# Comment on "Inference with minimal Gibbs free energy in information field theory"

D. Iatsenko, A. Stefanovska, and P. V. E. McClintock

*Department of Physics, Lancaster University, Lancaster LA1 4YB, United Kingdom*

(Received 17 December 2011; published 20 March 2012)

Enßlin and Weig [Phys. Rev. E **82**, 051112 (2010)] have introduced a "minimum Gibbs free energy" (MGFE) approach for estimation of the mean signal and signal uncertainty in Bayesian inference problems: it aims to combine the maximum *a posteriori* (MAP) and maximum entropy (ME) principles. We point out, however, that there are some important questions to be clarified before the new approach can be considered fully justified, and therefore able to be used with confidence. In particular, after obtaining a Gaussian approximation to the posterior in terms of the MGFE at some temperature $T$, this approximation should always be raised to the power of $T$ to yield a reliable estimate. In addition, we show explicitly that MGFE indeed incorporates the MAP principle, as well as the MDI (minimum discrimination information) approach, but not the well-known ME principle of Jaynes [E.T. Jaynes, Phys. Rev. **106**, 620 (1957)]. We also illuminate some related issues and resolve apparent discrepancies. Finally, we investigate the performance of MGFE estimation for different values of $T$, and we discuss the advantages and shortcomings of the approach.

DOI: 10.1103/PhysRevE.85.033101 PACS number(s): 89.70.−a, 11.10.−z, 98.80.Es, 95.75.Mn

In a recent paper, Enßlin and Weig [1] introduced what they refer to as a "minimum Gibbs free energy" (MGFE) method for Bayesian signal reconstruction. The MGFE approach is clearly valuable, and it has already been applied successfully to the general problem of signal reconstruction [1,2]. We comment, however, that there are some important issues that need to be exposed, and clarified, to enable this new approach to be applied with confidence.

In the MGFE approach [1], the functional

$$G[\tilde{P}(s|d)] = -\int \mathcal{D}s\,\tilde{P}(s|d)\ln P(s,d) + T\int \mathcal{D}s\,\tilde{P}(s|d)\ln \tilde{P}(s|d) \quad (1)$$

is minimized for a fixed value of the "temperature" $T$ with respect to the parameters of some chosen approximation $\tilde{P}(s|d)$ to the actual posterior $P(s|d)$ of the signal $s$, given the data $d$. The approximate posterior $\tilde{P}(s|d)$ is usually taken to be Gaussian,

$$\tilde{P}(s|d) = \frac{1}{\sqrt{\det(2\pi D)}}\exp\left\{-\frac{(s-m)^{+}D^{-1}(s-m)}{2}\right\} \equiv \mathcal{G}(s-m,D). \quad (2)$$

So the minimization of (1) is taken with respect to estimates of the mean signal $m$ and the uncertainty matrix $D$, which can be formulated mathematically as

$$\left.\frac{\delta G[\mathcal{G}(s-m',D')]}{\delta m'}\right|_{m'=m,D'=D} = 0, \qquad \left.\frac{\delta G[\mathcal{G}(s-m',D')]}{\delta D'}\right|_{m'=m,D'=D} = 0. \quad (3)$$

The Gaussian approximation is, of course, the most convenient approach because it allows one to calculate explicitly most of the path integrals that are commonly encountered. Finding such an approximation may be especially helpful if the posterior obtained is to be used for further Bayesian inference.

However, the use of the temperature in Eq. (1), as introduced in Ref. [1], is liable to create confusion. For example, in Ref. [1] the authors introduce temperature $T$ and a generating source $J$ into Bayes' theorem $[P(s|d) = P(d|s)P(s)/P(d)]$ and consider

$$P(s|d,T,J) = \frac{[P(s,d)e^{J^{+}s}]^{1/T}}{\int \mathcal{D}s[P(s,d)e^{J^{+}s}]^{1/T}}. \quad (4)$$

This expression coincides with Bayes' theorem only when $J = 0$, $T = 1$: $P(s|d) = P(s|d,1,0)$. Procedures of this kind can sometimes be useful, as one can take derivatives with respect to these parameters and then set $T = 1, J = 0$, which may make derivations faster. However, in the case of Bayesian inference, $T$ and $J$ are just part of the mathematical formalism; in particular, for $T \neq 1, J \neq 0$, Bayes' theorem is evidently violated if one is claiming that $P(s|d,T,J)$ can be used as a posterior. To avoid logical inconsistencies, these parameters must be set to their true values $T = 1, J = 0$ at the end of the calculation.

