Appendix to
Ignoramus, Ignorabimus? On Fundamental Uncertainty in Ecological Inference

Martin Elff

Faculty of Social Sciences, University of Mannheim, 68131 Mannheim, Germany

e-mail: elff@sowi.uni-mannheim.de (corresponding author)

Thomas Gschwend

MZES, University of Mannheim, 68131 Mannheim, Germany

e-mail: Thomas.Gschwend@mzes.uni-mannheim.de

Ron J. Johnston

School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK

e-mail: R.Johnston@bristol.ac.uk

A Entropy and the Kullback-Leibler information criterion

A.1 General Remarks

The Kullback-Leibler information criterion, or directed Kullback-Leibler divergence, has been widely used as a measure of dissimilarity between two probability distributions F and G with densities or probability mass functions $f(x)$ and $g(x)$. It is introduced by Kullback and Leibler (1951) and further developed in Kullback (1959). In case of continuous distributions, where $f(x)$ and $g(x)$ are densities, it is defined as

$$K[F : G] = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \, \mathrm{d}\, x \tag{1}$$

where $\mathcal{X}$ is the set of all x for which $f(x) > 0$ and $g(x) > 0$.

In case of discrete distributions, where $f(x)$ and $g(x)$ are probability mass functions, $f(x) = \Pr_F(X = x)$ and $g(x) = \Pr_G(X = x)$, it is defined as

$$K[F : G] = \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)} \tag{2}$$

If there exists a uniform density U defined on $\mathcal{X}$, for example, if $\mathcal{X}$ has only a finite number of elements or if $\mathcal{X}$ is an interval of real numbers, then the Shannon entropy of a distribution F, which is defined as

$$H[F] = -\operatorname{E}_F(\log f(x)) = -\int_{\mathcal{X}} f(x) \log f(x) \, \mathrm{d}\, x, \tag{3}$$

for the continuous case, or

$$H[F] = -\operatorname{E}_F(\log f(x)) = -\sum_{x \in \mathcal{X}} f(x) \log f(x), \tag{4}$$

for the discrete case, is, up to a constant, equal to the negative of the directed Kullback-Leibler divergence of F relative to this uniform distribution. Since the density of a Uniform distribution

is constant, say, allways equal to c if x is in $\mathcal{X}$, then the directed Kullback-Leibler divergence of F relative to U is, in the continuous case,

$$\begin{aligned} K[F:U] &= \int_{\mathcal{X}} f(x) \log \frac{f(x)}{c} \,\mathrm{d}\, x = \int_{\mathcal{X}} f(x) \log f(x) \,\mathrm{d}\, x + \log c \int_{\mathcal{X}} f(x) \,\mathrm{d}\, x \\ &= -H[F] + \log c \end{aligned} \tag{5}$$

and in the discrete case

$$\begin{aligned} K[F:U] &= \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{c} = \sum_{x \in \mathcal{X}} f(x) \log f(x) + \log c \sum_{x \in \mathcal{X}} f(x) \\ &= -H[F] + \log c \end{aligned} \tag{6}$$

Now if there are two distributions $F(\theta_1)$ and $F(\theta_2)$ which are members of a parametric family $\mathcal{F}$ of distributions and which are indexed by the parameter values θ_1 and θ_2 and if all members of the family have the same support $\mathcal{X}$ and a uniform distribution U on $\mathcal{X}$ exists, then, trivially, $H[F(\theta_1)] > H[F(\theta_2)]$ implies $K[F(\theta_1) : U] < K[F(\theta_2) : U]$, that is, the higher the entropy of a member $F(\theta)$ of $\mathcal{F}$ the more similar it is to the uniform distribution U.

The choice of the logarithm in this definition is immaterial. In statistics it is common to use the natural logarithm for convenience.

A.2 The Entropy of Multinomial Distributions

The probability mass function of a multinomial distribution $M(p_1, \ldots, p_I; n)$ with cell probabilities $p_1, \ldots, p_I$ and index n is

$$f(x_1, \ldots, x_I) = \frac{n}{x_1! \cdots x_I!} p_1^{x_1} \cdots p_I^{x_I} \tag{7}$$

where $(x_1, \ldots, x_I)$ is any I-tuple of positive integer numbers that sum to $\sum_i x_i = n$. Consequently, the entropy of this distribution is

$$\begin{aligned} H[\mathrm{M}(p_1, \ldots, p_I; n)] &= -\sum_{\substack{x_1,\ldots,x_I \\ \sum_i x_i = n}} f(x_1, \ldots, x_I) \ln f(x_1, \ldots, x_I) \\ &= -\sum_{\substack{x_1,\ldots,x_I \\ \sum_i x_i = n}} \frac{n}{x_1! \cdots x_I!} p_1^{x_1} \cdots p_I^{x_I} \ln \left(\frac{n}{x_1! \cdots x_I!} p_1^{x_1} \cdots p_I^{x_I} \right), \end{aligned} \tag{8}$$

where the number of summands is $\binom{n+k-1}{k-1}$.

There is no neat way to simplify this expression, except for the case $n = 1$, in which all x_i are either zero or one and the number of summands is I, thus

$$H[\mathrm{M}(p_1, \ldots, p_I; 1)] = -\sum_{i=1}^{I} p_i \ln p_i. \tag{9}$$

A.3 Maximum Entropy Multinomial Models

In the following we show that the Johnston-Pattie model Johnston and Pattie (2000) identifies the distribution of maximum entropy subject to the constraint that the expectations of the counts are equal to given (observed) marginal tables $\mathfrak{n}_1 = (n_{.jk})$, $\mathfrak{n}_2 = (n_{i.k})$, and $\mathfrak{n}_3 = (n_{ij.})$. We show that this is a special case of a distribution that minimizes the (directed) Kullback-Leibler information divergence relative to a given reference distribution in order to show how the Johnston-Pattie model can be generalized. We further show how the information criterion and maximum likelihood estimation are related (the 'template' for such derivations is given in Good 1963).

Lemma 1 *Let* $\mathrm{M}(\mathfrak{p}^*; n)$ *be a given multinomial distribution with cell probabilities* $\mathfrak{p}^* = (p^*_{ijk})$ *and size index* n*. If* $\mathfrak{p} = (p_{ijk})$ *are the cell probabilities of a multinomial distribution* $\mathrm{M}(\mathfrak{p}; n)$ *that minimizes the Kullback-Leibler information divergence*

$$K\left[\mathrm{M}(\mathfrak{p}; n) : \mathrm{M}(\mathfrak{p}^*; n)\right] = \sum_{i,j,k} p_{ijk} \ln \frac{p_{ijk}}{p^*_{ijk}} \tag{10}$$

subject to $\sum_k p_{ijk} = n_{ij.}/n$, $\sum_j p_{ijk} = n_{i.k}/n$, *and* $\sum_i p_{ijk} = n_{.jk}/n$ *then they can be expressed as*

$$p_{ijk} = \frac{p^*_{ijk} \exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} p^*_{rst} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st})} \tag{11}$$

and be found by maximizing the likelihood function

$$\begin{aligned} \sum_{i,j,k} x_{ijk} \ln p_{ijk} = \sum_{i,j} n_{ij}\alpha_{ij} + \sum_{i,k} n_{ik}\beta_{ik} + \sum_{j,k} n_{jk}\gamma_{jk} + \sum_{i,j,k} x_{ijk} \ln p^*_{ijk} \\ -n \ln(\sum_{i,j,k} p^*_{ijk} \exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk})) \end{aligned} \tag{12}$$

for α_{ij}, β_{ik}, *and* γ_{jk}.

