To: *C. Caves*
From: *C. M. Caves*
Subject: **Laplace's rule of succession and beta functions**

2000 June 25; second equality in Eq. (5) and first equality in Eq. (30) corrected 2001 June 5

The following is taken from E. T. Jaynes, "Monkeys, kangaroos, and N," in *Maximum Entropy and Bayesian Methods in Applied Statistics*, edited by J. H. Justice (Cambridge University Press, Cambridge, England, 1986), pages 26–58 (Proceedings of the Fourth Maximum Entropy Workshop, University of Calgary, August 1984).

Consider a quantity with N alternatives and prior probability on probabilities,

$$P(\mathbf{p}) = P(p_1, \ldots, p_N) = A p_1^{k_1-1} \cdots p_N^{k_N-1} \,. \tag{1}$$

Normalization of $P(\mathbf{p})$ on the probability simplex gives

$$1 = \int d\mathbf{p}\, P(\mathbf{p}) = A\sqrt{N} \int_0^\infty dp_1 \cdots \int_0^\infty dp_N\, p_1^{k_1-1} \cdots p_N^{k_N-1} \delta\!\left(\sum_j p_j - 1\right) , \tag{2}$$

where the integration measure on the simplex is

$$d\mathbf{p} = \sqrt{N} dp_1 \cdots dp_N \delta\!\left(\sum_j p_j - 1\right) , \tag{3}$$

and where the δ-function restricts the integral to the simplex even though the integrals run from 0 to ∞. Thus the normalization constant is given by

$$A = \frac{1}{\sqrt{N} B(\mathbf{k})} \,, \tag{4}$$

where $B(\mathbf{k})$ is the beta function

$$\begin{aligned} B(\mathbf{k}) &\equiv \int_0^\infty dp_1 \cdots \int_0^\infty dp_N\, p_1^{k_1-1} \cdots p_N^{k_N-1} \delta\!\left(\sum_j p_j - 1\right) \\ &= \int_0^1 dp_1\, p_1^{k_1-1} \int_0^{1-p_1} dp_2\, p_2^{k_2-1} \cdots \int_0^{1-p_1-\cdots-p_{N-3}} dp_{N-2}\, p_{N-2}^{k_{N-2}-1} \\ &\qquad \times \int_0^{1-p_1-\cdots-p_{N-2}} dp_{N-1}\, p_{N-1}^{k_{N-1}-1} (1-p_1-\cdots-p_{N-1})^{k_N-1} \,. \end{aligned} \tag{5}$$

To evaluate the beta function, consider the integral

$$B(\mathbf{k}; a) \equiv \int_0^\infty dx_1 \cdots \int_0^\infty dx_N\, x_1^{k_1-1} \cdots x_N^{k_N-1} \delta\!\left(\sum_j x_j - a\right) . \tag{6}$$

Making the change of variables $p_j = x_j/a$ and defining $K \equiv k_1 + \cdots k_n$, we can write

$$
\begin{aligned}
B(\mathbf{k}; a) &= a^K \int_0^\infty dp_1 \cdots \int_0^\infty dp_N \, p_1^{k_1-1} \cdots p_N^{k_N-1} \delta\left(a\left(\sum_j p_j - 1 \right) \right) \\
&= a^{K-1} \int_0^\infty dp_1 \cdots \int_0^\infty dp_N \, p_1^{k_1-1} \cdots p_N^{k_N-1} \delta\left(\sum_j p_j - 1 \right) \\
&= a^{K-1} B(\mathbf{k}) \; .
\end{aligned}
\tag{7}
$$

Integrating over e^{-a} gives

$$
\int_0^\infty da \, e^{-a} B(\mathbf{k}; a) = B(\mathbf{k}) \underbrace{\int_0^\infty da \, a^{K-1} e^{-a}}_{= \Gamma(K)} \; . \tag{8}
$$

But now notice that

$$
\begin{aligned}
\int_0^\infty da \, e^{-a} B(\mathbf{k}; a) &= \int_0^\infty dx_1 \cdots \int_0^\infty dx_N \, x_1^{k_1-1} \cdots x_N^{k_N-1} \int_0^\infty da \, e^{-a} \delta\left(\sum_j x_j - a \right) \\
&= \int_0^\infty dx_1 \, x_1^{k_1-1} e^{-x_1} \cdots \int_0^\infty dx_n \, x_n^{k_n-1} e^{-x_n} \\
&= \Gamma(k_1) \cdots \Gamma(k_N) \; .
\end{aligned}
\tag{9}
$$

