Generalized Maximum Entropy

Peter Cheeseman * and John Stutz †

*NICTA, Locked Bag 6016, University of New South Wales, Kensington, NSW, 1466, Australia
†MS 269-1, NASA-Ames Research Center, Moffet Field, CA

Abstract.
A long standing mystery in using Maximum Entropy (MaxEnt) is how to deal with constraints whose values are uncertain. This situation arises when constraint values are estimated from data, because of finite sample sizes. One approach to this problem, advocated by E.T. Jaynes [1], is to ignore this uncertainty, and treat the empirically observed values as exact. We refer to this as the classic MaxEnt approach. Classic MaxEnt gives point probabilities (subject to the given constraints), rather than probability densities. We develop an alternative approach that assumes that the uncertain constraint values are represented by a probability density (e.g. a Gaussian), and this uncertainty yields a MaxEnt posterior probability density. That is, the classic MaxEnt point probabilities are regarded as a multidimensional function of the given constraint values, and uncertainty on these values is transmitted through the MaxEnt function to give uncertainty over the MaxEnt probabilities. We illustrate this approach by explicitly calculating the generalized MaxEnt density for a simple but common case, then show how this can be extended numerically to the general case. This paper expands the generalized MaxEnt concept introduced in a previous paper [2].

INTRODUCTION

A mystery in using Maximum Entropy (MaxEnt) inference in practice is: "Where do the constraints come from?". The normalization constraint (1) comes from the logical requirement that the sum of all probabilities must equal 1, since some event i out of a set of events must occur.

$$\forall i,\ 0 \leq P_i \leq 1, \qquad \sum_i P_i = 1, \tag{1}$$

However, other constraints, such as the mean value constraint discussed by Jaynes in [3] are just asserted, without specifying where they come from and how their constraint values are found. Two possible sources of constraints are:

1. By Definition: Such as the normalization constraint, or constraints derived from logical requirements, physical laws, or by assumption.
2. By Measurement: The value of these constraints is inherently uncertain.

In the first type of constraint, there is no uncertainty associated with the constraint values. However, for constraints based on measurements, the observed value can only be known with a precision dictated by the sample size. The Classic MaxEnt (CME) approach has no obvious mechanism for taking this uncertainty into account. Jaynes was well aware of this problem, and attempted to address it in his paper [1], at the end of section B, where he offered three approaches. The first is to ignore it–just use the

empirically determined constraint values as if they are exact. The second solution is to generalize the classic maxent approach to accommodate the constraint uncertainty–the approach we take in this paper, although Jaynes dismisses this approach as *ad hoc*. The third approach is to introduce constraint uncertainty by adding extra variance constraints. We show that the first and last approaches are untenable, while the second approach is dictated by the laws of probability.

THE CLASSIC MAXENT SOLUTION (CME)

The principle of Maximum Entropy (MaxEnt) is a method for using constraint information to find a set of point probability values, $\vec{P}$, that assumes the least (Shannon) information consistent with the given constraints. When the given (linear) constraints are insufficient to uniquely constrain $\vec{P}$ to particular point values, MaxEnt picks out the unique distribution that satisfies all the constraints and also maximizes the entropy.

In the case of a finite set of I mutually exclusive and exhaustive events, described by discrete probabilities P_i, the entropy is defined as:

$$\mathrm{H}(\vec{P}) = -\sum_{i=1}^{i=I} P_i \times LogP_i. \tag{2}$$

Given a set of J independent linear constraints, including the normalization (1), each of the form:

$$\vec{A}_j \cdot \vec{P} = C_j, \tag{3}$$

with $J < I$, the maximum entropy distribution may be found by the following procedure [3]: define the partition function:

$$\mathrm{Z}(\vec{\lambda}) \equiv \sum_{i=1}^{I} \exp(-\sum_{j=1}^{J} \lambda_j A_{ji}), \tag{4}$$

with the Lagrange multipliers $\vec{\lambda}$ determined by the set of J simultaneous equations:

$$\frac{\partial}{\partial \lambda_j} \log(\mathrm{Z}(\vec{\lambda})) + C_j = 0. \tag{5}$$

