Choice of Basis for Laplace Approximation

David J.C. MacKay
Cavendish Laboratory, Cambridge, CB3 0HE,
United Kingdom. mackay@mrao.cam.ac.uk

Submitted to Machine Learning October 14th 1996
Accepted pending minor modifications February 23rd 1998
Revised version completed May 11th 1998
Published Volume 33, No. 1, October 1998

Abstract

Maximum a posteriori optimization of parameters and the Laplace approximation for the marginal likelihood are both basis-dependent methods. This note compares two choices of basis for models parameterized by probabilities, showing that it is possible to improve on the traditional choice, the probability simplex, by transforming to the 'softmax' basis.

1 Introduction

Laplace's method approximates the integral of a function $\int \mathrm{d}^k \mathbf{w} f(\mathbf{w})$ by fitting a Gaussian at the maximum $\hat{\mathbf{w}}$ of $f(\mathbf{w})$, and computing the volume under that Gaussian:

$$\int \mathrm{d}^k \mathbf{w} f(\mathbf{w}) \simeq f(\hat{\mathbf{w}}) (2\pi)^{k/2} \left| -\nabla\nabla \log f(\mathbf{w}) \right|^{-1/2} . \tag{1}$$

This method is widely used in probabilistic modelling to approximate the value of marginal likelihoods, which are of interest for model comparison (Ripley, 1996; Lindley, 1980; Smith and Spiegelhalter, 1980; MacKay, 1992; Chickering and Heckerman, 1996). In this paper I consider the case of models whose parameters are probabilities, for example, hidden Markov models, mixture models, belief networks and certain language models. I examine the neglected issue of the choice of *basis* in which the Laplace approximation is made.

It is well known that the location of the maximum of a density f is not invariant with respect to a non-linear reparameterization $\theta(\mathbf{w})$ of the parameters $\mathbf{w}$. Clearly the Laplace integral is not invariant either. Thus the choice of basis is important, but in many areas of statistical modelling, people rarely consider changing from the most obvious basis. Intuitively, if we are going to use Laplace's approximation we should try to reparameterize our model so that the density f is as near as possible to a Gaussian.

1.1 Models to be discussed

The likelihood function for a simple belief network with no hidden nodes is a product of factors, one for each unknown probability vector $\mathbf{p}$, of the form

$$P(\mathbf{F}|\mathbf{p}) = \prod_i p_i^{F_i}, \tag{2}$$

where $\mathbf{p}$ is a probability vector with I components, and $\mathbf{F}$ is a vector of counts F_i, the number of times that outcome i occurred when we sampled from the distribution $\mathbf{p}$. We will discuss the very simplest case of a network with one node which can take on I possible values, so that the likelihood is given by the one factor above. An example system to have in mind is a bent I–sided die whose probability of rolling an i is p_i. The unknown vector $\mathbf{p}$ is to be inferred from the outcome of F rolls in which face i came up F_i times, and we wish to calculate the marginal likelihood, which depends on the prior distribution over $\mathbf{p}$.

A popular prior for a probability vector $\mathbf{p}$ is the Dirichlet distribution (O'Hagan, 1994) parameterized by a measure $\mathbf{u}$ (a vector with all coefficients $u_i > 0$):

$$P(\mathbf{p}|\mathbf{u}) = \frac{1}{Z_{\rm Dir}(\mathbf{u})} \prod_{i=1}^{I} p_i^{u_i - 1} \delta\left(\textstyle\sum_i p_i - 1\right) \equiv \mathrm{Dirichlet}^{(I)}(\mathbf{p}|\mathbf{u}). \tag{3}$$

The function $\delta(x)$ is the Dirac delta function which simply restricts the distribution to the simplex such that $\mathbf{p}$ is normalized, *i.e.*, $\sum_i p_i = 1$; the distribution is restricted to non–negative p_is. The normalizing constant of the Dirichlet distribution is:

$$Z_{\rm Dir}(\mathbf{u}) = \prod_i \Gamma(u_i) \,/\, \Gamma(u) \;, \tag{4}$$

where we define $u = \sum_i u_i$. We will similarly define $F = \sum_i F_i$. The hyperparameter vector $\mathbf{u}$ controls how compact the prior distribution over $\mathbf{p}$ is. If $\mathbf{u}$ is large then the distribution over $\mathbf{p}$ is concentrated around the mean of the distribution, u_i/u. If all the components of $\mathbf{u}$ are small then extreme large and small probabilities are expected.

