Inference from multinomial data based on a MLE-dominance criterion

Alessio Benavoli and Cassio P. de Campos

Dalle Molle Institute for Artificial Intelligence
Manno, Switzerland
{alessio,cassio}@idsia.ch

Abstract. We consider the problem of inference from multinomial data with chances $\boldsymbol{\theta}$, subject to the a-priori information that the true parameter vector $\boldsymbol{\theta}$ belongs to a known convex polytope $\boldsymbol{\Theta}$. The proposed estimator has the parametrized structure of the conditional-mean estimator with a prior Dirichlet distribution, whose parameters $(s, \mathbf{t})$ are suitably designed via a *dominance* criterion so as to guarantee, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, an improvement of the *Mean Squared Error* over the *Maximum Likelihood Estimator* (MLE). The solution of this MLE-dominance problem allows us to give a different interpretation of: (1) the several Bayesian estimators proposed in the literature for the problem of inference from multinomial data; (2) the *Imprecise Dirichlet Model* (IDM) developed by Walley [13].

1 Introduction

An important estimation problem that has been treated extensively in the literature is the problem of inference from multinomial data, as the number of potential applications is huge. Accuracy of results relies on the quality of model parameters. Ideally, with enough data, it is possible to learn by standard statistical analysis like maximum likelihood estimation. However, the amount of training data may be small, for example, because of the cost of acquisition or natural conditions. In spite of that, domain knowledge through constraints is available in many real applications and can improve estimations.

The problem is as follows. Consider an infinite population which can be categorized in k categories or types from the set $C = \{c_1, \ldots, c_k\}$. The proportion of units of type c_j is denoted θ_j and called the chance of c_j. The population is thus characterized by the vector of chances $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_k]' \in \mathcal{S}_{\boldsymbol{\theta}}$, where $\mathcal{S}_{\boldsymbol{\theta}} = \{\theta_j : \ 0 \leq \theta_j \leq 1 \text{ for all } j \text{ and } \boldsymbol{\theta}'\mathbf{1} = 1\}$. The observed data consist in a sample of size N from the population, summarized by the counts $\mathbf{n} = [n_1, n_2, \ldots, n_k]'$, where n_j is the number of units of type c_j and $\mathbf{n}'\mathbf{1} = N$. The chances $\boldsymbol{\theta}$ are unknown parameters and the goal is to construct an estimator $\hat{\boldsymbol{\theta}}$ of the true chances $\boldsymbol{\theta}$, from the observations $\mathbf{n}$, that is close to $\boldsymbol{\theta}$ in some sense.

A popular measure of estimator's performance is the expected value of the quadratic loss function which is also called Mean-Squared Error (MSE) and is defined as

$$E_{\mathbf{n}}[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})']=(E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}]-\boldsymbol{\theta})(E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}]-\boldsymbol{\theta})'+(\hat{\boldsymbol{\theta}}-E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}])(\hat{\boldsymbol{\theta}}-E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}])' \quad (1)$$

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

With respect to the problem of inference from multinomial data, the MLE estimate $\hat{\boldsymbol{\theta}}_{MLE}$ has two properties: (1) it is unbiased, which means that $E_{\mathbf{n}}[\hat{\boldsymbol{\theta}}]=\boldsymbol{\theta}$; (2) it 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. Minimum variance and unbiasedness are suitable properties, but this does not imply that MLE always provides a small MSE, especially for small data samples. In fact, exploiting the relationship "MSE=variance + squared bias" and trading-off bias for variance, estimators may exist which provide a MSE lower than the CRLB for unbiased estimators.

Ranking the estimators in terms of MSE is not obvious (the MSE depends on the unknown $\boldsymbol{\theta}$) and an important practical question is how to decide which estimator to use. Although in general this question is hard to answer, some estimators may be uniformly better than others in terms of MSE. An estimator $\hat{\boldsymbol{\theta}}$ is said to *dominate* a given estimator $\hat{\boldsymbol{\theta}}_0$ on a 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 that is not dominated by any other estimator is said to be *admissible* on $\boldsymbol{\Theta}$ [5]. Hence, a desirable property of an estimator is to be admissible: otherwise it is dominated by some other estimator that have smaller MSE for all choices of $\boldsymbol{\theta}$. It can be proved that the MLE is admissible w.r.t. the MSE criterion [10]. However, estimators that dominate MLE may exist if a subregion $\boldsymbol{\Theta}\subseteq\mathcal{S}_{\boldsymbol{\theta}}$ of the parameters space is considered.

In this paper, we derive a procedure for determining estimators of a particular structure which dominates MLE on a polytopic membership set $\boldsymbol{\Theta}$. We consider an estimator $\hat{\boldsymbol{\theta}}=(\mathbf{n}+s\mathbf{t})/(N+s)$, which has the shape of the conditional-mean estimator obtained by assuming a prior Dirichlet distribution with parameters s and $\mathbf{t}$ for $\boldsymbol{\theta}$ and we design parameters s and $\mathbf{t}$, based on the knowledge $\boldsymbol{\Theta}$, so as to guarantee the MLE-dominance for the MSE. This proposal may be somehow interpreted as an intermediate approach between Bayesian and frequentist views. We assume that the unknown parameter θ is deterministic, which is in fact a frequentist approach. However our estimator would yield the optimal MMSE estimate under a Bayesian approach in the case the unknown vector $\boldsymbol{\theta}$ is actually Dirichlet distributed. This and the fact that the MLE is admissible w.r.t. the MSE are our motivations for choosing the MSE as risk function.

