Comments on "Maximum Entropy Theory in Ecology"

15 April 2016

1. For statisticians, any distribution of the form

$$p(x;\theta,h) = \frac{h(x)\exp(\theta \cdot t(x))}{z(\theta,h)}$$

is an exponential family.

2. The function $h(x)$ is called the "reference measure", "dominating measure" or "base measure"[1] . Notice that I say "measure" rather than "distribution", because $h(x)$ isn't necessarily normalized. I have never found this called a "prior"; in statistical jargon, "prior" and "posterior" always refer to distributions before and after conditioning on some event, and indeed "prior" almost always refers to a distribution over parameters, i.e., functionals of the whole distribution. It is well-known that maximizing the entropy under expectation-value constraints is *not* equivalent to conditioning on those constraints (Friedman and Shimony, 1971).

3. Certain results about exponential families simplify when the base measure is uniform, but others are indifferent to h. In particular:

   (a) $\mathbb{E}_{\theta,h}[T] = \nabla_\theta \log z(\theta,h)$, and

   (b) The MLE is the $\hat{\theta}$ where $\mathbb{E}_{\theta,h}[T] = t$, i.e., the MLE equates the observed value of the sufficient statistic and its expected value.

   (c) Therefore the maximum entropy solution is just the MLE in the exponential family where the dominating measure is uniform,

$$p(x;\theta,1) = \frac{\exp(\theta \cdot t(x))}{z(\theta,1)}$$

4. What you're talking about the special case where mechanistic models are also exponential families, but where the mechanistic model implies

[1] If we want to be pedantic, $h(x)$ is the density of those measures with respect to some other reference measure, typically Lebesgue measure (for continuous x) or counting measure (for discrete x).

a parameter-free dominating measure which is not the uniform measure. The latter of course don't solve the maximum entropy problem, but they *do* solve the minimum relative entropy problem, of minimizing $D(p\|h)$, *if* the dominating measure h is a normalized probability distribution. (I remember the proof in Cover and Thomas being especially clear.) — Notice the direction of the divergence, here, by the way.

5. You thus are really talking about two distinct, though related, statistical models, call them U for the uniform base measure and N for the non-uniform one. There is also a true data-generating distribution Q, which is not necessarily in either U or N. Each model has a "least false" or "pseudo-true" parameter value, which minimizes the KL divergence from P. In the one case it's the

$$\theta_N = \operatorname*{argmin}_{\theta} D(Q\|P(\theta, h))$$

and in the other it's

$$\theta_U = \operatorname*{argmin}_{\theta} D(Q\|P(\theta, 1))$$

In general, not only will θ_N and θ_U be different, but they will lead to different distributions over X and even over T. They only *have* to agree in the expected values of the sufficient statistics, which will be $\mathbb{E}_Q[T]$. That is, the least-false parameter values in each exponential family are the ones where the expected values of the sufficient statistics match the true (population / ensemble / process) expectations.

6. For IID observations, the MLE within each family with converge on the least-false parameter value, i.e., on θ_N or θ_U. This is *not* necessarily the case with dependent data; it is perfectly possible for the MLE of an exponential family to be an in-consistent estimator. Shalizi and Rinaldo (2013) gives separate necessary and sufficient conditions for this consistency. The necessary condition is an algebraic one, about how to update the sufficient statistic when more data becomes available. The sufficient condition is more probabilistic; it's that as the sample size S grows, $f(S)\log z(\theta) \to a(\theta)$, for some shrinking function f and some size-independent a. Generally speaking, when the sufficient condition is met, the MLE will have Gaussian fluctuations around the least-false parameter value, of order $n^{-1/2}$.

   For the rest of this note, I will presume that the MLE is consistent, and has the usual asymptotics. But it needs to be very clearly understood that this is *not* guaranteed by the existence of a sufficient statistic or the use of principle of maximum entropy, etc.

7. Your equation 9 is just the log likelihood ratio between two families of statistical models, which you're evaluating, as usual, at their MLEs. If

you normalized by the sample size and took the limit as the sample grew to infinity, this will generally converge to the difference in KL divergence from the source distribution Q, i.e., to

$$D(Q\|P(\theta_U, 1)) - D(Q\|P(\theta_N, h))$$

The finite-sample behavior of your eq. 9 is a well-studied topic in statistics. It depends crucially on whether the two least-false distributions are the same.

