MAXIMUM ENTROPY COMPETES WITH MAXIMUM LIKELIHOOD

Anonymous authors
Paper under double-blind review

ABSTRACT

Maximum entropy (MAXENT) method has a large number of applications in theoretical and applied machine learning, since it provides a convenient non-parametric tool for estimating unknown probabilities. The method is a major contribution of statistical physics to probabilistic inference. However, a systematic approach towards its validity limits is currently missing. Here we study MAXENT in a Bayesian set-up, i.e. assuming that there exists a well-defined prior Dirichlet density for unknown probabilities. This allows to employ the average Kullback-Leibler (KL) distance in evaluating the relevance of various MAXENT constraints, checking its general applicability, and comparing MAXENT with estimators having various degrees of dependence on the prior, *viz.* the regularized maximum likelihood (ML) and the Bayesian estimators. We show that MAXENT applies in sparse data regimes and can even outperform the optimally regularized ML provided that there are prior rank correlations between the estimated random quantity and its probabilities.

1 INTRODUCTION

The maximum entropy (MAXENT) method was proposed within statistical physics (Jaynes, 1957; 1978; Balian, 2007; Pressé et al., 2013), and later on got a wide range of inter-disciplinary applications in data science, probabilistic inference, biological data modeling *etc*; see e.g. (Erickson & Smith, 2013). MAXENT estimates unknown probabilities (that generated data) via maximizing the Boltzmann-Gibbs-Shannon entropy under certain constraints which can be derived from the observed data (Erickson & Smith, 2013). MAXENT leads to *non-parametric* estimators whose form does not depend on the underlying mechanism that generated data (i.e. prior assumptions). Also, MAXENT avoids the zero-probability problem, i.e. when operating on a sparse data—so that certain values of the involved random quantity may not appear due to a small, but non-zero probability—MAXENT still provides a controllable non-zero estimate for this small probability.

Though for MAXENT has a rich background, some of its problems stay open. These are not academic problems of providing yet another *formal* justification of MAXENT; cf. (Jaynes, 1957; 1978; Chakrabarti & Chakrabarty, 2005; Baez et al., 2011; Van Campenhout & Cover, 1981; Topsøe, 1979; Shore & Johnson, 1980; Paris & Vencovská, 1997). The following open problems are basic for MAXENT, because their insufficient understanding prevents its *valid* applications. *(i)* Which constraints of entropy maximization are to be extracted from data? *(ii)* When and how these constraints can lead to overfitting, where —due to a *noisy* data—involving more constraints leads to poorer results? *(iii)* How predictions of MAXENT compare with those of the (regularized) maximum likelihood?

Here we approach these open problems via tools of Bayesian decision theory (Cox & Hinkley, 1979). We assume that the data is given as an i.i.d. sample of a *finite* length M from a random quantity with n outcomes and unknown probabilities. Following the tenets of the Bayesian approach we shall assume that those probabilities are instanced from a non-informative prior Dirichlet density, or a mixture of such densities. Focusing on the sparse data regime $M < n$ we calculate average distances between real probabilities and their estimates, decide on the quality of MAXENT under various constraints, and compare it with the (regularized) maximum-likelihood (ML) estimator. Our main results is that MAXENT does not apply to sparse data generated in the fully non-informative way (i.e. via a single homogeneous Dirichlet prior). However, such an application is secured under

prior rank correlations between the inferred random quantity and its probabilities. Moreover, in the latter case the non-parametric MAXENT estimator is better in terms of the average KL-distance than the *optimally* regularized ML (parametric) estimator.

Some of above questions were already studied in literature. (Good, 1970; Zhu et al., 1997; Pandey & Dukkipati, 2013) applied formal principles of statistics (related to or directly based on the Minimum Description Length) to the selection of constraints (question *(i)*). Our approach to studying this question will be more direct, since the Bayesian statistics allows to state the prior assumptions unambiguously, and compare all predictions with the Bayesian estimator. The latter is normally not available in practice due to insufficient knowledge of prior details, but it still does provide an important theoretical benchmark. Note that (Thomas, 1979; Lebanon & Lafferty, 2002; Kazama & Tsujii, 2005; Altun & Smola, 2006; Dudik, 2007.; Rau, 2011; Campbell, 1999; Friedlander & Gupta, 2005) studied soft constraints that allow incorporation of prior assumptions into the MAXENT estimator making it effectively parametric. Here MAXENT will be taken in its original meaning as providing non-parametric estimators.

The rest of this paper is organized as follows. Section 2 describes the data-generation set-up and introduces prior densities. Section 3 introduces and motivates the Bayesian estimator and the regularized ML estimator. Section 4 recalls the basic formulas of MAXENT, and applies them to the studied set-up, and discusses their symmetry features. Section 5 compares predictions of MAXENT with the regularized ML. We close in the last section with discussing open problems.