Then

$$\mathrm{H}_{\max} = \log(\mathrm{Z}(\vec{\lambda})) + \vec{\lambda} \cdot \vec{C}, \tag{6}$$

and the corresponding probability distribution is:

$$P_i = \mathrm{Z}(\vec{\lambda})^{-1} \exp(-\sum_{j=1}^{J} \lambda_j A_{ji}) = \mathrm{Z}(\vec{\lambda})^{-1} \prod_{=1}^{J} j \exp(-\lambda_j A_{ji}). \tag{7}$$

Explicit solutions for the dice case are given below. These CME values are different from the Maximum A Posterior (MAP), Maximum Likelihood (ML) and Posterior Mean probability estimators, reflecting the different assumptions built-in to each estimator (see [2] for a discussion of these assumptions).

GENERALIZED MAXENT

Equations (2)–(7) show that the maximum entropy point probabilities P_i are a *function* of the constraint values C_j. If these C_js are estimated from a sample, then Bayes implies that our knowledge of their values is approximate and can be expressed as pdfs (e.g., Gaussians), whose width deceases with increasing sample size. Using the Jacobian determinant of the function relating the maxent P_is to the C_js, the joint posterior pdf on the C_js can in principle be transformed into a joint pdf on the P_is through the maximum entropy constraint/function. We illustrate this process by a simple example using a three-faced dice with an experimentally determined mean value. For the three face dice, there are three unknown probabilities $\vec{P} = P_1, P_2, P_3$, one for each of the three faces. These probabilities must satisfy the following linear constraints (i.e, the normalization and mean values constraints):

$$P_1 + P_2 + P_3 = 1, \qquad \sum_{i=1}^{i=3} iP_i = P_1 + 2P_2 + 3P_3 = \mu, \tag{8}$$

where the value μ is only known approximately from data. Since all face values are either 1, 2 or 3, μ must be in the range 1 to 3. Here, the set of linear constraint equations (3) reduces to equations (8). In this simple case there is only one degree of freedom left, and this is removed when the additional maxent constraint (2) is also imposed. Using (7), the resulting P_is are:

$$P_i = \exp(-\kappa - i\lambda), \tag{9}$$

where κ is fixed by the normalization constraint and λ is fixed by the mean value constraint. Let $x = \exp(-\lambda)$ and $z = \exp(-\kappa)$. Substituting into (8), we have (after eliminating z):

$$x^2(\mu - 3) + x(\mu - 2) + \mu - 1 = 0, \tag{10}$$

whose only positive solution is:

$$x(\mu) = \frac{\mu - 2 + \sqrt{-3\mu^2 + 12\mu - 8}}{2(3 - \mu)}. \tag{11}$$

The normalizing constant (or partition function) $z(\mu)$ is found by substituting $x(\mu)$ from (11) into the normalizing equation to give:

$$z(\mu) = \frac{2\,(3-\mu)^3}{-14 + 14\,\mu - 3\,\mu^2 + (4-\mu)\,\sqrt{-8 + 12\,\mu - 3\,\mu^2}} \tag{12}$$

Putting these equations together gives the desire maxent probabilities:

$$P_i(\mu) = z(\mu)x(\mu)^i, \tag{13}$$

which in turn reduce to:

$$\begin{aligned} P_1(\mu) &= (10 - 3\,\mu - \sqrt{-8 + 12\,\mu - 3\,\mu^2})\ /\ 6, \\ P_2(\mu) &= (-1 + \sqrt{-8 + 12\,\mu - 3\,\mu^2})\ /\ 3, \\ P_3(\mu) &= (-2 + 3\,\mu - \sqrt{-8 + 12\,\mu - 3\,\mu^2})\ /\ 6. \end{aligned} \tag{14}$$

