THE PRIOR-PREDICTIVE VALUE:

A PARADIGM OF NASTY MULTI-DIMENSIONAL INTEGRALS

W. VON DER LINDEN, R. PREUSS, AND V. DOSE
Max-Planck-Institut für Plasmaphysik, EURATOM Association
D-85740 Garching b. München, Germany ¶

Abstract. At the heart of Bayesian model comparison lies the so-called prior-predictive value. In the important class of Quantified-MaxEnt applications analytic approximations are routinely used which often give rise to noise-fitting and ringing. We present an improved analytic expression which overcomes these shortcomings. In most interesting real-world problems, however, standard approximations and straight forward application of Markov-Chain Monte Carlo are hampered by the complicated structure of the likelihood in parameter space. At the Maxent workshop 1997 in Boise John Skilling suggested to employ a formalism, borrowed from statistical physics, to compute the prior-predictive value. We have scrutinized his suggestion: IT WORKS!

Key words: Prior-predictive value, Bayes factor, MCMC, Model comparison

1. Introduction

Bayesian probability theory provides a general frame for model evaluation and hypothesis testing [1] irrespective of the complexity of the models involved, the structure of the underlying parameter space and the number of data points or model parameters. Bayes theorem relates the probability for model/hypothesis/theory M in the light of the data and background information $\mathcal{I}$ to probabilities which are at our disposal

$$P(M|D,\mathcal{I}) = \frac{P(M|\mathcal{I})p(D|M,\mathcal{I})}{\mathcal{Z}} \ . \tag{1}$$

The quantity $\mathcal{Z}$, which ensures normalization, is irrelevant as it drops out in model comparison. $P(M|\mathcal{I})$ stands for the prior probability, i.e. the probability we assign to M prior to our knowing the data D. The centerpiece of the Bayesian approach is the so-called prior-predictive value (ppv)

$$p(D|M,\mathcal{I}) = \int_V p(D|x,M,\mathcal{I})p(x|M,\mathcal{I})d\mu(x) \ , \tag{2}$$

¶e-mail: wvl@ipp-garching.mpg.de

i.e. the evidence for the data under model M given all background information $\mathcal{I}$. The terms under the integral are likelihood, prior and measure $\mu(x)$ in the parameter space spanned by the allowed model parameters x in volume V. Quite generally, the ppv is a multi-dimensional integral of the form $\int f(x)\pi(x)d\mu(x)$, where $\pi(x)$ is a probability density and $f(x)$ an arbitrary function. This integral plays an important role in many areas of science. Several techniques have been proposed to compute the ppv, including Laplace-type of approximations which are asymptotically exact [2].

Quantified Maximum Entropy (QME) [3,4] applications are an important class of problems for which the Laplace approximation, also known as *steepest descent* approximation, is routinely used [3,4]. It is required to determine the regularization parameter α in the so-called evidence approximation [3,4]. The probability for α given the data is proportional to the ppv which has the specific form

$$p(D|\alpha,\mathcal{I}) \propto \int_0^\infty e^{-\frac{1}{2}\chi^2+\alpha S} \prod_i x_i^{-1/2} d^N x \; . \tag{3}$$

For concreteness we have chosen a Gaussian likelihood although the following considerations are valid for arbitrary argument of the exponential $\phi = -\frac{1}{2}\chi^2 + \alpha S$. In the Laplace approximation, in which ϕ is expanded to second order in x_i and the lower integration limits are moved to $-\infty$, the ppv reads

$$p(D|\alpha,\mathcal{I}) \propto e^{\phi^*} |H^0|^{-\frac{1}{2}} \prod_i \hat{x}_i^{-1/2} \; . \tag{4}$$

Here ϕ^* stands for the maximum value of ϕ and $H^0 = \nabla\nabla^T\phi|_{\hat{x}}$ is the hessian of ϕ at the maximum a-posteriori (MAP) value $\hat{x}$. In the case of an ill-posed inversion problem, e.g. deconvolution of a broad point spread function, the Laplace approximation causes the following problem: For small α, i.e. weak regularization, the MAP-image $\hat{x}$ exhibits pronounced scatter (ringing) which along with the positivity constraint produces many image values close to zero. But particularly for $\hat{x}_i \to 0$ the Laplace approximation strongly overestimates the integral (3) due to the modification of the lower integration limits. Consequently, the probability for α, which is proportional to $p(D|\alpha,\mathcal{I})$, will be largely overestimated for small α which results in noise-fitting and ringing. It is important to note that the origin for the problems is not the Gaussian approximation of e^ϕ but rather the modification of the integration volume [5].