2 DATA GENERATION

Consider a random quantity Z with values $(z_1, ..., z_n)$ and respective probabilities $q = (q_1, ..., q_n) = (q(z_1), ..., q(z_n))$. We look at an i.i.d. sample of length M:

$$(z_{i_1}, ..., z_{i_M}), \qquad m = \{m_k\}_{k=1}^n, \qquad M \equiv \sum_{k=1}^n m_k, \tag{1}$$

where m_k is the number of appearances of z_k in (1). This sample will be an instance of our data, e.g. constraints of MAXENT will be determined from it. The conditional probability of m reads

$$P(m_1, ..., m_n | q_1, ..., q_n) = M! \prod_{k=1}^n \frac{q_k^{m_k}}{m_k!}. \tag{2}$$

To check the performance of various inference methods, the inferred probabilities $\hat{q}(m) = \{\hat{q}(z_k; m)\}_{k=1}^n$ (whose forms are given below and can depend on both $(z_1, ..., z_n)$ and $(m_1, ..., m_n)$) are compared with true probabilities $q = \{q(z_k)\}_{k=1}^n$ via the (non-symmetric) KL-distance

$$K[q, \hat{q}(m)] = \sum_{k=1}^n q_k \ln \frac{q_k}{\hat{q}_k(m)}. \tag{3}$$

The choice of distance (3) is motivated below, where we recall that it implies the global optimality of the standard (posterior-mean) Bayesian estimator. Another possible choice of distance is the squared (symmetric) Hellinger distance: $\mathrm{dist}_{\mathrm{H}}[q, \hat{q}] \equiv 1 - \sum_{k=1}^n \sqrt{q_k \, \hat{q}_k}$. In our situation, it frequently leads to the same qualitative results as (3).

Given an inference method $\hat{q}(m)$, we shall average (3) over samples (1) with fixed length M:

$$\overline{K} = \sum_{m_1, ..., m_n} P(m_1, ..., m_n | q_1, ..., q_n) K[q, \hat{q}(m)]. \tag{4}$$

Following the Bayesian approach (Cox & Hinkley, 1979), we shall assume that the probabilities of $(z_1, ..., z_n)$ are generated from a known probability density $\mathcal{D}(q_1, ..., q_n)$ that encapsulates the prior information about the situation. Since $\overline{K}$ does depend on $q = \{q(z_k)\}_{k=1}^n$, we average it over the prior density $\mathcal{D}(q_1, ..., q_n)$:

$$\langle \overline{K} \rangle = \int \prod_{k=1}^n \mathrm{d}q_k \, \mathcal{D}(q_1, ..., q_n) \, \overline{K}. \tag{5}$$

Hence $\langle \overline{K} \rangle$ eventually determines the average performance (quality) of the inference method $\hat{q}(m)$.

For the prior density of $q = \{q_k\}_{k=1}^n$ we choose the Dirichlet density (or a mixture of such densities as seen below) (Frigyik et al., 2010; Schafer, 1997):

$$\mathcal{D}(q_1, ..., q_n|\alpha_1, ..., \alpha_n) = \frac{\Gamma[\sum_{k=1}^n \alpha_k]}{\prod_{k=1}^n \Gamma[\alpha_k]} \prod_{k=1}^n q_k^{\alpha_k - 1}\, \delta(\sum_{k=1}^n q_k - 1), \tag{6}$$

where $\Gamma[x] = \int_0^\infty \mathrm{d}y\, y^{x-1}\, e^{-y}$ is Euler's Γ-function and delta-function $\delta(\sum_{k=1}^n q_k - 1)$ ensures the normalization of probabilities. Parameters $\alpha_k > 0$ determine the prior weight of q_k (Frigyik et al., 2010; Schafer, 1997):

$$\langle q_k \rangle \equiv \int_0^\infty \prod_{l=1}^n \mathrm{d}q_l\, q_k\, \mathcal{D}(q_1, ..., q_n|\alpha_1, ..., \alpha_n) = \frac{\alpha_k}{A}, \qquad A \equiv \sum_{k=1}^n \alpha_k. \tag{7}$$

Dirichlet density (6) is unique in holding several desired features of non-informative prior density over unknown probabilities; see (Frigyik et al., 2010; Schafer, 1997) for reviews. An important feature of density (6) is that it is conjugate to the multinomial conditional probability (2)

$$P(q_1, ..., q_n|m_1, ..., m_n) = \mathcal{D}(q_1, ..., q_n|\alpha_1 + m_1, ..., \alpha_n + m_n). \tag{8}$$

Eq. (8) is convenient when studying i.i.d. samples (1) of discrete random quantities.

