COMPARISON OF NUMERICAL METHODS FOR EVIDENCE CALCULATION

R. Preuss and U. von Toussaint
Max-Planck-Institut für Plasmaphysik
EURATOM Association, D-85748 Garching, Germany
preuss@ipp.mpg.de

Abstract

Model comparison requires the determination of integrals over the posterior probability function. We present a variety of numerical methods for this calculation. As working examples serve in various dimensions a single Gaussian peak as well as two Gaussian peaks with equal and different width and height, with and without infinitesimal integrands in between the two peaks.

Keywords: Evidence, Prior predictive value, Monte Carlo methods, Nested sampling

1 Introduction

Bayesian model comparison is inevitably associated with the calculation of the prior predictive value or evidence which involves integration over the posterior probability density function. Most of the time this integral has no analytical solution and one is referred to either approximative or numerical approaches or a mixture of both. Then employed integration methods have to cope with all the cumbersome features of a function which could sparsely populate a large parameter space, consisting of broad and narrow peaks, involving large and small scales and finally be spread such that the

integral weight between the structures is zero (at least according to numerical means). We present a comparison of a variety of numerical methods featuring Laplace approximation, trapezoidal rule, importance sampling, VEGAS (from Numerical Recipes [Press et al.(2001)]), thermodynamic integration scheme (thin-MCMC), and nested sampling [Skilling (????)]. This choice is far from being complete and simply arises from the fact that we have long-time experience with most of the methods (apart from [Press et al.(2001)] and [Skilling (????)]). For a further, much more sophisticated method for the integration of ill-conditioned problems (Perfect Tempering) we refer to the paper of M. Daghofer, published in the proceedings of the 2004 MaxEnt conference [Daghofer and von der Linden(2004a)].

2 Statement of the problem

The posterior probability function is composed of likelihood and prior probability density function. If the data is normally distributed the likelihood will be of Gaussian shape. Moreover, if the information gain from an experiment is large the likelihood will be much more structured than the prior and be of dominating role in the posterior. So the posterior will be of Gaussian-like character as well. The exercises of this paper shall therefore consist of the integration of Gaussian peaks in K-dimensions. We look for the integral

$$\boldsymbol{I} = \int p(\boldsymbol{x}|\sigma)\mathrm{d}\boldsymbol{x} \quad , \tag{1}$$

with the following choice of integrands: The simplest case shall consist of a single Gaussian peak. This problem should be feasible for any method.

$$p_1(\boldsymbol{x}|\sigma) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^K} \exp\left\{-\frac{\boldsymbol{x}^T\boldsymbol{x}}{\sigma^2}\right\} \quad . \tag{2}$$

Without loss of generality we set σ=0.3. To simulate multi-modal posteriors, the next case consists of two Gaussians of equal height and width.

$$p_2(\boldsymbol{x}) = \frac{1}{2}\frac{1}{(\sqrt{2\pi}\sigma)^K}\left[\exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{d}_1)^T(\boldsymbol{x}-\boldsymbol{d}_1)}{\sigma^2}\right\} + \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{d}_2)^T(\boldsymbol{x}-\boldsymbol{d}_2)}{\sigma^2}\right\}\right] . \tag{3}$$

The σ is set to 0.3 and 0.03, respectively. With the first setting, the integrand is still of small weight between the two peaks and therefore would allow "path

following" methods to pass from one mode to the other. For σ=0.03 this is not possible since the peaks are fully separated according to numerical means.

The Monte Carlo method cannot decide between weight originating in one peak or the other and therefore gives a correct result even if the samples are coming erroneously from the same peak. To disclose such failure we investigate another setup with two peaks of different height (10:1) and width (1:2).

$$p_3(\boldsymbol{x}) = \frac{1}{10+2^K}\frac{1}{\cdot(\sqrt{2\pi}\sigma)^K}\left[10\exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{d}_1)^T(\boldsymbol{x}-\boldsymbol{d}_1)}{\sigma^2}\right\}\right. \\ \left. + \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{d}_2)^T(\boldsymbol{x}-\boldsymbol{d}_2)}{(2\sigma)^2}\right\}\right] . \qquad (4)$$

σ is again set to 0.3 and 0.03, respectively. For the rest of the paper $\boldsymbol{d}_1$ consist of K numbers 2, $\boldsymbol{d}_2$ of numbers -2.