In the case of our bent die model, assuming a Dirichlet prior, the posterior probability of $\mathbf{p}$ given the data $\mathbf{F}$ is:

$$\begin{aligned} P(\mathbf{p}|\mathbf{F},\mathbf{u}) &= \frac{P(\mathbf{F}|\mathbf{p})P(\mathbf{p}|\mathbf{u})}{P(\mathbf{F}|\mathbf{u})} && (5)\\ &= \frac{\prod_i p_i^{F_i} \prod_i p_i^{u_i - 1}\, \delta(\sum_i p_i - 1)/Z_{\rm Dir}(\mathbf{u})}{P(\mathbf{F}|\mathbf{u})} && (6)\\ &= \mathrm{Dirichlet}^{(I)}(\mathbf{p}|\mathbf{F}+\mathbf{u}). && (7) \end{aligned}$$

The predictive distribution, that is, the probability that the next outcome will be an i, is given by

$$P(i|\mathbf{F},\mathbf{u}) = \int \mathrm{Dirichlet}^{(I)}(\mathbf{p}|\mathbf{F}+\mathbf{u})\, p_i \, \mathrm{d}^I\mathbf{p} \;= \frac{F_i + u_i}{F + u}. \tag{8}$$

Belief networks with hidden variables and mixture models have a likelihood function obtained by summation over the hidden variables; Dirichlet priors are also widely used for such models. The traditional MAP method for such models (Lee and Gauvain, 1993) is to maximize the posterior probability of the parameters $\mathbf{p}$, and the traditional Laplace method for such a model is, after maximizing in the $\mathbf{p}$ basis, to make the Gaussian approximation in the same basis (Chickering and Heckerman, 1996).

There is an obvious difficulty, namely that if a prior with $u_i < 1$ is used, then it is possible for the posterior density to diverge at the edge of the parameter space where at least one parameter p_i is equal to zero, if the sum of u_i and the effective count F_i for outcome i is less than 1. The Laplace approximation is only valid at the maximum of a smooth hump, so if this happens the

traditional method is in trouble. The traditional solution to this problem is to forbid the use of Dirichlet priors with any $u_i \leq 1$. However, as argued in (Jeffreys, 1939; MacKay and Peto, 1995; Gelman, 1996), there may be good reasons for expecting priors with $u_i < 1$ to be appropriate for many problems. I would argue that the '-1' terms in the traditional posterior probability are artefacts of the choice of basis.

2 A change of basis

I suggest that maximum *a posteriori* parameter estimation and Laplace approximations would be better conducted in the 'softmax' representation (widely used in neural networks (Bridle, 1989)) in which the parameters $\mathbf{p}$ are replaced by parameters $\mathbf{a}$:

$$p_i(\mathbf{a}) = \frac{\exp(a_i)}{\sum_{i'} \exp(a_{i'})}. \tag{9}$$

[Please do not confuse $\mathbf{p}(\mathbf{a})$, the function defined in equation (9), with the probability density $P(\mathbf{a})$.] The probability vector $\mathbf{p}$ has I components but only $I-1$ degrees of freedom; the sum of p_i must be 1. Similarly, $\mathbf{a}$ has I components, but it has a redundant degree of freedom: addition of an arbitrary multiple of $\mathbf{n} = (1, 1, 1, \ldots 1)^T$ to $\mathbf{a}$ leaves $\mathbf{p}(\mathbf{a})$ unchanged. We are free to constrain this degree of freedom however we wish.

In the softmax basis the Dirichlet prior may be written as

$$P(\mathbf{a}|\mathbf{u}) = \frac{1}{Z_S(\mathbf{u})} \prod_i p_i(\mathbf{a})^{u_i} g(\mathbf{a} \cdot \mathbf{n}), \tag{10}$$

where $g(\mathbf{a} \cdot \mathbf{n})$ is an arbitrary density constraining the redundant degree of freedom, and $Z_S(\mathbf{u})$ is the appropriate normalizing constant. This prior no longer has any '-1' terms in its exponents, because the Jacobian of the transformation from p-space to a-space is proportional to $\prod_i p_i$ (see appendix A).

