Sketch of replies to N. Comp. reviews

16 March 2020; updated 16 March 2020

Reviewer 1

(1). First, I would like to direct attention to the paper by Amari, Nakahara, Wu and Sakai in Neural Computation 2003, which the authors also cited in this paper. In this paper, Amari et al showed that "weak higher-order interactions of almost all orders are required for realizing a widespread activity distribution." (Theorem 3). As shown in the case of pairwise model (Eq.3.19 in their paper), if the fixed moments are limited and do not have higher-order, a distribution will be concentrated when we increase the number of neurons.

The suggested ME model by this paper constrains a few moments as opposed to larger number of neurons, which can not result in the wide-spread distributions according to the above theory. In other words, I speculate that the bimodal structure is an artifact of constraining fewer number of moments for a model with such large number of neurons. This could happen even if we use up to all 65th moments. I would like to know how the authors think of Amari's paper in the light of the current manuscript.

We think that the arguments of Amari et al. (2003) do not apply to the inference in our manuscript, for several reasons:

i. We did perform the calculations with an increasing number of moments as sufficient statistics, eventually using the full sample distribution. As mentioned on p. 9 and shown in § 4, from about 5 moments upwards the final result doesn't change appreciably.

   Most important, as we discuss in §§ 6 and 7, our results are the same *without* using any sufficient statistics – that is, *no maximum-entropy* model at all – under a fully Bayesian nonparametric approach. In fact, the fully nonparametric approach leads to slightly wider distributions (we hope to publish these results soon).

ii. As N increases, the two peaks we infer do become more pronounced, as suggested by our Fig. 1. So our inference is not really in contradiction with Amari et al.. The results of that paper hold for the $N \to \infty$, and no explicit form for the limit-remainder is given. It is therefore unclear whether at $N = 10\,000$ we should already observe a delta. Our findings i., above, suggest that we should not. Limit-reminders such as '$\mathrm{O}(1/N^k)$' as given in Amari et al. do not help unless we know their precise form. As an extreme example, $\frac{10^{10}}{N}$ is $\mathrm{O}(1/N)$, and in this case the $1/N$ term cannot really be neglected unless $N \gg 10^{10}$.

iii. The results of Amari et al. assume that the probability of the neuronal activity factorizes over time. In our manuscript we assume, as explained in § 6, that our degree of belief is *exchangeable* under time-bin permutations, but this exchangeability *does not imply factorizability* over time, as is clear from de Finetti's theorem. So we are not sure whether Amari et al.'s results apply at all to our inference.

(2) [we can reply by applying to real data.]

(3) I was not satisfied with the section of the Bayes factor. The Bayes factor (Eq.10) was calculated for models with different dimensionality. Then, a larger model should better account for the data. How do you determine whether you include the higher-order or not, in other words how do you determine the significance of the delta? If the models have the same dimension, one may use heuristics, for example given by Kass and Raftery (JASA 1995). If you have nesting models, one may resort to chi-squared test? Just computing the BF without such a verification method, providing the quantity BF adds only marginally to the visual inspection of the distributions.

Let us emphasize, first of all, that the purpose of our paper is to make inferences about the total activity of the larger population, not to decide among different sufficient statistics or correlation orders. Section 4 is meant to show the implications of our method for that kind of discussions. In particular we want to show that *conclusions based on a sample and conclusions based on the larger population can be very different*. Our method can bypass that difficulty, if one wants to use it for such comparisons.

We partly agree with the reviewer's concern, but for different reasons. The simple comparison of the log-evidences is not a problem in itself, as explained below. The problem is what is meant by 'model' and what is being compared. Under one interpretation, our calculation of the log-evidence is missing a factor, which comes from the full Bayesian analysis. This factor, however, should be roughly the same in the calculations from the sample and from the larger population, so our point that the two calculations lead to different conclusions still holds. Under another interpretation our calculation is correct as it is. We explain this below.