Nonetheless, there is a way to make use of (4) at different temperatures: one can express actual posterior $P(s|d)$ in terms of the tempered one $P(s|d,T,0)$ as

$$P(s|d) = A[P(s|d,T,0)]^{T}, \quad (5)$$

where $A^{-1} \equiv \int \mathcal{D}s[P(s|d,T,0)]^{T}$ is the normalization multiplier. Indeed, substituting Eqs. (4) into (5), one will recover Bayes' theorem. In this context, we note that, in [1], the MGFE principle is derived from (4). Thus MGFE, as proposed, finds an approximation to the tempered posterior $P(s|d,T,0)$, not to the actual posterior. So if we want to use MGFE at different temperatures (not only at $T = 1$), then, after finding the Gaussian approximation by (3), we should raise it to the power of $T$ (preserving normalization), which is equivalent to setting $D \to D/T$ in Eq. (2). Otherwise such estimation will contradict Bayes' theorem for all temperatures except $T = 1$.

This can readily be understood by considering the limit $T \to 0$. It is claimed that MGFE at $T = 0$ corresponds to maximum *a posteriori* (MAP) estimation. MAP estimates the mean signal $m$ as being the most probable one, i.e., that for which the posterior $P(s|d)$ is maximal. Mathematically this principle can be formulated as $\frac{\delta \ln P(s|d)}{\delta s}|_{s=m} = 0$. If one also wants to find a MAP-based uncertainty matrix estimate,

$D^{-1} \approx -\frac{\delta^2 \ln P(s|d)}{\delta s \delta s^+}|_{s=m}$ can be used. At the same time, under the Gaussian approximation (2) at $T = 0$, the functional (1) is $G = -\int \mathcal{D}s\mathcal{G}(s-m,D)\ln P(s|d)$, so that its minimization with respect to $m$ and $D$ yields

$$\int \mathcal{D}s \frac{\delta \mathcal{G}(s-m',D')}{\delta D'} \ln P(s|d)\bigg|_{m'=m,D'=D} = 0$$
$$\Rightarrow D = 0 \Rightarrow \mathcal{G}(s-m',D') = \delta(s-m') \qquad (6)$$

and

$$\frac{\delta}{\delta m'}\int \mathcal{D}s\mathcal{G}(s-m',D')\ln P(s|d)\bigg|_{m'=m,D'=D} = 0$$
$$\Rightarrow \frac{\delta \ln P(m'|d)}{\delta m'}\bigg|_{m'=m} = 0, \qquad (7)$$

where we have taken into account the result of (6) that $\mathcal{G}(s-m',D') = \delta(s-m')$. In effect, (7) is giving a MAP mean signal estimate, but (6) results in a confusing approximation of the posterior by a $\delta$-function. Such an approximation will therefore correspond only partially to MAP estimation, because $D = 0$ is being used rather than the correct form $D^{-1} = -\frac{\delta^2 \ln P(s|d)}{\delta s \delta s^+}|_{s=m}$.

However, if we take as the uncertainty estimate the limiting value of $D/T$ as $T \to 0$, we will obtain the correct MAP estimate. Indeed, using the explicit form of (2), the second term of (1) can be shown to be

$$\int \mathcal{D}s\mathcal{G}(s-m',D')\ln\mathcal{G}(s-m',D') = -\frac{1}{2}\ln\det(D') + \cdots, \qquad (8)$$

where "$\cdots$" denote terms that are independent of $m'$ and $D'$. Substituting (8) into the condition for the minimum of $G$ (3) and taking into account that $\frac{\delta\mathcal{G}(s-m',D')}{\delta D'} = \frac{1}{2}\frac{\delta^2\mathcal{G}(s-m',D')}{\delta s\delta s^+}$, we obtain the general equations

$$\int \mathcal{D}s\frac{\delta^2\mathcal{G}(s-m',D')}{\delta s\delta s^+}\ln P(s|d)\bigg|_{m'=m,D'=D} + TD^{-1} = 0,$$
$$\frac{\delta}{\delta m'}\int\mathcal{D}s\mathcal{G}(s-m',D')\ln P(s|d)\bigg|_{m'=m,D=D'} = 0. \qquad (9)$$