It is as straightforward to evaluate derivatives and curvatures with respect to $\mathbf{a}$ as it is with respect to $\mathbf{p}$. Consider the log likelihood $L = \log \prod_i p_i^{F_i} = \sum_i F_i \log p_i$ for example. Defining $z = \sum_i \exp(a_i)$, we have

$$L = \sum_i F_i \left[a_i - \log z \right] = \sum_i F_i a_i - F \log z, \tag{11}$$

where $F = \sum_i F_i$. The derivative is

$$\frac{\partial L}{\partial a_i} = F_i - F p_i, \tag{12}$$

and the curvature is

$$\frac{\partial^2 L}{\partial a_i \partial a_j} = -F \frac{\partial p_i}{\partial a_j} = -F \left[\delta_{ij} p_i - p_i p_j \right]. \tag{13}$$

We note that the most probable $\mathbf{a}$ in this basis under the above likelihood and a Dirichlet prior $P(\mathbf{a}|\mathbf{u})$ is $\mathbf{a} = \mathbf{a}_{\rm MP}$ such that

$$p_i(\mathbf{a}_{\rm MP}) = \frac{F_i + u_i}{F + u}, \tag{14}$$

which we recognize as being equivalent to the predictive distribution (8). So in the softmax basis, unlike the traditional basis, the MAP parameter vector $\mathbf{p}$ is, conveniently, equal to the mean of $\mathbf{p}$.

