Inference with Multinomial Data: Why to Weaken the Prior Strength

Cassio P. de Campos and Alessio Benavoli
Dalle Molle Institute for Artificial Intelligence
Manno-Lugano, Switzerland
{cassio,alessio}@idsia.ch

Abstract

This paper considers inference from multinomial data and addresses the problem of choosing the strength of the Dirichlet prior under a mean-squared error criterion. We compare the Maximum Likelihood Estimator (MLE) and the most commonly used Bayesian estimators obtained by assuming a prior Dirichlet distribution with "non-informative" prior parameters, that is, the parameters of the Dirichlet are equal and altogether sum up to the so called strength of the prior. Under this criterion, MLE becomes more preferable than the Bayesian estimators at the increase of the number of categories k of the multinomial, because non-informative Bayesian estimators induce a region where they are dominant that quickly shrinks with the increase of k. This can be avoided if the strength of the prior is not kept constant but decreased with the number of categories. We argue that the strength should decrease at least k times faster than usual estimators do.

1 Introduction

In this paper we consider the problem of inference from multinomial data with chances $\boldsymbol{\theta} = [\theta_1, \dots, \theta_k]'$. We compare the *Maximum Likelihood Estimator* (MLE) and the most commonly used Bayesian estimators obtained by assuming a prior Dirichlet distribution with "non-informative" prior parameters such as Laplace, Perks, Jeffreys, and Haldane. Inference in a multinomial-Dirichlet model is a recurrent problem in Artificial Intelligence and Statistics. For instance, it appears in parameter learning of probabilistic graphical models (such as Bayesian networks and some variations) [Koller and Friedman, 2009, Ch. 17], in smoothing methods in information retrieval [Zhai and Lafferty, 2001] and topic models [Mimno and McCallum, 2008], in Bayesian reliability analysis [Somerville *et al.*, 1997], etc.

Consider $c_1, \dots, c_k$ categories and θ_j the chance of c_j to be observed, for $j = 1, \dots, k$. Inference about the vector of k parameters $\boldsymbol{\theta} = [\theta_1, \dots, \theta_k]' \in \mathcal{S}_{\boldsymbol{\theta}}$ is desired, where $\mathcal{S}_{\boldsymbol{\theta}} = \{\theta_j : 0 \leq \theta_j \leq 1 \text{ for all } j \text{ and } \boldsymbol{\theta}'\mathbf{1} = 1\}$ ($\mathbf{1}$ denotes a row of 1s, i.e., $\mathbf{1} = [1, 1, \dots, 1]' \in \mathbb{R}^k$). The observed data consists in a vector of counts $\mathbf{n} = [n_1, n_2, \dots, n_k]'$, where n_j is the number of times in which the j-th category is observed and $N = \mathbf{n}'\mathbf{1}$ is the total number of observations.

The goal is thus to estimate the parameter vector $\boldsymbol{\theta}$ based on the vector of observations $\mathbf{n}$. Assuming that the probability of observing $\mathbf{n}$, conditionally on $\boldsymbol{\theta}$, can be represented as a multinomial distribution: $P(\boldsymbol{\theta}, \mathbf{n}) = N!/(n_1!n_2! \cdots n_k!) \prod_{j=1}^{k} \theta_j^{n_j}$, a point estimate of $\boldsymbol{\theta}$ can then be obtained by computing the Maximum Likelihood Estimator (MLE), i.e., by maximizing the likelihood $P(\boldsymbol{\theta}, \mathbf{n})$ w.r.t. $\boldsymbol{\theta}$ subject to the constraint $\boldsymbol{\theta}'\mathbf{1} = 1$, which gives: $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \mathbf{n}/N$.

An alternative way is to follow a Bayesian approach. The multinomial is a member of the exponential family and its natural conjugate prior is the Dirichlet distribution. Hence, assuming a Dirichlet prior over $\boldsymbol{\theta}$ and applying Bayes' rule to the multinomial-Dirichlet conjugate model, the following posterior density function is obtained:

$$p(\boldsymbol{\theta}|\mathbf{n}) \propto L(\boldsymbol{\theta}, \mathbf{n}) D(\boldsymbol{\alpha}, \boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{n_j + \alpha_j - 1}, \qquad (1)$$

where $D(\boldsymbol{\alpha}, \boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$ is the Dirichlet prior with parameters $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_k]'$ and $\alpha_j > 0$ for $j = 1, \dots, k$. In the following we introduce the notation used in this paper. We assume $\alpha_j = st_j$ and $s = \sum_{j=1}^{k} \alpha_j$, with $\mathbf{0} < \mathbf{t} < \mathbf{1}$, $\mathbf{t}'\mathbf{1} = 1$, $\mathbf{t} = [t_1, t_2, \dots, t_k]'$. Notice that s is the strength of the prior information (equivalent sample size or number of pseudo-counts) and t_j is the prior mean. The posterior expectation of $\boldsymbol{\theta}$ given $\mathbf{n}$ is then given by:

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{n} + \boldsymbol{\alpha}}{N + \sum_{j=1}^{k} \alpha_j} = \frac{\mathbf{n} + s\mathbf{t}}{N + s}, \qquad (2)$$

which gives a point estimate of $\boldsymbol{\theta}$.