Proof: In order to prove the proposition we start with the method of Lagrange multipliers. Minimizing the Kullback-Leibler information divergence subject to the constraints specified in the proposition is equivalent to minimizing the following function containing Lagange multipliers:

$$\begin{aligned} L = &\sum_{i,j,k} p_{ijk} \ln p_{ijk} - \sum_{i,j,k} p_{ijk} \ln p^*_{ijk} \\ &+ \sum_{i,j} \alpha_{ij} \left(\frac{n_{ij.}}{n} - \sum_k p_{ijk} \right) + \sum_{i,k} \beta_{ik} \left(\frac{n_{i.k}}{n} - \sum_j p_{ijk} \right) + \sum_{j,k} \gamma_{jk} \left(\frac{n_{.jk}}{n} - \sum_i p_{ijk} \right) + \tau \left(1 - \sum_{i,j,k} p_{ijk} \right). \end{aligned} \tag{13}$$

The first three sets of Lagrange multipliers assure that the given constraints are met, while the last Lagrange multiplier assures that the cell probabilities sum to one.

A necessary condition for L to be at maximum is

$$0 = \frac{\partial L}{\partial p_{ijk}} = \ln p_{ijk} + 1 - \ln p^*_{ijk} - \alpha_{ij} - \beta_{ik} - \gamma_{jk} - \tau \tag{14}$$

from which follows

$$\ln p_{ijk} = \ln p^*_{ijk} - 1 + \alpha_{ij} + \beta_{ik} + \gamma_{jk} + \tau \tag{15}$$

Since p_{ijk} are probabilities summing to one we my write

$$p_{ijk} = \frac{p_{ijk}}{\sum_{r,s,t} p_{rst}} = \frac{\exp(\ln p^*_{ijk} - 1 + \alpha_{ij} + \beta_{ik} + \gamma_{jk} + \tau)}{\sum_{r,s,t} \exp(\ln p^*_{rst} - 1 + \alpha_{rs} + \beta_{rt} + \gamma_{st} + \tau)}. \tag{16}$$

Because $\exp(\tau - 1)$ drops out of nominator and denominator, this directly leads to equation (11) so that the first part of the proof is complete.

For proving the second claim in the proposition we substitute the left hand side of equation (15) into equation (13). We then obtain

$$\begin{aligned} L &= \sum_{i,j,k} p_{ijk} \left(\ln p^*_{ijk} - 1 + \alpha_{ij} + \beta_{ik} + \gamma_{jk} + \tau \right) - \sum_{i,j,k} p_{ijk} \ln p^*_{ijk} \\ &+ \sum_{i,j} \alpha_{ij} \left(\frac{n_{ij.}}{n} - \sum_k p_{ijk} \right) + \sum_{i,k} \beta_{ik} \left(\frac{n_{i.k}}{n} - \sum_j p_{ijk} \right) + \sum_{j,k} \gamma_{jk} \left(\frac{n_{.jk}}{n} - \sum_i p_{ijk} \right) + \tau \left(1 - \sum_{i,j,k} p_{ijk} \right) \\ &= \sum_{i,j} \alpha_{ij} \frac{n_{ij.}}{n} + \sum_{i,k} \beta_{ik} \frac{n_{i.k}}{n} + \sum_{j,k} \gamma_{jk} \frac{n_{.jk}}{n} + \tau - 1 \end{aligned} \tag{17}$$

Because of

$$1 = \sum_{i,j,k} p_{ijk} = \exp(\tau - 1) \sum_{i,j,k} \exp\left(\ln p^*_{ijk} + \alpha_{ij} + \beta_{ik} + \gamma_{jk} \right), \tag{18}$$

from which follows

$$\tau - 1 = -\ln \left(\sum_{i,j,k} \exp\left(\ln p^*_{ijk} + \alpha_{ij} + \beta_{ik} + \gamma_{jk} \right) \right), \tag{19}$$

we obtain

$$\begin{aligned} L &= \sum_{i,j} \alpha_{ij} \frac{n_{ij.}}{n} + \sum_{i,k} \beta_{ik} \frac{n_{i.k}}{n} + \sum_{j,k} \gamma_{jk} \frac{n_{.jk}}{n} - \ln \left(\sum_{i,j,k} p^*_{ijk} \exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk}) \right) \\ &= n^{-1} \left(\sum_{i,j} n_{ij} \alpha_{ij} + \sum_{i,k} n_{ik} \beta_{ik} + \sum_{j,k} n_{jk} \gamma_{jk} - n \ln \left(\sum_{i,j,k} \exp(\ln p^*_{ijk} + \alpha_{ij} + \beta_{ik} + \gamma_{jk}) \right) \right). \end{aligned} \tag{20}$$

Now consider an array (x_{ijk}) of counts with $\sum_k x_{ijk} = n_{ij.}$, $\sum_j x_{ijk} = n_{i.k}$, $\sum_i x_{ijk} = n_{.jk}$. The log-likelihood of x_{ijk} under model (11) then is

$$\begin{aligned} \ell &= \sum_{i,j,k} x_{ijk} \ln p_{ijk} = \sum_{i,j,k} x_{ijk} \ln \frac{p^*_{ijk} \exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} p^*_{rst} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st})} \\ &= \sum_{i,j,k} x_{ijk} \ln p^*_{ijk} + \sum_{i,j} n_{ij} \alpha_{ij} + \sum_{i,k} n_{ik} \beta_{ik} + \sum_{j,k} n_{jk} \gamma_{jk} - n \ln \left(\sum_{i,j,k} \exp(\ln p^*_{ijk} + \alpha_{ij} + \beta_{ik} + \gamma_{jk}) \right) \end{aligned} \tag{21}$$

Since p^*_{ijk} is given, $\sum_{i,j,k} x_{ijk} \ln p^*_{ijk}$ does not depend on the α-, β- and γ-parameters. Now the derivatives of both ℓ and L are $\frac{\partial \ell}{\partial \alpha_{ij}} = n \frac{\partial L}{\partial \alpha_{ij}} = n \sum_k p_{ijk} - n_{ij.}$, $\frac{\partial \ell}{\partial \beta_{ik}} = n \frac{\partial L}{\partial \beta_{ij}} = n \sum_j p_{ijk} - n_{i.k}$, and $\frac{\partial \ell}{\partial \gamma_{jk}} = n \frac{\partial L}{\partial \gamma_{jk}} = n \sum_i p_{ijk} - n_{.jk}$, that is, both ℓ and L are at maximum if the constraints specified in the proposition are satisfied. This concludes the proof. Some trivial simplifications lead the following corollary:

Corollary 1 *If a multinomial distribution* $\mathrm{M}(\mathfrak{p}; n)$ *with probability parameters* $\mathfrak{p} = (p_{ijk})$ *has maximum entropy subject to the constraints* $\sum_k p_{ijk} = n_{ij.}/n$, $\sum_j p_{ijk} = n_{i.k}/n$, *and* $\sum_i p_{ijk} = n_{.jk}/n$, *then the probability parameters have the form*

$$p_{ijk} = \frac{\exp(\alpha_{ij} + \beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st})} \tag{22}$$

and can identified by maximizing

$$\ell = \sum_{i,j} n_{ij} \alpha_{ij} + \sum_{i,k} n_{ik} \beta_{ik} + \sum_{j,k} n_{jk} \gamma_{jk} - n \ln(\sum_{r,s,t} \exp(\alpha_{rs} + \beta_{rt} + \gamma_{st})). \tag{23}$$

for α_{ij}, β_{ik}, *and* γ_{jk}.