We have shown in previous publications that the QME results improve significantly if either [6] the α-marginalization is performed exactly – hence avoiding the Laplace approximation altogether – or [5] if the Laplace approximation is improved such as to allow for the correct integration limits. The latter had the advantage that existing QME programs could be used with only minor modifications but it had the disadvantage that the integral correction was still approximate. There is, however, a simple way to avoid additional approximations by using the representation $x_i = \exp(\rho_i)$. The integration limits for ρ_i are now really $\pm\infty$. On top of that this is the natural representation of the MaxEnt solution, see e.g. [7]. The ppv transforms into

$$p(D|\alpha,\mathcal{I}) \propto \int_{-\infty}^{\infty} e^{\phi+\frac{1}{2}\sum_i \rho_i}\, d^N\rho \tag{5}$$

As pointed out above, it is now legitimate to employ the Laplace approximation

$$p(D|\alpha,\mathcal{I}) \propto e^{\phi^*}\,|H^0 + \frac{1}{2}\mathrm{diag}(\hat{x}_i^{-2})|^{-\frac{1}{2}} \prod_i \hat{x}_i^{-1/2} \tag{6}$$

Our experience to date with this expression for the ppv or rather α-evidence is very positive: less ringing and the results are comparable to what one gets in the case of α-marginalization.

For PARAMETER-ESTIMATION problems, however, the approximate schemes are often to crude since the posterior probability is far from being a Gaussian and Monte Carlo integration methods are the only approaches furnishing reliable results. Straight forward Markov-Chain Monte Carlo (MCMC) [8,9] is, however, ineligible since sampling from the prior $\pi(x|\mathcal{I})$[1] would in general result in huge variances and extremely long Markov chains. The reason is that the likelihood is generally much more structured and localized than the prior. At the Maxent workshop 1997 in Boise John Skilling suggested to employ the thermodynamic integration method[10] (thin-MCMC), borrowed from statistical physics, to compute the ppv. There is a wide variety of different MCMC methods for calculating Bayes factors which are discussed by Neal [10]. Some of them seem promising but have not been applied in really demanding problems yet. We have analyzed thin-MCMC in great detail and found that it works also in challenging problems, though it still is computationally demanding. Here we will present some general considerations why and when it works. In another paper of these proceedings [11] it is applied to a challenging real-world application.

2. The thermodynamic integration scheme

In order to make contact to statistical physics, we consider $p(D|M,\mathcal{I})$ as the partition function $Z(\beta)$ of a canonical ensemble with a fictitious energy function equal to the log-likelihood $E(x) = -\ln(p(D|M,x,\mathcal{I}))$ and a fictitious inverse temperature β. The probability for the system being in micro-state volume dx corresponds to the posterior probability

$$\begin{aligned} P(dx|M,D,\beta,\mathcal{I}) &= \frac{1}{Z(\beta)} e^{-\beta E(x)}\pi(x)d\mu(x) = p(x|M,D,\beta,\mathcal{I})d\mu(x) \quad (7)\\ Z(\beta) &= \int e^{-\beta E(x)}\pi(x)d\mu(x)\ . \quad (8) \end{aligned}$$

[1] The prior is the only part for which the normalization is known.