The parameters s and $\mathbf{t}$ represent the a-priori information. In case no prior information is available, the common approach is to select these parameters to represent a non-informative prior. The most used non-informative priors select $t_j = 1/k$ for $j = 1, 2, \dots, k$ but differ in the choice of the value of s. Bayes and Laplace suggest to use a uniform prior $s = k$, Perks suggests $s = 1$, Jeffreys suggests $s = k/2$, and Haldane suggests $s = 0$. Nevertheless, the analysis we conduct is general and applies to other choices of s too.

To compare the goodness of different point estimates, a measure of estimation performance must be defined. A popular measure of performance is the matrix *mean-squared error* (MSE), which is defined as

$$E_{\mathbf{n}}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'] = (E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta})(E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta})' + (\hat{\boldsymbol{\theta}} - E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}} - E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}])', \quad (3)$$

where the first term of the summation is the "squared-bias" of the estimator and the second term is its variance matrix.[1] Here the unknown parameter vector $\boldsymbol{\theta}$ is assumed to be deterministic and, thus, the expectation is only over the data.

In the problem of inference from multinomial data, it is well known that the estimate $\hat{\boldsymbol{\theta}}_{MLE}$ is unbiased, which means that $E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$; and achieves the *Cramer-Rao Lower Bound* (CRLB) for unbiased estimators, i.e. $E_{\mathbf{n}}[(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta})'] = \boldsymbol{\Sigma}_{MLE}$, where $\boldsymbol{\Sigma}_{MLE}$ is the inverse of the Fisher information matrix. These facts do not imply that MLE always provides a small MSE, especially for small data samples. In fact, since "MSE=variance + squared bias" and trading-off bias for variance, it is possible to design estimators that yield a lower MSE than the CRLB for unbiased estimators [Ghosh *et al.*, 1983; Stein, 1956].

Since the MSE depends on the unknown $\boldsymbol{\theta}$, it is not obvious how to compare estimators in terms of MSE. However some estimators may be uniformly better than others in terms of MSE, in other words, they can be better for all possible values of $\boldsymbol{\theta}$. For this purpose, we say that an estimator $\hat{\boldsymbol{\theta}}$ *dominates* another estimator $\hat{\boldsymbol{\theta}}_0$ on a convex set $\boldsymbol{\Theta}$ if its MSE is never greater than that of $\hat{\boldsymbol{\theta}}_0$ for all values of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$, and is strictly smaller for some $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$. An estimator is $\boldsymbol{\Theta}$-*admissible* if it is not dominated by any other estimator on $\boldsymbol{\Theta}$ [Berger, 1985]. Hence, it is reasonable to prefer admissible estimators. In the problem of inference from multinomial data, it can be shown that the MLE is admissible w.r.t. the MSE criterion if $\boldsymbol{\Theta} = \mathcal{S}_{\boldsymbol{\theta}}$ [Johnson, 1971]. However, MLE might not be admissible on $\boldsymbol{\Theta} \subset \mathcal{S}_{\boldsymbol{\theta}}$, because estimators that dominate MLE may exist if a proper subregion of the parameter space is considered.

If one assumes the estimator $\hat{\boldsymbol{\theta}} = (\mathbf{n} + s\mathbf{t})/(N + s)$, obtained by a prior Dirichlet distribution with parameters s and $\mathbf{t}$ for $\boldsymbol{\theta}$, then the values of s and $\mathbf{t}$ can be designed in order to dominate MLE on $\boldsymbol{\Theta}$. The answer to this question is partially given in [Benavoli and de Campos, 2009], where the authors determine a closed-form solution for the dominance. This solution is employed there with two aims. First, for the binomial case, the authors analyze the performance of Bayesian estimators with $\mathbf{t} = \mathbf{1}/2$ and different choices of s corresponding to the most used non-informative priors. In particular, they determine the region $\boldsymbol{\Theta}$ (an interval in the binomial case) where these estimators dominates MLE. Second, assuming that $\boldsymbol{\Theta}$ is given as prior information, they derive an ad-hoc criterion to choose an "optimal" value for s and $\mathbf{t}$ which guarantees the dominance. In this paper, we generalize the analysis to the multinomial case, where we show that the coverage of the set $\boldsymbol{\Theta}$ (that is, the ratio between volume of $\boldsymbol{\Theta}$ and the volume of the whole space $\mathcal{S}_{\boldsymbol{\theta}}$) on which the Bayesian estimator of Eq. (2) dominates MLE decreases at the increasing of the number of categories k. This means that, if s is kept constant, then the region where MLE is preferable to the estimator of Eq. (2) becomes larger with the increasing of k, and soon the MLE becomes the only admissable estimator for any practical scenario. However, this can be avoided if the strength s of prior is not kept constant but decreased with k. As it will be clear by the analysis, we argue that s should decrease at a rate proportional to k, or in other words, each Dirichlet parameter α_j should be further corrected by dividing it to k. This corrected version of the Bayesian estimator tends quickly to MLE as k increases, having almost no practical difference already for somewhat small values of k (15 or so), but are still preferable to MLE as they avoid problems with zero counts.