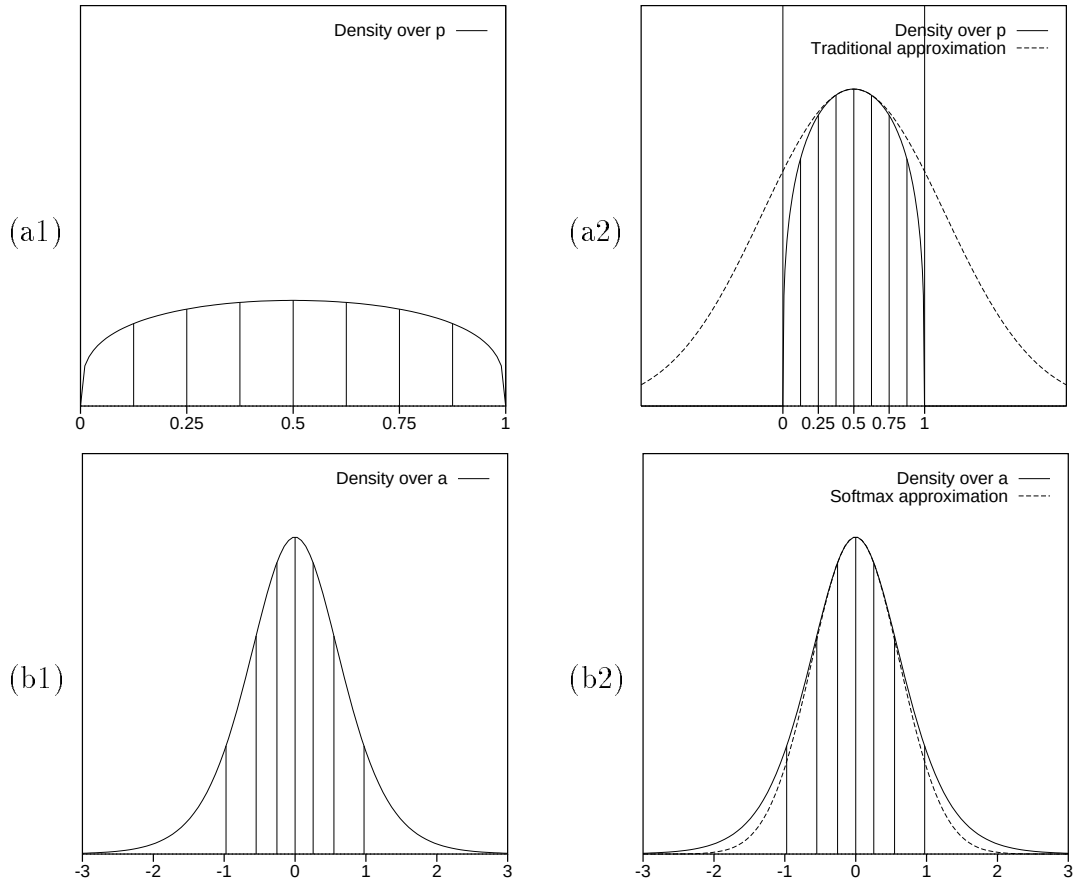

Figure 1: Intuition for the defect of the traditional Laplace approximation and the possible advantage of the softmax basis. This figure shows the case of a binary ($I = 2$) probability $(p, 1-p)$, which might alternatively be represented by the parameter a, where $p(a) = 1/(1 + e^{-2a})$. (a1) The density Dirichlet$^{(2)}(p|\mathbf{u} = (1.3, 1.3))$ in traditional p-space, $P(p) \propto p^{0.3}(1-p)^{0.3}$. (a2) The density, approximated by a Gaussian. Note how poorly the Gaussian tails match the true density. (b1) The same density transformed into a-space, $P(a) \propto p(a)^{1.3}(1-p(a))^{1.3}$. (b2) The Gaussian approximation in a-space. In both cases the Gaussian approximation is made by matching the first two derivatives of $\log P$ at the maximum. The Gaussians have variances $\sigma_p^2 = 1/(8(u_i - 1))$ and $\sigma_a^2 = 1/(2u_i)$ respectively, where $u_i = 1.3$. The vertical lines in (a) are equally spaced in p; the vertical lines in (b) show the corresponding values of a.

3 Comparison

Figure 1 gives an intuitive picture for why we might expect the softmax representation to be a superior representation in which to make Gaussian approximations. In the case $P[(p, 1-p)] = \text{Dirichlet}^{(2)}(\mathbf{p}|\mathbf{u} = (1.3, 1.3))$ the traditional Gaussian approximation puts more probability mass *outside* the interval $[0, 1]$ than inside. The curvature of the true density diverges at $p = 0$ and $p = 1$. As a function of a, in contrast, the density has no singularities. The Gaussian approximation in the softmax basis is still imperfect, however, in that the Dirichlet distribution falls exponentially for large $|a|$, so the softmax Gaussian approximation is too light-tailed.

We can make a quantitative comparison of the traditional Laplace approximation and the softmax Laplace approximation in a simple case where the exact marginal likelihood can be computed. This is the case, already discussed above, of inferring the probability vector $\mathbf{p}$ of a bent I-sided die from F rolls of the die. We will test the two approximations in cases where the prior is well matched to the source of the data and in cases where the prior and the data are at variance with each other.

3.1 Exact answer

If we assume a Dirichlet prior for $\mathbf{p}$ and observe F samples from $\mathbf{p}$, obtaining counts $\mathbf{F} = (F_1, F_2, \ldots F_I)$, the posterior probability of $\mathbf{p}$ is another Dirichlet distribution (equation (7)) and we can obtain the marginal likelihood $P(\mathbf{F}|\mathbf{u})$ from equation (4):

$$P(\mathbf{F}|\mathbf{u}) = \frac{Z_{\text{Dir}}(\mathbf{F}+\mathbf{u})}{Z_{\text{Dir}}(\mathbf{u})} = \frac{\prod_i \Gamma(F_i+u_i)}{\Gamma(F+u)} \frac{\Gamma(u)}{\prod_i \Gamma(u_i)}. \tag{15}$$

As in section 1, we define $u \equiv \sum_i u_i$.

3.2 Softmax Laplace approximation

Let us assume a Gaussian prior on the one unconstrained direction in $\mathbf{a}$ space so that the Dirichlet distribution on $\mathbf{a}$ is:

$$P(\mathbf{a}|\mathbf{u}, \epsilon) = \frac{1}{Z_S(\mathbf{u})} \prod_i p_i(\mathbf{a})^{u_i} \exp\left[-\epsilon \left(\mathbf{a}\cdot\mathbf{n}\right)^2/2\right]. \tag{16}$$

Let us approximate $Z_S(\mathbf{u})$ using Laplace's method. We know that the maximum is at $\hat{\mathbf{a}}$ such that the corresponding $\hat{p}_i$ is u_i/u. The curvature matrix $\mathbf{M}$ that we require is given by (c.f. equation (13))

$$M_{ij} = -\frac{\partial^2}{\partial a_i \partial a_j} \log P(\mathbf{a}|\mathbf{u}, \epsilon) = u \frac{\partial p_i}{\partial a_j} + \epsilon n_i n_j = u\left[\delta_{ij} p_i - p_i p_j\right] + \epsilon n_i n_j. \tag{17}$$

