Bayesian inference, entropy, and the multinomial distribution

Thomas P. Minka

February 7, 2001

Abstract

Instead of maximum-likelihood or MAP, Bayesian inference encourages the use of predictive densities and evidence scores. This is illustrated in the context of the multinomial distribution, where predictive estimates are often used but rarely described as Bayesian. By using an entropy approximation to the evidence, many Bayesian quantities can be expressed in information-theoretic terms. For example, testing whether two samples come from the same distribution or testing whether two variables are independent boils down to a mutual information score (with appropriate smoothing). The same analysis can be applied to discrete Markov chains to get a test for Markovianity.

1 Introduction

Let x be a discrete variable taking on values $1, 2, ..., K$. The set of probability distributions on x can be parameterized by a vector $\mathbf{p}$ where $p(x = k) = p_k$. Another way to write this is

$$p(x|\mathbf{p}) = \prod_{k=1}^{K} p_k^{\delta(x=k)} \tag{1}$$

where $\delta(x = k)$ is an indicator function. The joint probability of N IID samples $\mathbf{X} = \{x_1, ..., x_N\}$ is therefore

$$\begin{aligned} p(\mathbf{X}|\mathbf{p}) &= \prod_{k=1}^{K} p_k^{N_k} && (2) \\ N_k &= \sum_i \delta(x_i = k) && (3) \end{aligned}$$

The joint probability of a set of counts N_k, on the other hand, is

$$p(N_1..N_K|\mathbf{p}) = \binom{N}{N_1..N_K} \prod_{k=1}^{K} p_k^{N_k} \tag{4}$$

since we must account for all possible permutations. Both are multinomial distributions with parameter $\mathbf{p}$.

A conjugate prior for $\mathbf{p}$ is the Dirichlet distribution:

$$p(\mathbf{p}|\alpha) \sim \mathcal{D}(\alpha_1, ..., \alpha_K) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{\alpha_k - 1} \tag{5}$$

$$\text{where } p_k > 0 \tag{6}$$

$$\sum_k p_k = 1 \tag{7}$$

The Beta distribution is the special case when $K = 2$. The Dirichlet distribution has the properties

$$p(p_1|\alpha) \sim \mathcal{D}(\alpha_1, \alpha_2 + ... + \alpha_K) \tag{8}$$

$$E[p_1] = \frac{\alpha_1}{\sum_k \alpha_k} \tag{9}$$

$$E[\log p_1] = \Psi(\alpha_1) - \Psi(\sum_k \alpha_k) \tag{10}$$

$$\text{where } \Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} \tag{11}$$

The maximum of the density is at $p_k = (\alpha_k - 1)/((\sum_k \alpha_k) - K)$, when this value is valid. Otherwise, p_k corresponding to the smallest α_k must be zero. The remaining p's have a Dirichlet distribution among themselves which can be maximized recursively. If instead all α's are equal and less than 1, then the maxima consist of all binary $\mathbf{p}$'s.

Given a Dirichlet prior, the joint distribution of a set of samples $\mathbf{X}$ and $\mathbf{p}$ is

$$p(\mathbf{X}, \mathbf{p}|\alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k p_k^{N_k + \alpha_k - 1} \tag{12}$$

so the posterior is

$$p(\mathbf{p}|\mathbf{X}, \alpha) \sim \mathcal{D}(N_k + \alpha_k) \tag{13}$$

and the evidence is

$$p(\mathbf{X}|\alpha) = \int_{\mathbf{p}} p(\mathbf{X}, \mathbf{p}|\alpha) \tag{14}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(N_k + \alpha_k)}{\Gamma(\sum_k N_k + \alpha_k)} \int_{\mathbf{p}} \mathcal{D}(\mathbf{p}; N_k + \alpha_k) \tag{15}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(N_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{16}$$

which is the probability that this data all came from one multinomial distribution.