(a) If the least-false distributions are distinct, then your Eq. 9 has Gaussian fluctuations of order $1/\sqrt{n}$ around the limiting value I give above.

(b) If the least-false distributions coincide, one needs to do much more complicated math. It simplifies if one model is fully nested inside the other, in which case one gets back the usual χ^2 distribution.

On all of this, the best source is Vuong (1989).

— It is important to realize that this test is about *relative* fit of the two models to the true distribution P. It does not assess the absolute goodness of fit of either model.

8. The manuscript begins by asking why one should constrain just the statistics which Harte specifies. But the procedure you propose is quite incapable of answering this. It compares two exponential families with the same sufficient statistics but different dominating measures; it therefore presumes that the question of the correct sufficient statistics has been settled.

9. Suppose that one does find the non-uniform base measure is a better fit to the data than a uniform base measure. This is not necessarily very strong endorsement of a particular mechanism. For instance, the Poisson distribution and the geometric distribution share the same sufficient statistic and the same support, but the Poisson has a non-uniform dominating measure. But since there are tons of ways of getting Poisson distributions, or approximately Poisson distributions, the step from "the Poisson fits better than the geometric" to "this particular way of getting Poissons is a good model here" is a very large one. Indeed, a Poisson distribution could fit *better* than a geometric when even both fits are plainly horrible.

10. On the other hand, suppose we're agreed that about the sufficient statistics, and a uniform base measure fits better than a non-uniform one. Let's even suppose — this couldn't be deduced from the likelihood-ratio test — that the exponential family with a uniform base is correct. *This must have an explanation.* I cannot stress too much that "maximizing the entropy under constraints on expectation values" is simply not a principle with any independent logical or probabilistic basis (Seidenfeld, 1979, 1987; Csiszár, 1995; Kass and Wasserman, 1996; Uffink, 1995, 1996). There is no good

reason to treat it as a preferred reference point, deviations from which require explanation.

There are probabilistic accounts of why maximum entropy works so well in thermodynamics — I refer you specifically to Csiszár (1995) — but they rely on our only being *able* to interact with thermodynamic systems through selected, extensive functions of the microscopic state. Even if the exponential family distribution with a uniform base is exactly correct, there needs to be a mechanistic explanation of why those statistics are sufficient and why there should be a uniform distribution over microscopic states[2]. In thermodynamics the explanation for the uniform distribution is provided by Liouville's theorem, but in ecology that hardly applies.

11. It is not in fact true that statistical models with sufficient statistics must be exponential families. The Darmois-Koopman-Pitman theorem applies when (i) the sufficient statistic has finite dimension, (ii) the dimension of the sufficient statistic does not grow with the sample size, and (iii) the support of the distribution does not change with the parameters. Thus the uniform distributions on intervals, of the form $[\theta_1, \theta_2]$, have as their minimal sufficient statistics the sample minimum and maximum, but they cannot be put into exponential family form. Of course it doesn't agree with the maximum entropy distribution where the sample minimum and maximum match their expected values.

    Lauritzen (1984, 1988) has made an extensive and exhaustive study of families with sufficient statistics, showing that they can be written as multiplicative characters of semi-groups, where the semi-group is how the sufficient statistic gets updated as the sample size grows. Ordinary exponential families are special cases, with an algebraic characterization I won't try to reproduce here.

12. That being said, most statistical models — even very ordinary ones — do not have finite-dimensional sufficient statistics. An example in point is the Yule-Simon distribution, $p(x;\theta) = \theta B(x; \theta+1)$, where B is the beta function. This has power-law tails of exponent $\theta+1$, but it's easy to show that it isn't an exponential family, no matter what choice of base measure one might propose. It is a little harder to show, but still possible, that the minimal sufficient statistic for this family is the empirical distribution or empirical measure,

$$\frac{1}{n}\sum_{i=1}^{n}\delta(x-x_i)$$

    I bring up this example because preferential attachment does not, in fact, lead to a pure power law; it leads to a Yule-Simon distribution (Simon, 1955; Price, 1965; Bornholdt and Ebel, 2001).