In any case, please note that we avoid speaking about 'significance', also as recommended by the American Statistical Association in their recent official statements (ASA 2016; 2019 see especially § 2). When comparing hypotheses M_i we simply want their post-data probabilities $\mathrm{P}(M_i \mid D)$, with their continuum of cases. Other researchers are of course free to make dichotomies if they like; but we personally agree with the American Statistical Association that any threshold is unnecessary and arbitrary. If a *choice* among the hypotheses has to be made for some

purpose (say, hardware algorithm implementation in some automated device), such choice requires not only the probabilities $\mathrm{P}(M_i \mid D)$, but also the utility/loss matrix for the various choices conditional on the various hypotheses (Kadane & Dickey 1980). The choice is then made by maximization of the expected utility, a procedure dictated by some simple rational desiderata (Fishburn 1981; Bernardo & Smith 2000; Jaynes 2003 ch. 13). Some utility matrices can even lead to choosing the hypothesis having *lower* post-data probability; this frequently happens in medical decision problems, see e.g. Sox et al. (2013).

There are many differing, fragmented views on the questions of comparison, complexity, nesting, and 'penalization' of models. Our view is close to that expressed in several works, studies, and commentaries[1], and well-explained in MacKay (1992). According to this view, which we summarize below, the log-evidence automatically takes care of model complexity.

The main question is what we mean by 'model'. There are two main possibilities: model as a family, or model as a specific distribution.

(A) If by 'model' we mean a family of probability distributions, e.g. (apart from normalization)

$$\left\{ \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{d}x \;\middle|\; \mu \in \mathbf{R}, \sigma \in \left]0, \infty\right[\right\} ,$$

then model comparison is ill-defined. It may happen that the optimal distribution in family $\mathcal{F}_1$ assigns more probability to the observed data than the optimal distribution in family $\mathcal{F}_2$, and yet $\mathcal{F}_1$ assigns 'on average' lower probabilities to the data than $\mathcal{F}_2$. It is not clear how the two families should be compared in this case, or what $\mathrm{P}(D \mid \mathcal{F}_i)$ is. A hypothesis is ill-defined, or at least useless, if it does not allow us to assign a probability to the data. Such a probability can be obtained if we specify a pre-data probability distribution over the distributions in the family, e.g. $\mathrm{p}(\mu, \sigma \mid I)\,\mathrm{d}\mu\,\mathrm{d}\sigma$ in the example above. 'Model' must therefore be redefined as a family of distributions *and* a pre-data distribution over such family. This redefinition has several consequences:

First, two models are different if they have the same family but different distributions over it. In fact they lead to different probabilities on the data space. Two such models are also mutually exclusive, $\mathrm{P}(M_1 \wedge M_2 \mid I) = 0$,

[1] e.g. Good 1950; 1985; Kadane & Dickey 1980; Jones et al. 1986; Chatfield 1995; Copas et al. 1995; Draper 1995; Spiegelhalter et al. 1995; Raftery 1995; Gelman et al. 1995; MacKay 2003; Draper et al. 1996; Hoeting et al. 1999; Clyde et al. 1999; Browne & Draper 2006.

because we cannot use the one family distribution *and* the other family distribution at the same time. It's either-or. This consequence makes sense from a nonparametric perspective, where a parametric model (in the redefined sense) is just a delta-like nonparametric distribution, concentrated on a lower-dimensional manifold (the family above) of the full space of distributions (Draper et al. 1996 p. 761).

Second, the 'complexity' of a model cannot be judged from the dimensionality of its family. Model M_1 can have a family distribution concentrated on a very small region, with respect to model M_2 having the same family. M_1 is thus effectively using fewer parameters than M_2. Correspondingly the 'typical set' of data to which M_1 assigns appreciable probability is smaller than the typical set of M_2: model M_1 is less 'powerful'. The effective dimensionality or complexity of a model is thus best defined by the volume of data that is typical under that model. Most important, it may then happen that a model is more powerful than another even if it is defined on a lower-dimensional family.