Before proceeding, we point out that the analysis performed here assumes that no additional information is available to select the prior. It is obvious that better priors can be chosen if a-priori information, for example from domain knowledge or other data source, is available. For instance, this is mostly the case in language processing [Zhai and Lafferty, 2001]. Nevertheless, the argument that prior strength should be reduced with the increase of k might still have to be taken into account, however centering the analysis on the informative prior.

2 MLE-dominating priors

In this section we summarize the results from [Benavoli and de Campos, 2009] that are used in the rest of this paper. Consider an estimator with structure as in Eq. (2). The goal is to choose the free parameters s and $\mathbf{t}$ so as to guarantee that:[2]

$$E_{\mathbf{n}}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'] \leq E_{\mathbf{n}}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})'], \quad (4)$$

for each vector $\boldsymbol{\theta}$ in a convex set $\boldsymbol{\Theta}$. The right-hand side of Ineq. (4) is denoted by $\boldsymbol{\Sigma}_{MLE} = (\sigma_{ij})$, which represents the covariance matrix of the MLE whose elements are $\sigma_{ii} = \theta_i(1 - \theta_i)/N$ and $\sigma_{ij} = -\theta_i\theta_j/N$, for $i, j = 1, 2, \ldots, k$ and $i \neq j$. The matrix domination considered in Ineq. (4) guarantees a MSE reduction for all the components of the parameter vector $\boldsymbol{\theta}$ to be estimated and, thus, is stronger than a trace domination that would only guarantee an improvement for the sum of the MSEs of such components. This is the motivation behind the choice of the matrix MSE instead of the trace MSE that we have followed in this paper.

Manipulating $E_{\mathbf{n}}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})']$, it can be shown that:

$$E_{\mathbf{n}}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'] = \frac{s^2}{(N+s)^2}(\boldsymbol{\theta} - \mathbf{t})(\boldsymbol{\theta} - \mathbf{t})' + \frac{N^2}{(N+s)^2}\boldsymbol{\Sigma}_{MLE,} \quad (5)$$

[1] The MSE defined in Eq. (3) is a matrix and not a scalar. In the literature, sometimes the MSE is defined as the trace of the matrix in Eq. (3), however in this paper we adopt the matrix definition. The motivation for this choice will be clarified in Section 2.

[2] Notice that Ineq. (4) is a matrix inequality. For two matrices A, B of compatible dimensions, the inequality $A \leq B$ means that $B - A$ is nonnegative definite.

and Ineq. (4) becomes

$$(\boldsymbol{\theta}-\mathbf{t})(\boldsymbol{\theta}-\mathbf{t})' \leq (\tfrac{2}{s}+\tfrac{1}{N})N\boldsymbol{\Sigma}_{MLE}. \qquad (6)$$

The above inequality is satisfied if and only if

$$\sum_{i=1}^{k}\frac{(\theta_i-t_i)^2}{\theta_i} \quad\leq\quad (\tfrac{2}{s}+\tfrac{1}{N}) \qquad (7)$$

holds for each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Hence, an estimator $\hat{\boldsymbol{\theta}}$ has MSE lower than that of MLE for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ if s and $\mathbf{t}$ are chosen according to Ineq. (7). If $\boldsymbol{\Theta}$ is a convex polytope of vertices $\boldsymbol{\theta}^{v_1},\boldsymbol{\theta}^{v_2},\ldots,\boldsymbol{\theta}^{v_m}$, i.e. $\boldsymbol{\Theta}=Ch\{\boldsymbol{\theta}^{v_1},\boldsymbol{\theta}^{v_2},\ldots,\boldsymbol{\theta}^{v_m}\}$ ($Ch\{\cdot\}$ stands for *convex hull*), then Ineq. (7) can be further simplified. In fact, in this case, a necessary and sufficient condition for Ineq. (7) to be satisfied for each $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ is to hold on the vertices of the polytope $\boldsymbol{\Theta}$. The MLE-dominance is guaranteed if s and $\mathbf{t}$ are chosen such that:

$$\sum_{i=1}^{k}\frac{(\theta_i^{v_j}-t_i)^2}{\theta_i^{v_j}} \leq (\tfrac{2}{s}+\tfrac{1}{N}), \quad \text{for } j=1,2,\ldots,m, \qquad (8)$$

where $\theta_i^{v_j}$ denotes the i-th component of the j-th vertex. The above m-inequalities define all the values of s and $\mathbf{t}$ which guarantee the MLE-dominance. Using Ineq. (8), the binomial case can be analyzed by taking a Bayesian estimators with $\mathbf{t}=\mathbf{1}/2$ and different choices of s. In particular, the set $\boldsymbol{\Theta}$ becomes an interval $[\epsilon,1-\epsilon]$ and Ineq. (8) is satisfied if

$$0.5\left(1-\sqrt{1-\frac{1}{1+\dfrac{2}{s}+\dfrac{1}{N}}}\right) \leq \epsilon < 0.5. \qquad (9)$$