The idea of estimators that dominate MLE is not new (e.g. [1, 7, 9, 12]). A similar approach has also been followed in [1] for the problem of estimating an unknown parameter vector in an additive Gaussian-noise linear model by designing an estimator which dominates the least-square estimator. To our knowledge, the idea of designing the parameters s and $\mathbf{t}$, so as to guarantee the MLE-dominance inside $\boldsymbol{\Theta}$ and analysing other approaches under this perspective have not been explored in the literature.

2 Inference from multinomial data

Consider the problem of inference from multinomial data discussed at the beginning of Section 1. The objective is to compute an estimate of the parameter vector $\boldsymbol{\theta}$ based on the vector of observations $\mathbf{n}$. The probability of observing $\mathbf{n}$, conditionally on $\boldsymbol{\theta}$, is given by the multinomial distribution: $P(\boldsymbol{\theta}, \mathbf{n}) = \binom{N}{\mathbf{n}} \prod_{j=1}^{k} \theta_j^{n_j}$. The MLE can be obtained by maximizing the likelihood $L(\boldsymbol{\theta}, \mathbf{n}) \propto \prod_{j=1}^{k} \theta_j^{n_j}$ w.r.t. $\boldsymbol{\theta}$ subject to the constraint $\boldsymbol{\theta}'\mathbf{1} = 1$, which gives: $\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \mathbf{n}/N$. Another approach is to assume a Dirichlet model over $\boldsymbol{\theta}$, it generates a Dirichlet posterior density function:

$$p(\boldsymbol{\theta}|\mathbf{n}) \propto L(\boldsymbol{\theta}, \mathbf{n}) D(s, \mathbf{t}, \boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{n_j + st_j - 1} \tag{2}$$

where $D(s, \mathbf{t}, \boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{st_j - 1}$ is the prior, with $s > 0$, $\mathbf{t} = [t_1, t_2, \ldots, t_k]'$, $\mathbf{0} < \mathbf{t} < \mathbf{1}$ and $\mathbf{t}'\mathbf{1} = 1$. Using the posterior expectation of $\boldsymbol{\theta}$ given $\mathbf{n}$ as estimator, one gets:

$$\hat{\boldsymbol{\theta}} := E[\boldsymbol{\theta}|\mathbf{n}] = \frac{\mathbf{n} + s\mathbf{t}}{N + s} \tag{3}$$

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 $j = 1, 2, \ldots, k$ but differ in the choice of the value for s. Bayes and Laplace suggest to use a uniform prior $s = k$, Perks [11] suggests $s = 1$, Jeffreys $s = k/2$, and Haldane $s = 0$ [8].

3 MLE-dominance

We derive a procedure for choosing the values of the parameters s and $\mathbf{t}$ by using the MLE-dominance criterion. The idea is to design an estimator of structure as Equation (3) which dominates MLE on a polytopic membership set $\boldsymbol{\Theta}$. We choose the free parameters s and $\mathbf{t}$ so as to guarantee that:

$$E_{\mathbf{n}}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'] \leq E_{\mathbf{n}}[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MLE})'] := \boldsymbol{\Sigma}_{MLE} \tag{4}$$

for each vector $\boldsymbol{\theta}$ in the convex polytope $\boldsymbol{\Theta}$ of vertices $\boldsymbol{\theta}_{v_1}, \boldsymbol{\theta}_{v_2}, \ldots, \boldsymbol{\theta}_{v_m}$, i.e. $\boldsymbol{\Theta} = Co\{\boldsymbol{\theta}_{v_1}, \boldsymbol{\theta}_{v_2}, \ldots, \boldsymbol{\theta}_{v_m}\}$ where $Co\{\cdots\}$ stands for *convex hull*. $\boldsymbol{\Sigma}_{MLE} = (\sigma_{ij})$ 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 $i \neq j$. In order to compute the expectaction on the left-side of (4), it is convenient to rewrite the observation vector as $\mathbf{n} = N\boldsymbol{\theta} + \mathbf{v}$, where $\mathbf{v}$ is a random vector such that

$$E_{\mathbf{n}}[\mathbf{v}] = \mathbf{0}, \qquad E_{\mathbf{n}}[\mathbf{v}\mathbf{v}'] = N^2 E[(\mathbf{n}/N - \boldsymbol{\theta})(\mathbf{n}/N - \boldsymbol{\theta})'] = N^2 \boldsymbol{\Sigma}_{MLE} \tag{5}$$

where the vector of observations $\mathbf{n}$ is assumed to be unbiased, i.e. $E_{\mathbf{n}}[\mathbf{n}/N - \boldsymbol{\theta}] = \mathbf{0}$. The inequality (4) can be rewritten in the following way:

$$\begin{aligned} E_{\mathbf{n}}[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'] &= E_{\mathbf{v}}\left[\left(\boldsymbol{\theta} - \frac{N\boldsymbol{\theta}+\mathbf{v}+s\mathbf{t}}{N+s}\right)\left(\boldsymbol{\theta} - \frac{N\boldsymbol{\theta}+\mathbf{v}+s\mathbf{t}}{N+s}\right)'\right] \\ &= \frac{s^2}{(N+s)^2}(\boldsymbol{\theta}-\mathbf{t})(\boldsymbol{\theta}-\mathbf{t})' + \frac{N^2}{(N+s)^2}\boldsymbol{\Sigma}_{MLE} \leq \boldsymbol{\Sigma}_{MLE} \\ &\iff (\boldsymbol{\theta}-\mathbf{t})(\boldsymbol{\theta}-\mathbf{t})' \leq (\tfrac{2}{s}+\tfrac{1}{N})N\boldsymbol{\Sigma}_{MLE} \end{aligned} \tag{6}$$

The estimator $\hat{\boldsymbol{\theta}}$ has a MSE lower than that of MLE for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ if s and $\mathbf{t}$ are chosen to guarantee that (6) is satisfied.