Third, *the evidence* $\mathrm{P}(D \mid M_i)$ *of a model automatically takes care of its complexity*. This happens because the probability over the data space is normalized. If M_1 has a typical-data volume larger than M_2's, then it must also assign lower probability to each data point in that volume, to ensure normalization. As a result, if the observed data D^* lie in the typical volumes of both models, then $\mathrm{p}(D^* \mid M_1) < \mathrm{p}(D^* \mid M_2)$: *the evidence is automatically penalizing the larger model*. There is a continuous trade-off between how typical the data is for each model and each model's power. This is the so-called automatic Ockham razor of Bayes's theorem, extensively explained by MacKay (1992; also 2003 § 28.1). And this is the reason why we are personally satisfied with simply comparing the evidence for any two models.

In our calculations, however, we are using a zeroth-order Laplace approximation for calculating the evidence in § 4. A first-order approximation (containing the width of the distribution, influenced by the typical-set volume above) would correct for the model's power. We believe that the correction would nevertheless be the same in the calculation based on the larger population, eqs (11), and on the sample, eqs (12), because the data space is the same. Hence the difference in the two approaches – our main message – should persist. But it would be better to check this.

(B) If by 'model' we mean a specific distribution, e.g.

$$\exp\left[-\frac{(x-5)^2}{18}\right] \mathrm{d}x \ ,$$

then two models can of course be directly compared: $\mathrm{P}(D \mid M_i, I)$ is well-defined. In this case the 'dimensionality' of the model is undefined, however. The distribution above could be considered as belonging to the family $\left\{\exp\left[-\frac{(x-\mu)^2}{18}\right]\right\}$, or $\left\{\exp\left[-\frac{(x-5)^2}{2\sigma^2}\right]\right\}$, or $\{\exp[f(x-5)]\}$, and so on. A point in a manifold belongs to an infinity of submanifolds.

But the property of having a minimal sufficient statistic is still well-defined for a model in this sense, in an exchangeable context. The model above has $\{n, \bar{x}, \overline{x^2}\}$ (or any equivalent set) as minimal sufficient statistics, where the bar denotes the sample average. Thus two models having different sufficient statistics can be compared without reference to any family. Interpreted in this sense our calculations of § 4 are correct.

The question, in summary, is what exactly is being compared, whether families (with accompanying distribution) or single distributions. The literature is not at all clear. But, in either case, our calculations show that the approaches based on a population and a sample thereof can lead to different conclusions – in accordance with the non-preservation of sufficiency under sampling – and they provide a way to bypass this problem.

(4) I enjoyed somewhat philosophical arguments regarding the interpretation of the ME distribution (P8 and Appendix). However, I was not sure how these are relevant to the paper if it is just an interpretation. Can we consider different outcomes arising the different interpretations? If the authors are preparing another manuscript in this direction, these arguments may better be used for that paper? (This is just a weak suggestion.)

As stated on p. 8, the interpretation is important because it leads to numerically different results in first- and higher-order approximations; Appendix C explains where the differences appear. Also, it is important to know whether we are calculating the frequency of past results, or the probability for a new result, or the most probable frequency of future results. These three quantities are very different when the ratio $\frac{\#\text{ samples}}{\text{data-space dimension}}$ is high, and they are to be used in different ways.

So these are not philosophical matters, it's a matter of understanding what we are doing and stating it clearly in our work. There's also another important reason. We are generally concerned for master's and PhD students, who often learn concepts and methods through actual research, perusing works like the present one. Young researchers too often end up learning incorrect or confused ideas, gathered from unclear literature. They trust it because it's peer-reviewed. Sometimes the result, quoting Feynman (1989), is that 'they don't learn by understanding; they learn by some other way – by rote, or something. Their knowledge is so fragile!'. We

owe them clarity, precision, honesty, and at least a glimpse of the bigger picture.

Reviewer 2

2. The method requires that the measurements on n neurons are 'representative' of the full population of N neurons. The authors show one example, where they divide a dataset into two subsets of neurons with different statistics. They show that the extrapolations from these two subsets are significantly different. But this particular analysis is somewhat ad hoc. Are there any more systematic tests of procedures for understanding how sensitive this analysis is to choosing different subsets of neurons? Or different recoding periods?