[2] Cf. Dias and Shimony (1981) on the connection between maximum entropy and a uniform distribution over states.

13. In fact, for all models of IID without very special structure, the empirical measure *is* the minimal sufficient statistic, regardless of the dimension of the parameter space. (See Lauritzen again.) The empirical measure does reduce the data, because it throws away information about the order of observations, but, to repeat myself, even for very simple models, there is generally no further reduction possible. For models of dependent data, even less reduction is possible: e.g., the empirical pair measure, which puts a delta function on each pair of successive observations, is generally the minimal sufficient statistic for Markov models (Diaconis and Freedman, 1980).

    Constructing the maximum entropy distribution compatible with the expected empirical measure equaling its observed value is — instructive. This, however, is what the "least informative distribution compatible with the data" would actually be.

    The upshot of the last few points is that it's very easy to come up with very simple, plausible models for which "testing whether the exponential family has a uniform dominating measure" is quite inapplicable, for merely technical reasons.

14. Mechanistic models generally do not just predict an equilibrium or invariant distribution. They predict *dynamics*, and many aspects of them. Take preferential attachment as an example. If one has data at multiple time points, this makes predictions about the rate of growth in degree for each node. Even if one doesn't have data at that resolution, it predicts that the oldest nodes should (on average) have the highest degrees, and can even be solved for the age-degree relationship. Now, as I noted above, preferential attachment leads to a distribution which is only *approximately* a power law, but suppose it were exactly a power law. This might still be a superior model to the brute posit of "the degree distribution is a power law", which is what "the base measure is uniform and the sufficient statistic is mean log degree" is equivalent to. Preferential attachment could be a superior model, if those other predictions check out, because then preferential attachment is *explaining* why we should see a power law in cross-section, *and* a size-age relationship, *and* such-and-such growth rates.

15. Now, as it happens, there are not actually many networks with power-law degree distributions (Khanin and Wit, 2006; Clauset *et al.*, 2009); perhaps not any. There are lots of networks with heavy-tailed, highly right-skewed degree distributions, and even a respectable number where part of the distribution looks power-law-ish if you don't inspect it too closely. One reason to prefer a mechanistic model over a maximum entropy model, even when they predict exactly the same marginal distribution, is that the mechanistic model presents vastly more possibilities for explaining the departures from pure power laws in *principled* ways, by modifying the mechanisms. If the distribution to be explained is close to a power law, and our initial mechanistic model produces a power law, then if there is a

reasonable degree of "continuity of approximation" (Simon, 1963), a mechanistic model close to the original should work. Moreover, the search for that new mechanistic model can be principled, because we could, at least in principle, come up with independent evidence for the modifications, by examining the other predictions of the new model, beyond the headline distribution.

By contrast, the maximum entropy approach to matching deviations from the initial exponential family is either guess-and-hope, or mere curve-fitting (Barron and Sheu, 1991). At best, the curve-fitting might stylize our view of what is to be explained, or possibly suggestions as to what relevant sufficient statistics might be. I would not hold out too much hope for the last, though, because that approach ends up telling you "constrain not only the original variable of interest X, but the following m polynomials in X, and, by the way, let $m \to \infty$ as the sample size grows".

16. Again, if ecologists find that they've built a mechanistic model which implies an exponential family with a non-uniform base measure, and they want to test whether that's close to the true distribution than the same exponential family with a uniform base measure, a likelihood-ratio test is a perfectly reasonable way to go. If they find that the uniform base measure is better, then the mechanistic model is wrong, and they need to understand why.

References

Barron, Andrew R. and Chyong-Hwa Sheu (1991). "Approximation of Density Functions by Sequences of Exponential Families." *Annals of Statistics*, **19**: 1347–1369. URL `http://projecteuclid.org/euclid.aos/1176348252`. doi:10.1214/aos/1176348252.

Bornholdt, Stefan and Holger Ebel (2001). "World-Wide Web scaling exponent from Simon's 1955 model." *Physical Review E*, **64**: 035104. URL `http://arxiv.org/abs/cond-mat/0008465`.

Clauset, Aaron, Cosma Rohilla Shalizi and M. E. J. Newman (2009). "Power-law Distributions in Empirical Data." *SIAM Review*, **51**: 661–703. URL `http://arxiv.org/abs/0706.1062`.