Having established the correspondence to statistical physics we can borrow ideas to compute the partition function or equivalently the free energy F which is defined via $Z(\beta) = \exp(-\beta F)$. One of these ideas is the thermodynamic integration method. It proceeds from the derivative of $Z(\beta)$with respect to β

$$\frac{d\ln Z(\beta)}{d\beta} = \frac{1}{Z(\beta)} \frac{dZ(\beta)}{d\beta} = -\int E(x) p(x|M, D, \beta, \mathcal{I}) d\mu(x) = -\langle E(\check{x}) \rangle_\beta \ , \tag{9}$$

which corresponds to the negative internal energy. We recall, $Z(0) = 1$ by virtue of the normalization of the prior and $Z(1)$ is the desired ppv. In order to avoid confusion, we mark the variable, which has been marginalized (averaged) over, x in the present case, by a check $\check{x}$. The solution of the differential equation (9) readily furnishes the ppv as the thermodynamic integral of the internal energy

$$\ln[Z(1)] = -\int_0^1 \langle E(\check{x}) \rangle_\beta d\beta = \ln[p(D|M, \mathcal{I})] \ . \tag{10}$$

Now $\langle E(\check{x}) \rangle_\beta$ can be computed for a selection of inverse temperatures $\beta \in [0, 1]$ by MCMC using the sampling density $p(x|M, D, \beta, \mathcal{I})$. The problem appears to be more tractable as before since we need to compute the expectation value of the energy (log-likelihood) only and the probability density contains the structure of the likelihood as well. Both facts reduce the variance significantly. In the worst case $\beta \to 0$ the sampling density is the bare prior but the variation of the likelihood enters merely on a logarithmic scale. Provided, $\langle E(\check{x}) \rangle_\beta$ is a reasonably smooth function, the β-integral can easily be performed by standard numerical integration schemes.

3. Test Example

In order to assess the thermodynamic integration scheme we consider a simple but fairly realistic example. For likelihood and prior we use Gaussians

$$p(D|M, x, \omega, \mathcal{I}) = (\frac{\omega}{2\pi})^{\frac{N}{2}} e^{-\frac{\omega}{2}\|d - Ux\|^2} \ ; \qquad \pi(x) = (\frac{\theta}{2\pi})^{\frac{n}{2}} e^{-\frac{\theta}{2}\|x\|^2} \ , \tag{11}$$

with $x \in \mathbb{R}^{\mathrm{n}}$ and $d \in \mathbb{R}^{\mathrm{N}}$. U is a $N \times n$ matrix which represents the expansion in orthonormal functions, i.e. $U^T U = \mathbb{1}$. This is not really an example for which MC-techniques are required but it is fairly realistic as far as the MC-performance is concerned if we assume that we are able to generate uncorrelated samples from the posterior distribution. If this is the case, no matter how bizarre the structure of the posterior is it can always be re-arranged into a uni-modal convex function without modifying the MC performance. Of course, there is still the important topic of how to generate uncorrelated and ergodic samples from a multi-modal distribution. Neal [12] suggested a promising method based on periodically heating and annealing the system. Extensions of this idea are described by Skilling in another paper of these proceedings. Our experience to date suggests that moderate multi-modality can be overcome by a modified move-generation and acceptance schedules, like those based on Cauchy distributions [13] or Tsallis statistics [14,15]. For the Gaussian example the ppv can be determined analytically

$$p(D|M,\mathcal{I}) = \langle p(D|M,\check{x},\mathcal{I})\rangle_{\beta=0} = (\frac{\omega}{2\pi})^{\frac{N}{2}}(\frac{\theta}{\omega+\theta})^{\frac{n}{2}} e^{-\frac{\omega}{2} d^T\left(1\!\!1 - \frac{\omega}{\omega+\theta} UU^T\right) d} \ . \qquad (12)$$

In order to assess the performance of straight forward MCMC based on the sampling density $\pi(x)$ we need

$$\langle p(D|M,\check{x},\mathcal{I})^2\rangle_{\beta=0} \quad = \quad (\frac{\omega}{2\pi})^N(\frac{\theta}{2\omega+\theta})^{\frac{n}{2}} e^{-\omega d^T\left(1\!\!1 - \frac{2\omega}{2\omega+\theta} UU^T\right) d} \ . \qquad (13)$$