Thus we find that the beta function is given by

$$
B(\mathbf{k}) = \frac{\Gamma(k_1) \cdots \Gamma(k_N)}{\Gamma(K)} \; . \tag{10}
$$

The resulting normalized probability on probabilities is

$$
P(\mathbf{p}) = \frac{1}{\sqrt{N}} \frac{p_1^{k_1-1} \cdots p_N^{k_N-1}}{B(\mathbf{k})} = \frac{1}{\sqrt{N}} \frac{\Gamma(K)}{\Gamma(k_1) \cdots \Gamma(k_N)} \, p_1^{k_1-1} \cdots p_N^{k_N-1} \; . \tag{11}
$$

Notice that $P(\mathbf{p})$ is normalizable if and only if all the ks are positive. When all the ks are equal to 1, $P(\mathbf{p})$ is the uniform distribution on the simplex, and the normalization constant $A = (N-1)!/\sqrt{N}$ is the inverse of the volume of the probability simplex. It is useful to define

$$
g_j \equiv \frac{k_j}{K} \; , \quad \text{where} \quad \sum_{j=1}^N g_j = 1 \; . \tag{12}
$$

In L trials, the probability for occurrence numbers $n_1, \ldots n_N \equiv \mathbf{n}$, given probabilities $\mathbf{p}$, is the binomial distribution

$$P(\mathbf{n} \mid \mathbf{p}) = \frac{L!}{n_1! \cdots n_N!} \, p_1^{n_1} \cdots p_N^{n_N} \ . \tag{13}$$

Hence, the unconditioned probability for $\mathbf{n}$ is given by

$$\begin{aligned}
P(\mathbf{n}) &= \langle P(\mathbf{n} \mid \mathbf{p}) \rangle \\
&= \int d\mathbf{p} \, P(\mathbf{n} \mid \mathbf{p}) P(\mathbf{p}) \\
&= \frac{L!}{n_1! \cdots n_N!} \frac{\Gamma(K)}{\Gamma(k_1) \cdots \Gamma(k_N)} \\
&\qquad \times \underbrace{\int_0^\infty dp_1 \cdots \int_0^\infty dp_N \, p_1^{n_1+k_1-1} \cdots p_N^{n_N+k_N-1} \delta\left(\sum_j p_j - 1\right)}_{= B(\mathbf{n}+\mathbf{k})} \\
&= \frac{L!}{n_1! \cdots n_N!} \frac{\Gamma(K)}{\Gamma(k_1) \cdots \Gamma(k_N)} \frac{\Gamma(n_1 + k_1) \cdots \Gamma(n_N + k_N)}{\Gamma(L+K)} \ .
\end{aligned} \tag{14}$$

The probability for any sequence with occurrence numbers $\mathbf{n}$ is

$$\begin{aligned}
\langle p_1^{n_1} \cdots p_N^{n_N} \rangle &= \int d\mathbf{p} \, p_1^{n_1} \cdots p_N^{n_N} P(\mathbf{p}) \\
&= \frac{\Gamma(K)}{\Gamma(k_1) \cdots \Gamma(k_N)} \frac{\Gamma(n_1 + k_1) \cdots \Gamma(n_N + k_N)}{\Gamma(L+K)} \\
&= \frac{\prod_{j=1}^{N} (n_j + k_j - 1) \cdots k_j}{(L + K - 1) \cdots K} \ ,
\end{aligned} \tag{15}$$

where we use $\Gamma(x+1) = x\Gamma(x)$.