3 Description of the methods and results

In the following we give only a brief description of the employed numerical methods. Please refer to the literature for deeper insights. Some of these methods require a preceding MCMC run in order to determine the covariance of the parameters. The covariances are naturally provided if the expectation values of the parameters are needed anyway and determined beneficially with the Metropolis algorithm which does not require the norm. The MCMC samples are separated into so-called *bins* from which the expectation values and the variances are calculated. Each bin is preceded by burn-in sampling with randomly chosen starting values. The computer code was run on a 2.1GHz processor. The indicated running time is given for comparison reasons only.

3.1 Laplace approximation

This is also called steepest descent method or saddle-point approximation, where the term "Laplace approximation" is reserved for the real space. It constitutes a simple and powerful approximation to the integral of Eq. (1) if the integrand has a single mode only (regardless of dimension).

As mentioned above most posterior probability distributions resemble a Gaussian shape. It is therefore a good approximation to employ a Taylor

Dimension	1	2	4	8	16	32
1 peak, σ=0.3	0.99	0.99	1.02	1.00	0.99	0.94
time[s]	0.26	0.41	0.70	1.41	3.45	10.1

Table 1: Laplace approximation. The results for the other test cases containing two peaks were clearly senseless.

series at the maximum of the integrand (first order is zero)

$$p(\boldsymbol{x}|\sigma) \sim \exp\{\Phi(\boldsymbol{x})\} \quad . \tag{5}$$

$$\Phi(\boldsymbol{x}) = \Phi(\boldsymbol{x}_{max}) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_{max})^T \mathbf{H} (\boldsymbol{x} - \boldsymbol{x}_{max}) \quad , \tag{6}$$

with the Hessian matrix

$$H_{ij} = \frac{\partial^2 \ln p(\boldsymbol{x}|\sigma)}{\partial x_i \partial x_j} \quad . \tag{7}$$

The second order is just a Gaussian integral solvable analytically:

$$p(x|\sigma) = \frac{\text{const}}{\sqrt{\det \mathbf{H}}} \exp\{\Phi_0\} \quad . \tag{8}$$

One merely has to find the maximum of Φ in $\boldsymbol{x}$-space and take advantage of a previous MCMC run in determining the Hessian matrix during parameter estimation from the covariances of the parameters:

$$\mathbf{H} = \mathbf{C}^{-1} \quad , \tag{9}$$

with covariance matrix

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \quad . \tag{10}$$

The results are given in table 1. The desired value of 1 is reproduced as good as the previous Monte Carlo run was (10 times 2000 samples plus burn-in of altogether 2000 samples). The time given is therefore the running time of exactly this MCMC run. The time for the calculation of the Laplace approximation itself is negligible. So fast and satisfying the result for p_1 is, so utterly devastating it is for integrands with more than one mode (not shown). Nevertheless, for its simple use it should be part of every evidence calculation program, but be regarded as for diagnostic reasons only.

Dimension	1	2	4	8
1 peak, σ=0.3	1.00	0.99	0.98	0.97
time[s]	0.000	0.000	0.04	901
2 peaks, σ=0.3	1.00	1.00	1.00	1.00
time[s]	0.000	0.004	0.032	2041
2 peaks, σ=0.03	1.00	0.99	0.98	?
time[s]	0.000	0.024	432	$\approx$ 6935 years
2 different peaks, σ=0.3	1.00	1.00	1.00	0.99
time[s]	0.000	0.004	0.088	1995
2 different peaks, σ=0.03	1.00	0.99	0.99	?
time[s]	0.004	0.052	446	$\approx$ 7160 years

Table 2: Mesh integration. The number of points to calculate were N_{mesh}=15 for σ=0.3 and N_{mesh}=150 for σ=0.03. The question mark means "not calculated due to lack of sufficient time".

3.2 Trapezoidal rule (integration on a mesh)

We employ simple trapezoidal integration over the parameter space.

$$p(\boldsymbol{x}|\sigma) = \sum_{i=1}^{N} ... \sum_{j=1}^{N} p(x_{1i}, ..., x_{Kj}|\sigma)\Delta x_1 \cdot ... \cdot \Delta x_K \tag{11}$$

One can think of more sophisticated algorithms, refined in adjusting the integration grid automatically according to the integral weight. The accuracy of the result may be controlled by increasing the grid density and comparing the outcome with the step before. However, for larger numbers of parameters all these mesh integration techniques fail due to the curse of dimension (see table 2). In many cases it is possible to run a new programmed code for smaller dimensions where mesh integration still works. In order to detect errors in coding, the recommended procedure is to check the results with other evidence calculation methods and then proceed to the actual problem with its larger number of dimensions, i.e. parameters. Note: numerical problems may occur if the numbers in the exponent become too large (small), so we actually sum over $p(\boldsymbol{x}|\sigma) - p(\boldsymbol{x}_{max}|\sigma)$.