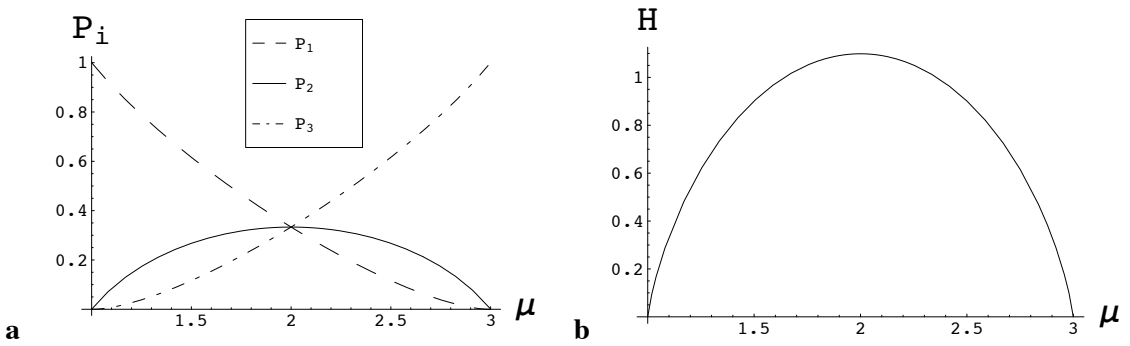

FIGURE 1. Maxent for the 3-faced die: **a** - Probabilities for faces as functions of μ. **b** - System entropy.

This is the maxent solution to the three-faced dice problem as a function of μ, and the resulting probabilities are shown in Fig. (1a). Note that for $\mu = 2, P_i = 1/3$ for all i, as expected, since $\mu = 2$ is the mean value for a "fair" dice, where all three faces are equally probable. Note also that the maxent solution precludes P_2 exceeding $1/3$, which has interesting consequences when a μ distribution is transformed to a set of P distributions.

The maxent probabilities (14) can now be substituted into the entropy formula (2) to give the three faced dice entropy as a function of μ, as shown in Fig. (1b). Note that the entropy has a maximum at $\mu = 2$, as expected, since this is the equiprobable entropy.

Extension to Distributions

Having found the maxent point probabilities $\vec{P}$ as a function of μ for the 3-faced dice problem, we now examine what happens if we do not know the value of μ exactly, but instead our knowledge is summarized as a pdf over μ–i.e. $f_\mu(\mu)$. From the vector calculus, when two coordinate systems $\bar{x}$ and $\bar{y}$ are related as $\bar{y} = \overline{Y}(\bar{x})$, integrals over any scalar function f are preserved if f is transformed as:

$$f(\bar{y}) = \pm |D| f(\overline{Y}(\bar{x})) = \pm \frac{\partial(x_1, x_2, \cdots, x_n)}{\partial(y_1, y_2, \cdots, y_n)} f(\overline{Y}(\bar{x})), \tag{15}$$

where $|D|$ is the Jacobian determinant of the inverse transformation $\bar{x} = \overline{X}(\bar{y})$, i.e.

$$|D| = \frac{\partial(x_1, x_2, \cdots, x_n)}{\partial(y_1, y_2, \cdots, y_n)} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \tag{16}$$

The sign must be chosen to preserve the sign of volume integrals, and in the case of a single variable, reduces to the absolute value.

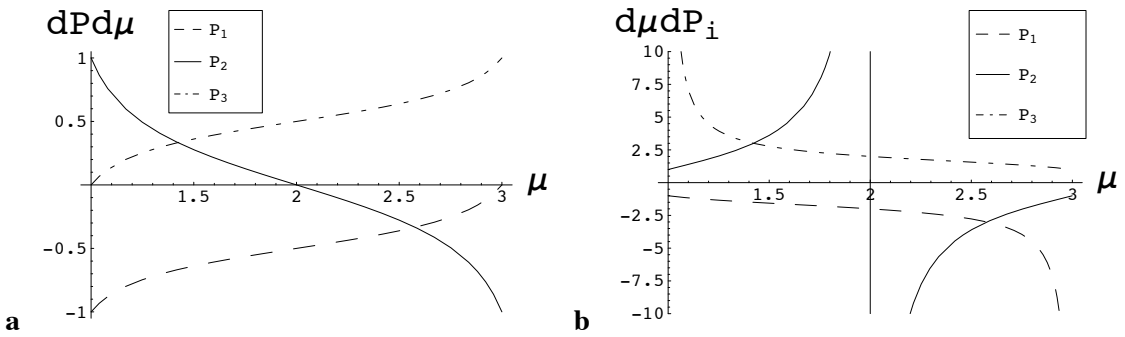

FIGURE 2. Partial derivatives for the 3-faced die: **a** - $\partial P_i/\partial\mu$. **b** - $\partial\mu/\partial P_i$.