If the vertices of the polytope $\boldsymbol{\Theta}$ satisfy $0 < \theta_{v_i} < 1$ for $i = 1, 2, \ldots, m$, then we can derive an alternative expression for (6)[1]. Since $N\boldsymbol{\Sigma}_{MLE} = \boldsymbol{\Lambda}_{\boldsymbol{\theta}} - \boldsymbol{\theta}\boldsymbol{\theta}'$, where $\boldsymbol{\Lambda}_{\boldsymbol{\theta}} = diag[\theta_1, \theta_2, \ldots, \theta_k]$, from (6) it follows that:

$$(\boldsymbol{\theta}-\mathbf{t})(\boldsymbol{\theta}-\mathbf{t})' \leq (\tfrac{2}{s}+\tfrac{1}{N})\left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}} - \boldsymbol{\theta}\boldsymbol{\theta}'\right) \tag{7}$$

and, thus, that:

$$\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\mathbf{t})(\boldsymbol{\theta}-\mathbf{t})'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1} \leq (\tfrac{2}{s}+\tfrac{1}{N})\left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}\right) \tag{8}$$

Equation (8) is matrix inequality $A \leq B$, which is verified if $v'(A-B)v \leq 0$ for all $v \in \mathbb{R}^k$. Then, since $v'(A-B)v \leq 0$ must hold for any v, it must also hold for $v = \boldsymbol{\theta} - \mathbf{t}$ and, thus, from $v'(A-B)v \leq 0$ and (8) one can derive:

$$\begin{aligned} &(\boldsymbol{\theta}-\mathbf{t})'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\mathbf{t})(\boldsymbol{\theta}-\mathbf{t})'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\mathbf{t}) \\ &\leq (\tfrac{2}{s}+\tfrac{1}{N})(\boldsymbol{\theta}-\mathbf{t})'\left(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}\right)(\boldsymbol{\theta}-\mathbf{t}) \end{aligned} \tag{9}$$

Since $\boldsymbol{\theta}'\mathbf{1} = 1$ and $\mathbf{t}'\mathbf{1} = 1$, it follows that $(\boldsymbol{\theta}-\mathbf{t})'(\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\boldsymbol{\theta}'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1})(\boldsymbol{\theta}-\mathbf{t}) = 0$. Manipulating (9), one finally gets

$$(\boldsymbol{\theta}-\mathbf{t})'\boldsymbol{\Lambda}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\mathbf{t}) = \sum_{i=1}^{k}\frac{(\theta_i - t_i)^2}{\theta_i} \leq (\tfrac{2}{s}+\tfrac{1}{N}) \tag{10}$$

By calculating the Hessian of the left-side of (10) and exploiting the fact that $\theta_k = 1 - \sum\limits_{i=1}^{k-1}\theta_i$, it can be verified that such function is convex on $\boldsymbol{\theta}$ (and also on $\mathbf{t}$). Therefore, (10) is satisfied for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ if the inequality (10) holds on the vertices of the polytope $\boldsymbol{\Theta}$. Thus, the MLE-dominance is guaranteed if s and $\mathbf{t}$ are chosen to satisfy the following m-inequalities

$$\sum_{i=1}^{k}\frac{(\theta^i_{v_j} - t_i)^2}{\theta^i_{v_j}} \leq (\tfrac{2}{s}+\tfrac{1}{N}), \quad \text{for } j = 1, 2, \ldots, m \tag{11}$$

[1] The problem of having vertices on the border of the probability simplex can be managed using (6) to define the constraints directly.

where $\theta^i_{v_j}$ denotes the i-th component of the j-th vertices.

The previous inequalities define all the values of s and $\mathbf{t}$ which guarantee the MLE-dominance. However, since the MSE of the admissible estimators depends on the true value of $\boldsymbol{\theta}$, there is no way to decide which admissible estimator is preferable in terms of MSE and, thus, to select one value for s and $\mathbf{t}$. Nevertheless, we can define an ad-hoc criterion to choose a single value for s and $\mathbf{t}$ [1]:

$$\begin{aligned}
&\max_{s,\mathbf{t}} s\\
&\text{subject to:}\\
&\begin{cases}\displaystyle\sum_{i=1}^{k}\frac{(\theta^i_{v_j}-t_i)^2}{\theta^i_{v_j}} \leq (\tfrac{2}{s}+\tfrac{1}{N}), \quad \text{for } j=1,2,\dots,m\\ \mathbf{t}\in\boldsymbol{\Theta}, \quad \mathbf{t}'\mathbf{1}=1, \quad s>0\end{cases}
\end{aligned}\tag{12}$$

Maximizing s means minimizing the prior uncertainty on $\boldsymbol{\theta}$. Notice that, since we know that $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ it is also natural to constrain $\mathbf{t}$ to be inside $\boldsymbol{\Theta}$. By denoting the solution of (12) with $(s_0,\mathbf{t}_0)$, it can be noticed that any pair $(s,\mathbf{t}_0)$ with $0<s<s_0$ is also a feasible solution of (12). Furthermore, notice that the optimal solution $(s_0,\mathbf{t}_0)$ of (12) depends on N. Given two values N_1 and N_2, such that $N_1\leq N_2$, from the previous remarks we have $s_0(N_2)\leq s_0(N_1)$. Therefore, $s_0(N_2)$ is a feasible value for s for any $N\leq N_2$.