For example:

$$p(\mathbf{X}|\alpha_k = 1/2) = \frac{\Gamma(K/2)}{\Gamma(N + K/2)\Gamma(1/2)^K} \prod_k \Gamma(N_k + 1/2) \qquad \text{(with Jeffreys' prior)} \tag{17}$$

$$
\begin{aligned}
p(\mathbf{X}|\alpha_k = 1) &= \frac{\Gamma(K)}{\Gamma(N+K)} \prod_k \Gamma(N_k + 1) \qquad \text{(with a uniform prior)} &(18)\\
&= \frac{1}{\binom{N}{N_1..N_K}\binom{N+K-1}{K-1}} &(19)
\end{aligned}
$$

The posterior predictive distribution is

$$
\begin{aligned}
p(x = k|\mathbf{X}, \alpha) &= E[p_k|\mathbf{X}] &(20)\\
&= \frac{N_k + \alpha_k}{N + \sum_k \alpha_k} &(21)\\
&= \frac{N_k + 1/2}{N + K/2} \qquad \text{(with Jeffreys' prior)} &(22)\\
&= \frac{N_k + 1}{N + K} \qquad \text{(with a uniform prior)} &(23)
\end{aligned}
$$

By contrast, the maximum of the posterior is

$$
\begin{aligned}
\hat{p}_k &= \frac{N_k + \alpha_k - 1}{N - K + \sum_k \alpha_k} &(24)\\
&= \frac{N_k - 1/2}{N - K/2} \qquad \text{(with Jeffreys' prior)} &(25)\\
&= \frac{N_k}{N} \qquad \text{(with a uniform prior)} &(26)
\end{aligned}
$$

This is one example, among many, where the maximum a posteriori estimate can be worse than the maximum likelihood estimate, even when the prior is correct. Using the posterior predictive distribution to represent our knowledge of $\mathbf{p}$ was the main argument of Bayes (1763). Formula (20) is sometimes mistakenly interpreted as saying that one should *always* use the expected value of a parameter as the "Bayesian estimate," which forgets that the posterior predictive distribution is not always an expectation as it is for multinomials.

For predicting multiple samples, we let M_k be the counts for the new sample and get

$$
\begin{aligned}
p(\mathbf{Y}|\mathbf{X}, \alpha) &= p(\mathbf{Y}, \mathbf{X}|\alpha)/p(\mathbf{X}|\alpha) &(27)\\
&= \frac{\Gamma(N + \sum_k \alpha_k)}{\Gamma(M + N + \sum_k \alpha_k)} \prod_k \frac{\Gamma(M_k + N_k + \alpha_k)}{\Gamma(N_k + \alpha_k)} &(28)
\end{aligned}
$$

i.e. it is just as if the prior were $\mathcal{D}(N_k + \alpha_k)$. Note that the new samples are not independent: we cannot use the posterior predictive distribution for one sample (equation (21)) as though it was the true distribution.

2 Samples vs. Counts

There is some confusion in the literature about when to use (2) versus (4). The preceding text used (2), which is the correct choice for most tasks. The confusion stems from the misconception that "the samples are equivalent to their counts." The samples are not equivalent to their counts because the samples occurred in some particular order. Every possible ordering has the same probability but that doesn't mean we should sum over orderings: the data only had one ordering.

For example, consider tossing a fair coin twice. The sample $< H, T >$ has probability 1/4, as does $< H, H >$. However, the count event $\{N_H = 1, N_T = 1\}$ has probability 1/2, which is greater than the probability of the count event $\{N_H = 2\}$. The discrepancy arises because we are considering a different event space. If someone told you that the sample had one head and one tail and asked for the probability of *that sample* (not its counts), then the answer would be 1/4, since every sample with those counts has probability 1/4.

As another example, suppose we have N samples from a univariate Gaussian distribution. The N-dimensional joint density of the samples only depends on the sample mean and sample variance of the sample; these are sufficient statistics. Every sample with the same sufficient statistics has the same probability. However, the joint density of the sufficient statistics has only two dimensions, with a different normalizing term. Asking for the probability of a sample with given sufficient statistics is not the same as asking for the probability *of those statistics*, which concerns a different event space.