In particular, notice that

$$\langle p_j \rangle = \frac{\Gamma(K)}{\Gamma(K+1)} \frac{\Gamma(k_j + 1)}{\Gamma(k_j)} = \frac{k_j}{K} = g_j \tag{16}$$

and that

$$\langle p_j p_k \rangle = \begin{cases} \dfrac{\Gamma(K)}{\Gamma(K+2)} \dfrac{\Gamma(k_j+1)\Gamma(k_k+1)}{\Gamma(k_j)\Gamma(k_k)} = \dfrac{k_j k_k}{K(K+1)} = \dfrac{g_j g_k K}{K+1} \ , & j \neq k, \\ \dfrac{\Gamma(K)}{\Gamma(K+2)} \dfrac{\Gamma(k_j+2)}{\Gamma(k_j)} = \dfrac{k_j(k_j+1)}{K(K+1)} = \dfrac{g_j(g_j K + 1)}{K+1} \ , & j = k. \end{cases} \tag{17}$$

Thus the covariance matrix of the probabilities is given by

$$\langle \Delta p_j \Delta p_k \rangle = \langle p_j p_k \rangle - \langle p_j \rangle \langle p_k \rangle = \begin{cases} -\dfrac{g_j g_k}{K+1}\;, & j \neq k, \\ \dfrac{g_j(1-g_j)}{K+1}\;, & j = k. \end{cases} \tag{18}$$

Of particular interest are the limits where K goes to zero and infinity. When K goes to zero [even though $P(\mathbf{p})$ can't be normalized in this limit, we can, nonetheless, extract meaningful results], we use the fact that $\Gamma(z) \to 1/z$ when $z \to 0$ to find

$$\langle p_1^{n_1} \cdots p_N^{n_N} \rangle \underset{K\to 0}{=} \sum_{j=1}^{N} g_j \delta_{n_j L} = P(\mathbf{n})\;. \tag{19}$$

This means that the only sequences that have nonzero probability are those in which all the trials yield exactly the same result, the jth of these possibilities occurring with probability g_j. Thus, when $K \to 0$, $P(\mathbf{p})$ describes a mixture of N probability distributions, the jth of which, occurring with probability g_j, gives alternative j with certainty. The first trial determines which alternative applies, and all subsequent trials yield the same result as the first.

When K goes to infinity, we find

$$P(\mathbf{n}) = \frac{L!}{n_1! \cdots n_N!} \langle p_1^{n_1} \cdots p_N^{n_N} \rangle \underset{K\to\infty}{=} \frac{L!}{n_1! \cdots n_N!} g_1^{n_1} \cdots g_N^{n_N}\;, \tag{20}$$

which means that $P(\mathbf{p})$ becomes a δ-function centered at $\mathbf{p} = \mathbf{g}$. It is useful to exhibit explicitly the asymptotic behavior of $P(\mathbf{p})$ when K becomes large. Using the Stirling formula, we find that

$$P(\mathbf{p}) \underset{K\to\infty}{\sim} \frac{1}{\sqrt{N}} \sqrt{\left(\frac{K}{2\pi}\right)^{N-1}} \frac{\sqrt{g_1 \cdots g_N}}{p_1 \cdots p_N}\, e^{-KH(\mathbf{g}||\mathbf{p})}\;. \tag{21}$$

Assuming that $\mathbf{g}$ and $\mathbf{p}$ are nearly the same, we can make the further Gaussian approximation, giving

$$P(\mathbf{p}) \underset{K\to\infty}{\simeq} \frac{1}{\sqrt{N}} \frac{1}{\sqrt{(2\pi/K)^{N-1} g_1 \cdots g_N}} \exp\left(-\frac{K}{2} \sum_{j=1}^{N} \frac{(p_j - g_j)^2}{g_j} \right)\;. \tag{22}$$

The Gaussian approximation is valid everywhere that the argument of the exponent is small, a region that, because K is large, contains essentially all of the probability. It is easy to verify that the Gaussian is normalized to unity on the probability simplex, provided we are allowed to extend the integration to negative values of each p_j on the grounds that the Gaussian is so localized that this extension makes a negligible difference. This allows us to say that

$$\sqrt{N}\delta\left(\sum_j p_j - 1 \right) P(\mathbf{p}) \underset{K\to\infty}{\longrightarrow} \delta(\mathbf{p} - \mathbf{g})\;. \tag{23}$$

We now turn to what happens when we use the results of the first L trials to update the probability on probabilities. Bayes's theorem gives

$$P(\mathbf{p} \mid \mathbf{n}) = \frac{P(\mathbf{n} \mid \mathbf{p})P(\mathbf{p})}{P(\mathbf{n})} = \frac{1}{\sqrt{N}} \frac{\Gamma(L+K)}{\Gamma(n_1+k_1)\cdots\Gamma(n_N+k_N)} p_1^{n_1+k_1-1}\cdots p_N^{n_N+k_N-1} \; . \tag{24}$$