The proof of the following lemma is omitted, because it is completely analogous to the one given above:

Lemma 2 *Let* $\mathrm{M}(\mathfrak{p}^*; n)$ *be a given multinomial distribution with cell probabilities* $\mathfrak{p}^* = (p^*_{ijk})$ *and size index* n. *If* $\mathfrak{p} = (p_{ijk})$ *are the cell probabilities of a multinomial distribution* $\mathrm{M}(\mathfrak{p}; n)$ *that minimizes the directed Kullback-Leibler information divergence*

$$K\left[\mathrm{M}(\mathfrak{p}; n) : \mathrm{M}(\mathfrak{p}^*; n)\right] = \sum_{i,j,k} p_{ijk} \ln \frac{p_{ijk}}{p^*_{ijk}} \tag{24}$$

subject to $\sum_j p_{ijk} = n_{i.k}/n$ *and* $\sum_i p_{ijk} = n_{.jk}/n$ *then they can be expressed as*

$$p_{ijk} = \frac{p^*_{ijk} \exp(\beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} p^*_{rst} \exp(\beta_{rt} + \gamma_{st})} \tag{25}$$

and be found by maximizing the likelihood function

$$\sum_{i,j,k} x_{ijk} \ln p_{ijk} + \sum_{i,k} n_{ik}\beta_{ik} + \sum_{j,k} n_{jk}\gamma_{jk} + \sum_{i,j,k} x_{ijk} \ln p^*_{ijk} \\ -n \ln(\sum_{i,j,k} p^*_{ijk} \exp(\beta_{ik} + \gamma_{jk})) \tag{26}$$

for β_{ik}, and γ_{jk}.

Corollary 2 *If a multinomial distribution $\mathrm{M}(\mathfrak{p}; n)$ with probability parameters $\mathfrak{p} = (p_{ijk})$ has maximum entropy subject to the constraints $\sum_j p_{ijk} = n_{i.k}/n$ and $\sum_i p_{ijk} = n_{.jk}/n$, then the probability parameters have the form*

$$p_{ijk} = \frac{\exp(\beta_{ik} + \gamma_{jk})}{\sum_{r,s,t} \exp(\beta_{rt} + \gamma_{st})} \tag{27}$$

and can identified by maximizing

$$\ell = \sum_{i,k} n_{ik}\beta_{ik} + \sum_{j,k} n_{jk}\gamma_{jk} - n \ln(\sum_{r,s,t} \exp(\beta_{rt} + \gamma_{st})). \tag{28}$$

for β_{ik}, and γ_{jk}.

Based on this model, for probability of falling in cell (i, j, k) conditional on falling in unit or slice k of the array we find:

$$p_{ij|k} = \frac{p_{ijk}}{\sum_{r,s} p_{rsk}} = \frac{\exp(\beta_{ik} + \gamma_{jk})}{\sum_{r,s} \exp(\beta_{rk} + \gamma_{sk})} \\ = \frac{\exp(\beta_{ik}) \exp(\gamma_{jk})}{\sum_r \sum_s \exp(\beta_{rk}) \exp(\gamma_{sk})} = \frac{\exp(\beta_{ik})}{\sum_r \exp(\beta_{rk})} \frac{\exp(\gamma_{jk})}{\sum_s \exp(\gamma_{sk})} = p_{i.|k} p_{.j|k}. \tag{29}$$

That is, conditional on having the property indicated by k — for example, on being in voting district k — the probability of having property j — for example to turn out or not to turn out to vote — is independent from property i — for example, to belong to ethnic group i. Substantially, this means that if only two marginal tables of aggregates about e.g. spatial units are observed, the maximum entropy principle leads to a model that does not imply any statistical relation between the properties the aggregates of which are observed.

All these models have in common that they take as point of departure the entropy of a multinomial distribution for which the index is $n = 1$ or the directed Kullback-Leibler divergence between two multinomial distributions with index $n = 1$. Of course, the ecological inference procedures so far discussed deal with count data, so that n is usually much larger than one. But the derivations of maximum entropy multinomial distributions discussed above can be justified by the fact that if counts (x_{ijk}) are generated by a multinomial distribution $\mathrm{M}(\mathfrak{p}; n)$, then they are the sum of n identically independent distributed random variables with multinomial distribution $\mathrm{M}(\mathfrak{p}; 1)$ (see also Vasicek 1980). But this consideration makes one thing clear: These maximum entropy models rely on the assumption that for all n individuals the cell probabilities (p_{ijk}) that describe their behavior are the same. This (implicit) assumption is lifted either by assuming that these cell probabilities are random themselves, as by the method proposed in our paper, or by moving from a (parametric) multinomial distribution to a general distribution of the cell counts (x_{ijk}) without a parametric specification, as done in the next section of this appendix.

A.4 A Non-parametric Maximum Entropy Approach

In the following we construct a single-step maximum entropy model of the unknown counts $\mathfrak{x} = (x_{ijk})$ in an $(I \times J \times K)$-array of which the marginal tables $\mathfrak{n}_1 = (n_{.jk})$, $\mathfrak{n}_2 = (n_{i.k})$, and $\mathfrak{n}_3 = (n_{ij.})$ are observed. We start with considering the $(I \times J \times K)$-dimensional random variable $\mathfrak{X} = (X_{ijk})$ of counts that satisfy $0 \leq X_{ijk} \leq n$ and $\sum_{i,j,k} X_{ijk} = n$. Such a random variable may take $S := \binom{n+IJK-1}{IJK-1}$ different values. Any probability measure on this random variable may thus be uniquely described by an S-dimensional vector $(p_1, \ldots, p_S)$, where each element is defined by $p_s = \Pr(\mathfrak{X} = \mathfrak{x}^{(s)})$ where $\sum_{s=1}^{S} p_s = 1$. (For the sake of the argument we presuppose that an ordering of the $S = \binom{n+IJK-1}{IJK-1}$ feasible values of $\mathfrak{X}$ has already been established.) Of course, in

the absence of a parametric specification of the distribution of the counts, there is no other way of describing such a distribution other than by this S-dimensional vector. As stated in the paper, if $I = 3$, $J = 2$, $K = 100$ and n equal to one million, the number $S = \binom{n+IJK-1}{IJK-1}$ of different values of the $(I \times J \times K)$-dimensional random variable is no less than approximately 10^{2189}.

Now consider the random variables that describe possible marginal tables: $\mathfrak{N}_1 = (N_{.jk}) = (\sum_i X_{ijk})$, $\mathfrak{N}_2 = (N_{i.k}) := (\sum_j X_{ijk})$, $\mathfrak{N}_3 = (N_{ij.}) := (\sum_k X_{ijk})$. For an arbitrary probability distribution of (X_{ijk}) these random variables have expectations

$$\mathrm{E}(\mathfrak{N}_1) = \sum_{s=1}^{S} p_s \mathfrak{n}_1^{(s)}, \mathrm{E}(\mathfrak{N}_2) = \sum_{s=1}^{S} p_s \mathfrak{n}_2^{(s)}, \text{ and, } \mathrm{E}(\mathfrak{N}_3) = \sum_{s=1}^{S} p_s \mathfrak{n}_3^{(s)}, \tag{30}$$

where $\mathfrak{n}_1^{(s)} = (n_{.ik}^{(s)}) = (\sum_i x_{ijk}^{(s)})$, $\mathfrak{n}_2^{(s)} = (n_{i.k}^{(s)}) = (\sum_j x_{ijk}^{(s)})$, $\mathfrak{n}_3^{(s)} = (n_{ij.}^{(s)}) = (\sum_k x_{ijk}^{(s)})$. Note that s is the "running number" of one of the $\binom{n+IJK-1}{IJK-1}$ possible cell configurations that satisfy $x_{ijk}^{(s)} = n$.