For the three faced dice case, this is a one dimensional transformation for each P_i. For example, for the probability function $P_2(\mu)$, $|D_2|$ is derived from (14) giving:

$$|D_2| = \frac{\partial\mu}{\partial P_2} = \frac{\sqrt{-3\mu^2+12\mu-8}}{2-\mu} = \frac{1+3P_2}{\sqrt{(1+P_2)(1-3P_2)}}, \tag{17}$$

and similarly for P_1 and P_3. Note that (17) has a singularity at $\mu = 2$ and $P_2 = 1/3$, and that the absolute value must be taken so that $|D_2|$ is always positive. Combining (15) and (17), we get finally:

$$f_{P_2}(P_2) = \frac{\partial\mu}{\partial P_2} f_\mu(\mu(P_2)) = \frac{1+3P_2}{\sqrt{(1+P_2)(1-3P_2)}} f_\mu(\mu(P_2)), \tag{18}$$

which has the desired effect of mapping the uncertainty about μ onto uncertainty about P_2. The singularity in (17) occurs at the maximum value for $P_2 = 1/3$, as can be seen in Fig. (1a), but the resulting pdf $f_{P_2}(P_2)$ is still normalized, despite the infinite value at this boundary. In the extreme case where $f_\mu(\mu)$ becomes a delta function, the corresponding $f(P_i)$s also become delta functions, and give the CME result.

The $\mu(P_i)$ are obtained by inverting Equations (14). This is easy for P_1 and P_3, meerly a matter of choosing the branch that matches the μ range. However $P_2(\mu)$ is inherently non-invertable, so both branches must be used. Thus our $f_\mu(\mu(P_i))$ are:

$$\begin{aligned} f_\mu(\mu(P_1)) &= f_\mu(\frac{1}{2}(6-3P1-\sqrt{4P1-3P1^2})) \\ f_\mu(\mu(P_2)) &= f_\mu(2-\sqrt{1-2P2-3P2^2}) + f_\mu(2+\sqrt{1-2P2-3P2^2}) \\ f_\mu(\mu(P_3)) &= f_\mu(\frac{1}{2}(2+3P3+\sqrt{4P3-3P3^2})) \end{aligned} \tag{19}$$

Given that μ is known to bounded by *min* and *max* values, a suitable $f_\mu(\mu)$ is given by the Log-Odds-Normal (LON) distribution generalized to arbitray bounds. This is just the Normal applied to a variable transformed by the log odds w.r.t. min and max values. The version used here largely preserves the Normal mean (μ) and variance (σ^2) by substituting $logOdds(\mu)$ for μ and rescaling σ by $\partial logOdds[x]/\partial x\,|_\mu$. This assures that

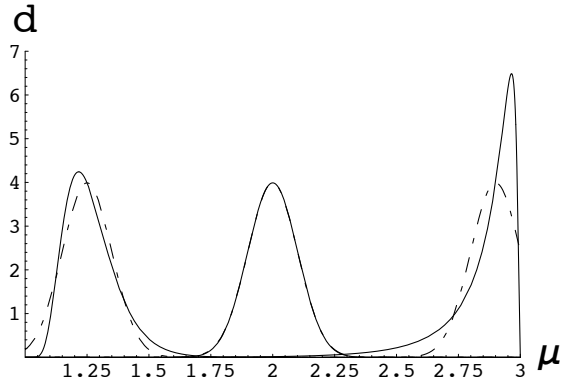

FIGURE 3. Superposition of three Log-Odds-Normal distributions (solid) with the equivalent Normals (dashed). $\mu \in \{1.25, 2.00, 2.90\}$, $\sigma = 0.1$, min = 1.0, max = 3.0. Note the long central tail for μ = 2.9. These are the three LON distributions used to generate figures (4 and 5).