The dependence of s on N violates the coherence principle, as it makes the prior information depending on the size of the data. However, taking the limit $N\to\infty$, the dependency on N can be removed from (12) and the resulting value for s will still be a feasible solution for any finite value of N. Another approach is to fix s and consider all the prior distributions defined in (11) by the inequalities in $\mathbf{t}$ and deal with a set of distributions like in the IDM [13]. Coherence and set of distributions will be discussed in Section 5.

4 Binomial data

Consider the case where only two categories ($k=2$) are distinguished. Since in this case $\boldsymbol{\theta}=[\theta_1,\theta_2]$ and $\theta_2=1-\theta_1$, there is only one degree of freedom, i.e. only one parameter to be estimated. It can easily be verified that the matrix-inequality (6) is satisfied if and only if:

$$\theta_1^2(1+\tfrac{2}{s}+\tfrac{1}{N})+\theta_1(-2t_1-\tfrac{2}{s}-\tfrac{1}{N})+t_1^2\leq 0 \tag{13}$$

Given t_1 and s, this inequality must be satisfied for each $\theta_1\in\boldsymbol{\Theta}=[\theta_a,\theta_b]$ with $0\leq\theta_a<\theta_b\leq 1$, i.e. in the binomial case the polytope is just an interval. Because of the convexity of the left side of (13), we can guarantee that the inequality is satisfied for all $\theta_1\in[\theta_a,\theta_b]$ if it is satisfied in the extremes of the interval:

$$\begin{cases}\theta_a^2(1+\tfrac{2}{s}+\tfrac{1}{N})+\theta_a(-2t_1-\tfrac{2}{s}-\tfrac{1}{N})+t_1^2\leq 0\\ \theta_b^2(1+\tfrac{2}{s}+\tfrac{1}{N})+\theta_b(-2t_1-\tfrac{2}{s}-\tfrac{1}{N})+t_1^2\leq 0\end{cases}\tag{14}$$

The values of the parameters $\theta_a \leq t_1 \leq \theta_b$ and $s > 0$ which satisfy both inequalities in (14) define all the admissible estimators (3) which dominate MLE in $[\theta_a, \theta_b]$. Solving the previous inequalities w.r.t t_1, one gets:

$$\begin{cases} \theta_a \leq \; t_1 \; \leq \theta_a + \sqrt{\alpha_s \theta_a (1-\theta_a)} \\ \theta_b - \sqrt{\alpha_s \theta_b (1-\theta_b)} \leq \; t_1 \; \leq \theta_b \end{cases} \tag{15}$$

where $\alpha_s = \frac{2}{s} + \frac{1}{N}$. If $\theta_a = 0$ and $\theta_b = 1$ then (15) yields that it does not exist any value of t_1 which satisfies all the inequalities and no estimator exists which dominates MLE (apart from the obvious case $s = 0$ in which (3) reduces to MLE). In general there is a feasible solution if

$$\theta_b - \sqrt{\alpha_s \theta_b (1-\theta_b)} \leq \theta_a + \sqrt{\alpha_s \theta_a (1-\theta_a)} \tag{16}$$

By solving (16) w.r.t to s one gets:

$$s \leq \frac{2N\left(\sqrt{\theta_a(1-\theta_a)} + \sqrt{\theta_b(1-\theta_b)}\right)^2}{N(\theta_b-\theta_a)^2 - \left(\sqrt{\theta_a(1-\theta_a)} + \sqrt{\theta_b(1-\theta_b)}\right)^2} \tag{17}$$

Notice that when the denominator is not greater than zero, any value of $s > 0$ satisfies (16). It can be seen that both s and the inequalities on t_1 depend on N. However, since in (17) the upper bound of s is a monotone decreasing function of N, the dependence on N can be dropped by taking the limit $N \to \infty$ and, thus, considering the most conservative bound

$$s \leq \frac{2\left(\sqrt{\theta_a(1-\theta_a)} + \sqrt{\theta_b(1-\theta_b)}\right)^2}{(\theta_b-\theta_a)^2} \tag{18}$$

For instance in the case $\theta_a = 0.1$ and $\theta_b = 0.9$, the previous bound states that $s \leq 1.125$. In this case, for large values of N, the right-side member of (18) and $t_1 = 1/2$ is the optimal solution of the problem (12).

An interesting result can be found by interpreting the dominance conditions (14) under the point of view of the Bayesian approach in the case of the non informative priors discussed in Section 2. Consider for the example the case $\theta_a = \epsilon, \;\; \theta_b = 1 - \epsilon$ with $\epsilon < 0.5$. In this case, by selecting $t_1 = 1/2$, the inequalities in (14) are satisfied if

$$0.5\left(1 - \sqrt{1 - \frac{1}{1 + \dfrac{2}{s} + \dfrac{1}{N}}}\right) \leq \epsilon < 0.5 \tag{19}$$

The values of the true θ_1 for which the MLE-dominance condition is satisfied when $N \to \infty$ are: Haldane ($s = 0$) needs $0 \leq \theta_1 \leq 1$; ($s = 0.5$) needs $0.05 \leq \theta_1 \leq 0.95$; Jeffreys, Perks ($s = 1$) needs $0.1 \leq \theta_1 \leq 0.9$; Bayes/Laplace ($s = 2$) needs $0.15 \leq \theta_1 \leq 0.85$.