A distribution of $\mathfrak{X}$ that maximizes the entropy $H(p_1, \ldots, p_S) = -\sum_s p_s \log p_s$ subject to the constraints $\mathrm{E}(\mathfrak{N}_1) = \mathfrak{n}_1^*$, $\mathrm{E}(\mathfrak{N}_2) = \mathfrak{n}_2^*$, and $\mathrm{E}(\mathfrak{N}_3) = \mathfrak{n}_3^*$, where $\mathfrak{n}_1^*$, $\mathfrak{n}_2^*$, $\mathfrak{n}_3^*$ may be observed marginal tables, equivalently maximizes the Lagrangian

$$L = -\sum_s p_s \log p_s + \sum_{i,j} \alpha_{ij} (n_{ij.}^* - \sum_s p_s n_{ij.}^{(s)}) + \sum_{i,k} \beta_{ik} (n_{i.k}^* - \sum_s p_s n_{i.k}^{(s)}) + \sum_{j,k} \gamma_{jk} (n_{.jk}^* - \sum_s p_s n_{.jk}^{(s)}) + \tau (1 - \sum_s p_s). \tag{31}$$

A necessary condition for the maximum of L is

$$\frac{\partial L}{\partial p_s} = -\log p_s - 1 - \sum_{i,j} \alpha_{ij} n_{ij.}^{(s)} - \sum_{i,k} \beta_{ik} n_{i.k}^{(s)} - \sum_{j,k} \gamma_{jk} n_{.jk}^{(s)} - \tau = 0 \tag{32}$$

for all $s = 1, \ldots, S$, which is equivalent to (because of $\sum_s p_s = 1$)

$$p_s = \frac{p_s}{\sum_t p_t} = \frac{\exp\left(\sum_{i,j} \alpha_{ij} n_{ij.}^{(s)} + \sum_{i,k} \beta_{ik} n_{i.k}^{(s)} + \sum_{j,k} \gamma_{jk} n_{.jk}^{(s)}\right)}{\sum_t \exp\left(-1 \sum_{i,j} \alpha_{ij} n_{ij.}^{(t)} + \sum_{i,k} \beta_{ik} n_{i.k}^{(t)} + \sum_{j,k} \gamma_{jk} n_{.jk}^{(t)}\right)}. \tag{33}$$

Substituting the expression for p_s into the equation for L leads to the reduced equation

$$\begin{aligned} L = & \sum_{i,j} \alpha_{ij} n^*_{ij.} + \sum_{i,k} \beta_{ik} n^*_{i.k} + \sum_{j,k} \gamma_{jk} n^*_{.jk} \\ & - \log \left(\sum_s \exp \left(\sum_{i,j} \alpha_{ij} n^{(s)}_{ij.} + \sum_{i,k} \beta_{ik} n^{(s)}_{i.k} + \sum_{j,k} \gamma_{jk} n^{(s)}_{.jk} \right) \right) \end{aligned} \tag{34}$$

The minimum conditions of L now become

$$\frac{\partial L}{\partial \alpha_{ij}} = \sum_s p_s n^{(s)}_{ij.} - n^*_{ij.} = 0, \; \frac{\partial L}{\partial \beta_{ik}} = \sum_s p_s n^{(s)}_{i.k} - n^*_{i.k} = 0, \text{ and } \frac{\partial L}{\partial \gamma_{jk}} = \sum_s p_s n^{(s)}_{.jk} - n^*_{.jk} = 0. \tag{35}$$

That is, finding the maximum entropy distribution of the array variable $\mathfrak{X} = (X_{ijk})$ just means finding values for the $IJ + IK + JK$ Lagrange-Multipliers that minimize the reduced form equation (34) of L. With respect to the number of parameters for which to optimize L, this problem is hardly more complex than the Johnston-Pattie model. Like in the case of the latter, there is no closed form solution for the minimum of L, so that the minimum has to be approximated iteratively. But in this case, since we cannot rely on the parametric assumption that the counts come from a multinomial distribution, it will be necessary to compute several sums with $S = \binom{n+IJK-1}{IJK-1}$ summands. If, for example, $I = 3$, $J = 2$, $K = 100$, and n equal to one million, there are approximately 10^{2189} summands. A (still hypothetical) quantum supercomputer that could perform an addition in only one Plank time ($\approx 10^{-43}$ seconds) would need approximately 10^{2146} seconds to compute such a sum, that is approximately 10^{2138} years. Although the problem of finding a non-parametric maximum entropy distribution of the $(I \times J \times K)$-dimensional random variable $\mathfrak{X} = (X_{ijk})$ is formally simple, its solution for non-trivial values of I, J, K, and n is infeasible in a world bound by the laws of physics.

B Some Properties of Dirichlet, Dirichlet-Multinomial, and Beta-binomial Distributions

B.1 Basic Properties of the Dirichlet Distribution

Here we give some basic facts about Dirichlet distributions which are relevant for the argumentation of the main text. Most of these facts were derived by by Mosimann (1962) who also refers to Dirichlet distributions as multivariate Beta distributions.

The family of Dirichlet distributions is *conjugate* to the family of multinomial distributions: If the distribution of I-tuples of counts has a multinomial distribution $\mathrm{M}(p_1, \ldots, p_I; n)$ with probability mass function

$$f_{\mathrm{M}}(x_1, \ldots, x_I) = \frac{n!}{\prod_i x_i!} \prod_i p_i^{x_i} \tag{36}$$

then a Dirichlet distribution $\mathrm{Dt}(\theta_1, \ldots, \theta_I)$ with density function

$$f_{\mathrm{Dt}}(p_1, \ldots, p_I) = \frac{\Gamma\left(\sum_i \theta_i\right)}{\prod_i \Gamma(\theta_i)} \prod_i p_i^{\theta_i - 1}, \tag{37}$$

for $p_1 > 0, \ldots, p_I > 0, \sum_i p_i = 1$ and parameters $\theta_1 > 0, \ldots, \theta_I > 0$ is a conjugate distribution (where $\Gamma(.)$ is the Gamma function, see Abramovitz and Stegun 1964, 255). A comparison of these two densities illustrates the way in which Dirichlet distributions are conjugate to multinomial distributions: In the formula of the multinomial probability mass function $p_1, \ldots, p_I$ are the parameters, while for the density function of a corresponding Dirichlet distribution they are the arguments, that is, the 'data'.

If a random variable $\boldsymbol{P} = (P_1, \ldots, P_I)$ has a Dirichlet distribution with parameters $\theta_1, \ldots, \theta_I$ then the expectation of each of its components P_i is (Mosimann 1962)

$$\mathrm{E}(P_i) = \frac{\theta_i}{\sum_i \theta_i}. \tag{38}$$

Using the notation $\pi_i := \mathrm{E}(P_i)$ and $\theta_0 := \sum_i \theta_i$, we have $\theta_i = \theta_0 \pi_i$ and the variances and covariances of the components of $\boldsymbol{P}$ are

$$\mathrm{Var}(P_i) = \frac{\theta_i(\theta_0 - \theta_i)}{\theta_0^2(\theta_0 + 1)} = \frac{\pi_i(1 - \pi_i)}{(\theta_0 + 1)}, \qquad \mathrm{Cov}(P_{i_1}, P_{i_2}) = \frac{-\theta_{i_1}\theta_{i_2}}{\theta_0^2(\theta_0 + 1)} = \frac{-\pi_{i_1}\pi_{i_2}}{(\theta_0 + 1)} \quad (\text{if } i_1 \neq i_2). \tag{39}$$

The marginal distribution of individual components P_i is a Beta distribution with shape parameters $\phi_i = \theta_i$ and $\phi_2 = \theta_0 - \theta_i$ and density function

$$f_{\mathrm{B}}(p) = \frac{p^{\phi_1}(1-p)^{\phi_2}}{\mathrm{B}(\phi_1, \phi_2)} = \frac{\Gamma(\theta_0)}{\Gamma(\theta_i)\Gamma(\theta_0 - \theta_i)} p^{\theta_i}(1-p)^{\theta_0 - \theta_i} \tag{40}$$

(where $\mathrm{B}(.,.)$ is the Beta function, see Abramovitz and Stegun 1964, 258). Thus the relation between the Dirichlet distribution and the Beta distribution parallels that between the multinomial distribution and the binomial distribution.

A notable special case of a Dirichlet distribution occurs when all $\theta_1 = \cdots = \theta_I = 1$. The density function then reduces to:

$$f_{\mathrm{Dt}}(p_1, \ldots, p_I) = \frac{\Gamma\left(\sum_i \theta_i\right)}{\prod_i \Gamma(\theta_i)} \prod_i p_i^{\theta_i - 1} = \frac{\Gamma(I)}{\prod_i \Gamma(1)} \prod_i p_i^0 = \Gamma(I) = (I-1)!, \tag{41}$$

that is, such a Dirichlet distribution is a *uniform* distribution over the set of tuples $\{(p_1, \ldots, p_I) : p_1 > 0, \ldots, p_I > 0, \sum_i p_i = 1\}$.