Dimension	1	2	4	8	16	32
1 peak, σ=0.3	1.00	1.00	1.00	1.00	1.00	1.00
time[s]	0.084	0.124	0.192	0.344	0.552	0.808
2 peaks, σ=0.3	0.99	0.98	0.94	##	##	##
time[s]	0.156	0.208	0.276	##	##	##
2 peaks, σ=0.03	1.04	0.50	##	##	##	##
time[s]	0.444	0.676	##	##	##	##
2 different peaks, σ=0.3	1.00	0.29	0.62	0.96	0.99	##
time[s]	0.156	0.208	0.284	0.436	0.548	##
2 different peaks, σ=0.03	1.00	0.28	##	##	##	##
time[s]	0.224	0.848	##	##	##	##

Table 3: Importance sampling. The entry "##" means an obviously erroneous result.

3.3 Importance sampling

The idea is to generate samples from a simpler function easy to sample from.

$$\boldsymbol{I} = \int p(\boldsymbol{x}|\sigma)d\boldsymbol{x} = \int \frac{p(\boldsymbol{x}|\sigma)}{g(\boldsymbol{x})} g(\boldsymbol{x})d\boldsymbol{x} \quad . \tag{12}$$

For the function $g(\boldsymbol{x})$ we employ a Gaussian with widths from the covariances generated by the preceding MCMC run already utilized for the Laplace approximation. 100000 sampling steps are performed to create each entry to table 3. While being excellent for the one peak problem, importance sampling fails rapidly for the other problems. It is simply harmed from the fact that already the MCMC run for the determination of the covariances produces wrong results getting stuck in a particular peak.

3.4 VEGAS algorithm

The algorithm invented by Peter Lepage is freely available from e.g. Numerical Recipes [Press et al.(2001)] and reportedly "widely used for multidimensional integrals that occur in elementary particle physics". It works accordingly to the importance sampling scheme, however separates the target function into a multidimensional weight function g

$$p(\boldsymbol{x}) = g_1(x_1)g_2(x_2)...g_K(x_K) \quad . \tag{13}$$

The implementation as an additional integration method is simple and straight forward. Table 4 shows the problems of Monte Carlo methods with peaky structures as in the importance sampling case, however it scores somewhat better. Expecially for the one peak problem it would be possible to get reasonable results for dimensions larger than 4 if the integration integral would be confined to a smaller range around the maximum. However, since we pretend to have no knowledge about the structure of the integrand, we stay with the range of { -4,4 } as for all other case in this paper.

3.5 Thermodynamic integration scheme

At the Maxent workshop 1997 in Boise John Skilling suggested to employ a formalism, borrowed from statistical physics, to compute the prior-predictive value, the so-called 'thermodynamic integration' scheme[Neal(1993)]: Define the function

$$Z(\lambda) = \int \Lambda^{\lambda}(\boldsymbol{x})\Pi(\boldsymbol{x})\ d\boldsymbol{x} \quad , \tag{14}$$

with $Z(\lambda{=}0) = 1$ and $Z(\lambda{=}1)$ as the desired quantity. Commonly the function Λ comprises terms from the likelihood. Π is the normalized prior. Here we chose $\Lambda = p(\boldsymbol{x}|\sigma)$ and $\Pi{=}1/8$ within $[-4, 4]$ and 0 otherwise. The derivative with respect to λ gives

$$\begin{aligned} \frac{\partial \ln Z(\lambda)}{\partial \lambda} &= \int \ln \Lambda(\boldsymbol{x})\rho_{\lambda}(\boldsymbol{x})\ d\boldsymbol{x} \\ &= \langle \ln \Lambda(\boldsymbol{x}) \rangle_{\lambda} \quad , \end{aligned} \tag{15}$$

with

$$\rho_{\lambda}(\boldsymbol{x}) = \frac{\Lambda^{\lambda}(\boldsymbol{x})\Pi(\boldsymbol{x})}{\int \Lambda^{\lambda}(\boldsymbol{x}')\Pi(\boldsymbol{x}')\ d\boldsymbol{x}'} \tag{16}$$

as the new sampling density. Both sides of Eq. (15) are integrated over λ:

$$\int_0^1 \langle \ln \Lambda(\boldsymbol{x}) \rangle_{\lambda}\ d\lambda = \int_0^1 \frac{\partial \ln Z(\lambda)}{\partial \lambda}\ d\lambda \tag{17}$$

$$= \ln Z(\lambda = 1) - \ln Z(\lambda = 0) \tag{18}$$

$$= \ln \boldsymbol{I} \quad . \tag{19}$$

To obtain the prior predictive value one therefore has to calculate the integral on the l.h.s. in (17) where the expectation value $\langle \ln \Lambda(\boldsymbol{x}) \rangle_{\lambda}$ is accessible by Markov chain Monte Carlo techniques[von der Linden et al.(1999)]. A plot