3 Bayesian and regularized maximum likelihood (ML) estimators

We minimize (5) over inferred probabilities $\hat{q}(m)$ to find the best estimator:

$$\min[\, \langle \overline{K} \rangle \,] = \sum_m P(m) \min_{\hat{q}(m)} \left[\int \prod_{k=1}^n \mathrm{d}q_k\, P(q|m) K[q, \hat{q}(m)] \right], \tag{9}$$

where $P(m)P(q|m) = P(m|q)\mathcal{D}(q)$ with $P(q|m)$ read from (2, 6). Implementing the constraint $\sum_{k=1}^n \hat{q}_k(m) = 1$ via a Lagrange multiplier, we get from (9):

$$\mathrm{argmin}_{\hat{q}(m)} \left[\int \prod_{k=1}^n \mathrm{d}q_k\, P(q|m)\, K[q, \hat{q}(m)] \right] = \left\{ \int \prod_{k=1}^n \mathrm{d}q_k\, q_l\, P(q|m) \right\}_{l=1}^n. \tag{10}$$

We emphasize that the posterior average in (10) emerged as the best Bayesian estimator only for the KL distance $K[q, \hat{q}(m)]$. The best estimator will be different upon using another distance in (9), e.g. $K[\hat{q}(m), q]$, or the Hellinger distance.

If the prior is a single Dirichlet density (6) we get from (8, 10) for the Bayesian estimator:

$$p(z_k) = \frac{m_k + \alpha_k}{M + A}. \tag{11}$$

The average KL-distance (5) for the estimator (11) reads from (8, 4, 2) (denoting $\psi[x] \equiv \frac{\mathrm{d}}{\mathrm{d}x} \ln \Gamma[x]$):

$$\begin{aligned} \langle \overline{K[q,p]} \rangle &= \frac{1}{A} \sum_{k=1}^n \alpha_k \psi(1 + \alpha_k) - \psi(1 + A) + \ln(M + A) \\ &- \frac{\Gamma[M+1]\,\Gamma[A]}{\Gamma[M+A+1]} \sum_{k=1}^n \sum_{m=0}^M \frac{\Gamma[m+1+\alpha_k]\,\Gamma[M-m+A+\alpha_k]\, \ln(m+\alpha_k)}{\Gamma[\alpha_k]\,\Gamma[A-\alpha_k]\,\Gamma[m+1]\,\Gamma[M-m+1]}. \end{aligned} \tag{12}$$

If the prior density is given by mixture of Dirichlet densities with weights $\{\pi_a\}_{a=1}^L$:

$$\sum_{a=1}^L \pi_a \mathcal{D}(q_1, ..., q_n|\alpha_1^{[a]}, ..., \alpha_n^{[a]}), \qquad \sum_{a=1}^L \pi_a = 1, \tag{13}$$

then instead of (7) and (11) we have from (10)

$$\langle q_k \rangle = \sum_{a=1}^{L} \pi_a \frac{\alpha_k^{[a]}}{A^{[a]}}, \qquad A^{[a]} \equiv \sum_{k=1}^{n} \alpha_k^{[a]}, \tag{14}$$

$$p(z_k) = \frac{\sum_{a=1}^{L} \pi_a \; \Phi^{[a]} \; \frac{m_k+\alpha_k^{[a]}}{M+A^{[a]}}}{\sum_{a=1}^{L} \pi_a \, \Phi^{[a]}}, \qquad \Phi^{[a]} \equiv \frac{\Gamma[A^{[a]}]}{\Gamma[M+A^{[a]}]} \prod_{k=1}^{n} \frac{\Gamma[m_k+\alpha_k^{[a]}]}{\Gamma[\alpha_k^{[a]}]}. \tag{15}$$

For a mixture prior density, the Bayesian estimator (15) depends on *all* numbers $\{m_k; \alpha_k^{[1]}, ..., \alpha_k^{[1]}\}$ not just on m_k. Below we show on concrete examples that not knowing precisely details of the prior mixture can lead to serious losses when applying Bayesian estimators.

It is interesting (both conceptually and practically) to have a simple estimator, where the dependence on the prior is reduced to a single parameter. A good candidate is the regularized maximum likelihood (ML) estimator (see (Hausser & Strimmer, 2009) for a review):

$$p_{\rm ML}(z_k) \equiv \frac{m_k+b}{M+nb} = \lambda \frac{m_k}{M} + (1-\lambda)\frac{1}{n}, \quad \lambda = \frac{M}{M+nb}, \qquad b \geq 0, \quad 0 < \lambda < 1, \tag{16}$$

where the regularizer b (or λ) takes care of the fact that for a finite sample (1) not all values z_k had a chance to appear (i.e. $m_k = 0$ for them). Then (16) avoids to claim a zero probability due to $b > 0$. Eq. (16) is a shrinkage estimator, where the proper ML estimator $\frac{m_k}{M}$ is shrunk towards uniform distribution $\frac{1}{n}$ by the shrinkage factor λ. The proper ML estimator $p_{\rm ML}(z_k)|_{b=0}$ will be shown to be a meaningless estimator for not very long samples (1) producing results that are worse than $\{q(z_k) = \frac{1}{n}\}_{k=1}^{n}$. Moreover, for such samples the correct choice of b (based on the prior information) is crucial, i.e. (16) is a parametric estimator (Hausser & Strimmer, 2009). The estimator (16) recovers true probabilities for $M \to \infty$ (Cox & Hinkley, 1979), where n and b are fixed, hence $\lambda \to 1$ in (16).