The Dirichlet family of distributions has a relation to the Gamma family of distributions that makes the generation of random numbers straightforward (Mosimann 1962): If the random numbers $X_1, \ldots, X_I$ have a Gamma distribution with shape parameters $\theta_1, \ldots, \theta_I$, respectively, and common scale parameter 1, then the joint distribution of the random variables $P_i = \dfrac{X_i}{\sum_k X_k}$ is a Dirichlet distribution with parameter vector $(\theta_1, \ldots, \theta_I)'$. The generation of Gamma random numbers is provided for by statistical packages like *R* (R Development Core Team 2007), so that random numbers with Dirichlet distribution can be generated easily.

B.2 The Entropy of the Dirichlet Distribution

While the basic properties of the Dirichlet distribution presented above are already well known in the statistical literature, the entropy of Dirichlet distributions is not yet commonly known. Since we have not found a derivation of the entropy of Dirichlet distributions in the literature, we prove the following lemma, which generalizes a formula that can be found, without proof however, in Lindley (1956, 1957):

Lemma 3 *The the entropy of a Dirichlet distribution with parameters* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)$ *is*

$$H[\mathrm{Dt}(\boldsymbol{\theta})] = \sum_i \ln \Gamma(\theta_i) - \ln \Gamma(\theta_0) + (\theta_0 - I)\psi(\theta_0) - \sum_i (\theta_i - 1)\psi(\theta_i), \tag{42}$$

where $\theta_0 = \sum_i \theta_i$.

Here $\psi(x) := \frac{\mathrm{d}}{\mathrm{d}\,x} \ln \Gamma(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}\,x}\Gamma(x)}{\Gamma(x)}$ denotes the digamma function (Abramovitz and Stegun 1964, 258).

Proof: First note that the entropy is defined as

$$H[\mathrm{Dt}(\boldsymbol{\theta})] = -\int \ln(f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}))\, f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} \tag{43}$$

where the integral ranges over all values of $\boldsymbol{p} = (p_1, \ldots p_I)'$ with $\sum_i p_i = 1$.

In the subsequent derivations we use, for notational brevity, the following shorthands:

$$g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) := \prod_i p_i^{\theta_i - 1} \qquad G_{\mathrm{Dt}}(\boldsymbol{\theta}) := \frac{\prod_i \Gamma(\theta_i)}{\Gamma(\sum_i \theta_i)} = \int \prod_i p_i^{\theta_i - 1}\, \mathrm{d}\, p_1 \cdots \mathrm{d}\, p_I, \tag{44}$$

so that

$$f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}) = \frac{g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \text{ and } \ln f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}) = \ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) - \ln G_{\mathrm{Dt}}(\boldsymbol{\theta}) \tag{45}$$

and we can write

$$\begin{aligned}
-H[f_{\mathrm{Dt}}(\cdot;\boldsymbol{\theta})] &= \int \ln(f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}))f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\,\mathrm{d}\,\boldsymbol{p} \\
&= \int \left[\ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) - \ln G_{\mathrm{Dt}}(\boldsymbol{\theta})\right] \frac{g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})}{G_{\mathrm{Dt}}(\boldsymbol{\theta})}\,\mathrm{d}\,\boldsymbol{p} \\
&= \frac{1}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \int g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})\,\mathrm{d}\,\boldsymbol{p} - \frac{\ln G_{\mathrm{Dt}}(\boldsymbol{\theta})}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \int g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})\,\mathrm{d}\,\boldsymbol{p} \\
&= \frac{1}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \int g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})\,\mathrm{d}\,\boldsymbol{p} - \ln G_{\mathrm{Dt}}(\boldsymbol{\theta}).
\end{aligned} \tag{46}$$

Note that

$$\frac{\partial}{\partial\theta_i} p_i^{\theta_i-1} = \frac{\partial}{\partial\theta_i} \mathrm{e}^{(\theta_i-1)\ln p_i} = \mathrm{e}^{(\theta_i-1)\ln p_i} \frac{\partial}{\partial\theta_i}\left((\theta_i-1)\ln p_i\right) = p_i^{\theta_i-1}\ln p_i \tag{47}$$

and therefore

$$\begin{aligned}
\sum_i (\theta_i-1)\frac{\partial g_{\mathrm{Dt}}}{\partial\theta_i}(\boldsymbol{\theta};\boldsymbol{p}) &= \sum_i (\theta_i-1) g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln p_i \\
&= g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln\left(\prod_i p_i^{\theta_i-1}\right) \\
&= g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}).
\end{aligned} \tag{48}$$

Therefore we find

$$\begin{aligned}
\int g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})\,\mathrm{d}\,\boldsymbol{p} &= \int \sum_i (\theta_i-1)\frac{\partial g_{\mathrm{Dt}}}{\partial\theta_i}(\boldsymbol{\theta};\boldsymbol{p})\,\mathrm{d}\,\boldsymbol{p} \\
&= \sum_i (\theta_i-1)\frac{\partial}{\partial\theta_i} G_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}).
\end{aligned} \tag{49}$$

From

$$\begin{aligned}
\frac{\partial}{\partial\theta_i} G_{\mathrm{Dt}}(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta_i}\frac{\prod_k \Gamma(\theta_k)}{\Gamma(\sum_k \theta_k)} &= \frac{\prod_k \Gamma(\theta_k)}{\Gamma(\sum_k \theta_k)}\frac{\Gamma'(\theta_i)}{\Gamma(\theta_i)} - \frac{\prod_k \Gamma(\theta_k)}{\Gamma(\sum_k \theta_k)}\frac{\Gamma'(\sum_k \theta_k)}{\Gamma(\sum_k \theta_k)} \\
&= G_{\mathrm{Dt}}(\boldsymbol{\theta})\left(\psi(\theta_i) - \psi\Big(\sum_k \theta_k\Big)\right)
\end{aligned} \tag{50}$$

we conclude

$$\begin{aligned}
-H[f_{\mathrm{Dt}}(\cdot;\boldsymbol{\theta})] =& \frac{1}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \int_{\Omega} g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p})\, d\boldsymbol{p} - \ln G_{\mathrm{Dt}}(\boldsymbol{\theta}) \\
=& \frac{1}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \sum_i (\theta_i - 1) \frac{\partial}{\partial \theta_i} G_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) - \ln G_{\mathrm{Dt}}(\boldsymbol{\theta}) \\
=& \sum_i (\theta_i - 1) \left(\psi(\theta_i) - \psi\Big(\sum_k \theta_i\Big) \right) - \ln \frac{\prod_i \Gamma(\theta_i)}{\Gamma(\sum_i \theta_i)} \\
=& \sum_i (\theta_i - 1)\psi(\theta_i) - (\theta_0 - I)\psi(\theta_0) + \ln \Gamma(\theta_0) - \sum_i \ln \Gamma(\theta_i),
\end{aligned} \tag{51}$$

from which our proposition follows directly.