$LON(x,\mu,\sigma,min,max) \simeq N(x,\mu,\sigma)$, so long as their common μ is between and several σ away from min and max.

$$LON(x,\mu,\sigma,min,max) = \tag{20}$$
$$\frac{(max-\mu)\,(min-\mu)}{\sqrt{2\pi}\sigma\,(max-x)\,(min-x)} \times$$
$$Exp(-\frac{(max-\mu)^2(\mu-min)^2}{2(max-min)^2\sigma^2}\log\left[\frac{(max-\mu)\,(min-x)}{(min-\mu)\,(max-x)}\right]^2)$$

Figure (3) illustrates the LON with three superimposed pairings of a LON and the corresponding Normal. The pairs are nearly indistinguishable when well away from the bounds. Unlike the Normal, The LON places no probability mass outside the bounds, but develops a long central tail when the mean approaches either bound. The match with the Normal breaks down everywhere when σ is a significant fraction of the range, and the LON distribution turns bimodal as σ approaches half the range.

For the 3-faced die, the distributions plotted in Figures (4 and 5) illustrate the results of transforming the Log-Odds-Normal μ distributions of Fig.(3) into P distributions via the maxent probability functions. Figure (4) shows most graphically the non-intuitive consequences of variable transformation using the maxent $P_2(\mu)$ function: (1) P_2 never has any mass density above $P_2 = 1/3$, and (2) any μ-space mass near $\mu = 2$ is transformed to a spike at and below $P_2 = 1/3$. This delta spike is always present, so long as the μ-space distribution has any density at $\mu = 2$, as the Log-Odds-Normal always does. A low mass spike is visible for the $\mu = 2.9$ case of Fig. (5b). Numerical integration, using Mathematica, confirms that these distributions are normalized to mass 1. While non-intuitive, these results are a straight forward consequence of the transformational calculus, applied to the $P_2(\mu)$ function, and were predictable from the shape of $P_2(\mu)$ in Fig. (1a).

The $\partial\mu/\partial P_1$ and $\partial\mu/\partial P_3$, see Fig. (2), are each finite everywhere except one end point. The P_1 and P_2 distributions in Fig. (5) show that these infinities are dominated by the Log-Odds-Normal zeros at min and max. This may not be the case with other

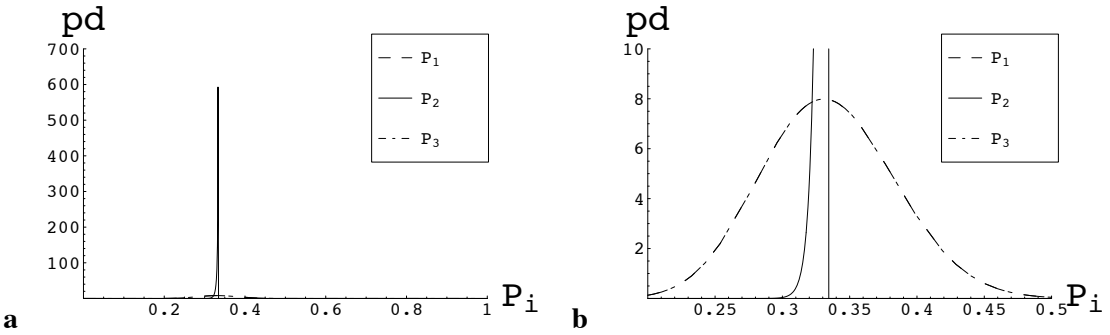

FIGURE 4. Two plots of the P_i distributions for mean(μ) = 2.00, and σ = 0.10. The full version (**a**) shows P_2's near delta function, generated by the large μ-space density at $\mu = 2$, while (**b**) shows that P_2 has no mass above $P_2 = 1/3$. The P_1 and P_3 distributions are effectively indistinguishable.