For the optimal estimator (16), the value of b is found by minimizing the average KL-distance (5). When the prior is given by a Dirichlet density (6), the average KL-distance amounts to (12), where we need to replace $\ln(M+A) \to \ln(M+nb)$ and $\ln(m+\alpha_k) \to \ln(m+b\,)$. Now (10, 11) imply that for a homogeneous Dirichlet prior, i.e. for (6) with $\alpha_k = \alpha$, we have $b_{\rm opt} = \alpha$ for the optimal value of b, i.e. the regularized ML estimator coincides with the Bayesian estimator: $p_{\rm ML}(z_k) = p(z_k)$. This does not anymore hold for the mixture of Dirichlet prior densities; see below. Then b is to be found numerically. For $M \gg n$ the precise value of b is not essential.

4 THE MAXIMUM ENTROPY (MAXENT) METHOD

MAXENT infers probabilities from maximizing the Boltzmann-Gibbs-Shannon entropy

$$S[q] = -\sum_{k=1}^{n} q(z_k) \ln q(z_k), \tag{17}$$

under constraints taken from the sample (1). The rationale of maximizing (17) is that a larger S means a smaller bias (or information) according to several axiomatic schemes (Jaynes, 1957; Pressé et al., 2013; Chakrabarti & Chakrabarty, 2005; Baez et al., 2011; Van Campenhout & Cover, 1981; Topsøe, 1979; Shore & Johnson, 1980; Paris & Vencovská, 1997).

First we can apply no constraint and maximize the entropy:

$$q^{[0]}(z_k) = 1/n. \tag{18}$$

The calculation of the average distance is straightforward from (12, 18) both for a single Dirichlet prior and a mixture of such priors. We examplify the single Dirichlet case (6):

$$\langle \overline{K[q, q^{[0]}]} \rangle = \sum_{k=1}^{n} \langle q_k \ln q_k \rangle + \ln n = \frac{1}{A} \sum_{k=1}^{n} \alpha_k \psi(1+\alpha_k) - \psi(1+A) + \ln n. \tag{19}$$

Now $\langle \overline{K[q, q^{[0]}]} \rangle$ plays an important role: once (18) is completely data-independent and simply reproduced the prior expectation on the unbiased probabilities, estimators that provide the average KL-distance larger than (19) are meaningless; see below for examples.

Next, we employ the empiric mean of (1) as a constraint for the expected value of Z:

$$\mu_1 = \frac{1}{M}\sum_{u=1}^{M} z_{i_u} = \frac{1}{M}\sum_{k=1}^{n} z_k m_k = \sum_{k=1}^{n} q_k z_k. \tag{20}$$

Maximizing (17) leads to the famous Gibbs formula:

$$q^{[1]}(z_k) = \frac{e^{-\beta z_k}}{\sum_{l=1}^{n} e^{-\beta z_l}}, \tag{21}$$

where the Lagrange multiplier β is found from (20) (Jaynes, 1957). We order the values of Z as $z_1 < ... < z_n$, and note a specific feature of (21): depending on the sign of β, we either get

$$q^{[1]}(z_1) \leq ... \leq q^{[1]}(z_n) \quad \text{or} \quad q^{[1]}(z_1) \geq ... \geq q^{[1]}(z_n). \tag{22}$$

One can try to acquire further information from (1) by fixing also the second moment:

$$\mu_2 = \frac{1}{M}\sum_{u=1}^{M} z_{i_u}^2 = \frac{1}{M}\sum_{k=1}^{n} z_k^2 m_k = \sum_{k=1}^{n} q_k z_k^2, \tag{23}$$

and maximize (17) under two constraints (20) and (23):

$$q^{[1+2]}(z_k) = \frac{e^{-\beta_1 z_k - \beta_2 z_k^2}}{\sum_{l=1}^{n} e^{-\beta_1 z_l - \beta_2 z_l^2}}, \tag{24}$$

where β_1 and β_2 are determined from solving (20) and (23) simultaneously. Eqs. (21, 24) make obvious the result of fixing other (e.g. fractional) moments. Such maximization results are unique, because (17) is a convex function of $\{p_k\}_{k=1}^n$, while the moment constraints are linear.