The Bayesian updating simply updates the prior ks to new values $k_j' = n_j + k_j$. Defining the observed frequencies

$$f_j \equiv \frac{n_j}{L} \; , \tag{25}$$

one can define updated gs by

$$g_j' = \frac{n_j + k_j}{L+K} = \frac{L}{L+K} f_j + \frac{K}{L+K} g_j \; . \tag{26}$$

The updated g_j' is a weighted mean of the prior g_j and the observed frequency f_j. The weights $L/(L+K)$ and $K/(L+K)$ determined literally how much weight to put on the prior information and how much on the observed frequencies. The value of K characterizes how many trials are necessary so that the data starts to have an impact on the prior distribution. If $k_j \gg n_j$, $j = 1, \ldots, N$, then $P(\mathbf{p} \mid \mathbf{n})$ is essentially unchanged from the prior beta distribution, whereas if $n_j \gg k_j$, $j = 1, \ldots, N$, then $P(\mathbf{p} \mid \mathbf{n})$ is a beta distribution determined by the observed occurrence numbers n_j.

It is easy to calculate the probability for occurrence numbers $m_1, \ldots, m_N \equiv \mathbf{m}$ in M trials, conditioned on the occurrence numbers $\mathbf{n}$ for the first L trials:

$$\begin{aligned} P(\mathbf{m} \mid \mathbf{n}) &= \int d\mathbf{p}\, P(\mathbf{m} \mid \mathbf{p}) P(\mathbf{p} \mid \mathbf{n}) \\ &= \frac{M!}{m_1! \cdots m_N!} \frac{\Gamma(L+K)}{\Gamma(n_1+k_1)\cdots\Gamma(n_N+k_N)} \\ &\qquad \times \frac{\Gamma(m_1+n_1+k_1)\cdots\Gamma(m_N+n_N+k_N)}{\Gamma(M+L+K)} \; . \end{aligned} \tag{27}$$

The probability for any sequence with occurrence numbers $\mathbf{m}$, given $\mathbf{n}$ in the first L trials, is

$$\begin{aligned} \langle p_1^{m_1} \cdots p_N^{m_N} \rangle_{\mathbf{n}} &= \int d\mathbf{p}\, p_1^{m_1} \cdots p_N^{m_N} P(\mathbf{p} \mid \mathbf{n}) \\ &= \frac{\Gamma(L+K)}{\Gamma(n_1+k_1)\cdots\Gamma(n_N+k_N)} \frac{\Gamma(m_1+n_1+k_1)\cdots\Gamma(m_N+n_N+k_N)}{\Gamma(M+L+K)} \\ &= \frac{\displaystyle\prod_{j=1}^{N} (m_j+n_j+k_j-1)\cdots(n_j+k_j)}{(M+L+K-1)\cdots(L+K)} \; . \end{aligned} \tag{28}$$

In particular, notice that

$$\langle p_j \rangle_{\mathbf{n}} = \frac{n_j + k_j}{L + K} = g'_j = \frac{L}{L+K} f_j + \frac{K}{L+K} g_j \; . \tag{29}$$

Since $\langle p_j \rangle_{\mathbf{n}}$ is the probability to obtain alternative j in the $(L+1)$th trial, given the results of the first L trials, Eq. (29) is a generalized version of the famous Laplace rule of succession. Laplace's original formulation assumed a uniform prior, i.e., all the ks equal to 1, so $K = N$ and $g_j = 1/N$, in which case we have

$$\langle p_j \rangle_{\mathbf{n}} = \frac{n_j + 1}{L + N} = \frac{L f_j + 1}{L + N} \; . \tag{30}$$