Csiszár, Imre (1995). "Maxent, Mathematics, and Information Theory." In *Maximum Entropy and Bayesian Methods: Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods* (Kenneth M. Hanson and Richard N. Silver, eds.), pp. 35–50. Dordrecht: Kluwer Academic.

Diaconis, Persi and David Freedman (1980). "De Finetti's Theorem for Markov Chains." *Annals of Probability*, **8**: 115–130. URL `http://projecteuclid.org/euclid.aop/1176994828`. doi:10.1214/aop/1176994828.

Dias, Penha Maria Cardoso and Abner Shimony (1981). "A Critique of Jaynes' Maximum Entropy Principle." *Advances in Applied Mathematics*, **2**: 172–211. doi:10.1016/0196-8858(81)90003-8.

Friedman, Kenneth and Abner Shimony (1971). "Jaynes' Maximum Entropy Prescription and Probability Theory." *Journal of Statistical Physics*, **3**: 381–384. doi:10.1007/BF01008275.

Kass, Robert E. and Larry Wasserman (1996). "The Selection of Prior Distributions by Formal Rules." *Journal of the American Statistical Association*, **91**: 1343–1370. URL `http://www.stat.cmu.edu/~kass/papers/rules.pdf`. doi:10.1080/01621459.1996.10477003.

Khanin, Raya and Ernst Wit (2006). "How Scale-Free Are Biological Networks?" *Journal of Computational Biology*, **13**: 810–818. URL `http://iwi.eldoc.ub.rug.nl/root/2006/JCompBiolKhanin/`. doi:10.1089/cmb.2006.13.810.

Lauritzen, Steffen L. (1984). "Extreme Point Models in Statistics." *Scandinavian Journal of Statistics*, **11**: 65–91. URL `http://www.jstor.org/pss/4615945`. With discussion and response.

— (1988). *Extremal Families and Systems of Sufficient Statistics*. Berlin: Springer-Verlag.

Price, Derek J. de Solla (1965). "Networks of Scientific Papers." *Science*, **149**: 510–515.

Seidenfeld, Teddy (1979). "Why I Am Not an Objective Bayesian: Some Reflections Prompted by Rosenkrantz." *Theory and Decision*, **11**: 413–440. URL `http://www.hss.cmu.edu/philosophy/seidenfeld/relating%20to%20other%20probability%20and%20statistical%20issues/Why%20I%20Am%20Not%20an%20Objective%20B.pdf`. doi:10.1007/BF00139451.

— (1987). "Entropy and Uncertainty." In *Foundations of Statistical Inference* (I. B. MacNeill and G. J. Umphrey, eds.), pp. 259–287. Dordrecht: D. Reidel. URL `http://www.hss.cmu.edu/philosophy/seidenfeld/relating%20to%20other%20probability%20and%20statistical%20issues/Entropy%20and%20Uncertainty%20(revised).pdf`.

Shalizi, Cosma Rohilla and Alessandro Rinaldo (2013). "Consistency Under Sampling of Exponential Random Graph Models." *Annals of Statistics*, **41**: 508–535. URL `http://arxiv.org/abs/1111.3054`. doi:10.1214/12-AOS1044.

Simon, Herbert A. (1955). "On a Class of Skew Distribution Functions." *Biometrika*, **42**: 425–440. URL `http://www.jstor.org/pss/2333389`.

— (1963). "Problems of Methodology — Discussion." *American Economic Review*, **53**: 229–231. URL `http://www.jstor.org/stable/1823866`.

Uffink, Jos (1995). "Can the Maximum Entropy Principle be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Modern Physics*, **26B**: 223–261. URL `http://www.phys.uu.nl/~wwwgrnsl/jos/mepabst/mepabst.html`. doi:10.1016/1355-2198(95)00015-1.

— (1996). "The Constraint Rule of the Maximum Entropy Principle." *Studies in History and Philosophy of Modern Physics*, **27**: 47–79. URL `http://www.phys.uu.nl/~wwwgrnsl/jos/mep2def/mep2def.html`. doi:10.1016/1355-2198(95)00022-4.

Vuong, Quang H. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica*, **57**: 307–333. URL `http://www.jstor.org/pss/1912557`.