As we saw, the second equation of (9) at $T \to 0$ gives the MAP mean signal estimate, so we are now interested in obtaining an expression for $D(T)/T$ in this limit. Integrating by parts in the first equation of (9), and noting that $\mathcal{G}(s-m',D')$ tends to zero exponentially as $s \to \infty$, we can rewrite as

$$\int\mathcal{D}s\frac{\delta^2\mathcal{G}(s-m',D')}{\delta s\delta s^+}\ln P(s|d)\bigg|_{m'=m,D=D'}$$
$$= \int\mathcal{D}s\mathcal{G}(s-m',D')\frac{\delta^2\ln P(s|d)}{\delta s\delta s^+}\bigg|_{m'=m,D=D'}$$
$$\to \frac{\delta^2\ln P(s|d)}{\delta s\delta s^+}\bigg|_{s=m} \quad \text{as} \quad T\to 0, \qquad (10)$$

where we have taken into account that $\mathcal{G}(s-m',D') \to \delta(s-m')$ as $T \to 0$. Substituting (10) into (9), we will have $D = T(-\frac{\delta^2\ln P(s|d)}{\delta s\delta s^+}|_{s=m})^{-1}$, which coincides with the uncertainty estimate based on MAP if we move to $D \to \lim_{T\to 0}(D/T)$.

Thus, to obtain reliable results using MGFE, one should *always* set $D \to D/T$ at the end of the calculation. Only then does the MGFE principle fully reproduce MAP estimation in the limit $T \to 0$, for example. The authors failed to make this point clear, notwithstanding its crucial importance.

It is also stated in Ref. [1] that at $T \to \infty$, MGFE corresponds to maximum entropy (ME). The authors probably mean the unconstrained maximum of the entropy functional, and not the standard ME principle of Jaynes [3]. Namely, they state in the second paragraph that, "...maximum entropy alone cannot be the inference-determining criterion, since it favors states of a complete lack of knowledge, irrespective of the data." On the contrary, the ME principle does not always "favor states of a complete lack of knowledge" (which would correspond to a uniform distribution), and it especially does not favor any state "irrespective of the data." In reality, it singles out the particular distribution that can be realized in the largest number of ways [4] *with respect to all the given data* [3–7]. And in general, of course, the ME principle can indeed be the inference-determining criterion.

However, great care needs to be taken when employing principles like that of maximum entropy, as *all* available information must be taken into account. In Bayesian inference, the ME principle is usually used only for assigning prior probabilities [$P(s)$ in our case] [6,8], because finding the posterior fully in terms of ME is not appropriate to the case in which noise is present (see p. 949 of Ref. [4]).

Nonetheless, as a method of reasoning, the ME principle is very general and powerful. For illustrative purposes, let us try to apply it instead of Bayesian inference. Thus, following [5,7], for our case of finding $P(s|d)$ one can introduce the information entropy $S_I$ as

$$S_I = -\int\mathcal{D}s\,P(s|d)\ln\frac{P(s|d)}{Q(s)}, \qquad (11)$$

where $Q(s)$ is the invariant measure [5–7] which, roughly speaking, characterizes the density of points in $s$. The maximum entropy principle states that the most reasonable PDF ("uniquely determined as the one which is maximally noncommittal with regard to the missing information" [3]) that we can assume on the basis of the given information is the one for which $S_I$ (11) is maximal *subject to the constraints of all the given information*. The most exact, elegant, general, and understandable definition of the ME principle is given in the conclusion of Jaynes' celebrated paper [3]: "In the problem of prediction, the maximization of entropy is not an application of the law of physics, but merely a method of reasoning which ensures that no unconscious arbitrary assumptions have been introduced." Such a maximization involves all the information that we have and all the assumptions that we make, and thus it is dangerously easy to misuse. For the case considered in Ref. [1], the given information includes the data $d$, the model for the data $d = R[s] + n$, and all assumptions about the probability distributions $P(s)$, $P(d|s)$, and so on, as well as the normalization condition $\int\mathcal{D}s\,P(s|d) = 1$. If we are only given some average values $G_i = \int\mathcal{D}fg_i[f]P(f)$, such a maximization may be carried out through the use of Lagrange multipliers [3–7]. In our case, taking all the information into account is a highly nontrivial task, but one that is easily effected through Bayesian inference (moreover, the ME principle is in some sense incorporated within the Bayesian framework and related to MAP estimation; see pp. 949–950 of Ref. [4]).

It is important to note that, although Enßlin and Weig [1] *de facto* consider information entropy with a uniform invariant measure $Q(s) = 1$, it can be justified because, in the case under consideration, one can regard the signal $s$ as being defined on a uniformly continuous set of points. However, in the limit $T \to \infty$, the MGFE principle reduces to unconstrained maximization of $\int \mathcal{D}s\mathcal{G}(s-m,D)\ln\mathcal{G}(s-m,D)$ with respect to parameters $m$ and $D$ of the Gaussian approximation. Such an approach evidently takes no account at all of any of the given information, and thus does not correspond to the ME principle. Nevertheless, the MGFE principle (setting $D \to D/T$ after minimization) can still be valuable for $T > 1$ for some specific forms of posterior, but a more detailed investigation of this case is required.