Hence, the values of the true θ_1 (notice that $\theta_2 = 1-\theta_1$) for which the MLE-dominance condition is satisfied when $N\to\infty$ are: Haldane ($s=0$) needs $0\leq\theta_1\leq 1$; Jeffreys ($s=0.5$) needs $0.05\leq\theta_1\leq 0.95$; Perks ($s=1$) needs $0.1\leq\theta_1\leq 0.9$; Bayes/Laplace ($s=2$) needs $0.15\leq\theta_1\leq 0.85$; For instance, if $s=6$ the Bayesian estimator has a lower MSE than MLE if the true θ is in $[0,25,0.75]$, and thus it is preferable in half of the parameter space. Assuming $N\to\infty$ leads to two properties: (i) the choice of s (if one wants to base their choice in this analysis) does not depend on the sample size; (ii) the obtained value of s is a tighter bound than that of using any finite N, which implies that such value is also a feasible choice for any finite N (just slightly smaller than it could be if the finite N was used).

3 Multinomial data

Hereafter we extend the analysis of the end of Section 2 to the multinomial case. In particular, we aim to show that the most used non-informative Bayesian estimators do have a region where they are MLE-dominant, but such region quickly reduces in size with the increase in the number of categories of the multinomial. Before performing this analysis we must define the meaning of "size" of a MLE-dominance region. There are two criteria that are defined and explored here: *coverage* and *fitness*. In the following, we assume that $\boldsymbol{\theta}=[\theta_1,\ldots,\theta_k]'$, with $\boldsymbol{\theta}'\mathbf{1}=1$ and $\boldsymbol{\theta}\in\boldsymbol{\Theta}^k$ (from now on we use the superscript k on sets to indicate the dimension on which the set is embedded).

3.1 Fitness of a region

A first way to characterize MLE-dominance regions is *fitness*. The left-hand side of Ineq. (8) can be seen as a chi-square distributed statistics, with $\mathbf{t}$ being the observed frequency and $\boldsymbol{\theta}$ the true distribution:

$$\sum_{i=1}^{k}\frac{(\theta_i-t_i)^2}{\theta_i}\approx\mathcal{X}^2(k-1), \qquad (10)$$

where $\mathcal{X}^2(k)$ is a chi-square with k degrees of freedom. Let $\boldsymbol{\Theta}^k$ be the subset of $\mathcal{S}_{\boldsymbol{\theta}}{}^k$ where the Bayesian estimator dominates MLE. Eq. (10) suggests that the "significance" of information encoded by a set $\boldsymbol{\Theta}^k\subseteq\mathcal{S}_{\boldsymbol{\theta}}{}^k$ can be evaluated by a chi-square test with $k-1$ degrees of freedom. As described in Section 2, it not hard to see that the extremes of $\boldsymbol{\Theta}^k$ will generate the most extreme values of this statistics. Hence, we define the *fitness* of $\boldsymbol{\Theta}^k$ (w.r.t. prior mean $\mathbf{t}$) by comparing

$$X^2(\boldsymbol{\Theta}^k,\mathbf{t})=\sup_{v_j}\sum_{i=1}^{k}\frac{(\theta_i^{v_j}-t_i)^2}{\theta_i^{v_j}}$$

to the chi-square distribution $\mathcal{X}^2(k-1)$. If CDF$_k$ is the cumulative distribution function of $\mathcal{X}^2(k)$, we have

$$F(\boldsymbol{\Theta}^k,\mathbf{t})=1-\text{CDF}_{k-1}(X^2(\boldsymbol{\Theta}^k,\mathbf{t}))$$

defined as the fitness measurement of the set $\boldsymbol{\Theta}^k$ w.r.t. $\mathbf{t}$. In fact this is the *p-value* of observing frequencies $\mathbf{t}$ if the true is the farthest extreme of $\boldsymbol{\Theta}^k$, and therefore a small value of $F(\boldsymbol{\Theta}^k,\mathbf{t})$ indicates that few information is encoded by $\boldsymbol{\Theta}^k$. For instance, $F(\mathcal{S}_{\boldsymbol{\theta}}{}^k,\mathbf{t})=0$ and $F(\boldsymbol{\Theta}^k,\mathbf{t})=1$ if $\boldsymbol{\Theta}^k=\{\mathbf{t}\}$.

3.2 Coverage of a region

Another way to characterize the information carried out by $\boldsymbol{\Theta}^k$ is through the ratio between its volume and the volume of the whole parameter space $\mathcal{S}_{\boldsymbol{\theta}}{}^k$. Assuming that these sets are regular $(k-1)$-simplices (corresponding to points in dimension k whose coordinates sum one) with side L, their volumes are given by $L^{k-1}\frac{\sqrt{k}}{2^{(k-1)/2}(k-1)!}$. Considering for instance $\mathcal{S}_{\boldsymbol{\theta}}{}^k$, which has side $\sqrt{2}$, its volume is:

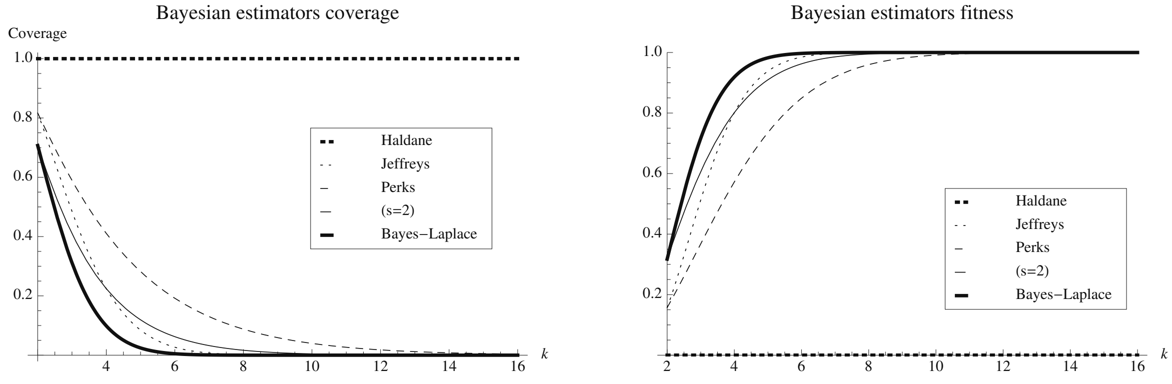
$$V(\mathcal{S}_{\boldsymbol{\theta}}{}^k)=(\sqrt{2})^{k-1}\frac{\sqrt{k}}{2^{(k-1)/2}(k-1)!}=\frac{\sqrt{k}}{(k-1)!}.$$

We thus define the proportion of coverage $\lambda(\boldsymbol{\Theta}^k)$ for a set $\boldsymbol{\Theta}^k\subseteq\mathcal{S}_{\boldsymbol{\theta}}{}^k$ to be equal to its volume divided by the volume of $\mathcal{S}_{\boldsymbol{\theta}}{}^k$, that is $\lambda(\boldsymbol{\Theta}^k)=V(\boldsymbol{\Theta}^k)/V(\mathcal{S}_{\boldsymbol{\theta}}{}^k)$. An interesting property of coverage is that, under the assumption that the true $\boldsymbol{\theta}$ has equal probability of being any point within $\mathcal{S}_{\boldsymbol{\theta}}$, $\lambda(\boldsymbol{\Theta}^k)$ can be viewed as the chance of $\boldsymbol{\theta}$ lying in $\boldsymbol{\Theta}^k$.

3.3 The ε-contaminated set

In order to analyze the dominance that is implied by the choice of different well-known priors, in the sequel we assume $\boldsymbol{\Theta}^k$ to be an ε-contaminated set defined as follows:

$$\boldsymbol{\Theta}_{\varepsilon}^k=Ch\left\{(1-\varepsilon)\boldsymbol{\theta}_{ext}+\varepsilon\frac{\mathbf{1}}{k}:\boldsymbol{\theta}_{ext}\in\text{ext}\{\mathcal{S}_{\boldsymbol{\theta}}{}^k\}\right\},$$



(a)

(b)

Figure 1: Bayesian estimators have less coverage and greater fitness as the number of categories k increases. MLE becomes the preferable estimator already for k around 4 (obviously apart from Haldane's, which equals to MLE here).

where
$\mathrm{ext}\{\mathcal{S}_{\boldsymbol{\theta}}{}^{k}\} = \{[\theta_1, \dots, \theta_k]' \in \mathcal{S}_{\boldsymbol{\theta}}{}^{k} : \theta_i = 1, i \in \{1, \dots, k\}\}$
are the extreme points (vertices) of the simplex. Note that this set is symmetric on $\mathcal{S}_{\boldsymbol{\theta}}{}^{k}$ and has k vertices, namely

$$\left[1 - \frac{k-1}{k}\varepsilon, \frac{\varepsilon}{k}, \frac{\varepsilon}{k}, \frac{\varepsilon}{k}, \dots\right], \left[\frac{\varepsilon}{k}, 1 - \frac{k-1}{k}\varepsilon, \frac{\varepsilon}{k}, \frac{\varepsilon}{k}, \dots\right],$$
$$\left[\frac{\varepsilon}{k}, \frac{\varepsilon}{k}, 1 - \frac{k-1}{k}\varepsilon, \frac{\varepsilon}{k}, \dots\right], \dots, \left[\dots, \frac{\varepsilon}{k}, \frac{\varepsilon}{k}, \frac{\varepsilon}{k}, 1 - \frac{k-1}{k}\varepsilon\right].$$

The set $\boldsymbol{\Theta}_{\varepsilon}^{k}$ has several useful properties for the analysis of MLE-dominance: (i) it is symmetric w.r.t. $\mathbf{t} = \frac{\mathbf{1}}{k}$; (ii) it is a regular simplex of side $\sqrt{2}(1-\varepsilon)$; (iii) its sides are equally distant from the border of the simplex $\mathcal{S}_{\boldsymbol{\theta}}{}^{k}$ and touch it only if $\varepsilon \to 0$. This latter property is very important for the MLE-dominance analysis. Consider Ineq. (8): if any coordinate θ_i of $\boldsymbol{\theta}$ is zero, then the left-hand side of the inequality will go to infinity (because the denominator is zero and the numerator is approximately $1/k$), forbidding the inequality to hold for any s except zero. As the coordinates θ_i get farther from zero as the left-hand side of the inequality makes it easier to be satisfied.