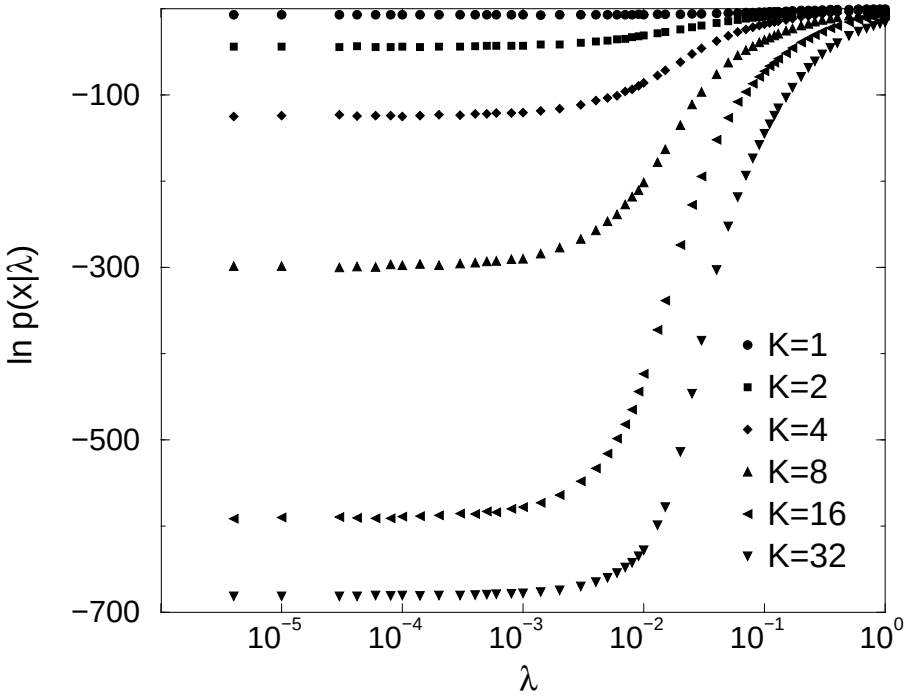

Figure 1: $\langle \ln \Lambda(\boldsymbol{x}) \rangle_\lambda$ for the case of two equal peaks with σ=0.3 in various dimensions.

of the latter for the integrand of a two peak case in various dimensions is given in Fig. 1. The λ-axis is on the logarithmic scale. As can be seen, one has to be careful to approach steadily $\lambda \to 0$.

The results are shown in table 5. In comparison with the other methods, the thermodynamic integration scheme works best. However, for the most difficult exercise with the peaky structures completely separated, it fails as well.

3.6 Nested sampling

Nested sampling, a recently proposed method [Skilling (????), Skilling(2004), Skilling(2006)] for evidence computation tabulates the likelihood function in a probabilistic 'sorted' manner. It samples a collection of n objects $\boldsymbol{x}$ from the prior distribution subject to the constraint that only objects with a likelihood value above an evolving threshold $L(\boldsymbol{x}) > L^*$ are accepted. For

the threshold L^* the lowest likelihood value of the collection is used, then the respective object is discarded and a new object is sampled from the prior within the restriction of the constrained likelihood. The worst object of this new sample gives the next threshold L^* and the process is iterated until convergence. The key idea is that the sequence of iterated objects contains (probabilistic) information about the enclosed prior volume $\delta\chi$, which in turn allows the estimate of the corresponding contribution to the total evidence $Z \approx \sum_k z_k$, with $z_k = \delta\chi_k \times L_k$. At the same time the random samples allow the computation of all interesting posterior distributions. Nested sampling is a new sampling method different from the standard techniques. However it does not solve the curse of dimensions. If there is a small hidden likelihood peak in some corner of the prior space the probability of detecting it is low, like in all other MCMC approaches. Nevertheless the different characteristics of this approach makes nested sampling an important tool complementing the conventional suite of MCMC techniques. Taking into account the simple structure (no implementation issues) it is highly recommended to be used as a standard tool for evidence computations.

