl} \sim \mathcal{P}(\lambda)$.
- The element $g_{1l}, g_{k1}, g_{1L}, g_{K1}$ have been generated independently using a random number generator $\mathcal{N}(0, 1)$, but others are generated by $g_{kl} \sim \mathcal{N}(\bar{g}_{kl}, 1)$ where $\bar{g}_{kl} = \frac{1}{4}[g_{k-1,l} + g_{k,l-1} + g_{k+1,l} + g_{k,l+1}]$.

Let consider only the case of independent Gaussian $g_{kl} \sim \mathcal{N}(\lambda, \lambda)$ and $g_{kl} \sim \mathcal{P}(\lambda)$ where we may be able to do all the computations.

Gaussian case: $g_{kl} \sim \mathcal{N}(\lambda, \lambda)$**:**

$$p(g_{kl}) = (\frac{1}{2\pi\lambda})^{\frac{1}{2}} \exp\left\{-\frac{1}{2\lambda}(g_{kl} - \lambda)^2\right\} \longrightarrow \ln p(g_{kl}) = -\frac{1}{2\lambda}[(g_{kl} - \lambda)^2 + \ln(2\pi)]$$

$$g_{kl} \sim \mathcal{N}(\lambda, \lambda), \quad r_k = \sum_l g_{kl} \sim \mathcal{N}(r_k | L\lambda, L\lambda), \quad c_l = \sum_k g_{kl} \sim \mathcal{N}(c_l | K\lambda, K\lambda),$$

$$p(\boldsymbol{r}) = \prod_k \mathcal{N}(r_k | L\lambda, L\lambda) = (\frac{1}{2\pi L})^{\frac{K}{2}} \exp\left\{-\frac{1}{2L\lambda}\sum_k (r_k - L\lambda)^2\right\}$$

$$p(\boldsymbol{c}) = \prod_l \mathcal{N}(c_l | K\lambda, K) = (\frac{1}{2\pi K})^{\frac{L}{2}} \exp\left\{-\frac{1}{2K\lambda}\sum_l (c_l - K\lambda)^2\right\}$$

$$\begin{aligned}
p(g_{kl}|\boldsymbol{r}, \boldsymbol{c}, \lambda) &\propto P(g_{kl}, \boldsymbol{r}, \boldsymbol{c}|\lambda) = \exp\left\{-\frac{1}{2\lambda}(g_{kl} - \lambda)^2\right\} \\
&\times \exp\left\{-\frac{1}{2L\lambda}\sum_k (r_k - L\lambda)^2\right\} \times \exp\left\{-\frac{1}{2K\lambda}\sum_l (c_l - K\lambda)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\lambda}\left[(g_{kl} - \lambda)^2 + \frac{1}{K}\sum_k (r_k - L\lambda)^2 + \frac{1}{L}\sum_l (c_l - K\lambda)^2\right]\right\} \\
&\quad \text{with} \quad r_k = \sum_l g_{kl} \ \text{ and } \ c_l = \sum_k g_{kl}
\end{aligned}$$

It is then easily seen that

$$\begin{aligned}
p(g_{kl}|\boldsymbol{r}, \boldsymbol{c}, \lambda) &\propto \exp\left\{-\frac{1}{2\lambda}J(g_{kl})\right\} \\
\text{with } J(g_{kl}) &= (g_{kl} - \lambda)^2 + \frac{1}{K}\sum_k (\sum_l g_{kl} - L\lambda)^2 + \frac{1}{L}\sum_l (\sum_k g_{kl} - K\lambda)^2
\end{aligned}$$

is Gaussian and we can easily compute its mean and variance. To obtain the mean values, we can compute the derivative of

$$J = (g_{kl} - \lambda)^2 + \frac{1}{K}\sum_k (r_k - L\lambda)^2 + \frac{1}{L}\sum_l (c_l - K\lambda)^2$$

which is

$$\partial J/\partial g_{kl} \quad = \quad 2(g_{kl} - \lambda) + \frac{2}{K}\sum_k (r_k - L\lambda) + \frac{2}{L}\sum_l (c_l - K\lambda)$$

and equal it to zero to obtain

$$g_{kl} = \lambda - \frac{1}{K}\sum_k (r_k - L\lambda) - \frac{1}{L}\sum_l (c_l - K\lambda) = \lambda(1 + L + K) - [\frac{1}{K}\sum_k r_k + \frac{1}{L}\sum_l c_l] \tag{70}$$