The determinant of $\mathbf{M}$ is readily evaluated using the identities

$$\left[\mathbf{A} + \beta \mathbf{g}\mathbf{g}^{\mathrm{T}}\right]^{-1} = \mathbf{A}^{-1} - \frac{\beta \mathbf{A}^{-1}\mathbf{g}\mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}}{1+\beta \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}} \qquad \text{and} \qquad \det\left[\mathbf{A} + \beta \mathbf{g}\mathbf{g}^{\mathrm{T}}\right] = (\det \mathbf{A})(1+\beta \mathbf{g}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{g}), \tag{18}$$

and we obtain

$$\begin{aligned} Z_S(\mathbf{u}) &\simeq \prod_i \hat{p}_i^{u_i} (2\pi)^{I/2} \det{}^{-1/2}\mathbf{M} && (19)\\ &= (2\pi)^{I/2} \prod_i \hat{p}_i^{u_i} \Big/ \left(\epsilon I^2 u^{I-1} \prod_i \hat{p}_i\right)^{1/2} && (20)\\ &= \kappa \frac{\prod_i \hat{p}_i^{u_i - \frac{1}{2}}}{u^{(I-1)/2}} = \kappa \frac{\prod_i u_i^{u_i - \frac{1}{2}}}{u^{u - \frac{1}{2}}}, && (21) \end{aligned}$$

where $\kappa = (2\pi)^{I/2}/(\epsilon^{1/2} I)$ is independent of $\mathbf{u}$.

We are now ready to approximate the marginal likelihood:

$$P(\mathbf{F}|\mathbf{u}) \simeq Z_S(\mathbf{F}+\mathbf{u})/Z_S(\mathbf{u}) \tag{22}$$

$$= \frac{\prod_i (F_i+u_i)^{F_i+u_i-\frac{1}{2}}}{(F+u)^{F+u-\frac{1}{2}}} \frac{u^{u-\frac{1}{2}}}{\prod_i u_i^{u_i-\frac{1}{2}}}. \tag{23}$$

3.3 Traditional Laplace approximation

We need the version of the Laplace approximation in which the density is multiplied by a delta function:

$$\int \mathrm{d}^k\mathbf{x}\, \delta(\mathbf{x}\cdot\mathbf{n}) e^{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}} = (2\pi)^{\frac{k-1}{2}} \frac{\det{}^{-\frac{1}{2}}\mathbf{A}}{\sqrt{\mathbf{n}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{n}}}. \tag{24}$$

We now approximate $P(\mathbf{F}|\mathbf{u})$:

$$P(\mathbf{F}|\mathbf{u}) = \int \mathrm{d}^I\mathbf{p}\; P(\mathbf{F}|\mathbf{p}) P(\mathbf{p}|\mathbf{u}) \tag{25}$$

$$\simeq \prod_i \hat{p}_i^{F_i+u_i-1} (2\pi)^{(I-1)/2} \frac{\det{}^{-1/2}\mathbf{A}}{\sqrt{\mathbf{n}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{n}}} \frac{1}{Z_{\mathrm{Dir}}(\mathbf{u})} \tag{26}$$

where $\hat{p}_i = (F_i+u_i-1)/(F+u-I)$ and

$$A_{ij} = -\frac{\partial^2}{\partial p_i \partial p_j} \sum_i (F_i+u_i-1)\log p_i \Big|_{\mathbf{p}=\hat{\mathbf{p}}} \tag{27}$$

$$= \delta_{ij}(F_i+u_i-1)\frac{1}{\hat{p}_i^2} \tag{28}$$

$$= (F+u-I)\delta_{ij}\frac{1}{\hat{p}_i} \tag{29}$$

so that

$$\mathbf{n}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{n} = 1/(F+u-I) \tag{30}$$

and

$$\det{}^{-\frac{1}{2}}\mathbf{A} = (F+u-I)^{-I/2} \prod_i \hat{p}_i^{1/2}. \tag{31}$$

Thus the traditional Laplace method gives

$$P(\mathbf{F}|\mathbf{u}) \simeq \prod_i \hat{p}_i^{F_i+u_i-\frac{1}{2}} \frac{1}{(F+u-I)^{(I-1)/2}} \frac{(2\pi)^{(I-1)/2}}{Z_{\mathrm{Dir}}(\mathbf{u})} \tag{32}$$

$$= \frac{\prod_i (F_i+u_i-1)^{F_i+u_i-\frac{1}{2}}}{(F+u-I)^{F+u-\frac{1}{2}}} \frac{(2\pi)^{(I-1)/2}}{Z_{\mathrm{Dir}}(\mathbf{u})}. \tag{33}$$

3.4 Results

Four simple experiments were performed, with $I = 20$ in all cases.