Now consider a given distribution F_h with density $h(\boldsymbol{p})$, where $h(\boldsymbol{p}) > 0$ if $f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}) > 0$. Then the Kullback-Leibler information divergence of $f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})$ from $h(\boldsymbol{p})$ can be written as:

$$\begin{aligned}
K[\mathrm{Dt}(\boldsymbol{\theta}) : F_h] &= \int \ln \left(\frac{f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})}{h(\boldsymbol{p})} \right) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} \\
&= \int \left(\ln f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}) - \ln h(\boldsymbol{p}) \right) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} \\
&= \int \ln f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} - \int \ln h(\boldsymbol{p}) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p}
\end{aligned} \tag{52}$$

that is, the entropy of the Dirichlet distribution differs from the negative of the Kullback-Leibler information divergence by the integral $-\int \ln h(\boldsymbol{p}) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p}$. If $h(\boldsymbol{p})$ is the density of a Dirichlet distribution with parameters $\boldsymbol{\theta}^*$, that is $h(\boldsymbol{p}) = f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}^*)$, then

$$\begin{aligned}
\int \ln(h(\boldsymbol{p})) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} &= \int \ln(f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}^*)) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} \\
&= \frac{1}{G_{\mathrm{Dt}}(\boldsymbol{\theta})} \int g_{\mathrm{Dt}}(\boldsymbol{\theta};\boldsymbol{p}) \ln g_{\mathrm{Dt}}(\boldsymbol{\theta}^*;\boldsymbol{p})\, \mathrm{d}\,\boldsymbol{p} - \ln G_{\mathrm{Dt}}(\boldsymbol{\theta}^*).
\end{aligned} \tag{53}$$

Computations similar to the above proof lead to

$$\int \ln f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta}^*) f_{\mathrm{Dt}}(\boldsymbol{p};\boldsymbol{\theta})\, \mathrm{d}\,\boldsymbol{p} = \sum_i (\theta_i^* - 1) \left(\psi(\theta_i) - \psi\Big(\sum_k \theta_k\Big) \right) + \ln \Gamma\Big(\sum_i \theta_i^*\Big) - \sum_i \ln \Gamma(\theta_i^*) \tag{54}$$

so that the corollary follows immediately:

Corollary 3 *The Kullback-Leibler information divergence of a Dirichlet distribution with parameters* $\theta_1, \ldots, \theta_r$

relative to a Dirichlet distribution with parameters $\theta_1^*, \ldots, \theta_r^*$ and the same support is

$$\begin{aligned} K[\mathrm{Dt}(\boldsymbol{\theta}) : \mathrm{Dt}(\boldsymbol{\theta}^*)] = \sum_i (\theta_i - \theta_i^*) \left(\psi(\theta_i) - \psi\Big(\sum_k \theta_k\Big) \right) \\ + \ln\Gamma\Big(\sum_i \theta_i\Big) - \sum_i \ln\Gamma(\theta_i) - \ln\Gamma\Big(\sum_i \theta_i^*\Big) + \sum_i \ln\Gamma(\theta_i^*) \end{aligned} \tag{55}$$

B.3 Properties of the Dirichlet-multinomial distribution

In the main text we make use of Dirichlet-multinomial distributions, also known as compound-multinomial distributions. We therefore present some of the properties of these distributions relevant for the argument of our paper.

Most of the basic properties of these distributions are known from Mosimann (1962). The probability mass function of a Dirichlet-multinomial distribution with parameters $\theta_1 > 0, \ldots, \theta_I > 0$ and size index n is derived as

$$\begin{aligned} f_{\mathrm{DtM}}(x_1, \ldots, x_I) &= \int \frac{n!}{\prod_i x_i!} \prod_i p_i^{x_i} \frac{\Gamma(\theta_0)}{\prod_i \Gamma(\theta_i)} \prod_i p_i^{\theta_i - 1} \,\mathrm{d}\, p_1 \cdots \mathrm{d}\, p_I \\ &= \frac{n!\Gamma(\theta_0)}{\Gamma(n + \theta_0)} \prod_i \frac{\Gamma(x_i + \theta_i)}{x_i!\Gamma(\theta_i)} \end{aligned} \tag{56}$$

where, again, $\theta_0 := \sum_i \theta_i$ and the integral is over all $p_1 > 0, \ldots, p_I > 0$ with $\sum_i p_i = 1$.

The Expectations of the components of a random variable $(X_1, \ldots, X_I)$ with such a distribution are:

$$\mathrm{E}(X_i) = n \frac{\theta_i}{\theta_0} = n\pi_i \tag{57}$$

where, again $\pi_i := \frac{\theta_i}{\theta_0}$. This is a property that helps to justify the method proposed in our paper: If we obtain some $\hat{p}_{ijk}$ from an ecological inference procedure that satisfy constraints like

$$\sum_i \mathrm{E}(X_{ijk}) = n \sum_i \hat{p}_{ijk} = n_{\cdot jk}, \quad \sum_j \mathrm{E}(X_{ijk}) = n \sum_j \hat{p}_{ijk} = n_{i\cdot k}, \quad \text{and } \sum_k \mathrm{E}(X_{ijk}) = n \sum_k \hat{p}_{ijk} = n_{\cdot jk} \tag{58}$$

then these $\hat{p}_{ijk}$ can be interpreted as estimates of the cell probabilities of a multinomial distribution, as in case of the conditional independence or the Johnston-Pattie model, or they can be interpreted as estimates $\hat{\pi}_{ijk}$ of the means π_{ijk} of a Dirichlet distribution: The corresponding Dirichlet-multinomial distribution then has means $\mathrm{E}(X_{ijk}) = n\hat{\pi}_{ijk}$ so that

$$\sum_i \mathrm{E}(X_{ijk}) = n\sum_i \hat{\pi}_{ijk} = n_{.jk}, \quad \sum_j \mathrm{E}(X_{ijk}) = n\sum_j \hat{\pi}_{ijk} = n_{i.k}, \quad \text{and } \sum_k \mathrm{E}(X_{ijk}) = n\sum_k \hat{\pi}_{ijk} = n_{.jk}. \tag{59}$$

The variances and covariances of components of a random variable $(X_1, \ldots, X_I)$ with a Dirichlet-multinomial distribution are:

$$\mathrm{Var}(X_i) = n\pi_i(1-\pi_i)\frac{n+\theta_0}{1+\theta_0} \qquad \mathrm{Cov}(X_{i_1}, X_{i_2}) = -n\pi_{i_1}\pi_{i_2}\frac{n+\theta_0}{1+\theta_0} \quad \text{for } i_1 \neq i_2. \tag{60}$$

That is, while the mean of a Dirichlet-multinomial distribution is the same as the mean of a multinomial distribution whose cell probabilities are equal to the mean of the Dirichlet mixing distribution, the variance is proportional to the variance of such a multinomial distribution. However, while the variance of the proportions $F_i := X_i/n$ that correspond to multinomial distributed counts (with cell probabilities $p_1, \ldots, p_I$) approaches zero as n approaches infinity:

$$\lim_{n\to\infty} \mathrm{Var}_{\mathrm{M}}(F_i) = \lim_{n\to\infty} \frac{1}{n^2}\,\mathrm{Var}_{\mathrm{M}}(X_i) = \lim_{n\to\infty} \frac{1}{n^2}\,(np_i(1-p_i)) = \lim_{n\to\infty} \frac{p_i(1-p_i)}{n} = 0 \tag{61}$$

this is not the case for the Dirichlet-multinomial distribution. Instead, the variance of the proportions $F_i := X_i/n$ approaches that of the mixing Dirichlet distribution:

$$\lim_{n\to\infty} \mathrm{Var}_{\mathrm{DtM}}(F_i) = \lim_{n\to\infty} \frac{1}{n^2}\left(n\pi_i(1-\pi_i)\frac{n+\theta_0}{1+\theta_0}\right) = \lim_{n\to\infty} \pi_i(1-\pi_i)\frac{1+\theta_0/n}{1+\theta_0} = \frac{\pi_i(1-\pi_i)}{1+\theta_0}. \tag{62}$$

This explains why in the simulation study of our paper prediction intervals based on a Dirichlet-multinomial distribution, in contrast to intervals based on a multinomial distribution, do not show a deteriorating coverage performance as n increases.

Also in case of the Dirichlet-multinomial distribution, it is a notable special case if $\theta_1 = \cdots = \theta_I = 1$. Then, the probability mass function, because of $\Gamma(x+1) = x!$ reduces to

$$f_{\mathrm{DtM}}(x_1, \ldots, x_I) = \frac{n!\Gamma(I)}{\Gamma(n+I)} \prod_i \frac{\Gamma(x_i+1)}{x_i!\Gamma(1)} = \frac{n!(I-1)!}{(n+I-1)!} = 1 \Big/ \binom{n+I-1}{I-1} \tag{63}$$

that is, the distribution becomes a uniform distribution for all I-tuples of integers $(x_1, \ldots, x_I)$ with $\sum_i x_i = n$.