This result is interesting, because $\frac{1}{K}\sum_k r_k + \frac{1}{L}\sum_l c_l$ is what is called the backprojection in computed tomography.

We can generalize these results, if we work with the vectors $\boldsymbol{g}$, $\boldsymbol{r} = \boldsymbol{A}_1\boldsymbol{g}$, $\boldsymbol{c} = \boldsymbol{A}_2\boldsymbol{g}$ and $\boldsymbol{d} = \begin{bmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}\boldsymbol{g} = \boldsymbol{A}\boldsymbol{g}$:

$$\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{g}_0, \boldsymbol{R}_g), \quad \boldsymbol{d} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{g}_0, \boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t), \quad \begin{bmatrix}\boldsymbol{g}\\ \boldsymbol{d}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{g}_0\\ \boldsymbol{A}\boldsymbol{g}_0\end{bmatrix}, \begin{bmatrix}\boldsymbol{R}_g & \boldsymbol{R}_g\boldsymbol{A}^t\\ \boldsymbol{A}\boldsymbol{R}_g & \boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t\end{bmatrix}\right)$$

and thus

$$\boldsymbol{g}|\boldsymbol{d} \sim \mathcal{N}(\widehat{\boldsymbol{g}}, \widehat{\boldsymbol{R}}_g), \quad \text{with} \quad \begin{cases} \widehat{\boldsymbol{g}} = \boldsymbol{g}_0 + \boldsymbol{R}_g\boldsymbol{A}^t(\boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t)^+(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{g}_0) \\ \widehat{\boldsymbol{R}}_g = \boldsymbol{R}_g - \boldsymbol{R}_g\boldsymbol{A}^t(\boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t)^+\boldsymbol{A}\boldsymbol{R}_g \end{cases}$$

where $(\boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t)^+$ is the generalized inverse of $\boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t$. Note that when $\boldsymbol{A}\boldsymbol{R}_g\boldsymbol{A}^t$ is invertible, we have $\widehat{\boldsymbol{g}} = \boldsymbol{A}^{-1}\boldsymbol{d}$ and $\widehat{\boldsymbol{R}}_g = 0$. For the particular case of $\boldsymbol{R}_g = \lambda\boldsymbol{I}$ we have

$$\begin{cases} \widehat{\boldsymbol{g}} = \boldsymbol{g}_0 + \boldsymbol{A}^t(\boldsymbol{A}\boldsymbol{A}^t)^+(\boldsymbol{d} - \boldsymbol{A}\boldsymbol{g}_0) \\ \widehat{\boldsymbol{R}}_g = \lambda(\boldsymbol{I} - \boldsymbol{A}^t(\boldsymbol{A}\boldsymbol{A}^t)^+\boldsymbol{A}) \end{cases} \tag{71}$$

and for the particular case of $\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_1\\ \boldsymbol{A}_2\end{bmatrix}$ we have $\boldsymbol{A}\boldsymbol{A}^t = \begin{bmatrix}\boldsymbol{A}_1\boldsymbol{A}_1^t & \boldsymbol{A}_1\boldsymbol{A}_2^t\\ \boldsymbol{A}_2\boldsymbol{A}_1^t & \boldsymbol{A}_2\boldsymbol{A}_2^t\end{bmatrix} = \begin{bmatrix}K\boldsymbol{I} & \boldsymbol{1}\\ \boldsymbol{1} & L\boldsymbol{I}\end{bmatrix}$, where $\boldsymbol{1}$ is a matrix with all its elements equal to 1. We may note that $\boldsymbol{A}\boldsymbol{A}^t$ is singular and its rank is $K + L - 1$. We can however compute numerically $\widehat{\boldsymbol{g}}$ and $\widehat{\boldsymbol{R}}_g$. Note also that, even if *a priori* g_{kl} were independent, *a posteriori* they are correlated.