4 Conclusion

In conclusion we have to admit that for integrands in dimensions larger than 10 with sparsely distributed multi-modal structures showing no integral weight in-between, one is left with *Perfect Tempering* [Daghofer and von der Linden(2004b)] as seemingly the only method capable of performing such integrations. However, as it is with any sophisticated method in the numerical analysis business, *Perfect Tempering* needs a lot of experience to sail around the pitfalls of Monte Carlo methods (among them omission of important contributions to the integral from undiscovered parameter space, correlated samples or erroneous bookkeeping). Apart from that, on 'second place', the thermodynamic integration scheme and nested sampling work best, though showing their problems with peaky structures as well. Still, the bunch of the presented methods can come close and should be used as standard diagnostic tools monitoring the outcome all the time.

References

[Press et al.(2001)] Press, W., et al., *Numerical Recipes in Fortran 77: The Art of Scientific Computing (Vol. 1 of Fortran Numerical Recipes), 2nd edition*, Cambridge University Press, Cambridge, 2001.

[Skilling (????)] Skilling, J., http://www.inference.phy.cam.ac.uk/bayesys/.

[Daghofer and von der Linden(2004a)] Daghofer, M., and von der Linden, W., "Perfect Tempering," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP, Melville, N.Y., 2004a, vol. 735 of *AIP Conference proceedings*, p. 355.

[Neal(1993)] Neal, R. M., *Probabilistic inference using markov chain monte carlo methods*, Dept. of Computer Science, University Toronto, 1993.

[von der Linden et al.(1999)] von der Linden, W., Preuss, R., and Dose, V., "The prior predictive value," in *Maximum Entropy and Bayesian Methods*, edited by W. von der Linden et al., Kluwer Academic, Dordrecht, 1999, p. 319.

[Skilling(2004)] Skilling, J., "Nested Sampling," in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, edited by R. Fischer, R. Preuss, and U. von Toussaint, AIP, Melville, N.Y., 2004, vol. 735 of *AIP Conference proceedings*, p. 395.

[Skilling(2006)] Skilling, J., *Bayesian Analysis*, **1**, 833 (2006).

[Daghofer and von der Linden(2004b)] Daghofer, M., and von der Linden, W., Perfect tempering (2004b), these proceedings, and references therein.

Dimension	1	2	4	8	16	32
1 peak, σ=0.3	1.00	1.00	1.00	##	##	##
time[s]	0.33	0.55	0.66	##	##	##
2 peaks, σ=0.3	1.00	1.00	0.99	##	##	##
time[s]	0.50	0.77	1.21	##	##	##
2 peaks, σ=0.03	1.00	1.00	##	##	##	##
time[s]	0.50	0.90	##	##	##	##
2 different peaks, σ=0.3	1.00	1.00	0.99	0.95	##	##
time[s]	0.50	0.68	1.08	2.68	##	##
2 different peaks, σ=0.03	1.00	1.00	##	##	##	##
time[s]	0.64	0.70	##	##	##	##

Table 4: Results obtained with the VEGAS algorithm.

Dimension	1	2	4	8	16	32
1 peak, σ=0.3	1.00	0.97	0.96	0.95	0.90	0.83
time[s]	5	11	25	64	185	598
2 peaks, σ=0.3	0.99	0.97	0.96	0.95	1.27	449
time[s]	9	20	45	122	365	1133
2 peaks, σ=0.03	1.70	##	##	##	##	##
time[s]	12	##	##	##	##	##
2 different peaks, σ=0.3	1.04	0.64	0.69	0.93	0.94	##
time[s]	9	20	44	118	337	##
2 different peaks, σ=0.03	2.92	##	##	##	##	##
time[s]	12	##	##	##	##	##

Table 5: Thermodynamic integration.

Dimension	1	2	4	8	16	32
1 peak, σ=0.3	1.01	0.94	0.95	1.02	1.03	0.7
time[s]	90	160	210	270	490	920
2 peaks, σ=0.3	0.96	0.85	0.53	0.32	0.13	0.012
time[s]	144	360	410	566	680	1200
2 peaks, σ=0.03	0.96	0.64	0.32	0.13	0.04	##
time[s]	200	256	270	400	700	1200
2 different peaks, σ=0.3	0.99	0.86	0.65	0.55	0.25	0.06
time[s]	205	246	290	560	740	1150
2 different peaks, σ=0.03	0.85	0.73	0.52	0.11	0.05	##
time[s]	265	305	350	560	700	1200

Table 6: Nested sampling.