In the first experiment, all u_i were set to 1, and a probability vector $\mathbf{p}$ was drawn from the corresponding Dirichlet distribution using the method described by Gelman et al., 1995. The vector $\mathbf{F}$ was then set to $N\mathbf{p}$ for a range of values of N, the effective number of data points (this fake data set thus has non-integer 'counts'). The three methods of evaluating $P(\mathbf{F}|\mathbf{u})$ were compared

as a function of N. This first experiment tests the ability of the methods to evaluate the marginal likelihood when all u_i are 1 and this prior is well matched to the data source. In the second experiment we keep the same data source, but change the model's prior to $u_i = 0.05$ for all i. This tests the accuracy of the methods for the case when the assumed prior expects a $\mathbf{p}$ that is more 'spiky' than the data source. The third experiment reverses the situation: the probability vector $\mathbf{p}$ is drawn from the Dirichlet distribution with $u_i = 0.05$, and we find the marginal likelihoods when the prior has $u_i = 1$ and thus expects an overly 'smooth' $\mathbf{p}$. Finally, the fourth experiment looks at the case $u_i = 0.05$ (source) and $u_i = 0.05$ (model).

The results are shown in figure 2. In the cases where the assumed $u_i = 0.05$, the traditional method is not plotted because it fails utterly on account of some components p_i having negative exponents. The traditional method is at its best when u_i is large (*e.g.*, 1), and the data come from a source that matches this prior, and the amount of data is large. The softmax method is superior for most parameter settings, but it is not globally superior: in the limit of large amounts of data in all bins i (where both methods perform well) the traditional approximation can be a little more accurate.

4 Discussion

This paper's aim is not to advocate the use of Laplace approximations; indeed a good case can be made for using other methods such as Markov chain Monte Carlo (see, for example, (Neal, 1992)). And deterministic Bayesian approximations that are *basis independent* are under development (MacKay, 1997). But if MAP methods *are* used, this paper offers a way of evaluating marginal likelihoods which satisfies these two desiderata:

1. We can make a Laplace approximation for any Dirichlet priors and any amount of data. Singularities will not be encountered, in contrast to the traditional MAP method.

2. In simple cases where the predictive distribution $\int \mathrm{d}\mathbf{p}\, P(\mathbf{p})\mathbf{p}$ is easy to compute, the maximum *a posteriori* parameters $\hat{\mathbf{p}}$ are equal to the predictive distribution.

In the cases where exact results are available this 'softmax' Laplace approximation is often much more accurate than the traditional approximation. It seems plausible that these advantages will carry over to the case of models with hidden variables.

The softmax parameterization also has the convenient property that any setting of the parameters $\mathbf{a}$ is valid; there are no constraints. This makes it easy to integrate models into computational packages such as optimizers.

Acknowledgements

I thank David Heckerman, Radford Neal, Virginia de Sa and Graeme Mitchison for helpful discussions.

A The Dirichlet prior in the softmax basis

We sketch the proof that the density over $\mathbf{a}$ given in equation (10),

$$P(\mathbf{a}|\mathbf{u}) = \frac{1}{Z_S(\mathbf{u})} \prod_i p_i(\mathbf{a})^{u_i} g(\mathbf{a}\cdot\mathbf{n}),$$