Poisson case: $g_{kl} \sim \mathcal{P}(\lambda)$**:**

$$P(g_{kl}) = \lambda^{g_{kl}} \exp\{-\lambda\} / (g_{kl}!) \longrightarrow \ln P(g_{kl}) = (\ln \lambda) g_{kl} - \ln(g_{kl}!) - \lambda$$

$$g_{kl} \sim \mathcal{P}(\lambda), \quad r_k = \sum_l g_{kl} \sim \mathcal{P}(L\lambda), \quad c_l = \sum_k g_{kl} \sim \mathcal{P}(K\lambda),$$

$$P(\boldsymbol{r}) = \prod_k (L\lambda)^{r_k} \exp\{-L\lambda\} / (r_k!), \quad P(\boldsymbol{c}) = \prod_l (K\lambda)^{c_l} \exp\{-K\lambda\} / (c_l!)$$

$$\begin{aligned} P(g_{kl}|\boldsymbol{r}, \boldsymbol{c}, \lambda) &\propto (\lambda)^{g_{kl}} / (g_{kl}!) \\ &\times \prod_k (L\lambda)^{r_k} / (r_k!) \times \prod_l (K\lambda)^{c_l} / (c_l!) \\ &\text{with} \quad r_k = \sum_l g_{kl} \ \text{ and } \ c_l = \sum_k g_{kl} \end{aligned}$$

It is then possible to show that $P(g_{kl}|\boldsymbol{r}, \boldsymbol{c}, \lambda)$ is also a Poisson law, but it is not easy to find an explicite expression for its mean value. However, using again the Striling formula when working with $\ln P(g_{kl}|\boldsymbol{r}, \boldsymbol{c}, \lambda)$ one can obtain approximate expression for it

$$P(g_{kl}|\{g_{k' \neq k, l' \neq l}\}, \boldsymbol{r}, \boldsymbol{c}, \lambda) = \mathcal{P}(\lambda(1 + L\exp\{c_l\} + K\exp\{r_k\}))$$

and thus we have

$$\begin{aligned} \mathrm{E}\{g_{kl}|\{g_{k' \neq k, l' \neq l}\}, \boldsymbol{r}, \boldsymbol{c}, \lambda\} &= \lambda(1 + L\exp\{c_l\} + K\exp\{r_k\}) \\ &= KL\lambda(1/(KL) + (1/K)\exp\{c_l\} + (1/L)\exp\{r_k\}). \end{aligned}$$

This is interesting, because $(1/K)\exp\{c_l\} + (1/L)\exp\{r_k\}$ corresponds to the famous backprojection operation in computed tomography.

CONCLUSIONS

This paper was another analysis of dice problems trying to answer some of the questions about the situations where we can use the Bayesian or the Maximum Entropy approaches. Through this paper, we distinguished three approaches: Bayesian, classical MaxEnt and MaxEnt on the mean. I showed some of the situations where we can use these approaches.

The Bayesian approach can be used when we can write explicitely a probabilistic model relating the data to the unknown parameters from which we can deduce the expression of the likelihood and can assign an *a priori* law to those parameters, we can then use the Bayesian approach to compute the *a posteriori* from which we can infer about the parameters.

The classical MaxEnt can be used in cases where we have a set of data which can be considered as linear constraints on a set of parameters which are themselves a probability distribution. Then the classical MaxEnt give the possibility to find a unique solution to the underdetermined problem.

The MaxEnt on the mean can be used in cases where we have a set of data which can be considered as linear constraints on the expected values of a set of parameters which are the elements of a convex set on which we can define a reference measure. Then, we can use the MaxEnt on the mean approach to compute a probability law on that set such that the expected values of the parameters satisfy exactly the data. We can then compute those expected values which depend on the choice of the reference measure. We showed also that there are strong relation between the two MaxEnt approaches.