Considering the fitness and coverage of $\boldsymbol{\Theta}_{\varepsilon}^{k}$, we have:

$$X^2(\boldsymbol{\Theta}_{\varepsilon}^{k}, \frac{\mathbf{1}}{k}) = (k-1)\frac{(\frac{\varepsilon}{k} - \frac{1}{k})^2}{\frac{\varepsilon}{k}} + \frac{(1 - \varepsilon\frac{(k-1)}{k} - \frac{1}{k})^2}{1 - \varepsilon\frac{(k-1)}{k}}$$
$$= \frac{(1-\varepsilon)^2(k-1)}{k(\varepsilon + k - \varepsilon \cdot k)}, \tag{11}$$

and

$$\lambda(\boldsymbol{\Theta}_{\varepsilon}^{k}) = \frac{(\sqrt{2}(1-\varepsilon))^{k-1}\frac{\sqrt{k}}{2^{(k-1)/2}(k-1)!}}{\frac{\sqrt{k}}{(k-1)!}} = (1-\varepsilon)^{k-1}. \tag{12}$$

Because of the symmetry of $\boldsymbol{\Theta}_{\varepsilon}^{k}$, Ineq. (8), which has to hold for each vertex of the given set, reduces to:

$$\frac{(1-\varepsilon)^2(k-1)}{k(\varepsilon + k - \varepsilon \cdot k)} = X^2(\boldsymbol{\Theta}_{\varepsilon}^{k}, \frac{\mathbf{1}}{k}) \leq (\frac{2}{s} + \frac{1}{N}). \tag{13}$$

As larger ε as faster the coverage of $\boldsymbol{\Theta}_{\varepsilon}^{k}$ decreases, which implies that the farthest possible $\boldsymbol{\theta}$ in the set becomes quickly close to $\mathbf{1}/k$ (the sets are shrinking). Moreover, a quick analysis of Ineq. (13) shows that ε and the strength s of a prior whose MLE-dominance region equals to $\boldsymbol{\Theta}_{\varepsilon}^{k}$ are directly correlated. For any given k and N, at the increase of s there is an increase of ε (and vice-versa).

3.4 Analysis of estimators

Figure 1 presents the estimators of Haldane, Jeffreys, Perks and Bayes/Laplace, as well as an estimator with $s = 2$ (all of them use $\mathbf{t} = \frac{\mathbf{1}}{k}$). We see that the coverages of all estimators (apart Haldane, which gives the same estimate as MLE) quickly drop with the number of categories, meaning that the size (relative to the size of the simplex) of the region where they are preferred quickly approaches zero. At the same pace, their fitness increases, again showing their reduction in size. For these reasons, MLE becomes the preferred estimator (in the sense of better MSE on more than half of the parameter space) already with $k = 3$ and greater, because the coverage of the Bayesian estimators drastically reduces with k (Figure 1 shows that coverage is already small even for $k = 4$).

By Eq. (12) we clearly see that the coverage of $\boldsymbol{\Theta}_{\varepsilon}^{k}$ considerably decreases when k increases (ε is kept fixed on k – this is equivalent to s kept fixed). Hence, a natural approach to maintain the *quality* of the estimators is to keep the coverage $\lambda(\boldsymbol{\Theta}_{\varepsilon}^{k})$ constant over k, which implies that ε (and thus the strength s) has to vary with k. If λ_0 denotes the desired coverage, we obtain:

$$(1-\varepsilon)^{k-1} = \lambda_0 \iff \varepsilon_k(\lambda_0) = 1 - \lambda_0^{\frac{1}{k-1}}. \tag{14}$$

Figure 2 shows the value of $\varepsilon_k(\frac{1}{2})$, which keeps a coverage of one half of the parameter space for every k. It also presents the (dashed) curve with the value of ε_k to keep the fitness measure $F(\boldsymbol{\Theta}_{\varepsilon}^{k})$ constant instead. We see that both curves have similar slopes, indicating that coverage and fitness of $\boldsymbol{\Theta}_{\varepsilon}^{k}$ react similarly to the increase of k. At first this is slightly surprising, because chi-square distribution that is used by the fitness measure has a correction for the degrees of freedom of

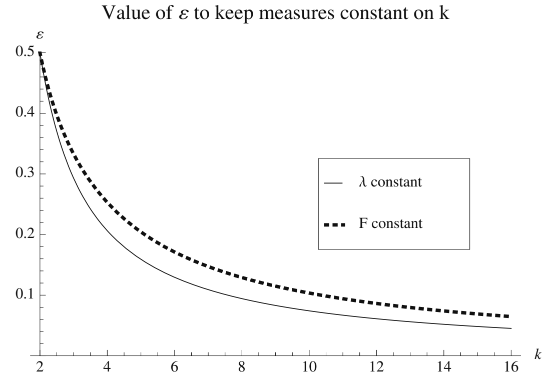


Figure 2: Value of ε to keep half of the parameter space within the membership set Θ_ε^k. The dashed curve keeps the fitness constant over k.