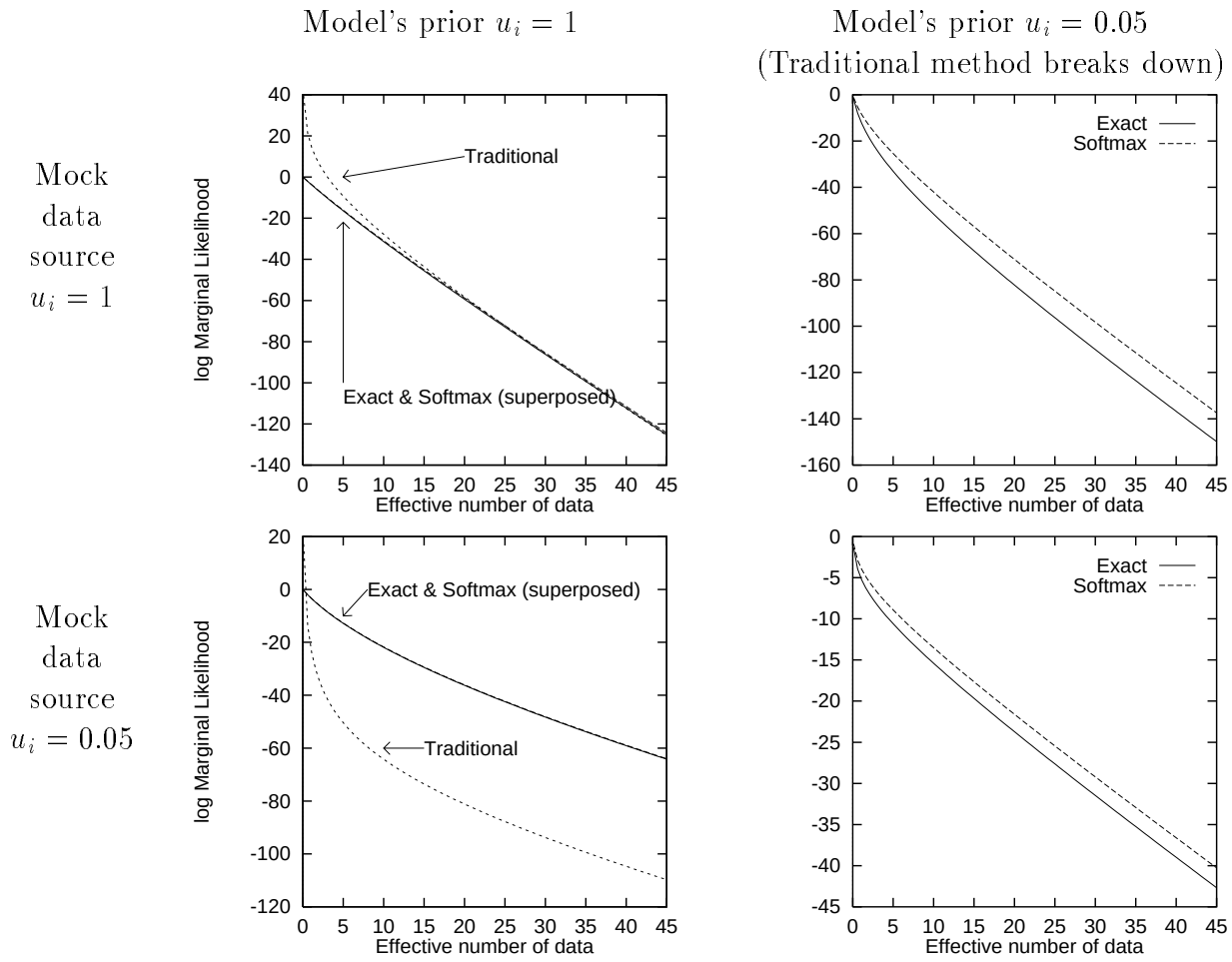

Figure 2: Comparison of the two Laplace approximations with the correct answer in four experimental situations (two data sources, and two priors.)
The probability vectors produced by the two sources are (with their components in rank order) for $u_i = 1$: $\mathbf{p} = [0.23, .17, .17, .074, .064, .040, .034, .034, .032, .026, .026, .025, .017, .016, .015, .010, .0082, .0038, .0027, .000057]$ and for $u_i = 0.05$, $\mathbf{p} = [.69, .29, .012, .00095, .00030, 7.5e\text{-}5, 7.5e\text{-}5, 5.2e\text{-}5, 3.9e\text{-}5, 1.0e\text{-}5, 9.1e\text{-}6, 2.8e\text{-}7, 8.0e\text{-}9, 1.2e\text{-}13, 1.7e\text{-}15, 6.3e\text{-}18, 6.2e\text{-}19, 6.6e\text{-}21, 7.7e\text{-}24, 5.3e\text{-}26]$.