3 Evidence vs. Entropy

When the parameter vector $\mathbf{p}$ is uniform, it is well-known (Bishop, 1995) that the log-probability of a set of counts can be approximated by the entropy of the ML estimate, i.e.

$$\begin{aligned}
\log \Gamma(x+1) &\approx x \log(x) - x & (29)\\
\log p(N_1..N_K | p_k = 1/K) &= \log \binom{N}{N_1..N_K} - N \log K & (30)\\
&\approx N \log N - N - \sum_k (N_k \log N_k - N_k) - N \log K & (31)\\
&= -\sum_k N_k \log \frac{N_k}{N} - N \log K & (32)\\
&= N\mathcal{H}(N_k/N) - N \log K & (33)\\
\text{where } \mathcal{H}(p) &= -\sum_k p_k \log p_k & (34)
\end{aligned}$$

supporting the intuition that more entropic counts are more probable. Here it is crucial that we compute the probability of counts; the probability of any particular sample is constant.

There is a corresponding result for the *evidence* of a sample, when the parameter *prior* is uniform ($\alpha_k = 1$). For a large sample, the evidence is approximately the *negative* entropy of the ML estimate:

$$
\begin{aligned}
\frac{\Gamma(K)}{\Gamma(N+K)} &\approx \frac{\Gamma(K)}{\Gamma(N+1)N^{K-1}} \approx \frac{1}{\Gamma(N+1)} && (35)\\
\log p(\mathbf{X}|\alpha_k = 1) &\approx -N\log N + N + \sum_k (N_k \log N_k - N_k) && (36)\\
&= \sum_k N_k \log \frac{N_k}{N} && (37)\\
&= -N\mathcal{H}(N_k/N) && (38)
\end{aligned}
$$

That is, less entropic samples are more probable. This is also clear from (19), which has the multinomial coefficient $\binom{N}{N_1..N_K}$ in the denominator rather than in the numerator. The reason is that virtually all parameter vectors $\mathbf{p}$ exert a bias toward some particular outcome; a homogeneous sample which contains only that outcome will be the most probable sample from that parameter vector. So when we integrate over all $\mathbf{p}$, we are left with a huge preference for homogeneous samples. This didn't happen before because we deliberately excluded all of these biased parameter vectors.

Here it is crucial that we compute the evidence of the particular sample; the counts have constant evidence (which is clear from (19)).

4 Mutual Information

Another information-theoretic quantity is mutual information, which arises in hypothesis testing. Suppose we want to know the probability that two discrete random variables are independent. We are given N (x, y) pairs, where x has J possible values and y has K possible values. The data can be arranged in a table as in figure 1.

	$y=1$	$y=2$	$y=3$
$x=1$	15	29	14
$x=2$	46	83	56

Figure 1: Joint outcomes of two random variables x and y. Each cell reports the number of times we saw the pair $(x = j, y = k)$. What is the probability that the two variables are independent?

Under the hypothesis of independence, the probability of the data is

$$p(D|\text{indep}) = p(\mathbf{X}|\alpha_{j.})p(\mathbf{Y}|\alpha_{.k}) \quad (39)$$

($\alpha_{j.}$ and $\alpha_{.k}$ are two different prior vectors). The other hypothesis is that the variables are dependent and arise from a multinomial distribution on pairs (x, y). This distribution has JK possible values. Under this hypothesis, the probability of the data is

$$p(D|\text{dep}) = p((x_1, y_1), ..., (x_N, y_N)|\alpha_{jk}) \tag{40}$$

(α_{jk} is yet a third prior). What we want to know is

$$p(\text{indep}|D) = \frac{p(D|\text{indep})p(\text{indep})}{p(D|\text{indep})p(\text{indep}) + p(D|\text{dep})p(\text{dep})} \tag{41}$$

$$= \frac{1}{1 + \frac{p(D|\text{dep})}{p(D|\text{indep})}\frac{p(\text{dep})}{p(\text{indep})}} \tag{42}$$