Another very interesting case arises for $T = 1$ when the MGFE approximation to the posterior is based on the minimum discrimination information (MDI) principle [9,10], as shown by the authors. Note that the MDI principle can be regarded mathematically as an extension or generalization of the ME principle [11], as it is equivalent to maximization of the same quantity (11), but with some given probability $P^*(s)$ instead of the invariant measure $Q(s)$ (the latter can be regarded as the most ignorant probability distribution on a given continuous set of points: see p. 16 of Ref. [6]). In this case, MGFE amounts to finding the Gaussian approximation closest to the actual posterior in terms of MDI. It corresponds to a kind of MDI estimator (see, e.g., Ref. [12]), which can be useful as we will see below.

The authors of [1] showed that their MGFE principle includes both the MDI approach and MAP mean signal estimation. In this Comment, we have shown that, after setting $D \to D/T$, it also provides the correct MAP uncertainty estimate in the corresponding limit, thus fully reproducing the MAP approach. The fact that the Gibbs free energy used for Bayesian inference effectively incorporates both the full MAP and MDI estimations is indeed remarkable. It is instructive to investigate the relative quality of the MGFE-based approximation to the posterior at different $T$, together with other such approximations. To avoid complications, let us compare the MAP (MGFE at $T = 0$), MDI (MGFE at $T = 1$), and minimum mean-squared error [MMSE, $m_{\text{MMSE}} = \arg\min_{m} \int \mathcal{D}s(s-m)^2 P(s|d) = \langle s\rangle$] estimators. It is difficult to compare their performance in general: in some cases, one may perform better, in other cases, a different one may perform better. This is illustrated in Fig. 1, where the original posterior is compared with Gaussian approximations based on MDI, MAP, and MMSE for the (oversimplified) case of a signal consisting of a single point $x$. The parameters of these approximations for the two cases considered are given in Table I, together with characteristics of the original posteriors. From Fig. 1(a) we see that, in the case of simple symmetric posteriors, all estimators give identically good (true) mean signal estimates, while MDI gives much better uncertainty for the signal (Table I), resulting in a much closer approximation, as shown by the dotted line in Fig. 1(a). The situation changes when we move to the asymmetric posterior shown by the full curve in Fig. 1(b). In this case, the estimators considered give completely different approximations: MMSE gives the best mean signal (as it should), but too large an estimate of $D$; MAP gives a significantly incorrect $m$ (Table I), but a reasonable signal uncertainty; while MDI estimation results in an almost true $m$, but an underestimated $D$. An MGFE estimation at, e.g., $T = 0.5$ produces an average between MAP and MDI (not shown) in both cases.

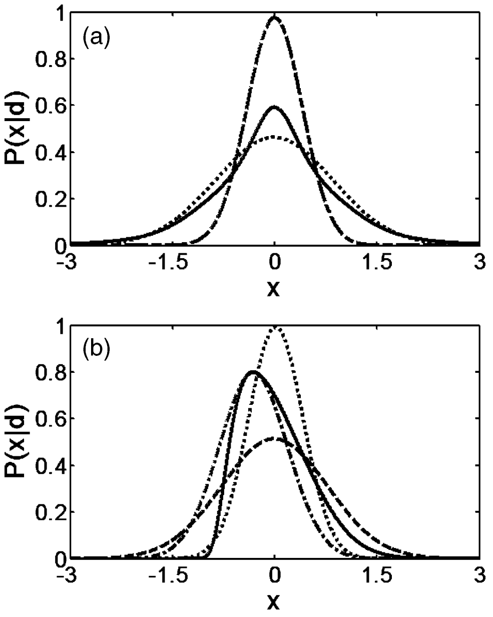

FIG. 1. The posterior $P(x|d)$ (full curves) compared with Gaussian approximations (2) based on MMSE (dashed), MAP (dash-dotted), and MDI (dotted) for (a) symmetric and (b) asymmetric posteriors. In (a) $P(x|d) \sim \frac{1}{1+3x^2e^{-x^2}+x^4}$ (and the MAP and MMSE approximations coincide); in (b) $P(x|d) \sim \exp[-x^2(1+e^{-3x})]$. Both original posteriors are shifted in $x$ to give zero mean $\langle x\rangle = 0$. Parameters of corresponding approximations are given in Table I.