the multinomial. On the other hand, volumes used to compute the coverage have no such adaptive parameter corresponding to the increase in dimensionality. In spite of that, we analyze how fitness and coverage are correlated (Figure 3). We point out that this correlation is almost linear for small values of k, but becomes non-linear with its increase. More over, with the increase of the number of categories, we see that fitness goes faster and faster to zero. Still, both measures lead to similar conclusions w.r.t. the correction that has to be applied to ε (or strength s) when one varies k within practical settings – considerably large k suggests $\varepsilon \to 0$ (i.e. $s \to 0$) anyway.

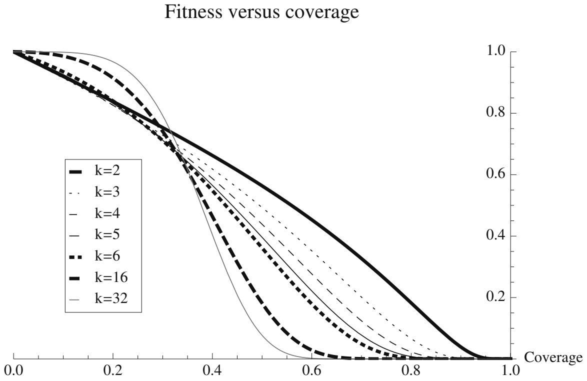


Figure 3: Comparison between fitness and coverage for various values of k.

Applying Eq. (14) on Ineq. (13) and assuming that $N \to \infty$ (which does not considerably affect the analysis, as we discuss in the final part of this section), we have

$$s_k(\lambda_0) = \frac{-2\varepsilon_k(\frac{1}{\lambda_0})(k - \varepsilon_k(\frac{1}{\lambda_0}))}{k-1}. \qquad (15)$$

(When λ_0 is omitted, then it is assumed that $\lambda_0 = 1/2$ so as to separate the parameter space in two equal parts.)

The main goal of this study is to devise rules to smartly choose the strength s of the prior distribution. If the estimator under analysis dominates MLE on (at least) half of $\mathcal{S}_{\boldsymbol{\theta}}{}^k$, then such an estimator is preferred to MLE. Hence, using Eq. (15) with $\lambda_0 = 1/2$, we obtain the maximum values that are admissible for s in settings with different values of k (Table 1). Figure 5 presents the graph of s_k for $\lambda_0 = 1/2$ (and also the dashed curve to keep the fitness measure constant).

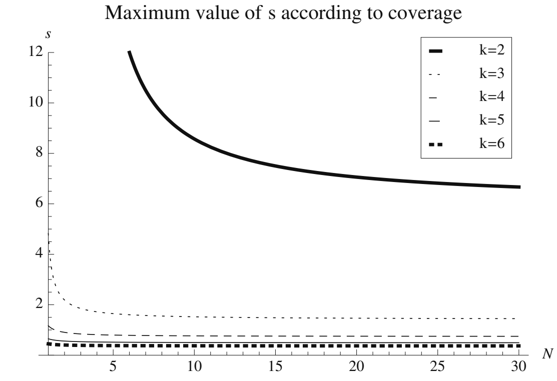


Figure 4: Maximum value of s for distinct k such that the Bayesian estimator is preferred to MLE (for true $\boldsymbol{\theta}$ uniformly generated over the parameter space).

N	$k=2$	3	4	5	6	8	16
1	any	4.83	1.17	0.65	0.45	0.27	0.11
2	any	2.19	0.91	0.56	0.40	0.26	0.10
4	24	1.72	0.81	0.52	0.38	0.25	0.10
10	8.57	1.52	0.77	0.50	0.37	0.24	0.10
100	6.19	1.42	0.74	0.49	0.37	0.24	0.10
∞	6	1.41	0.74	0.49	0.37	0.24	0.10

Table 1: Maximum value of the prior strength s for different number of categories and sample size (rounded to two digits of precision).

We point out that the numerator of Eq. (15) approaches $2\log(\frac{1}{\lambda})$ when $k \to \infty$, and it, together with the denominator $(k-1)$, justifies the reduction of s by a magnitude of $O(k)$ when k increases. Figure 6 shows that an adjustment of $O(k)$ is enough to make the coverage of Perks and the estimator with $s = 2$ almost constant, but still insufficient for the estimators of Jeffreys and Bayes/Laplace, which use $s = k/2$ and $s = k$, respectively. In fact any estimator using strength $s = c/O(k)$, for any constant c and a properly chosen linear function $O(k)$, will have its coverage kept constant with the increase of k. Therefore our suggestion of using estimators with $s = c/k$ follows.