The crucial quantity here is $\frac{p(D|\text{indep})}{p(D|\text{dep})}$, the evidence ratio in favor of independence. Let

$$N_{jk} = \sum_{i=1}^{N} \delta(x_i = j)\delta(y_i = k) \tag{43}$$

$$N_{j.} = \sum_{i=1}^{N} \delta(x_i = j) = \sum_{k=1}^{K} N_{jk} \qquad N_{.k} = \sum_{i=1}^{N} \delta(y_i = k) = \sum_{j=1}^{J} N_{jk} \tag{44}$$

$$\alpha_{j.} = \sum_{k=1}^{K} \alpha_{jk} \qquad \alpha_{.k} = \sum_{j=1}^{J} \alpha_{jk} \tag{45}$$

then

$$\frac{p(\mathbf{X}|\alpha)p(\mathbf{Y}|\alpha)}{p((x_1, y_1), ..., (x_N, y_N)|\alpha)} = \frac{\Gamma(\sum_{jk} \alpha_{jk})}{\Gamma(N + \sum_{jk} \alpha_{jk})} \prod_{j=1}^{J} \frac{\Gamma(N_{j.} + \alpha_{j.})}{\Gamma(\alpha_{j.})} \prod_{k=1}^{K} \frac{\Gamma(N_{.k} + \alpha_{.k})}{\Gamma(\alpha_{.k})} \prod_{jk} \frac{\Gamma(\alpha_{jk})}{\Gamma(N_{jk} + \alpha_{jk})} \tag{46}$$

Using the entropy approximation (38), the logarithm of this ratio is

$$\log \frac{p(D|\text{indep})}{p(D|\text{dep})} \approx -N\mathcal{H}(\frac{N_{j.}}{N}) - N\mathcal{H}(\frac{N_{.k}}{N}) + N\mathcal{H}(\frac{N_{jk}}{N}) \tag{47}$$

$$= -N\mathcal{D}(\frac{N_{jk}}{N} \mid\mid \frac{N_{j.}}{N} \times \frac{N_{.k}}{N}) \tag{48}$$

$$= -N\mathcal{I}(x, y) \tag{49}$$

$$\text{where } \mathcal{D}(p \mid\mid q) = \sum_k p_k \log \frac{p_k}{q_k} \tag{50}$$

It measures the information we gain about y from knowing x; if we gain a lot of information, then the variables are dependent. This result was obtained in a similar way by Wolf (1994); he used an event space of counts instead of samples but the answers are the same. Unfortunately, the mutual information score (49) is biased toward the hypothesis of dependence; it can never be positive and reaches zero only if the empirical joint distribution factors precisely into the marginals, a case in which (46) would strongly favor independence. Mutual information can still be used for testing; it's just that the result cannot be interpreted as a probability. The orthodox literature also uses mutual information to test for independence; see e.g. the "G-test" of Sokal & Rohlf (1969; sec 16.4). But the orthodox derivation and interpretation is quite different. In fact, many books consider the mutual information score to be arbitrary and prefer instead the χ^2 score, which is the same as (48) but using

$$\mathcal{D}(p \mid\mid q) \approx \sum_k \frac{(p_k - q_k)^2}{2q_k} \tag{51}$$

With $\alpha_{jk} = 1$ and $p(\text{indep}) = 1/2$ on the table in figure 1, the logarithm of (46) is 3.28, which favors independence by $26:1$. The mutual information and χ^2 tests yield values of -0.43 and -0.42, respectively.

A closely related problem is the test for homogeneity: we want to know the probability that two samples $\mathbf{X}$ and $\mathbf{Y}$ came from the same multinomial distribution vs. different multinomial distributions. That is, we want

$$\begin{aligned} p(\text{same}|\mathbf{X},\mathbf{Y}) &= \frac{p(\mathbf{X},\mathbf{Y}|\text{same})p(\text{same})}{p(\mathbf{X},\mathbf{Y}|\text{same})p(\text{same}) + p(\mathbf{X},\mathbf{Y}|\text{different})p(\text{different})} && (52) \\ &= \frac{1}{1 + \frac{p(\mathbf{X},\mathbf{Y}|\text{different})}{p(\mathbf{X},\mathbf{Y}|\text{same})}\frac{p(\text{different})}{p(\text{same})}} && (53) \end{aligned}$$

The quantity $\frac{p(\mathbf{X},\mathbf{Y}|\text{different})}{p(\mathbf{X},\mathbf{Y}|\text{same})}$ is the evidence ratio in favor of difference, which is

$$\frac{p(\mathbf{X}|\alpha)p(\mathbf{Y}|\alpha)}{p(\mathbf{X},\mathbf{Y}|\alpha)} = \frac{\Gamma(\sum_k \alpha_k)\Gamma(M+N+\sum_k \alpha_k)}{\Gamma(M+\sum_k \alpha_k)\Gamma(N+\sum_k \alpha_k)} \prod_k \frac{\Gamma(M_k+\alpha_k)\Gamma(N_k+\alpha_k)}{\Gamma(\alpha_k)\Gamma(M_k+N_k+\alpha_k)} \tag{54}$$