Let the values $(z_1, ..., z_n)$ of Z be subject to affine transformation

$$\widetilde{z}_k = \mathcal{F}(z_k), \quad \mathcal{F}(z) = gz + h, \quad k = 1, ..., n. \tag{25}$$

Hence as a result of transformation (25): $\mu_1 \to \widetilde{\mu}_1 = g\mu_1 + h$ and $\mu_2 \to \widetilde{\mu}_2 = g^2\mu_2 + h^2 + 2gh\mu_1$; see (20, 23). These relations show that (25) leaves the inferred probabilities (21, 24) invariant, because the resulting set of equations for the unknowns in (21, 24) are identical for both the original $(z_1, ..., z_n)$ and transformed values (25). Likewise, fixing first p moments $\mu_1, ..., \mu_p$ produces affine-invariant probabilities. Note that fixing only (23) [without fixing (20)] will lead to the invariance of the probabilities with respect to a limited affine-symmetry, where $h = 0$ in (25). Another example of limited affine symmetry is fixing the fractional moment $\sum_{k=1}^{n} q_k\sqrt{z_k}$ (for $z_k \geq 0$ and instead of (20, 23)). Then the probabilities $q^{[1/2]}(z_k) \propto e^{-\beta_{1/2}\sqrt{z_k}}$ will stay intact only under $h = 0$ and $g > 0$ in (25). Note in this context that the ML estimator (16) is invariant with respect (25) with an arbitrary bijective $\mathcal{F}$, which keeps the values of $\widetilde{z}_k$ different.

The symmetry features of various estimators are clearly important, though we so far have no analytical results that would relate them to the estimation quality quantified by (5). But we noted from numerical comparison of MAXENT estimators based on various constraints, that estimators with the largest affine symmetry—i.e. (25) with arbitrary g and h—tend to be better in terms of the average KL-distance (5). Intuitively, higher (affine) symmetry should be related to higher susceptibility with respect to noises. Thus based on numerical evidence, we below never employ MAXENT estimators with fractional fixed moments or with the second moment alone.

5 Numerical results

Recall that maximization of entropy (17) can be applied if there is no prior information that distinguishes one probability from another. If such information is present, MAXENT is generalized to the minimum relative entropy method (Shore & Johnson, 1980). We shall not study this generalization here. Hence to ensure applicability of MAXENT, we always choose prior densities such that $\langle q_k \rangle = \frac{1}{n}$; i.e. all n values are equally likely to be generated, on average. As seen from (7), for a single prior Dirichlet density (6) condition $\langle q_k \rangle = \frac{1}{n}$ implies:

$$\alpha_k = \alpha, \qquad k = 1, ..., n, \tag{26}$$

Table 1: For $n = 60$ and $z_k = k$ ($k = 1, ..., n$) we show the average KL-distance (5) for various estimators. The full affine symmetry (25) holds for all shown probabilities. M is the length of sample (1). The initial prior Dirichlet density (6) holds $\alpha_k = 2$; cf. (26). Eq. (19) equals $\langle \overline{K[q, q^{[0]}]} \rangle = 0.225$, i.e. values of the average KL-distance larger than 0.225 are *meaningless*; cf. the discussion after (19). $\langle \overline{K}_{\rm Bayes} \rangle$ is the averaged KL-distance for the Bayes estimator (11) that for this case coincides with the optimally regularized ML estimator. $\overline{K}_1$ and $\overline{K}_{1+2}$ are defined (resp.) via (21, 20) and (24, 23). The averages are found numerically (applies to all Tables): first we generate 10^2 instances of $\{q_k\}_{k=1}^n$ from the Dirichlet density, and then for each instance we generate 10^2 samples (1). Such parameters lead to 3-digit precision, as reported.

M	$\langle \overline{K}_{\rm Bayes} \rangle$	$\overline{K}_1$	$\overline{K}_{1+2}$
35	0.177	0.236	0.247
25	0.188	0.240	0.260
15	0.202	0.259	0.301

Now recall from (16, 11) that under (26) the Bayesian and the regularized ML coincide. I.e. we conclude that the regularized ML is a better estimator than MAXENT (under any constraint).

How meaningful are predictions of MAXENT under (26)? Table 1 compares predictions of (16) with those of MAXENT solutions (21) and (24) for the Dirichlet prior (6) holding (26). It is seen that the very applicability of MAXENT demands a sufficiently large sample length M. Otherwise, MAXENT is meaningless, because the trivial estimator (18) provides a smaller average KL-distance; cf. (19). For parameters of Table 1 predictions of MAXENT are meaningless for $M \leq 35$ with $n = 60$. In contrast, for the Bayesian estimator even $M = 1$ leads to a meaningfull prediction; e.g. for parameters of Table 1 we have: $\langle \overline{K}_{\rm Bayes} \rangle|_{M=1} = 0.224 < 0.225$.