Finally, if we do not assume $N \to \infty$, the same values derived for s are valid, because finite values of N can only help (in the sense that finite N can only increase the upper bound of s and the maximum strength s devised by Eq. (15) would still suffice). Figure 4 shows the actual maximum value of s such that at least half of the parameter space is covered, while N varies between 1 and 30. For $k = 2$, s can be chosen as high as 24 if the sample size is very small, and converges to 6 as the sample size increases. For greater values of k, the convergence to the limit value as if $N \to \infty$ is much faster, and differences in the maximum admissible value of s only

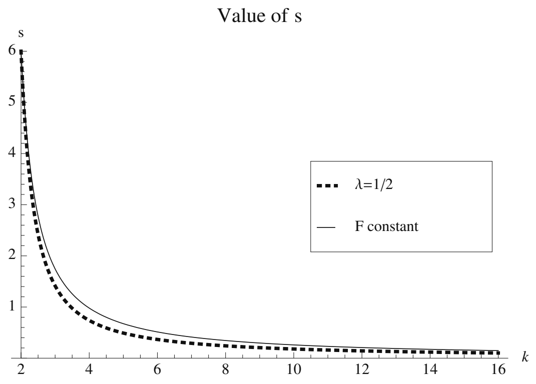


Figure 5: Maximum admissible value of s for varying number of categories k such that coverage remains $1/2$ and fitness remains constant (equal to the fitness for $k = 2$).

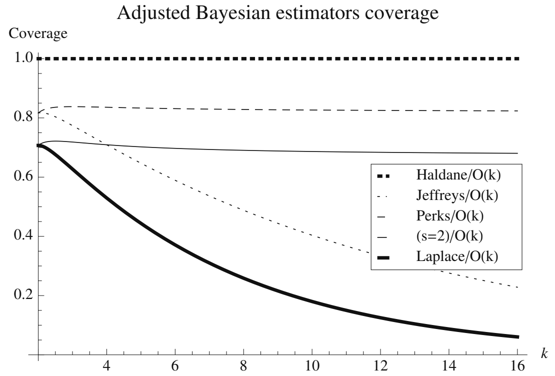


Figure 6: Comparison between Bayesian estimators with s decreasing at a rate of $O(k)$ as k increases.

occur for very small values of N (Table 1).

4 Conclusion

This paper discusses the problem of inference from multinomial data and address the problem of choosing the strength of the Dirichlet prior under a MLE-dominance criterion. This approach consists of designing free parameters of the estimator so as to guarantee, for any value of the unknown parameter vector to be estimated, an improvement of the mean-squared error with respect to MLE. Given that the true parametrization is equally probable to be any vector of the parameter space, desirable priors are those that lead to MLE-dominance in at least half of the parameter space. We show that non-informative Bayesian estimators have a region of MLE-dominance that shrinks with the increase in the number of categories of the multinomial. After a careful analysis, we devise formulas that suggest how one should select the strength of their prior to avoid such problem. They are are based on the coverage of the parameter space and the fitness of the underlying MLE-dominance region. We conclude that priors must have their strength reduced by a factor proportional to the number of categories of the multinomial, otherwise the MLE becomes a preferred estimator. We emphasize that this regards even the estimators that are already "smoothed" by k, such as Laplace (which would receive s/k^2 for each Dirichlet parameter α_j). Finally, we point out that if one has additional information and can choose a better prior than the non-informative, then the analysis of this paper does not directly apply. Yet, the additional information could be integrated into the analysis, and the general conclusion would be similar: strengths of priors have to react to the number of categories of the multinomial. As future work, we intend to investigate other measures of quality for the estimators, such as Kullback-Leibler divergence and mean absolute error instead of mean squared error, as well as analyze the case of informative priors.

Acknowledgments

This work has been partially supported by the *Computational Life Sciences – Ticino in Rete* project and the Swiss NSF grant n. 200020-121785/1. The authors thank the anonymous reviewers and the program chair for their suggestions.

References

[Benavoli and de Campos, 2009] A. Benavoli and C. P. de Campos. Inference from multinomial data based on a MLE-dominance criterion. In *Proc. of ECSQARU*, pages 22–33. Springer, 2009.

[Berger, 1985] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, New York, 1985.

[Ghosh *et al.*, 1983] M. Ghosh, J.T. Hwang, and K.W Tsui. Construction of improved estimators in multiparameter estimation for discrete exponential families. *Ann. Statist.*, pages 351–376, 1983.

[Johnson, 1971] B.M. Johnson. On admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist.*, 42:1579–1587, 1971.

[Koller and Friedman, 2009] D. Koller and N. Friedman. *Probabilistic Graphical Models*. MIT press, 2009.

[Mimno and McCallum, 2008] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proc. of UAI*, pages 411–418. AUAI Press, 2008.

[Somerville *et al.*, 1997] I. F. Somerville, D. L. Dietrich, and T. A. Mazzuchi. Bayesian reliability analysis using the Dirichlet prior distribution with emphasis on accelerated life testing run in random order. *Nonlinear Analysis*, 30(7):4415–4423, 1997.

[Stein, 1956] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Prob.*, pages 197–206. Univ. Calif. Press, 1956.

[Zhai and Lafferty, 2001] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. ACM SIGIR Conference on Research and development in information retrieval*, pages 334–342, New York, NY, USA, 2001. ACM.