does indeed transform into the Dirichlet distribution over $\mathbf{p}$. We start with the special case where $\mathbf{a}$ is confined to an $I-1$ dimensional subspace satisfying $\mathbf{a}\cdot\mathbf{n}=\alpha$, so that $g()$ above is a delta function. In this case we can represent $\mathbf{a}$ by an $I-1$ dimensional vector $\mathbf{b}$ thus:

$$\begin{aligned} a_i &= b_i \qquad\qquad i=1,2,\ldots,I-1 \\ a_I &= \alpha-\textstyle\sum_{i=1}^{I-1} b_i, \end{aligned} \tag{34}$$

and similarly we can represent $\mathbf{p}$ by an $I-1$ dimensional vector $\mathbf{q}$:

$$\begin{aligned} p_i &= q_i \qquad\qquad i=1,2,\ldots,I-1 \\ p_I &= 1-\textstyle\sum_{i=1}^{I-1} q_i, \end{aligned} \tag{35}$$

then we can find the density over $\mathbf{q}$ (which is proportional to the required density over $\mathbf{p}$) from the density over $\mathbf{b}$ (which is proportional to the given density over $\mathbf{a}$) by finding the determinant of the $(I-1)\times(I-1)$ Jacobian $\mathbf{J}$ given by

$$J_{ik}=\frac{\partial q_i}{\partial b_k}=\sum_{j=1}^{I}\frac{\partial p_i}{\partial a_j}\frac{\partial a_j}{\partial b_k}=\delta_{ik}p_i-p_ip_k+p_ip_I=p_i(\delta_{ik}-(p_k-p_I)). \tag{36}$$

Defining the $I-1$ dimensional vectors $p_k^+\equiv p_k-p_I$ and $n_k\equiv 1$, and using $\det\left[\mathbf{I}-\mathbf{x}\mathbf{y}^{\mathrm{T}}\right]=1-\mathbf{x}\cdot\mathbf{y}$, which is easily proved by changing basis such that $\mathbf{x}$ is aligned with one of the coordinates, we find

$$\det\mathbf{J}=\prod_{i=1}^{I-1}p_i\times\det\left[\mathbf{I}-\mathbf{n}\mathbf{p}^{+\mathrm{T}}\right]=\prod_{i=1}^{I-1}p_i\times(1-\mathbf{n}\cdot\mathbf{p}^+)=\prod_{i=1}^{I-1}p_i\times(1-\sum_k p_k^+)=I\prod_{i=1}^{I}p_i. \tag{37}$$

So, for a delta function $g()$ in equation (10), neglecting constant factors which can be incorporated into the normalizing constants, we find that

$$P(\mathbf{p})=P(\mathbf{a}(\mathbf{p}))/|\det\mathbf{J}|\propto\prod_{i=1}^{I}p_i^{u_i-1}\delta\left(\textstyle\sum_{i=1}^{I}p_i-1\right). \tag{38}$$

This result is true for any α; so we can integrate over any normalized distribution $g()$, and the probability over $\mathbf{p}$ will be the same.

References

Bridle, J. S. (1989). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In Fougelman-Soulie, F. and Hérault, J., editors, *Neuro-computing: algorithms, architectures and applications*. Springer–Verlag.

Chickering, D. M. and Heckerman, D. (1996). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Microsoft Research Technical Report MSR-TR-96-08.

Gelman, A. (1996). Bayesian model–building by pure thought: Some principles and examples. *Statistica Sinica*, 6:215–232.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman and Hall, London.

Jeffreys, H. (1939). *Theory of Probability*. Oxford Univ. Press. 3rd edition reprinted in paperback 1985.

Lee, C. H. and Gauvain, J. L. (1993). Speaker adaptation based on MAP estimation of HMM parameters. In *IEEE Proceedings*, pages II–558–561.

Lindley, D. V. (1980). Approximate Bayesian methods. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics*, pages 223–237. Valencia University Press, Valencia.

MacKay, D. J. C. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

MacKay, D. J. C. (1997). Ensemble learning for hidden Markov models. Available from `http://wol.ra.phy.cam.ac.uk/`.

MacKay, D. J. C. and Peto, L. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3):1–19.

Neal, R. M. (1992). Bayesian mixture modelling. In Smith, C., Erickson, G., and Neudorfer, P., editors, *Maximum Entropy and Bayesian Methods, Seattle 1991*, pages 197–211, Dordrecht. Kluwer.

O'Hagan, A. (1994). *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Edward Arnold.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge.

Smith, A. and Spiegelhalter, D. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society B*, 42(2):213–220.

August 25, 1998 — Version 3.6