Uniformly distributed random arrays (x^r_{ijk}) of counts with the property $\sum_{i,j,k} x_{ijk} = n$, as they were used in the simulation study of our paper, can therefore be generated as follows: First, random numbers $p^{(r)}_{ijk}$ are generated from a Dirichlet distribution with all parameters $\theta_{ijk} = 1$, second, random numbers are generated from a multinomial distribution with cell probabilities $p^{(r)}_{ijk}$. The generation of Dirichlet random numbers is described in section B.1 above, the generation of multinomial random numbers is provided for by a statistical package like *R* (R Development Core Team 2007).

If a random variable $(X_1, \ldots, X_I)$ has a Dirichlet-multinomial distribution with parameters $\theta_1 > 0, \ldots, \theta_I > 0$ and size index n, then the individual components $X_1, \ldots, X_I$ have individually a Beta-binomial distribution with shape parameters $\phi_1 = \theta_i$ and $\phi_2 = \theta_0 - \theta_i$ and probability mass function

$$f_{\mathrm{Bb}}(x) = \binom{n}{x} \frac{\mathrm{B}(x+\phi_1, n-x+\phi_2)}{\mathrm{B}(\phi_1, \phi_2)} \tag{64}$$

that is, parallel to the way a binomial distribution is related to a multinomial distribution and a Beta distribution is related to a Dirichlet distribution, a Beta-binomial distribution is related to a Dirichlet-multinomial distribution.

We use Beta-binomial distributions to construct prediction intervals for the unknown cell counts x_{ijk} for which only marginal summaries $n_{.jk}$, $n_{i.k}$, and sometimes $n_{ij.}$ are observed. Since asymptotic normality does not hold for Beta-binomial distributions as it does for binomial distributions

if n approaches infinity, some other methods have to employed for the construction of these prediction intervals, which are explained in the following section.

B.4 Cumulative Probabilities of Beta-binomial Distributions

In order to construct prediction intervals for a discrete random variable X with probability mass function $\Pr(X = x) = f(x)$ and cumulative probability distribution $F(x) := \Pr(X \leq x) = \sum_{k=0}^{x} \Pr(X = k) = \sum_{k=0}^{x} f(k)$ one needs to compute the quantile function $F^{-1}(\alpha) := \sup\{x : F(x) < \alpha\}$ for given α_{lower} and α_{upper}. If the random variable X has a Beta-binomial distribution with parameters ϕ_1 and ϕ_2 and size index n, its cumulative probability function is

$$F_{\text{Bb}}(x) = \Pr_{\text{Bb}}(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} \frac{\text{B}(k + \phi_1, n - k + \phi_2)}{\text{B}(\phi_1, \phi_2)}. \tag{65}$$

Since the cumulative probability function involves several binomial coefficients and Beta functions, its computation in this form can be very costly. On the other hand, asymptotic approximations (of which we will discuss one further below) will not work well if n, x, or $n - x$ is small so that an exact computation will be needed in such cases. The effort for computing the cumulative distribution function can be drastically reduced if some recurrence relations of binomial coefficients and the gamma function are exploited.

First we define $\text{B}^*(x, n, \phi_1, \phi_2) := \binom{n}{x} \text{B}(x + \phi_1, n - x + \phi_2)$ so that the Beta-binomial probability mass function becomes

$$f_{\text{Bb}} = \frac{\text{B}^*(x, n, \phi_1, \phi_2)}{\text{B}(\phi_1, \phi_2)} \tag{66}$$

For $x = 0$ and $x = 1$ we have

$$\text{B}^*(0, n, \phi_1, \phi_2) = \text{B}(\phi_1, n + \phi_2) \tag{67}$$

and

$$\mathrm{B}^*(1, n, \phi_1, \phi_2) = n\mathrm{B}(1 + \phi_1, n - 1 + \phi_2), \tag{68}$$

For integer numbers $x > 1$ we can exploit the following recurrence relation of the Gamma function

$$\Gamma(x + \phi) = \Gamma(x - 1 + \phi)(x - 1 + \phi) \tag{69}$$

and of the binomial coefficient

$$\binom{n}{x} = \frac{n - x + 1}{x}\binom{n}{x - 1} \tag{70}$$

and also the relation between the Beta and the Gamma functions

$$\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}. \tag{71}$$

Therefore we find for $x > 1$

$$\begin{aligned}
\frac{\mathrm{B}^*(x, n, \phi_1, \phi_2)}{\mathrm{B}^*(x - 1, n, \phi_1, \phi_2)} &= \frac{n - x + 1}{x}\frac{\mathrm{B}(x + \phi_1, n - x + \phi_2)}{\mathrm{B}(x - 1 + \phi_1, n - x + 1 + \phi_2)} \\
&= \frac{n - x + 1}{x}\frac{\Gamma(x + \phi_1)\Gamma(n - x + \phi_2)\Gamma(x - 1 + \phi_1 + n - x + 1 + \phi_2)}{\Gamma(x + \phi_1 + n - x + \phi_2)\Gamma(x - 1 + \phi_1)\Gamma(n - x + 1 + \phi_2)} \\
&= \frac{n - x + 1}{x}\frac{\Gamma(x + \phi_1)}{\Gamma(x - 1 + \phi_1)}\frac{\Gamma(n - x + \phi_2)}{\Gamma(n - x + 1 + \phi_2)} \\
&= \frac{n - x + 1}{x}\frac{x - 1 + \phi_1}{n - x + \phi_2}.
\end{aligned} \tag{72}$$

So $F_{\mathrm{Bb}}(x)$ can be computed efficiently by using the following steps:

1. $a = \mathrm{B}(\phi_1, \phi_2)$
2. $b_0 = \mathrm{B}^*(0, n, \phi_1, \phi_2) = \mathrm{B}(\phi_1, n + \phi_2)$
3. $b_1 = \mathrm{B}^*(1, n, \phi_1, \phi_2) = n\mathrm{B}(1 + \phi_1, n - 1 + \phi_2)$
4. $b_k = \frac{n - k + 1}{k}\frac{k - 1 + \phi_1}{n - k + \phi_2}\, b_{k-1}$ for $1 < k \leq x$
5. $F_{\mathrm{BB}}(x) = \frac{1}{a}\sum_{k=0}^{x} b_k$

Further efficiency gains can be obtained by using

$$\Pr_{\mathrm{BB}}(X \leq x; \phi_1, \phi_2, n) = \Pr_{\mathrm{BB}}(X \geq x; \phi_2, \phi_1, n). \tag{73}$$

In the search of $\sup\{x : F_{\mathrm{BB}}(x) < \alpha\}$ the updating step

$$F_{\mathrm{BB}}(x) = \frac{n-x+1}{x} \frac{x-1+\phi_1}{n-x+\phi_2} \frac{b_{x-1}}{a} + F_{\mathrm{BB}}(x-1) \tag{74}$$

can be employed for further gains in computational efficiency.

If n is very large, these exact methods may still be computationally too costly. In this case we employ an approximation to Beta-binomial probability mass and cumulative probability functions developed by Hald (1968). These approximations rely on the fact that as n goes to infinity, the distribution of X/n approaches a Beta distribution. Hald proposes the following finite-n corrections:

$$\Pr_{\mathrm{BB}}(X = x) \approx \frac{f_{\mathrm{B}}(h; \phi_1, \phi_2)}{n} \left(1 + \frac{1}{n} b_1(h; \phi_1, \phi_2) + \frac{1}{n^2} b_2(h; \phi_1, \phi_2)\right) \tag{75}$$

$$\Pr_{\mathrm{B}}(X \leq x) \approx F_{\mathrm{B}}(h; \phi_1, \phi_2) + \frac{1}{n} B_1(h; \phi_1, \phi_2) + \frac{1}{n^2} B_2(h; \phi_1, \phi_2) \tag{76}$$

where $h = x/n$, while $b_1(h; \phi_1, \phi_2)$, $b_2(h; \phi_1, \phi_2)$, $B_1(h; \phi_1, \phi_2)$, and $B_2(h; \phi_1, \phi_2)$ are rational functions of h, ϕ_1, and ϕ_2 the details of which we omit here. Both the exact method and the Hald approximation were implemented in *R* (R Development Core Team 2007) in order to conduct the simulation studies of our paper.