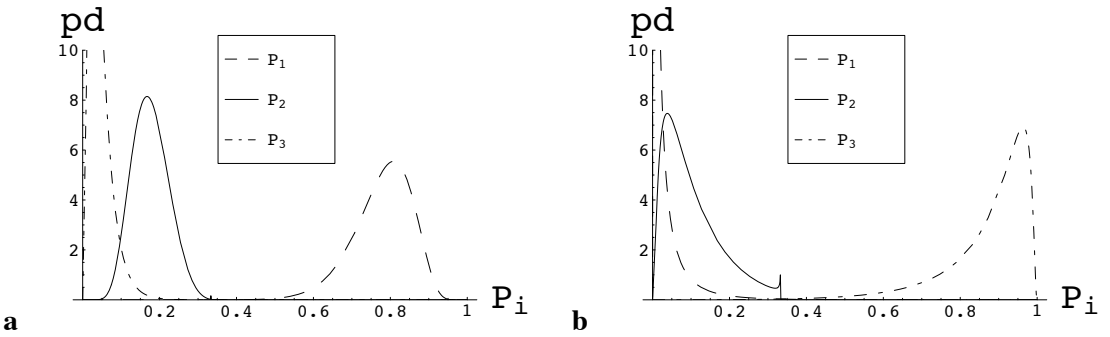

FIGURE 5. P_i distributions for mean(μ) of 1.25 **a**, and 2.90 **b**, with $\sigma = 0.1$. The small spike in **b** below $P_2 = 1/3$ is real, due to appreciable μ-space probability mass from the long LON tail, at $\mu = 2$.

μ-distributions. But here the P_1 and P_2 distributions are everywhere finite, numerical integration confirms their normality, and severe distortions occur only where the mean of μ approaches it bounds, and only for the P_i that is least likely.

For the above simple three faced dice case, it was possible to do the analysis explicitly. Higher order cases cannot be done explicitly because they involve analytic roots of high order polynomials. In such cases, it is relatively easy to approximate the Jacobian determinant with a Taylor expansion about the estimated constraint values. In the simplest case, this involves the linear/tangent plane approximation of the maxent probability functions as the constraint values are numerically varied around their maximum values. This should be a good approximation provided the maxent probability functions do not curve significantly over the range of constraint values. In the three faced die case, Fig. (1), the P_is can be seen to be approximately straight lines for small μ ranges, showing that the linear approximation would work well in this case.

DISCUSSION

We call this mapping of uncertainty in the constraint values into uncertainty in the MaxEnt probabilities (expressed as pdfs) the Generalized principle of Maximum Entropy (GME). We have not previously seen this generalization in the literature, but it is a direct result of applying probability theory to the situation. In the limit of sample size $N \to \infty$, the constraint values are given exactly, so GME becomes CME. Jaynes in [1] end of section B, briefly mentions an approach that resembles GME, but he calls this approach *ad hoc* and does not elaborate. Instead he advocates adding the uncertainty on constraint values as extra constraints, such as a variance (on the constraint values). In our three faced dice problem, this would require asserting a variance on μ. Jaynes did not develop this alternative approach, but if he had he would have discovered that it does not work, because the additional constraint(s) are on the wrong space! In CME, the constraints are on the space of possible probability values, P_i, not on the values of the constraints (such as μ), so his extra constraints would not have any direct effect on the P_is. Even if it was possible to translate constraints on constraint values into constraints on the underlying P_is, the resulting CME probabilities would be point probabilities, not pdfs, and so would not reflect the underlying uncertainty.

We stress that GME is only a solution to the problem of how to handle uncertain constraint values within the maximum entropy inference framework. The resulting pdfs avoid overly strong commitment compared with point probabilities. However, if the GME pdfs are used for prediction or decision making, it is important to remember that they embody other strong assumptions that may be incorrect in particular cases. The essential assumption is that the set of constraints used in either GME or CME is *complete*, meaning that no other significant constraints apply. Without good reason to believe in the constraint set's completeness, there is no good reason for believing the predictions of GME or CME. See [2] for further discussion on this point.

REFERENCES

1. Jaynes, E. T., "Where do We Stand on Maximum Entropy", in *The Maximum Entropy Formalism*, edited by R. D. Levine and M. Tribus, MIT Press, Cambridge, MA, USA, 1978, pp. 15–118, http://bayes.wustl.edu/.
2. Cheeseman, P. C., and Stutz, J. C., "On the Relationship between Bayesian and Maximum Entropy Inference", in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, American Insititute of Physics, Melville, NY, USA, 2004, pp. 445–461.
3. Jaynes, E. T., "Concentration of Distributions at Entropy Maxima", in *E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics*, edited by R. D. Rosenkrantz, D. Reidel, Dordrecht, 1983, chap. 11, pp. 315–336, http://bayes.wustl.edu/.