In the worst case, $n = N$ and $\omega \gg \theta$ the relative Monte Carlo error is to leading order in θ/ω

$$\epsilon_{\mathrm{MC}} := \sqrt{\frac{\langle \Delta^2 p(D|M,\check{x},\mathcal{I})\rangle_{\beta=0}}{N_{\mathrm{MC}}}} / p(D|M,\mathcal{I}) = (\frac{\omega}{2\theta})^{\frac{N}{4}} e^{\frac{\theta}{4\omega}\omega N\overline{d^2}} / \sqrt{N_{\mathrm{MC}}} \ , \qquad (14)$$

where N_{MC} is the number of independent points in the Markov chain. For a set of realistic but still moderate parameters: signal-to-noise ratio $\sqrt{\omega\overline{d^2}} = 10$, prior width to likelihood width ratio $\omega/\theta = 100$, and $n = N = 128$ the relative MC error is

$$\epsilon_{\mathrm{MC}} \approx 10^{68}/\sqrt{N_{\mathrm{MC}}} \ . \qquad (15)$$

This quantifies our statement that straight-forward MCMC is simply not eligible in the case of weak prior-knowledge. We would need more than 10^{138} independent MC moves for 10% accuracy.

Next we will see whether the thermodynamic integration scheme can cope with this order of magnitude. To this end we determine the internal energy using (12)

$$\langle E(\check{x})\rangle_\beta = -\frac{N}{2}\ln[\frac{\omega}{2\pi}] + \frac{n\omega}{2(\beta\omega+\theta)} + \frac{\omega}{2} d^T(1\!\!1 - UU^T(1 - \frac{\theta^2}{(\beta\omega+\theta)^2}))d \quad . \quad (16)$$

In Fig.1 $\langle E(\check{x})\rangle_\beta$ is depicted. We see that the integrand is indeed a good-natured smooth function which can be integrated by standard numerical techniques with only a small number of inverse temperatures $\beta \in [0,1]$. From Fig.1 and from (16) it is obvious that the temperature scale is set by ω/θ. We therefore expect a more structured β-dependence if the likelihood has several scales. Next we determine the statistical error associated with the MC evaluation of $\langle E(\check{x})\rangle_\beta$ for which we need the variance

$$\langle \Delta^2 E(\check{x})\rangle_\beta = -\frac{d}{d\beta}\langle E(\check{x})\rangle_\beta = \frac{n\omega^2}{2(\beta\omega+\theta)^2} + \frac{\omega^2\theta^2}{(\beta\omega+\theta)^3} d^T UU^T d \ . \qquad (17)$$

The β-dependence of the variance (17) is plotted in Fig.1. Fortunately, the variance is good-natured even for $\beta \to 0$. The total variance is obtained upon integrating (17) with respect to β

$$-\int_0^1 \frac{d}{d\beta}\langle E(\check{x})\rangle_\beta \, d\beta = \langle E(\check{x})\rangle_\beta \Big|_0^1 = \frac{n\omega^2}{2\theta(\omega+\theta)} + \frac{\omega}{2}(1 - \frac{\theta^2}{(\omega+\theta)^2}) d^T UU^T d. \quad (18)$$

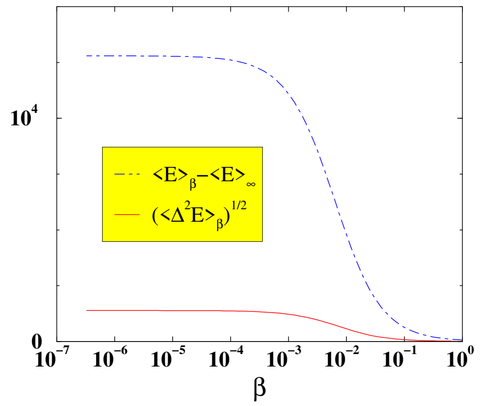

Figure 1. β*-dependence on a log-linear scale of the internal energy* $\langle E(\check{x})\rangle_\beta - \langle E(\check{x})\rangle_\infty$ *and the variance* $\sqrt{\langle \Delta^2 E(\check{x})\rangle_\beta}$ *for* $\omega\overline{d^2} = 100$*,* $\omega/\theta = 100$*, and* $N = 128$*.*

We consider only the interesting and realistic case in which the likelihood is significantly more informative (localized) than the prior, i.e. $\omega \gg \theta$. In addition, the model under consideration is only sensible if the basis U is large enough to ensure $d^T U U^T d \approx ||d||^2$ and we have the estimate

$$\Delta^2_{\mathrm{MC}} \approx \left(n\omega/\theta + N\omega\overline{d^2}\right)/2N_{\mathrm{MC}} . \tag{19}$$