This formula was also examined by Wolpert (1995). Using the entropy approximation (38), the logarithm of this ratio is

$$\begin{aligned} \log \frac{p(\mathbf{X}|\alpha)p(\mathbf{Y}|\alpha)}{p(\mathbf{X},\mathbf{Y}|\alpha)} &\approx -M\mathcal{H}(\frac{M_k}{M}) - N\mathcal{H}(\frac{N_k}{N}) + (M+N)\mathcal{H}(\frac{M_k+N_k}{M+N}) && (55) \\ &= M\mathcal{D}(\frac{M_k}{M} \mid\mid \frac{M_k+N_k}{M+N}) + N\mathcal{D}(\frac{N_k}{N} \mid\mid \frac{M_k+N_k}{M+N}) && (56) \\ \text{where } \mathcal{D}(p \mid\mid q) &= \sum_k p_k \log \frac{p_k}{q_k} && (57) \end{aligned}$$

This "average divergence to the mean" is known in information theory as Jensen-Shannon divergence (Lin, 1991) (El-Yaniv et al, 1997):

$$\begin{aligned} \mathcal{D}_\lambda(p_1,p_2) &= \lambda\mathcal{D}(p_1 \mid\mid q) + (1-\lambda)\mathcal{D}(p_2 \mid\mid q) && (58) \\ q &= \lambda p_1 + (1-\lambda)p_2 && (59) \end{aligned}$$

Another way to solve this problem is to write it as a test for independence. Let the random variable $c \in \{1, 2\}$ indicate which data set an observation x came from. If the two distributions are the same, then c and x are independent. So we can apply (46) to the following table:

	$x = 1$	$x = 2$	$\cdots$	$x = K$
$c = 1$	N_1	N_2	$\cdots$	N_K
$c = 2$	M_1	M_2	$\cdots$	M_K

The result is slightly different from (54), because of different Dirichlet priors. But the entropy approximation is the same and reduces to the mutual information between x and c.

5 Discrete Markov chains

A Markov chain is a sequence of random variables x_i such that the conditional distribution $p(x_i|x_{i-1}, ..., x_1)$ is actually $p(x_i|x_{i-1})$. The present is sufficient to determine the future. In a discrete Markov chain, the conditional distribution is multinomial. The probability of a sequence $\mathbf{X} = \ < x_1, ..., x_N >$ is therefore

$$p(\mathbf{X}|\mathbf{P}) = p(x_1) \prod_{i=2}^{N} p(x_i|x_{i-1}, \mathbf{P}) \tag{60}$$

where the matrix $\mathbf{P}$ denotes the parameters of the conditional distribution. Another way to write this involves partitioning the x_i according to the value that preceded them, i.e.

$$\mathbf{X} = \{x_1\} \cup \mathbf{X}^1 \cup \mathbf{X}^2 \cup ... \cup \mathbf{X}^K \tag{61}$$

$$\text{where } \mathbf{X}^k = \{x_i | x_{i-1} = k\} \tag{62}$$

Define $N_{.k}$ to be the cardinality of $\mathbf{X}^k$. Each member of $\mathbf{X}^k$ was independently chosen from the multinomial distribution $p(x_i|x_{i-1} = k, \mathbf{P})$. This distribution is parameterized by K numbers which we will store in $\mathbf{p}_k$, the kth column of $\mathbf{P}$. In other words, $p(x_i = j|x_{i-1} = k, \mathbf{P}) = p_{jk}$ by definition. Using this notation, we get

$$p(\mathbf{X}|\mathbf{P}) = p(x_1) \prod_{k=1}^{K} p(\mathbf{X}^k|\mathbf{p}_k) \tag{63}$$

$$= p(x_1) \prod_{k=1}^{K} \prod_{j=1}^{K} p_{jk}^{N_{jk}} \tag{64}$$

$$\text{where } N_{jk} = \sum_{i=2}^{N} \delta(x_i = j)\delta(x_{i-1} = k) \tag{65}$$