3. In particular, correlations tend to be stronger between nearby neurons than far-away neurons. This should be a consistent pattern that shows up in most experimental datasets and which should lead to systematic biases in inferring the properties of populations, N, which are large enough to result in lower average correlation (see Nonnenmacher et al., PLoS CB 2017). Is there any way to take this dependence into account?

Bibliography

Amari, S.-i., Nakahara, H., Wu, S., Sakai, Y. (2003): *Synchronous firing and higher-order interactions in neuron pool*. Neural Comp. **15**[1], 127–142.

ASA (American Statistical Association) (2016): *ASA statement on statistical significance and p-values*. Am. Stat. **70**[2], 131–133. Ed. by R. L. Wasserstein. See also introductory editorial in Wasserstein, Lazar (2016) and discussion in Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman, Altman, et al. (2016).

— (2019): *Moving to a world beyond "$p < 0.05$"*. Am. Stat. **73**[S1], 1–19. Ed. by R. L. Wasserstein, A. L. Schirm, N. A. Lazar.

Bernardo, J.-M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., eds. (1985): *Bayesian Statistics 2*. (Elsevier and Valencia University Press, Amsterdam and Valencia). https://www.uv.es/~bernardo/valenciam.html.

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). First publ. 1994.

Browne, W. J., Draper, D. (2006): *A comparison of Bayesian and likelihood-based methods for fitting multilevel models*. Complexity **1**[3], 473–513. See also comments and rejoinder in Gelman, Kass, Natarajan, Lambert, Browne, Draper (2006).

Chatfield, C. (1995): *Model uncertainty, data mining and statistical inference*. J. Roy. Stat. Soc. A **158**[3], 419–444. See also discussion in Copas, Davies, Hand, Lunneborg, Ehrenberg, Gilmour, Draper, Green, et al. (1995).

Clyde, M., Draper, D., George, E. I., Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999): *[Bayesian model averaging: a tutorial:] Comments and rejoinder*. Stat. Sci. 412–417. See Hoeting, Madigan, Raftery, Volinsky (1999).

Copas, J. B., Davies, N., Hand, D. J., Lunneborg, C. E., Ehrenberg, A. S., Gilmour, S. G., Draper, D., Green, P. J., et al. (1995): *Discussion of the paper by Chatfield [Model uncertainty, data mining and statistical inference]*. J. Roy. Stat. Soc. A **158**[3], 444–466. See Chatfield (1995).

Draper, D. (1995): *Assessment and propagation of model uncertainty*. J. Roy. Stat. Soc. B **57**[1], 45–70. See also discussion and reply in Spiegelhalter, Grieve, Lindley, Lindley, Copas, Cairns, Chatfield, Box, et al. (1995). https://classes.soe.ucsc.edu/ams206/Winter05/draper.pdf.

Draper, D., Hill, B. M., Kass, R. E., Wasserman, L., Lewis, S. M., Raftery, A. E., Rubin, D. B., Weerahandi, S., et al. (1996): *Comments. Utility, sensitivity analysis, and cross-validation in Bayesian model-checking. Posterior predictive assessment for data subsets in hierarchical models via MCMC. On posterior predictive p-values. rejoinder*. Stat. Sinica **6**[4], 760–808. See Gelman, Meng, Stern (1996).

Feynman, R. P. (1989): *"Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character*, repr. (Bantam, New York). 'As told to Ralph Leighton', ed. by Edward Hutchings. First publ. 1985.

Fishburn, P. C. (1981): *Subjective expected utility: a review of normative theories*. Theory Decis. **13**, 139–199.

Gelman, A., Kass, R. E., Natarajan, R., Lambert, P. C., Browne, W. J., Draper, D. (2006): *[Comments on article by Browne and Draper:] prior distributions for variance parameters in hierarchical models. A default conjugate prior for variance components in generalized linear mixed models. Comment. Rejoinder*. Complexity **1**[3], 515–549. See Browne, Draper (2006).