The MC error Δ_{MC} in the log-ppv corresponds to a relative error in the ppv, provided $\Delta_{\mathrm{MC}} < 1$. In this case Δ_{MC} corresponds roughly to the relative error in $p(D|M, \mathcal{I})$. To achieve a relative error $\epsilon = 10\%$ the number of uncorrelated samples for the previous set of parameters is dramatically reduced to $\underline{\boldsymbol{N_{\mathrm{MC}} > 10^6!}}$. By virtue of the smoothness of $\langle E(\check{x})\rangle_\beta$ as function of β, it takes roughly 10 separate MC computations to do the β-integration by standard numerical integration techniques.

4. Direct Computation of Bayes Factors

Instead of computing the absolute value of the ppv for each model separately, it would be expedient to head for the Bayes factor directly. This is indeed feasible if the two models cover the same state-space (micro-states) [10]. Generalization to models with different state-space are conceivable upon introduction of dummy variables. Let the un-normalized (posterior) probability for model M_α, $\alpha \in \{1, 2\}$ be $f_\alpha(x) = p(D|x, M_\alpha, \mathcal{I})p(x|M_\alpha, \mathcal{I})\mu(x, \alpha)$. We define a mixture probability

$$p(x, \beta) = f_1(x)^{1-\beta} f_2(x)^\beta / Z(\beta) \tag{20}$$

$$Z(\beta) = \int f_1(x)^{1-\beta} f_2(x)^\beta \, dx . \tag{21}$$

with mixing coefficient β and mixture-partition function $Z(\beta)$. Obviously, the individual partition functions for the two models are recovered for $\beta = 0$ and $\beta = 1$, respectively. As before we obtain

$$\frac{d\ln Z(\beta)}{d\beta} = \int \ln(\frac{f_2(x)}{f_1(x)}) p(x|M,\beta) dx = \langle \ln(\frac{f_2(\check{x})}{f_1(\check{x})}) \rangle_\beta \ . \tag{22}$$

Finally, the Bayes factor of the two models M_1 and M_2, respectively, reads

$$\ln(\frac{P(M_2|D,\mathcal{I})}{P(M_1|D,\mathcal{I})}) = \ln(\frac{Z(1)}{Z(0)}) = \int_0^1 \langle \ln(\frac{f_2(\check{x})}{f_1(\check{x})}) \rangle_\beta d\beta \ . \tag{23}$$

Like before, we want to assess this approach guided by the example given in (11). The two models M_α differ in the sets of orthonormal basis vectors U_α. This results in two Gaussian posterior distributions of identical width $\sigma^2 = 1/(\omega + \theta)$ located at $x_\alpha = \frac{\omega}{\omega+\theta} U_\alpha^T d$ and with amplitudes $A_\alpha \propto \exp(x_\alpha^2/2\sigma^2)$:

$$\begin{aligned} f_\alpha(x) &= A_\alpha e^{-(x-x_\alpha)^2/2\sigma^2} \\ p(x|\beta) &= (2\pi\sigma^2)^{-n/2} e^{-(x-[(1-\beta)x_1+\beta x_2])^2/2\sigma^2} \ . \end{aligned} \tag{24}$$

The integrand of (23) has a trivial β-dependence

$$\langle \ln(f_2(\check{x})/f_1(\check{x})) \rangle_\beta = (\beta - 1/2)(x_2 - x_1)^2/\sigma^2 \tag{25}$$

which presumably will not get much more complicated in real-world problems. Also the MC error has a good-natured parameter dependence

$$\Delta_{\text{MC}}^2 = n(x_2 - x_1)^2/(N_{\text{MC}}\sigma^2) \ . \tag{26}$$

In order to achieve a relative error for the Bayes factor of ϵ the number of independent MC steps has to obey $N_{\text{MC}} \geq n(x_2 - x_1)^2/(\sigma^2\epsilon^2)$. A pleasant feature is the weak (quadratic) dependence on the distance $(x_2 - x_1)$ between the posterior modes of the 2 models. To give some numbers we consider the previously discussed example of two linear models. The $N = 128$ dimensional data vector d should be described by the two models U_α with $n = 10$ expansion coefficients x for each model. Since the x_α-dependence of the Bayes factors are approximately given by $\exp(-(x_2 - x_1)^2/2\sigma^2)$ high accurate results for the Bayes factors are only needed for $(x_2 - x_1)^2/2\sigma^2 \lesssim 1$. The number of MC-steps for a 10% accurate result is therefore $N_{\text{MC}} = 10^5$, which is easily doable on present-days computers. It is also reassuring that the prior does not enter the expressions explicitly and weak priors do not deteriorate the performance.