Table 2: For $n = 60$ and $z_k = k$ ($k = 1, ..., n$) we show the average KL-distance for various estimators. M is the length of samples. The initial prior density is a Dirichlet mixture given by (13, 29) with $\alpha_0 = 0.3$ and $\epsilon = 1.1$. The average KL-distance $\langle \overline{K[q, q^{[0]}]} \rangle$ for the trivial estimator (18) equals 0.212, i.e. values of the average KL-distance larger than 0.212 are *meaningless*; cf. the discussion after (19). $\langle \overline{K}_{\rm Bayes} \rangle$ is the averaged KL-distance for the Bayes estimator (15). $\langle \overline{\mathbf{K}}_{\rm Bayes} \rangle$ is the averaged KL-distance for the perturbed estimator (31). $\langle \overline{K}_{\rm ML} \rangle_{b=b_{\rm opt}}$ and $\langle \overline{K}_{\rm ML} \rangle_{b=1}$ refer to regularized ML estimator (16) under $b = 1$ and the optimal value of b found from numerically minimizing $\langle \overline{K}_{\rm ML} \rangle$. The optimal value $b_{\rm opt}$ of b changes from 2.46 for $M = 35$ to 2.65 for $M = 1$. We also report the value of $\langle \overline{K}_{\rm ML} \rangle_{b=1}$ with a sensible value of b to confirm that if b is not chosen properly, then the corresponding (regularized) ML estimator (16) is meaningless. $\langle \overline{K}_1 \rangle$ is defined via (5, 21). $\langle \overline{K}_{1+2} \rangle$ is not shown, since $\langle \overline{K}_{1+2} \rangle > \langle \overline{K}_1 \rangle$ for $35 \geq M \geq 1$. We do not show $\langle \overline{K}_1 \rangle|_{M=1}$, since it is larger than the average KL-distance for all other estimators.

M	$\langle \overline{K}_{\rm Bayes} \rangle$	$\langle \overline{\mathbf{K}}_{\rm Bayes} \rangle$	$\langle \overline{K}_{\rm ML} \rangle_{b=b_{\rm opt}}$	$\langle \overline{K}_{\rm ML} \rangle_{b=1}$	$\overline{K}_1$
35	0.014	0.206	0.180	0.204	0.048
25	0.015	0.207	0.188	0.210	0.053
15	0.017	0.209	0.197	0.214	0.065
11	0.022	0.209	0.201	0.215	0.077
7	0.035	0.209	0.205	0.215	0.105
5	0.052	0.210	0.207	0.214	0.141
3	0.083	0.210	0.209	0.213	0.268
1	0.150	0.211	0.211	0.212	—

Thus the applicability range of MAXENT is to be restricted to having some prior knowledge about the unknown probabilities $\{q_k\}_{k=1}^n$. We still assume that prior expectations are equal: $\langle q_k \rangle = \frac{1}{n}$. As a prior density we shall now apply a mixture of Dirichlet densities such that there are definite *rank correlations* between the random variable $(z_1, ..., z_n)$, and probabilities $(q_1, ..., q_n)$. I.e. if

Table 3: The same as in Table 2, but for different values of M. Here $\overline{K}_{1+2}$ refers to MAXENT estimator (24) with first and second empiric moments fixed. The Bayesian estimator is found from (15, 29). For this range of sufficiently large M the MAXENT estimator (24) performs better than (24): $\overline{K}_{1+2} < \overline{K}_1 < \langle \overline{K}_{\rm ML} \rangle_{b=b_{\rm opt}}$.

M	$\langle \overline{K}_{\rm Bayes} \rangle$	$\langle \overline{K}_{\rm ML} \rangle_{b=b_{\rm opt}}$	$\langle \overline{K}_{\rm ML} \rangle_{b=1}$	$\overline{K}_1$	$\overline{K}_{1+2}$
45	0.015	0.172	0.196	0.045	0.042
65	0.014	0.157	0.180	0.042	0.035
85	0.014	0.145	0.164	0.040	0.031
241	0.013	0.087	0.091	0.038	0.024

we assume the ordering $(z_1 < ... < z_n)$, then $(q_1, ..., q_n)$ prefer to be ordered in the same way $(q_1 < ... < q_n)$ or in the opposite way: $(q_1 > ... > q_n)$; cf. (22). Such a prior knowledge can be modeled via a mixture (13) of two Dirichlet priors with $L = 2$, $\pi_1 = \pi_2 = \frac{1}{2}$, and

$$\alpha_1^{[1]} < ... < \alpha_n^{[1]}, \qquad \alpha_1^{[2]} > ... > \alpha_n^{[2]}, \tag{27}$$

$$\frac{\alpha_k^{[1]} - \alpha_l^{[1]}}{A^{[1]}} = \frac{\alpha_l^{[2]} - \alpha_k^{[2]}}{A^{[2]}}, \quad \text{for any} \quad k, l = 1, ..., n, \tag{28}$$

where (28) ensures the needed $\langle q_k \rangle = \frac{1}{n}$, as seen from (14). A simple case that leads to (27, 28) is

$$L = 2, \quad \pi_1 = \pi_2 = \frac{1}{2}, \quad \alpha_k^{[1]} = \alpha_0 + \epsilon(k-1), \quad \alpha_k^{[2]} = \alpha_0 + \epsilon(n-k), \qquad k = 1, ..., n, \tag{29}$$

where $A^{[1]} = A^{[2]} = \alpha_0 n + \frac{\epsilon n(n-1)}{2}$. Recall that for a mixture of Dirichlet densities the Bayes estimator (15) and the optimally regularized ML estimator (16) are different.