Gamma function and Stirling formulas

$$\Gamma(x) = \int du\, u^{x-1} e^{-u} \; , \qquad \Gamma(x) = (x-1)\Gamma(x-1) \; , \quad x > 1$$

$$\Gamma(n+1) = n!$$

$$\Gamma(1/2) = \sqrt{\pi} \; , \qquad \Gamma(n + 1/2) = \frac{(2n-1)!!}{2^n} \sqrt{\pi}$$

$$\Gamma(x+1) = x^x e^{-x} \sqrt{2\pi x}\, e^{\theta/12x} \; , \qquad x > 0, \quad 0 < \theta < 1$$

$$\ln \Gamma(x+1) = x \ln x - x + \frac{1}{2} \ln(2\pi x) + \frac{\theta}{12x}$$

$$\Gamma(z) \sim z^z e^{-z} \sqrt{\frac{2\pi}{z}} \left(1 + \frac{1}{12z} + \cdots \right)$$

$$\ln \Gamma(z) \sim z \ln z - z + \frac{1}{2} \ln\left(\frac{2\pi}{z} \right) + \frac{1}{12z} + \cdots$$

$$B(\mathbf{k}) = \frac{\Gamma(k_1) \cdots \Gamma(k_N)}{\Gamma(K)} \underset{K \to \infty}{\sim} \sqrt{\left(\frac{2\pi}{K} \right)^{N-1}} \, \frac{e^{-KH(\mathbf{g})}}{\sqrt{g_1 \cdots g_N}}$$

$$\ln B(\mathbf{k}) = \ln\left(\frac{\Gamma(k_1) \cdots \Gamma(k_N)}{\Gamma(K)} \right) \underset{K \to \infty}{\sim} -KH(\mathbf{g}) - \frac{1}{2} \ln\left[\left(\frac{K}{2\pi} \right)^{N-1} g_1 \cdots g_N \right]$$

Entropies

$$H(\mathbf{p}) = -\sum_j p_j \ln p_j$$

Relative entropy: $H(\mathbf{g}||\mathbf{p}) = -H(\mathbf{g}) - \sum_j g_j \ln p_j = \sum_j g_j \ln(g_j/p_j) \geq 0$

$$H(\mathbf{g}||\mathbf{p}) \simeq \frac{1}{2}\sum_j \frac{(p_j - g_j)^2}{g_j} \quad \text{for} \quad \left|\frac{p_j - g_j}{g_j}\right| \ll 1 \ , j = 1, \ldots, N$$

Fourier transform of a Gaussian

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} dx\, e^{-(x-a)^2/2\sigma^2} e^{ikx} = e^{-\sigma^2 k^2/2} e^{ika}$$

Multi-dimensional Gaussians

Let

$$M_{jk} = \sum_{l=1}^{N} \lambda_l O_{lj} O_{lk}$$

be an $N \times N$ real symmetric matrix, diagonalized by the orthogonal matrix O_{jk} and having positive eigenvalues λ_l. We have

$$\sqrt{\frac{\det M}{(2\pi)^N}} \int dx_1 \cdots dx_N \, \exp\left(-\frac{1}{2}\sum_{j,k}(x_j - a_j) M_{jk} (x_k - a_k)\right)$$
$$= \sqrt{\frac{\lambda_1 \cdots \lambda_N}{(2\pi)^N}} \int dv_1 \cdots dv_N \, \exp\left(-\frac{1}{2}\sum_l \lambda_l (v_l - b_l)^2\right) = 1 \ ,$$

where

$$v_l \equiv \sum_j O_{lj} x_j \ , \qquad b_l \equiv \sum_j O_{lj} a_j \ .$$

If we let

$$M_{jk} = \Lambda m_{jk} = \Lambda \sum_{l=1}^{N} \mu_l O_{lj} O_{lk} \qquad \Longleftrightarrow \qquad \lambda_l = \Lambda \mu_l \ ,$$

we can take the limit $\Lambda \to \infty$ to get a multi-dimensional δ function:

$$\lim_{\Lambda\to\infty} \sqrt{\left(\frac{\Lambda}{2\pi}\right)^N} \sqrt{\det m} \exp\left(-\frac{1}{2}\Lambda \sum_{j,k}(x_j - a_j) m_{jk} (x_k - a_k)\right)$$
$$= \lim_{\Lambda\to\infty} \sqrt{\left(\frac{\Lambda}{2\pi}\right)^N} \sqrt{\mu_1 \cdots \mu_N} \exp\left(-\frac{1}{2}\Lambda \sum_l \mu_l (v_l - b_l)^2\right)$$
$$= \delta(\mathbf{v} - \mathbf{b}) = \delta(\mathbf{x} - \mathbf{a}) \ .$$