A conjugate prior for $\mathbf{p}_k$ is

$$p(\mathbf{p}_k) \sim \mathcal{D}(\alpha_{1k}, ..., \alpha_{Kk}) = \mathcal{D}(\alpha_{jk}) \qquad \text{(as shorthand)} \tag{66}$$

The prior for the whole matrix $\mathbf{P}$ is therefore

$$
\begin{aligned}
p(\mathbf{P}) &= \prod_k p(\mathbf{p}_k) && (67)\\
&= \frac{\prod_k \Gamma(\sum_j \alpha_{jk})}{\prod_{jk} \Gamma(\alpha_{jk})} \prod_{jk} p_{jk}^{\alpha_{jk}-1} && (68)
\end{aligned}
$$

The joint distribution comes from combining (64) and (68):

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{P}|\alpha) &= p(x_1) \prod_{k=1}^{K} p(\mathbf{X}^k, \mathbf{p}_k|\alpha) && (69)\\
&= p(x_1) \frac{\prod_k \Gamma(\sum_j \alpha_{jk})}{\prod_{jk} \Gamma(\alpha_{jk})} \prod_{jk} p_{jk}^{N_{jk}+\alpha_{jk}-1} && (70)
\end{aligned}
$$

so the posterior is

$$
\begin{aligned}
p(\mathbf{P}|\mathbf{X}, \alpha) &= \prod_k p(\mathbf{p}_k|\mathbf{X}^k, \alpha) && (71)\\
&= \prod_k \mathcal{D}(N_{1k} + \alpha_{1k}, ..., N_{Kk} + \alpha_{Kk}) && (72)\\
&= \prod_k \mathcal{D}(N_{jk} + \alpha_{jk}) \qquad \text{(shorthand)} && (73)
\end{aligned}
$$

and the evidence is

$$
\begin{aligned}
p(\mathbf{X}|\alpha) &= \int_{\mathbf{P}} p(\mathbf{X}, \mathbf{P}|\alpha) && (74)\\
&= p(x_1) \prod_{k=1}^{K} \int_{\mathbf{p}_k} p(\mathbf{X}^k, \mathbf{p}_k|\alpha) && (75)\\
&= p(x_1) \prod_{k=1}^{K} p(\mathbf{X}^k|\alpha) && (76)\\
&= p(x_1) \prod_{k=1}^{K} \left(\frac{\Gamma(\sum_j \alpha_{jk})}{\Gamma(N_{.k} + \sum_j \alpha_{jk})} \prod_j \frac{\Gamma(N_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})} \right) && (77)
\end{aligned}
$$

which is a product of ordinary multinomial evidences.

The posterior predictive distribution is

$$
\begin{aligned}
p(x_{N+1} = j | x_N = k, \mathbf{X}, \alpha) &= E[p_{jk}|\mathbf{X}] && (78)\\
&= \frac{N_{jk} + \alpha_{jk}}{N_{.k} + \sum_j \alpha_{jk}} && (79)
\end{aligned}
$$

For predicting multiple samples, we let M_{jk} be the counts for the new sample and get

$$\begin{aligned}
p(\mathbf{Y}|\mathbf{X},\alpha) &= p(\mathbf{Y},\mathbf{X}|\alpha)/p(\mathbf{X}|\alpha) && (80)\\
&= \prod_k p(\mathbf{Y}^k|\mathbf{X}^k,\alpha_{jk}) && (81)\\
&= \prod_k \left(\frac{\Gamma(\sum_j N_{jk}+\alpha_{jk})}{\Gamma(\sum_j M_{jk}+N_{jk}+\alpha_{jk})} \prod_j \frac{\Gamma(M_{jk}+N_{jk}+\alpha_{jk})}{\Gamma(N_{jk}+\alpha_{jk})} \right) && (82)
\end{aligned}$$

which is a product of ordinary posterior predictive distributions.