Although we estimate the uncertainty matrix (which is scalar here) for MAP and MMSE as $D^{-1}_{\text{MAP,MMSE}} = -\frac{\partial^2 \ln P(x|d)}{\partial^2 x}|_{x=m_{\text{MAP,MMSE}}}$, it should be emphasized that MAP and MMSE originally estimate only the mean signal; estimation of the signal uncertainty, on the other hand, is an intuitive extension that does not necessarily work well. Such estimation fails when the peak (where the $m_{\text{MAP}}$ is applicable) is flat

TABLE I. Parameters $m$ and $D$ for MMSE, MAP, and MDI-based Gaussian approximations compared to the original values of mean signal $\langle x\rangle$ and signal uncertainty $\langle(x-\langle x\rangle)^2\rangle$ in the cases (a) and (b) of Fig. 1.

| Quantity | (a) | (b) |
|---|---|---|
| $\langle x\rangle$ | 0 | 0 |
| $m_{\text{MMSE}}$ | 0 | 0 |
| $m_{\text{MAP}}$ | 0 | −0.31 |
| $m_{\text{MDI}}$ | 0 | 0.02 |
| $\langle(x-\langle x\rangle)^2\rangle$ | 1.12 | 0.27 |
| $D_{\text{MMSE}}$ | 0.17 | 0.60 |
| $D_{\text{MAP}}$ | 0.17 | 0.25 |
| $D_{\text{MDI}}$ | 0.74 | 0.16 |

[consider, e.g., $P(x|d) \sim \frac{1}{1+x^4}$], so that the second derivative of the posterior is zero at $s = m_{\mathrm{MAP}}$, hence implying that $D_{\mathrm{MAP}} \to \infty$. The same concerns apply to MMSE-based uncertainty matrix estimation, for which $D$ can even reach negative values in rare cases. In general, therefore, it is not a good idea to base uncertainty estimation on either MAP or MMSE. In this sense MDI estimation is more general. MDI estimates of $m$ and $D$ maximize the resemblance of the Gaussian approximation to the original posterior, which makes such an estimator useful in cases when one needs to reuse the resultant posterior (e.g., for the next step of Bayesian inference). MMSE produces the best mean signal estimate (which is exact by definition), whereas MDI generally produces comparatively good estimates of both mean signal and signal uncertainty. The advantage of MAP is that it is the easiest approximation to use, although sometimes (e.g., for multiple-peaked posteriors) it even fails to find a reasonable estimate of the mean signal.

In summary, our comment is that the Gaussian approximation of the posterior based on the MGFE principle [1] should always be taken as $\mathcal{G}(s - m, D/T)$, where $m, D$ are the original MGFE estimates (3) at temperature $T$. Otherwise, such an estimation will contradict Bayes' theorem. This is a crucially important point. In particular, we have shown above that only in this form does MGFE fully incorporate the MAP principle as $T \to 0$. To avoid confusion, it also needs to be appreciated that MGFE does not in fact incorporate Jaynes' ME principle [3]. While this does not result in a significant disadvantage, it seems better to use MGFE in the range $0 \leqslant T \leqslant 1$, where it provides a compromise between the MAP and MDI approaches. We also investigated the effect of different temperature choices in MGFE, taking as examples $T = 0$ and 1. Although the optimal choice depends on the situation, and more detailed investigation is needed, we note that the choice of $T = 1$, corresponding to MDI estimation, performs quite well. This accounts for the good results obtained in the authors' subsequent work [2], where MGFE was used mainly for this particular temperature.

We thank T. A. Enßlin and C. Weig for the collegial way in which they discussed the science behind our Comment. Our research was supported by the Engineering and Physical Sciences Research Council (UK) (Grant No. EP/100999X1).

[1] T. A. Enßlin and C. Weig, Phys. Rev. E **82**, 051112 (2010).

[2] N. Oppermann, G. Robbers, and T. A. Ensslin, Phys. Rev. E **84**, 041118 (2011).

[3] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[4] E. T. Jaynes, Proc. IEEE **70**, 939 (1982).

[5] E. T. Jaynes, in *Statistical Physics*, Vol. 3, edited by K. Ford (Benjamin, New York, 1963), pp. 181–218.

[6] E. T. Jaynes, IEEE Trans. Syst. Sci. Cybernet. **SSC4**, 227 (1968).

[7] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).

[8] S. F. Gull, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, edited by G. J. Erickson and C. R. Smith (Kluwer, Dordrecht, 1988), pp. 53–74.

[9] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[10] S. Kullback and M. A. Khairat, Ann. Math. Statist. **37**, 279 (1966).

[11] I. J. Good, Ann. Math. Statist. **34**, 911 (1963).

[12] E. S. Soofi and D. V. Gokhale, Comput. Stat. Data Anal. **11**, 165 (1991).