A remark is that the length of the interval where the Bayesian estimators dominate MLE decreases at the increasing of s. Thus, under the MLE-dominance point of view, only the Haldane's prior is really non-informative. The other Bayesian estimators express preferences among subregions of the parameter space and these preferences are as stronger as higher is the value of s.

5 Imprecise Dirichlet Model

An important argument against the Bayesian approach is that, at least without a large amount of samples, inferences depend on the value of $\mathbf{t}$ to be fixed in advance, typically without having sufficient information to guide the choice. This problem is addressed by the imprecise Dirichlet model proposed by Walley [13, 14] as a model for prior ignorance about the chances $\boldsymbol{\theta}$. It avoids unjustifiable prior assumption by relying on the set of all Dirichlet distributions $D(s, \mathbf{t}, \boldsymbol{\theta})$ that can be obtained by varying the values of the vector of parameters $\mathbf{t}$. In the IDM, prior information is defined as the set $\mathcal{M}_0$ of all Dirichlet distributions on $\boldsymbol{\theta}$ with a fixed parameter $s > 0$,

$$\mathcal{M}_0 = \{D(s, \mathbf{t}, \boldsymbol{\theta}) \; \forall \, \mathbf{t} = [t_1, t_2, \dots, t_k]' \text{ s.t. } 0 < t_j < 1, \; j = 1, \dots, k, \; \mathbf{t}'\mathbf{1} = 1\}$$

After observing the data $\mathbf{n}$, each Dirichlet distribution in the set $\mathcal{M}_0$ is updated by Bayes' theorem as in (2). Under the IDM, the posterior lower and upper expectations are obtained by the minimization or maximization of $E[\boldsymbol{\theta}|\mathbf{n}]$ w.r.t. $\mathbf{t}$, which gives:

$$\underline{E}[\boldsymbol{\theta}|\mathbf{n}] = \inf_{\mathbf{0}<\mathbf{t}<\mathbf{1}} \frac{\mathbf{n} + s\mathbf{t}}{N + s} = \frac{\mathbf{n}}{N + s}, \qquad \overline{E}[\boldsymbol{\theta}|\mathbf{n}] = \sup_{\mathbf{0}<\mathbf{t}<\mathbf{1}} \frac{\mathbf{n} + s\mathbf{t}}{N + s} = \frac{\mathbf{n} + s\mathbf{1}}{N + s} \tag{20}$$

Notice that when no data are available, the lower and upper expectations reduce to $\underline{E}[\boldsymbol{\theta}] = \mathbf{0}$ and $\overline{E}[\boldsymbol{\theta}] = \mathbf{1}$ which is the vacuous probability model used to encode the initial lack of information on $\boldsymbol{\theta}$.

In the IDM, the parameter s determines how quickly upper and lower expectations converge as statistical data accumulate. The value of s must not depend on k to guarantee the *representation invariance principle* or the number of observations to guarantee the *coherence* [13, 14]. An important criterion for the choice of s is the requirement that the IDM should be cautious enough to encompass frequentist or Bayesian alternatives, but not too cautious to avoid too weak inferences. Several convincing arguments [3, 4, 14] lead to choosing $1 \leq s \leq 2$. Notice in fact that in the binomial case, Haldane ($s = 0$), Perks ($s = 1$) and uniform ($s = 2$) models are encompassed by the IDM with $s = 2$. Similar relationships have also been proved for other statistical tools such as frequentist p-values and Bayesian significance levels [3, 4, 14].

The degree of imprecision in the IDM is defined as $\overline{E}[\boldsymbol{\theta}|\mathbf{n}] - \underline{E}[\boldsymbol{\theta}|\mathbf{n}] = s/(N + s)$. This is precisely the weight of $(\boldsymbol{\theta} - \mathbf{t})(\boldsymbol{\theta} - \mathbf{t})'$ in the MSE see Equation (6). Hence, the degree of imprecision in the IDM specifies the trade-off between bias and variance in the MSE.

The IDM, like the mentioned Bayesian approaches, expresses some preference among subregions of the parameter space. Since $\mathbf{0} < \mathbf{t} < \mathbf{1}$, this preference is due to the choice of s. In the binomial case, for the IDM MLE-dominance is guaranteed for each value of t_1 if it holds for the extreme values $t_1 = 0$ and $t_1 = 1$ (because of the convexity of (13) w.r.t t_1):

$$\begin{cases} \theta_1^2(1+\frac{2}{s}+\frac{1}{N}) + \theta_1(-\frac{2}{s}-\frac{1}{N}) \leq 0 \\ \theta_1^2(1+\frac{2}{s}+\frac{1}{N}) + \theta_1(-2-\frac{2}{s}-\frac{1}{N}) + 1 \leq 0 \end{cases} \tag{21}$$