5.1 Testing for Markovianity

The hypothesis testing technique in section 4 can be used to determine if we have a true Markov chain or just a sequence of IID random variables. Under the hypothesis of independence, the sets $\mathbf{X}^k$ all came from the same distribution. Therefore this question is identical to the "same vs. different" question examined in section 4. Let the prior for the single distribution be Dirichlet with parameters β_j. Then the evidence for independence comes from applying (16):

$$\begin{aligned}
N_{j.} &= \sum_i \delta(x_i = j) = \sum_k N_{jk} && (83)\\
p(\mathbf{X}|\beta,\text{independence}) &= p(x_1)\frac{\Gamma(\sum_j \beta_j)}{\Gamma(N-1+\sum_j \beta_j)} \prod_j \frac{\Gamma(N_{j.}+\beta_j)}{\Gamma(\beta_j)} && (84)
\end{aligned}$$

If $\alpha_{jk} = \beta_j = 1$, then the evidence ratio in favor of independence is

$$\frac{p(\mathbf{X}|\beta_j = 1,\text{independence})}{p(\mathbf{X}|\alpha_{jk}=1)} = \frac{\prod_k \Gamma(N_{.k}+K)}{\Gamma(K)^{K-1}\Gamma(N-1+K)} \prod_j \frac{\Gamma(N_{j.}+1)}{\prod_k \Gamma(N_{jk}+1)} \qquad (85)$$

For example, suppose $\mathbf{X} = < 1, 2, 1, 2, ..., 1, 2 >$. The counts are

$$N_{jk} = \begin{bmatrix} 0 & N/2-1 \\ N/2 & 0 \end{bmatrix} \quad N_{j.} = \begin{bmatrix} N/2-1 \\ N/2 \end{bmatrix} \qquad (86)$$

and the evidence ratio is

$$\frac{\Gamma(N/2+K)\Gamma(N/2-1+K)}{\Gamma(K)^{K-1}\Gamma(N-1+K)} \qquad (87)$$

which is extremely small as K and N increase, implying the variables must be dependent.

Using the entropy approximation, we can approximate the logarithm of the evidence ratio as

$$-(N-1)\mathcal{H}(\frac{N_{j.}}{N-1}) + \sum_{k=1}^{K} N_{.k}\mathcal{H}(\frac{N_{jk}}{N_{.k}}) = \sum_{k=1}^{K} N_{.k}\mathcal{D}(\frac{N_{jk}}{N_{.k}} \mid\mid \frac{N_{j.}}{N-1}) \qquad (88)$$

which is again an "average divergence to the mean" that can be written as mutual information:

$$\log \frac{p(\mathbf{X}|\beta_j = 1, \text{independence})}{p(\mathbf{X}|\alpha_{jk} = 1)} \approx -(N-1)\mathcal{H}(x_i) + (N-1)\sum_{k=1}^{K} p(x_{i-1} = k)\mathcal{H}(x_i|x_{i-1} = k) \quad (89)$$
$$= -(N-1)\mathcal{H}(x_i) + (N-1)\mathcal{H}(x_i|x_{i-1}) \quad (90)$$
$$= -(N-1)\mathcal{I}(x_i, x_{i-1}) \quad (91)$$

If we gain a lot of information about x_i from knowing x_{i-1}, then it is probably a Markov chain.

References

[1] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc. London*, 53:370–418, 1763. Reprinted in *Biometrika*, 45:293–315 (1958), and in Press (1989).

[2] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press, 1995.

[3] Ran El-Yaniv, Shai Fine, and Naftali Tishby. Agnostic classification of markovian sequences. In *NIPS*, pages 465–471. MIT Press, 1997. `http://www.cs.huji.ac.il/labs/learning/Papers/MLT_list.html`.

[4] J. Lin. Divergence measures based on the Shannon entropy. *IEEE Trans Info Theory*, 37(1):145–151, 1991.

[5] R. R. Sokal and F. J. Rohlf. *Biometry; the principles and practice of statistics in biological research.* Freeman, 1969.

[6] David Wolf. Mutual information as a Bayesian measure of independence. `ftp://dino.ph.utexas.edu/wolf/Papers/MutInd.ps`, 1994.

[7] David H. Wolpert. Determining whether two data sets are from the same distribution. In *Maximum Entropy and Bayesian Methods*, 1995. `ftp://ftp.santafe.edu/pub/dhw_ftp/maxent.95.wv.ps.Z`.