Gelman, A., Meng, X.-L., Stern, H. (1996): *Posterior predictive assessment of model fitness via realized discrepancies*. Stat. Sinica **6**[4], 733–760. See also comments and rejoinder in Draper, Hill, Kass, Wasserman, Lewis, Raftery, Rubin, Weerahandi, et al. (1996).

Gelman, A., Rubin, D. B., Hauser, R. M., Raftery, A. E. (1995): *Avoiding model selection in Bayesian social research. Better rules for better decisions. Rejoinder: model selection is unavoidable in social research*. Sociol. Methodol. **25**, 165–195. See Raftery (1995).

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

— (1985): *Weight of evidence: a brief survey*. In: Bernardo, DeGroot, Lindley, Smith (1985), 249–270. With discussion by H. Rubin, T. Seidenfeld, and reply.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., Altman, D. G., Altman, N. S., et al. (2016): *Online supplement and discussion: ASA statement on statistical significance and p-values*. Am. Stat. **70**[2], 129. http://www.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108. See ASA (2016) and Wasserstein, Lazar (2016).

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T. (1999): *Bayesian model averaging: a tutorial*. Stat. Sci. **14**[4], 382–412. See also Clyde, Draper, George, Hoeting, Madigan, Raftery, Volinsky (1999).

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jones, B., Spiegelhalter, D. J., Kimber, G. R., Healy, M. J. R., Gore, S. M., Armitage, P., Ross, G. J. S., Barnard, G. A., et al. (1986): *Discussion of the paper by Mr. Racine* et al. Appl. Statist. **35**[2], 121–150. See Racine, Grieve, Flühler, Smith (1986).

Kadane, J. B., Dickey, J. M. (1980): *Bayesian decision theory and the simplification of models*. In: *Evaluation of econometric models*. Ed. by J. Kmenta, J. B. Ramsey (Academic Press, New York), 245–268. https://www.nber.org/chapters/c11704.

MacKay, D. J. C. (1992): *Bayesian interpolation*. Neural Comp. **4**[3], 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html.

MacKay, D. J. C. (2003): *Information Theory, Inference, and Learning Algorithms*. (Cambridge University Press, Cambridge). http://www.inference.phy.cam.ac.uk/mackay/itila/. First publ. 1995.

Racine, A., Grieve, A. P., Flühler, H., Smith, A. F. M. (1986): *Bayesian methods in practice: experiences in the pharmaceutical industry*. Appl. Statist. **35**[2], 93–120. See also discussion and reply in Jones, Spiegelhalter, Kimber, Healy, Gore, Armitage, Ross, Barnard, et al. (1986).

Raftery, A. E. (1995): *Bayesian model selection in social research*. Sociol. Methodol. **25**, 111–163. See also comments and rejoinder in Gelman, Rubin, Hauser, Raftery (1995). https://www.stat.washington.edu/raftery/Research/PDF/socmeth1995.pdf.

Sox, H. C., Higgins, M. C., Owens, D. K. (2013): *Medical Decision Making*, 2nd ed. (Wiley, New York). First publ. 1988.

Spiegelhalter, D. J., Grieve, A. P., Lindley, D. V., Lindley, D. V., Copas, J. B., Cairns, A. J. G., Chatfield, C., Box, G., et al. (1995): *Discussion of the paper by Draper [Assessment and propagation of model uncertainty]*. J. Roy. Stat. Soc. B **57**[1], 71–97. See Draper (1995). https://classes.soe.ucsc.edu/ams206/Winter05/draper.pdf.

Wasserstein, R. L., Lazar, N. A. (2016): *The ASA's statement on p-values: context, process, and purpose*. Am. Stat. **70**[2], 129–133. See ASA (2016) and discussion in Greenland, Senn, Rothman, Carlin, Poole, Goodman, Altman, Altman, et al. (2016). https://catalyst.harvard.edu/pdf/biostatsseminar/ASA_s_statement_on_p_values_context_process_and_purpose.pdf.