A value of θ_1 which satisfies both the inequalities exists if:

$$\frac{1}{1+\frac{2}{s}+\frac{1}{N}} \leq \frac{\frac{2}{s}+\frac{1}{N}}{1+\frac{2}{s}+\frac{1}{N}} \tag{22}$$

or, equivalently, if $1 \leq \frac{2}{s}+\frac{1}{N}$. Dropping the dependence of N by assuming that $N \to \infty$, it follows that $s \leq 2$. Therefore, in the IDM, $s \leq 2$ is a necessary and sufficient condition to the existence of one value of θ_1 for which all the IDM estimators dominate MLE. Still, when $s = 2$ only if the true value θ_1 is $1/2$ the IDM dominates MLE, while for $s = 1$ the true value θ_1 must be in $[1/3, 2/3]$ to have the dominance. It can be proved that, in the multinomial case, $s \leq 2$ is a necessary condition. When s is fixed, the MLE-dominance criterion proposed on Section 3 can also be interpreted in the imprecise probability formalism. However, IDM and the proposed approach are different. The set of distributions considered in this paper includes all the estimators that dominate MLE. Conversely, the IDM considers all the estimators consistent with the *near-ignorance* set of priors used to model the prior uncertainty on $\boldsymbol{\theta}$.

6 Numerical Examples

Consider a multinomial experiment with three categories. We assume that the available information is in the form of the convex polytope $\boldsymbol{\Theta}$, shown in Figure 1 and defined by the following vertices: $\theta_{v_1} = [0.15, 0.15, 0.7]'$, $\theta_{v_2} = [0.5, 0.15, 0.35]'$, $\theta_{v_3} = [0.5, 0.4, 0.1]'$, $\theta_{v_4} = [0.4, 0.5, 0.1]'$, and $\theta_{v_5} = [0.15, 0.5, 0.35]'$. The following estimators are compared: $\hat{\boldsymbol{\theta}}_{MLE} = \dfrac{\mathbf{n}}{N}$, $\hat{\boldsymbol{\theta}}_{MMSE_s} = \dfrac{\mathbf{n}+s\mathbf{t}}{N+s}$,

$$\hat{\boldsymbol{\theta}}_{MLE_h} = \arg\max_{\substack{\boldsymbol{\theta}\in\boldsymbol{\Theta} \\ \boldsymbol{\theta}'\mathbf{1}=1}} \sum_{i=1}^{k} n_i \log(\theta_i), \qquad \hat{\boldsymbol{\theta}}_{MMSE_h} = \frac{\int_{\boldsymbol{\theta}} \boldsymbol{\theta}\; p(\boldsymbol{\theta}|\mathbf{n})\; \mathcal{U}(\boldsymbol{\Theta})\; d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{n})\; \mathcal{U}(\boldsymbol{\Theta})\; d\boldsymbol{\theta}}$$

where $p(\boldsymbol{\theta}|\mathbf{n})$ is defined by Eq. (2), and $\mathcal{U}(\boldsymbol{\Theta})$ denotes the uniform distribution over $\boldsymbol{\Theta}$. These estimators are respectively the MLE, the MLE-dominance estimator (MMSE$_s$), the constrainted MLE (MLE$_h$), and the constrained Bayesian MMSE with $s = 1$ (MMSE$_h$), where the last two explicitly impose the hard constraint $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The parameters of the MLE-dominance estimator are $\mathbf{t} = [0.332, 0.332, 0.336]'$ and $s = 6.77$ for $N = 3$, and $s = 3.78$ for $N = 10$, obtained

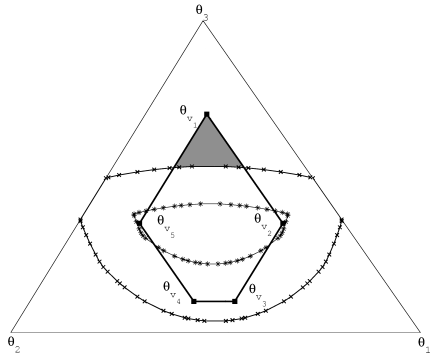

Fig. 1. Graphical representation of the a-priori information on $\boldsymbol{\theta}$. The polytope $\boldsymbol{\Theta}$ (square-marked line) is plotted inside the probability simplex $\mathcal{S}_{\boldsymbol{\theta}}$. The star-marked and plus-marked lines delimitate the set of values of $\mathbf{t}$ which guarantee the MLE-dominance over $\boldsymbol{\Theta}$ for $s = 2$ and, respectively, $s = 1$.

using the available a-priori information, as specified in (12). The estimators are compared by evaluating the MSE via Monte Carlo simulations. More specifically, 300 vectors of parameters $\boldsymbol{\theta}$ are randomly generated, uniformly over $\boldsymbol{\Theta}$, and for each vector 1000 independent trials are run by varying the realizations of $\mathbf{n}$.

The simulation results are reported in Table 1. More precisely, Table 1 reports the MSE of each component of the parameter vector of the MLE estimator as well as the % MSE reductions, w.r.t. the MLE estimator. Furthermore, the best-case ($\boldsymbol{\theta} = \mathbf{t}$) and the worst-case (average over the vertices $\boldsymbol{\theta}_{v_i}$ of the polytope) are also reported; these results are labelled as *best* and, respectively, *worst* in the table while the label *rand* refers to averages w.r.t randomly generated $\boldsymbol{\theta}$.