For numerical illustration we choose $\{z_k = k\}_{k=1}^n$. Prior probability densities generated via (29) will be employed with $z_1 < ... < z_n$. Now Tables 2 and 3 show that for $M \geq 5$ the MAXENT estimator (21, 20) is clearly better than the *optimally* regularized ML estimator (16) (where the optimal value of b is found from minimizing the averaged KL-distance (5)):

$$\langle \overline{K}_1 \rangle < \langle \overline{K}_{\rm ML} \rangle_{b=b_{\rm opt}}. \tag{30}$$

Moreover, for $M \geq 7$, we see that $\langle \overline{K}_1 \rangle$ is closer to the optimal $\langle \overline{K}_{\rm Bayes} \rangle$ than to $\langle \overline{K}_{\rm ML} \rangle_{b=b_{\rm opt}}$. Note that such threshold values for M do depend on the assumed prior density and on n.

For $M \to \infty$ the performance of the optimally regularized ML estimator (16) (for a fixed $b \sim 1$) will be better than MAXENT with any finite number of constraints, since the regularized ML converges to the true probabilities for $M \to \infty$ (Cox & Hinkley, 1979), while MAXENT under a few constraints does not. But as Table 3 shows, MAXENT with constraints (20) or (20)+(23) still performs better than the optimally regularized ML even for M as large as 241 (for $n = 60$).

Table 2 shows that MAXENT with two constraints (20)+(23) performs worse than the method under the single constraint (20) although the affine invariance (25) of probabilities holds. The situation is worsened since the second constraint (23) is again estimated from the data, and with too little data the estimated constraints may be be far from the true values, and *overfitting* these constraints leads to worse results. Naturally, this situation changes for larger values of M, i.e. $M \geq 45$, as Table 3 demonstrates.

We emphasize that the advantage (30) of MAXENT closely relates to the agreement between (27) and the ordering $(z_1 < ... < z_n)$ of Z. If the vector $(z_1 < ... < z_n)$ is randomly permuted, and employed for values of Z, predictions of MAXENT become devastating even for rather large values of M. To illustrate this point, let us take parameters of Table 2, but apply a random permutation to values $\{z_k = k\}_{k=1}^{n=60}$. Taking this new vector for values of Z, we noted that even for $M = 85 > n = 60$ predictions of MAXENT under constraint (20) is meaningless, since the average KL distance amounts to 0.215, which is larger than the prediction 0.212 of (19). For $M = 105$ this prediction 0.210 is not meaningless in this sense, but rather poor by itself. It does not substantially improve even for $M = 500$ (for $n = 60$), where it amounts to 0.207. Though such predictions are not meaningless, they are practically useless.

Recall that both the Bayesian (15) and the regularized ML estimator (16) are parametric, i.e. the very their form depends on the prior, which is frequently not available in practice. Hence we need to understand how strong is dependence. Let us assume that one has to built a Bayesian estimator without knowing the full form of the equal-weight mixture (29). Instead one knows the average values of $\alpha_k = \frac{1}{2}[\alpha_0 + \epsilon(k-1)] + \frac{1}{2}[\alpha_0 + \epsilon(n-k)]$ from (29), prescribes them to a single Dirichlet prior (6, 26) and builts up from (11) an estimator

$$\mathbf{p}(z_k) = \frac{m_k + \alpha_0 + \epsilon(n-1)/2}{M + n\alpha_0 + \epsilon n(n-1)/2}. \tag{31}$$

The performance of this perturbed Bayesian estimator deteriorates and gets worse than that of the MAXENT solution: $\langle \overline{K}_1 \rangle < \langle \overline{\mathbf{K}}_{\mathrm{Bayes}} \rangle$; see Table 2. Likewise, the choice of b in the regularized ML estimator (16) is important. If just some reasonable value is chosen instead of the optimal one—e.g. $b = 1$ instead of $b \sim 2.5$ in Table 2—then the ML estimator can turn meaningless; see Table 2 for $M \leq 15$. In this context, we emphasize that the Bayes estimator and the optimally regularized ML estimator are never meaningless even for $M = 1$. For parameters of Table 2, the MAXENT estimator (21) becomes meaningless for $M \leq 3$.

6 Summary and Discussion

The maximum entropy (MAXENT) method provides non-parametric estimators for inferring unknown probabilities (Erickson & Smith, 2013). MAXENT is widely applied both in statistical physics and probabilistic inference. However, its physical applications are mostly data-free and are based on additional principles (e.g. conservation laws (Balian, 2007; Pressé et al., 2013)) that are absent in inference. This fact is/was not sufficiently emphasized by creators and practitioners of MAXENT. Hence we needed a systematic approach towards understanding the validity limits of MAXENT as an inference tool.