C A Non-parametric Bayesian Approach

Like the non-parametric maximum-entropy model, a direct Bayesian model of the unknown cell counts will consider an $(I \times J \times K)$-dimensional random variable $\mathfrak{X} = (X_{ijk})$ of counts that satisfy $0 \leq X_{ijk} \leq n$ and $\sum_{i,j,k} X_{ijk} = n$. Like in section A.4 of this appendix, we consider arbitrary probability distributions of this random variable, which can be uniquely described by S-dimensional

vectors $(p_1, \ldots, p_S)$, where each element is defined by $p_s = \Pr(\mathfrak{X} = \mathfrak{x}^{(s)})$ with $\sum_{s=1}^{S} p_s = 1$ and $S = \binom{n+IJK-1}{IJK-1}$. We further consider the random variables that represent the marginal tables, $\mathfrak{N}_1 = (N_{\cdot jk}) = (\sum_i X_{ijk})$, $\mathfrak{N}_2 = (N_{i\cdot k}) := (\sum_j X_{ijk})$, $\mathfrak{N}_3 = (N_{ij\cdot}) := (\sum_k X_{ijk})$. Now, the joint distribution of $\mathfrak{N}_1$, $\mathfrak{N}_2$, and $\mathfrak{N}_3$ conditional on $\mathfrak{X}$ is fairly simple:

$$\begin{aligned}
&\Pr(\mathfrak{N}_1 = \mathfrak{n}_1 \wedge \mathfrak{N}_2 = \mathfrak{n}_2 \wedge \mathfrak{N}_3 = \mathfrak{n}_3 | \mathfrak{X} = \mathfrak{x}) \\
&\qquad = \begin{cases} 1 & \text{if } n_{ij\cdot} = \sum_k x_{ijk} \text{ and } n_{i\cdot k} = \sum_j x_{ijk} \text{ and } n_{\cdot ik} = \sum_i x_{ijk} \text{ for all } i, j, k \\ 0 & \text{if } n_{ij\cdot} \neq \sum_k x_{ijk} \text{ or } n_{i\cdot k} \neq \sum_j x_{ijk} \text{ or } n_{\cdot ik} \neq \sum_i x_{ijk} \text{for some } i, j, k \end{cases} \\
&\qquad = \delta(\mathfrak{x} : \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3})
\end{aligned} \tag{77}$$

where $\mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3} := \{(x_{ijk}) : n_{ij\cdot} = \sum_k x_{ijk} \wedge n_{i\cdot k} = \sum_j x_{ijk} \wedge n_{\cdot ik} = \sum_i x_{ijk} \text{ for all } i, j, k\}$ that is, the set of all possible arrays that have the given marginal sums and $\delta(x : \mathcal{S})$ is the *indicator function* of $\mathcal{S}$, which is equal to one if and only if its first argument is an element of $\mathcal{S}$ and null otherwise.

Bayes' theorem gives the posterior distribution of $\mathfrak{X}$ given the observed marginal tables $\mathfrak{n}_1$, $\mathfrak{n}_2$, and $\mathfrak{n}_3$:

$$\begin{aligned}
&\Pr(\mathfrak{X} = \mathfrak{x} | \mathfrak{N}_1 = \mathfrak{n}_1 \wedge \mathfrak{N}_2 = \mathfrak{n}_2 \wedge \mathfrak{N}_2 = \mathfrak{n}_3) \\
&\qquad = \frac{\Pr(\mathfrak{N}_1 = \mathfrak{n}_1 \wedge \mathfrak{N}_2 = \mathfrak{n}_2 \wedge (N_{\cdot jk}) = \mathfrak{n}_3 | \mathfrak{X} = \mathfrak{x}^{(s)}) \Pr(\mathfrak{X} = \mathfrak{x})}{\sum_s \Pr(\mathfrak{N}_1 = \mathfrak{n}_1 \wedge \mathfrak{N}_2 = \mathfrak{n}_2 \wedge \mathfrak{N}_3 = \mathfrak{n}_3 | \mathfrak{X} = \mathfrak{x}^{(s)}) \Pr(\mathfrak{X} = \mathfrak{x}^{(s)})} \\
&\qquad = \frac{\delta(\mathfrak{x} : \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3}) \Pr(\mathfrak{X} = \mathfrak{x})}{\sum_s \delta(\mathfrak{x}^{(s)} : \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3}) \Pr(\mathfrak{X} = \mathfrak{x}^{(s)})}
\end{aligned} \tag{78}$$

where the sum in the denominator runs over all $S = \binom{n+IJK-1}{IJK-1}$ possible values the random variable may take.

Now if the prior probability if $\mathfrak{X}$ is uniform with $p_s = S^{-1}$, then the posterior simplifies to

$$\begin{aligned}
&\Pr(\mathfrak{X} = \mathfrak{x} | \mathfrak{N}_1 = \mathfrak{n}_1 \wedge \mathfrak{N}_2 = \mathfrak{n}_2 \wedge \mathfrak{N}_3 = \mathfrak{n}_3) \\
&\qquad = \frac{\delta(\mathfrak{x} : \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3}) S^{-1}}{\sum_s \delta(\mathfrak{x}^{(s)} : \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3}) S^{-1}} = \frac{\delta(\mathfrak{x} : \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3})}{\# \mathcal{S}_{\mathfrak{n}_1, \mathfrak{n}_2, \mathfrak{n}_3}}
\end{aligned} \tag{79}$$

where $\#\mathcal{S}$ is the number of elements of the set $\mathcal{S}$. This posterior distribution has a very simple structure: The posterior probability, conditional on the observed marginal tables $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_3$, that the random array $\mathfrak{X}$ takes the value $\mathfrak{x}^*$ is $(\#\mathcal{S}_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{n}_3})^{-1}$ if $n_{ij.} = \sum_k x_{ijk}$, $n_{i.k} = \sum_j x_{ijk}$, and $n_{.ik} = \sum_i x_{ijk}$, and zero otherwise. Despite its simple structure, this probability is difficult to compute: In order to compute $\#\mathcal{S}_{\mathbf{n}_1,\mathbf{n}_2,\mathbf{n}_3}$ one will have to check for each of the possible arrays $\mathfrak{x}^{(s)}$, $s = 1, \ldots, S$, whether its marginal tables are equal to the observed marginal tables. As mentioned in section A.4 of this appendix, for array and population sizes one usually encounters in typical ecological inference applications, the computational cost will be prohibitive.

References

Abramovitz, Milton and Irene A. Stegun, eds. 1964. *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*. Washington, D.C.: National Bureau of Standards.

Good, I. J. 1963. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics* 34:911–934.

Hald, A. 1968. The mixed binomial distribution and the posterior distribution of p for a continuous prior distribution. *Journal of the Royal Statistical Society Series B (Methodological)* 30:359–367.

Johnston, Ron and Charles Pattie. 2000. Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting. *Political Analysis* 8:333–345.

Kullback, S. and R. A. Leibler. 1951. On information and sufficency. *Annals of Mathematical Statistics* 22:79–86.

Kullback, Solomon. 1959. *Information Theory and Statistics*. New York: Wiley.

Lindley, D. V. 1956. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics* 27:986–1005.

———. 1957. Binomial sampling schemes and the concept of information. *Biometrika* 44:179–186.

Mosimann, James E. 1962. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* 49:65–82.

R Development Core Team. 2007. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`. ISBN 3-900051-07-0.

Vasicek, Oldrich Alfonso. 1980. A conditional law of large numbers. *Annals of Probability* 8:142–147.