From Table 1, as expected, the MMSE$_h$ estimator provides the best results for *rand* and *best* cases, because it exploits the constraints in a hard way [2] (the same as the MLE$_h$ does, but it does not force the solution to be in the border as the latter). The drawbacks are: (i) it overweights distributions that are central to the polytope and its performance quickly degradates towards the border; (ii) it is computationally expensive, since it requires the solution of an integration that has no closed form, which is therefore tackled by numerical methods. Conversely, the estimator we propose, which is the second best overall, is much less time consuming since we have a convex quadratic programming problem that can be solved in polynomial time. Furthermore, the solution of this programming problem does not depend on the measurements but only on $\boldsymbol{\theta}$ and, therefore, can be calculated just once. Actually, MMSE$_s$ depends on the size of the data N, but it could be evaluated off-line for all desired N. Moreover, this dependency can be completely removed by taking the value of s corresponding to $N \to \infty$ (i.e. $s = 3.18$ in the example), which we call MMSE$_{s_\infty}$. Despite of that, its performance is still good as it can be seen in Table 1. Finally, the performance gain of MMSE$_s$ is specially relevant when size of the constrained region is small compared to the variance of the MLE (this is usually the case when the data set is small, i.e. $N = 3$ in the example). Figure 1 shows the set of all values of $\mathbf{t}$ that satisfy the MLE-dominance criterion for $N = 10$ in the $s = 1$ and $s = 2$ cases, i.e. the convex sets on $\mathbf{t}$ defined by (11) and $\mathbf{0} < \mathbf{t} < \mathbf{1}$. As discussed in Section 5, given s, this convex set defines the set of all estimators which dominate MLE on $\boldsymbol{\Theta}$. The upper and lower expectations of (3) w.r.t. the values of the MLE-dominating $\mathbf{t}$ are

Table 1. Simulation results.

		MLE		MLE$_h$		MMSE$_s$		MMSE$_{s_\infty}$		MMSE$_h$	
	MSE	N=3	N=10	N=3	N=10	N=3	N=10	N=3	N=10	N=3	N=10
rand	$\hat{\theta}_1$	0.0746	0.0221	-77%	-43%	-89%	-47%	-76%	-42%	-95%	-89%
	$\hat{\theta}_2$	0.0743	0.0224	-77%	-43%	-89%	-47%	-76%	-42%	-95%	-89%
	$\hat{\theta}_3$	0.0702	0.0213	-55%	-23%	-87%	-47%	-75%	-42%	-93%	-83%
best	$\hat{\theta}_1$	0.0740	0.0226	-77%	-41%	-91%	-48%	-77%	-43%	-99%	-95%
	$\hat{\theta}_2$	0.0735	0.0219	-77%	-41%	-91%	-48%	-77%	-43%	-99%	-95%
	$\hat{\theta}_3$	0.0721	0.0222	-55%	-23%	-91%	-49%	-77%	-43%	-96%	-93%
worst	$\hat{\theta}_1$	0.0660	0.0199	-72%	-56%	-70%	-38%	-67%	-35%	-26%	+5%
	$\hat{\theta}_2$	0.0654	0.0200	-72%	-56%	-70%	-38%	-67%	-35%	-26%	+5%
	$\hat{\theta}_3$	0.0563	0.0169	-50%	-35%	-47%	-27%	-54%	-26%	-24%	+90%

$$\underline{E}[\boldsymbol{\theta}|\mathbf{n}] = \frac{\mathbf{n} + s\mathbf{t_0}}{N + s} \qquad \overline{E}[\boldsymbol{\theta}|\mathbf{n}] = \frac{\mathbf{n} + s\mathbf{t_1}}{N + s} \tag{23}$$

where $\mathbf{t}_1 \approx [0.6625, 0.6625, 0.5325]'$ and $\mathbf{t}_0 \approx [0, 0, 0.0375]'$ for the $s = 1$ case and $\mathbf{t}_1 \approx [0.5125, 0.5125, 0.4125]'$ and $\mathbf{t}_0 \approx [0.1225, 0.1225, 0.22]'$ for the $s = 2$ case. These estimators can be compared with those defined for the IDM in (20). In the case $\mathbf{t}$ is constrained to be inside the polytope, the set of MLE-dominating $\mathbf{t}$ is given by the intersection of the previous set with the polytope. For instance, when $s = 1$ the resulting set will be the polytope excluding the shadowed area (which is exactly where IDM does not dominate MLE).