Here we presented a Bayesian approach that allows to decide on whether MAXENT is applicable at all, i.e. whether it is better than a random guess. It also allows to compare different estimators with each other (e.g. to compare MAXENT with the regularized maximum likelihood) and determine the relevance of various constraints employed in MAXENT. In particular, the approach allows to discuss the phenomenon of overfitting, where adding more constraints leads to poorer inference.

Our results can be briefly summarized as follows. MAXENT does not apply to a sparse data generated by homogeneous (i.e. maximally non-informative) Dirichlet density (26). Here sparse means $M < n$, i.e. the sample length M is smaller than the number of probabilities n to be inferred. We explored the simplest prior information that makes MAXENT applicable and superior over the optimally regularized maximum-likelihood estimator: prior rank correlations between the random quantity and its probabilities. We currently explore more general types of prior information that can make MAXENT applicable.

Several important problems were uncovered by this study and should be addressed in future together with more detailed understanding of the presented results. First of all, this concerns the applicability of MAXENT to a categorical data, where the values $(z_1, ..., z_n)$ of the random variable Z in sample (1) are not numerical, but instead refer to certain distinguishable categories. The major difference between maximum likelihood and MAXENT estimator is that the former freely applies to categorical data. In contrast, MAXENT does depend on the concrete numerical implementation (i.e. *encoding*) of data, though this dependence is somewhat weakened by the affine symmetry (25). Thus an open problem demands considering various encoding schemes in view of their applicability to MAXENT estimators. (In this paper we in fact assumed the simplest encoding via natural numbers; see Tables.) The second open problem relates to the influence of affine symmetries on the performance of various MAXENT estimators. We observed numerically that the contraints which produce affine-invariant probabilities produce better estimators; see after (25). We were not able to prove or disprove this statement in general, and hence it is left as an open problem.

ACKNOWLEDGMENTS

REFERENCES

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *International Conference on Computational Learning Theory*, pp. 139–153. Springer, 2006.

John C Baez, Tobias Fritz, and Tom Leinster. A characterization of entropy in terms of information loss. *Entropy*, 13(11):1945–1957, 2011.

Roger Balian. *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics. Volumes I, II*. Springer Science & Business Media, 2007.

L Lorne Campbell. Minimum cross-entropy estimation with inaccurate side information. *IEEE Transactions on Information Theory*, 45(7):2650–2652, 1999.

CG Chakrabarti and Indranil Chakrabarty. Shannon entropy: axiomatic characterization and application. *International Journal of Mathematics and Mathematical Sciences*, 2005, 2005.

David Roxbee Cox and David Victor Hinkley. *Theoretical statistics*. CRC Press, 1979.

Miroslav Dudik. Maximum entropy density estimation and modeling geographic distributions of species. phd dissertation presented to the princeton university, 2007.

Gary Erickson and C Ray Smith. *Maximum-Entropy and Bayesian Methods in Science and Engineering: Volume 2: Applications*, volume 31. Springer Science & Business Media, 2013.

Michael P Friedlander and Maya R Gupta. On minimizing distortion and relative entropy. *IEEE Transactions on Information Theory*, 52(1):238–245, 2005.

Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the dirichlet distribution and related processes. *Department of Electrical Engineering, University of Washignton, UWEETR-2010-0006*, (0006):1–27, 2010.

IJ Good. Some statistical methods in machine intelligence research. *Mathematical Biosciences*, 6: 185–208, 1970.

Jean Hausser and Korbinian Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(7), 2009.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

ET Jaynes. Where do we stand on maximum entropy, in the maximum entropy formalism, rd levine and m. tribus, editors, 1978.

Junichi Kazama and Junichi Tsujii. Maximum entropy models with inequality constraints: A case study on text categorization. *Machine Learning*, 60(1-3):159–194, 2005.

Guy Lebanon and John D Lafferty. Boosting and maximum likelihood for exponential models. In *Advances in neural information processing systems*, pp. 447–454, 2002.

Gaurav Pandey and Ambedkar Dukkipati. Minimum description length principle for maximum entropy model selection. In *2013 IEEE International Symposium on Information Theory*, pp. 1521–1525. IEEE, 2013.

Jeff Paris and Alena Vencovská. In defense of the maximum entropy inference process. *International Journal of approximate reasoning*, 17(1):77–103, 1997.

Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85(3):1115, 2013.

Jochen Rau. Inferring the gibbs state of a small quantum system. *Physical Review A*, 84(1):012101, 2011.

Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

Marlin U Thomas. A generalized maximum entropy principle. *Operations Research*, 27(6):1188–1196, 1979.

Flemming Topsøe. Information-theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

Jan Van Campenhout and T Cover. Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, 27(4):483–489, 1981.

Song Chun Zhu, Ying Nian Wu, and David Mumford. Minimax entropy principle and its application to texture modeling. *Neural computation*, 9(8):1627–1660, 1997.

A APPENDIX

You may include other additional sections here.