5. Concluding remarks

We have demonstrated that the prior-predictive value can be determined by thermodynamic integration MCMC. We studied an example where the required sample size could be reduced from more than 10^{138} down to $\approx 10^6$, which is a manageable order of magnitude. Moreover, we have shown that the numerical effort can be reduced even further if we directly determine Bayes factors.

There is another class of MCMC approaches to determine the free energy, which are closely related to the thin-MCMC, the so-called 'method of expanded ensembles' [16] or 'acceptance ratio method' [10]. These methods are described by Skilling in these proceedings. Besides the applications given there, a similar method has been applied successfully to an electrolyte model in physical chemistry [16]. In general one would expect thin-MCMC to be superior, since the number of intermediate 'temperatures' (β) is small and does not depend on the dimension of the problem, while in the other schemes the number of intermediate 'temperatures' will increase with the system size, since the free energy is an extensive quantity. On the other hand, these methods might be superior, if the posterior distribution has a pronounced multi-modal structure with isolated modes. In this case thin-MCMC might have ergodicity-problems at low temperatures.

References

1. H. Jeffreys, *Theory of Probability*, Oxford University Press, London, 1939.
2. R. Kass and A. E. Raftery, "Bayes factors and model uncertainty," *J. Amer. Statist. Assoc.*, **90**, pp. 773–795, 1995.
3. S. Gull, "Developments in maximum entropy data analysis," in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed., p. 53, Kluwer Academic Publishers, Dordrecht, 1989.
4. J. Skilling, "Quantified maximum entropy," in *Maximum Entropy and Bayesian Methods*, P. Fougère, ed., p. 341, Kluwer Academic Publishers, Dordrecht, 1990.
5. W. von der Linden, R. Fischer, and V. Dose, "Evidence integrals," in *Maximum Entropy and Bayesian Methods*, R. Silver and K. Hanson, eds., p. 443, Kluwer Academic Publishers, Dordrecht, 1996.
6. R. Fischer, W. von der Linden, and V. Dose, "On the importance of α marginalization in maximum entropy," in *Maximum Entropy and Bayesian Methods*, R. Silver and K. Hanson, eds., p. 229, Kluwer Academic Publishers, Dordrecht, 1996.
7. W. von der Linden, "Maximum-entropy data analysis," *Appl. Phys. A*, **60**, p. 155, 1995.
8. J. Hammersley and D. Handscomb, eds., *Monte Carlo Methods*, Chapman and Hall Ltd, London, 1964.
9. J. Ruanaidh and W. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer, New York, 1996.
10. R. Neal, "Probabilistic inference using markov chain monte carlo methods," Dept. of Computer Science, University Toronto, 1993.
11. R. Preuss, V. Dose, and W.v.d.Linden, "Model comparison in plasma energy confinement scaling," Kluwer Akademic, 1999.
12. R. Neal, "Sampling from multimodal distributions using tempered transitions," *Statistics and Computing*, **6**, pp. 353–366, 1996.
13. H. Szu and R. Hartley, "Fast simulated annealing," *Physics Letters A*, **122**, pp. 157–162, 1987.
14. C. Tsallis and D. Stariolo, "Generalized simulated annealing," cond-matter/9501047Jan95, 1995.
15. T. Penna, "Traveling salesman problem and tsallis statistics," *Phys.Rev.E*, **51**, pp. R1–3, 1995.
16. A. Lyubartsev, A. Martsinovski, S. Shevkunov, and P. Vorontsov-Velyaminov *J. Chem.Phys.*, **96**, p. 1776, 1992.