6.1 Learning Bayesian networks

Bayesian networks encode joint probability distributions using a compact representation based on a graph with nodes associated to random variables and conditional distributions specified for variables given parents in the graph. It can be defined by a triple $(\mathcal{G}, \mathcal{X}, \mathcal{P})$, where $\mathcal{G}$ is a directed acyclic graph with nodes associated to variables $\mathcal{X} = \{X_1, \ldots, X_n\}$ (which we assume to be discrete), and $\mathcal{P}$ is a collection of parameters $p(x_{ik}|\pi_{ij})$, with $\sum_k p(x_{ik}|\pi_{ij}) = 1$, where $x_{ik} \in \Omega_{X_i}$ is a category or state of X_i and $\pi_{ij} \in \times_{Y \in \pi_i} \Omega_Y$ a complete instantiation for the parents π_i of X_i in $\mathcal{G}$ (j is viewed as an index for each parent configuration). In a Bayesian network every variable is conditionally independent of its non-descendants given its parents. Hence, the joint probability distribution is obtained by $p(\mathcal{X}) = \prod_i p(X_i|\pi_i)$. We perform parameter learning in a Bayesian network where $\mathcal{G}$ is known in advance. Because of the decomposition properties of Bayesian networks, the local distributions $p(X_i|\pi_{ij}) \in \mathcal{P}$, for each X_i and configuration π_{ij} can be learned separately using the ideas previously discussed. Using three well known Bayesian network graphs (*Asia*, *Insurance* and *Alarm* networks), we generate true parameters for the distributions in $\mathcal{P}$. Using these parameters, datasets are randomly created with distinct sizes (10, 50 and 100 observations). Furthermore, constraints are generated such that true values certainly lie inside the constrained set (one interval constraint for each parameter). To compare the results, we work with five estimators: (unconstrained) MLE, hard constrained MLE, constrained maximum entropy (as described in [6]), Bayesian Dirichlet model with $s = 1$ and uniform $\mathbf{t}$ (named MMSE), and the MLE-dominant MMSE$_s$ estimator. Note that MMSE$_h$ is not included in the experiment because it is computationally too expensive, since there are hundreds of distributions to be learned. The bars in Fig. 2 represent the average Kullback–Leibler (KL) divergence for 30 runs of these methods. Size of data and nodes in the networks are presented in the labels. The number of local distributions is much higher, as it depends on the number of categories and states of the parents of each node: *Asia*, *Insurance* and *Alarm* have 21, 411 and 243 local distributions, respectively. The same set of constraints and data are applied to each method in each run. The results of MLE are not displayed because they are more than 5 times worse than the best estimator. MMSE$_s$ achieves the best results. It is clearly superior to MMSE and constrained MLE, and in general better than constrained maximum entropy.

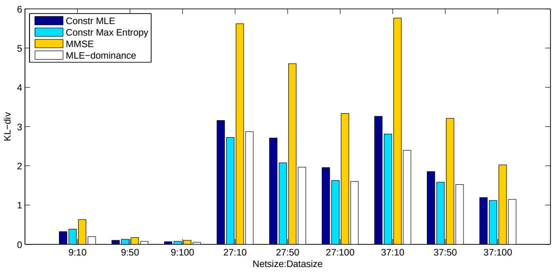

Fig. 2. Comparison between Bayesian network learning ideas.

7 Conclusion

This paper addresses the problem of inference from multinomial data under polytopic constraints on the parameter vector to be estimated, following a MLE-dominance approach. This approach consists of designing free parameters of the estimator so as to guarantee, for any admissible value of the unknown parameter vector to be estimated, an improvement of the MSE w.r.t. the standard MLE estimator. This allows us to define an objective method to choose the values of parameters s and $\mathbf{t}$ of the Dirichlet in contrast to ad-hoc practices. Using such method, we derive an estimator that is compared with existing estimators for constrained parameters, obtaining good performance. In fact, if we consider a trade-off between accuracy and computational time, our proposal surpasses the other analyzed methods. It is indeed inferior to constrained minimum mean squared error estimator (MMSE_h), but the latter cannot be run in many practical situations because of its computational cost. Besides that, the proposed estimator uses the constraints in a soft way, which makes it more robust than hard constrained estimators (like the MMSE_h) in case constraints are incorrect.

The relationship between the proposed method and the Imprecise Dirichlet Model (IDM) are briefly discussed, but some interesting conclusions already appear. For instance, according to the MLE-dominance criterion, s shall be less than 2 in the IDM. As future work, we intend to explore deeper connections of our approach with the IDM, for example, the design of a minimum value of s that guarantees MLE-dominance. As results look promising, we also intend to apply the method to large domains where data scarceness is one of the challenges.

Acknowledgements: this work has been partially supported by the Swiss NSF grant n. 200020-121785/1 (first author) and the “Ticino in Rete” project (second author). The authors are grateful to Marco Zaffalon for his suggestions.

References

1. A. Benavoli, L. Chisci, and A. Farina. Estimation of constrained parameters with guaranteed MSE improvement. *IEEE T. on Signal Processing*, 55:1264–1274, 2007.
2. A. Benavoli, L. Chisci, A. Farina, L. Ortenzi, and G. Zappa. Hard-constrained vs. soft-constrained parameter estimation. *IEEE Trans. on Aerospace and Electronic Systems*, 42:1224–1239, 2006.
3. J.M. Bernard. Bayesian interpretation of frequentist procedures for a bernoulli process. *Amer. Statist.*, 50:7–13, 1996.
4. J.M. Bernard. An introduction to the imprecise dirichlet model for multinomial data. *Int. J. of Approximate Reasoning*, pages 123–150, 2005.
5. G. Casella and E. L. Lehmann. *Theory of Point Estimation*. Springer Series in Statistics, New York, 1999.
6. C. P. de Campos and Q. Ji. Improving bayesian network parameter learning using constraints. In *Int. Conference on Pattern Recognition*, 2008.
7. 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.
8. J. B. S. Haldane. The precision of observed values of small frequencies. *Biometrika*, 35:297–300, 1948.
9. J.T. Hwang. Improving upon standard estimators in discrete exponential families with applications to poisson and negative binomial. *Ann. Statist.*, pages 857–867, 1982.

10. B.M. Johnson. On admissible estimators for certain fixed sample binomial problems. *Ann. Math. Statist.*, 42:1579–1587, 1971.
11. W. Perks. Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries*, 73:285–334, 1947.
12. C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. pages 197–206, 1956.
13. P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.
14. P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, pages 3–57, 1996.