In some cases, it may happens that we have both the moment data and the sampling data. Then we can first use the MaxEnt approach to assign the prior law using the moment data and then use it with the likelihood to compute the *a posteriori* law of the parameters from which we can infer about them.

Finally, even if I tried to answer to some of the questions, I also asked more questions to be answered. We thus still have a lot to do with all the three approaches. However, it seems that for practical applications the Bayesian approach seems to be the right and the easiest one.

REFERENCES

1. E. T. Jaynes, "Prior probabilities," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-4, no. 3, pp. 227–241, Sept. 1968.
2. E. T. Jaynes, "Where do we stand on maximum entropy ?," in *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus, Eds. M.I.T. Press, Cambridge (MA), 1978.
3. E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, Sept. 1982.
4. E. T. Jaynes, "Where do we go from here?," in *Maximum-Entropy and Bayesian Methods in Inverse Problems*, C. R. Smith and W. T. J. Grandy, Eds., pp. 21–58. 1985.
5. Frieden, "Dice, entropy and likelihood," *J. Opt. Soc. Amer.*, vol. 73, no. 12, pp. 1764–1770, 1985.
6. B. R. Frieden, "Maximum-probable restoration of photon-limited images," *Applied Optics*, vol. 26, no. 9, pp. 1755–1764, 1987.
7. J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. 26, pp. 26–37, Jan. 1980.
8. J. M. Van Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *IEEE Trans. Inf. Theory*, vol. 27, no. 4, pp. 483–489, July 1981.
9. P. Fougere, "Maximum entropy calculations on a discrete probability space," in *Maximum Entropy and Bayesian Methods in Science and Engineering*, G. J. Erickson and C. R. Smith, Eds., vol. 1, pp. 205–234. Kluwer Academic Publ., Dordrecht,, 1988.
10. C. Robert, *L'analyse statistique Bayésienne*, Economica, 1992.
11. C. P. Robert, *The Bayesian Choice. A Decision-Theoretic Motivation*, Springer Texts in Statistics. Springer-Verlag, New York, 1997.
12. S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.
13. S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.
14. R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
15. J. M. Borwein and A. S. Lewis, "Duality relationships for entropy-like minimization problems," *SIAM J. Control and Optimization*, vol. 29, no. 2, pp. 325–338, Mar. 1991.
16. A. Mohammad-Djafari, *A Matlab Program to Calculate the Maximum Entropy Distributions*, pp. 221–233, Kluwer Academic Publ., Laramie, WY, T.W. Grandy edition, 1991.
17. R. T. Rockafellar, "Lagrange multipliers and optimality," *SIAM Review*, vol. 35, no. 2, pp. 183–238, June 1993.
18. G. Le Besnerais, *Méthode du maximum d'entropie sur la moyenne, critères de reconstruction d'image et synthèse d'ouverture en radio-astronomie*, Phd thesis, Université de Paris-Sud, Orsay, Dec. 1993.
19. A. Mohammad-Djafari, "Maximum d'entropie et problèmes inverses en imagerie," *Traitement du Signal*, pp. 87–116, 1994.
20. J.-F. Bercher, *Développement de critères de nature entropique pour la résolution des problèmes inverses linéaires*, Phd thesis, Université de Paris–Sud, Orsay, Feb. 1995.
21. A. Mohammad-Djafari, "A comparison of two approaches: Maximum entropy on the mean (MEM) and Bayesian estimation (BAYES) for inverse problems," in *Maximum Entropy and Bayesian Methods*, Berg–en–Dal,, Aug. 1996, Kluwer Academic Publ.
22. G. Le Besnerais, J.-F. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1565–1578, July 1999.
23. A. Mohammad-Djafari, "Entropie en traitement du signal," *Traitement du signal*, vol. Num. spécial, volume 15, no. 6, pp. 545–551, 1999.