Article

A First Approach to Closeness Distributions

Jesus Cerquides [ID]

Instituto de Investigación en Inteligencia Artificial (IIIA-CSIC), Campus UAB, 08193 Cerdanyola, Spain; j.cerquides@csic.es; Tel.: +34-935809570 (ext. 228)

Abstract: Probabilistic graphical models allow us to encode a large probability distribution as a composition of smaller ones. It is oftentimes the case that we are interested in incorporating in the model the idea that some of these smaller distributions are likely to be similar to one another. In this paper we provide an information geometric approach on how to incorporate this information and see that it allows us to reinterpret some already existing models. Our proposal relies on providing a formal definition of what it means to be close. We provide an example on how this definition can be actioned for multinomial distributions. We use the results on multinomial distributions to reinterpret two already existing hierarchical models in terms of closeness distributions.

1. Introduction

Bayesian modeling [1] builds on our ability to describe a given process in probabilistic terms, known as probabilistic modeling. As stated in [2]: "Statistical methods and models commonly involve multiple parameters that can be regarded as related or connected in such a way that the problem implies dependence of the joint probability model for these parameters". Hierarchical modeling [3] is widely used for that purpose in areas such as epidemiological modeling [4] or to model oil or gas production [5]. The motivation for this paper comes from realizing that many hierarchical models can be understood, from a high level perspective, as defining a distribution over the *multiple parameters* that establishes that distributions which are closer to each other, are more likely. Thus, the main motivation is to start providing the mathematical tools that allow a probabilistic modeler to build hierarchical (and non-hierarchical) models starting from that geometrical concepts.

We start by introducing a simple example to illustrate the kind of problems we are interested in solving. Consider the problem of estimating a parameter θ using data from a small experiment and a prior distribution constructed from similar previous experiments. The specific problem description is borrowed from [2]:

> *In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents. For a particular study drawn from the statistical literature, suppose the immediate aim is to estimate θ, the probability of a tumor in a population of female laboratory rats of type 'F344' that receive a zero dose of the drug (a control group). The data show that 4 out of 14 rats developed endometrial stromal polyps (a kind of tumor). (...) Typically, the mean and standard deviation of underlying tumor risks are not available. Rather, historical data are available on previous experiments on similar groups of rats. In the rat tumor example, the historical data were in fact a set of observations of tumor incidence in 70 groups of rats (Table 1). In the ith historical experiment, let the number of rats with tumors be y_i and the total number of rats be n_i. We model the y_i's as independent binomial data, given sample sizes n_i and study-specific means θ_i.*

Example. Estimating the risk of tumor in a group of rats.

Table 1. Tumor incidence in 70 historical groups of rats and in the current group of rats (from [6]). The table displays the values of: (number of rats with tumors)/(total number of rats).

Previous experiments:									
0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24
Current experiment: 4/14									

We can depict our graphical model (for more information on the interpretation of the graphical models in this paper the reader can consult [7,8]) as shown in Figure 1, where current and historical experiments are a random sample from a common population, having **h** as hyperparameters, which follow f as prior distribution. Equationally, our model can be described as:

$$\mathbf{h} \sim f \tag{1}$$

$$\theta_i \sim g(\mathbf{h}) \qquad \forall i \in [1:71] \tag{2}$$

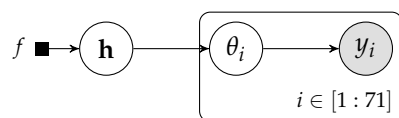$$y_i \sim Binomial(n_i, \theta_i) \qquad \forall i \in [1:71]. \tag{3}$$

Figure 1. General probabilistic graphical model for the rodents example.

The model used for this problem in [2] is the Beta-Binomial model, where g is taken to be the Beta distribution, hence $\mathbf{h} = (\alpha, \beta)$ (see Figure 2). Furthermore, in [2] the prior f over α, β is taken to be proportional to $(\alpha + \beta)^{-5/2}$, giving the model

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2} \tag{4}$$

$$\theta_i \sim Beta(\alpha, \beta) \qquad \forall i \in [1:71] \tag{5}$$

$$y_i \sim Binomial(n_i, \theta_i) \qquad \forall i \in [1:71]. \tag{6}$$
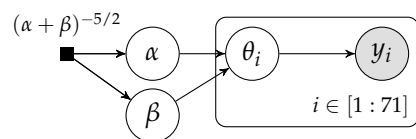

Figure 2. PGM for the rodents example proposed in [2].

The presentation of the model in [2] simply introduces the assumption that "the Beta prior distribution with parameters (α, β) is a good description of the population distribution of the θ_i's in the historical experiments" without further justification. In this paper we would like to show that a large part of this model can be obtained from the intuitive idea that the probability distributions for rats with tumors in each group are similar. To do that we develop a framework for encoding as a probability distribution the assumption that two probability distributions are close to each other, and rely on information geometric concepts to model the idea of closeness.

We start by introducing the general concept of closeness distribution in Section 2. Then, we analyze the particular case in which we choose to measure remoteness between distributions by means of the Kullback Leibler divergence in the family of multinomial distributions in Section 3. The results from Section 3 are used in Section 4 to reinterpret

the Beta Binomial model proposed in [2] for the rodents example, and in Section 5 to reinterpret the Hierarchical Dirichlet Multinomial model proposed by Azzimonti et al. in [9–11]. We are convinced that closeness distributions could play a relevant role in probabilistic modeling, allowing for more explicitly geometrically inspired probabilistic models. This paper is just a first step towards a proper definition and understanding of closeness distributions.

2. Closeness Distributions

We start by introducing the formal framework required to discuss the probability distributions. Then, we formalize what we mean by remoteness through a remoteness function, and we introduce closeness distributions as those that implement a remoteness function.

2.1. Probabilities over Probabilities

Information geometry [12] has shown us that most families of probability distributions can be understood as a Riemannian manifold. Thus, we can work with probabilities over probabilities by defining random variables which take values in a Riemannian manifold. Here, we only introduce some fundamental definitions. For a more detailed overview of measures and probability see [13], of Riemannian manifolds see [14]. Finally, Pennec provides a good overview of the probability on Riemannian manifolds in [15].

We start by noting that each manifold $\mathcal{M}$, has an associated σ-algebra, $\mathcal{L}_{\mathcal{M}}$, the Lebesgue σ-algebra of $\mathcal{M}$ (see Section 1 , chapter XII in [16]). Furthermore, the existence of a metric g induces a measure η_g (see Section 1.3 in [15]). The volume of $\mathcal{M}$ is defined as $Vol(\mathcal{M}) = \int_{\mathcal{M}} 1 d\eta_g$.

Definition 1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathcal{M}, g)$ be a Riemannian manifold. A random variable (referred to as a random primitive in [15]) $\mathbf{x}$ taking values in $\mathcal{M}$ is a measurable function from Ω to $\mathcal{M}$. Furthermore, we say that $\mathbf{x}$ has a probability density function (p.d.f.) $p_{\mathbf{x}}$ (real, positive, and integrable function) if:*

$$\forall \mathcal{X} \in \mathcal{L}_{\mathcal{M}} \quad P(\mathbf{x} \in \mathcal{X}) = \int_{\mathcal{X}} p_{\mathbf{x}} d\eta_g, \qquad \text{and} \qquad P(\mathcal{M}) = 1.$$

We would like to highlight that the density function $p_{\mathbf{x}}$ is intrinsic to the manifold. If $x' = \pi(x)$ is a chart of the manifold defined almost everywhere, we obtain a random vector $\mathbf{x}' = \pi(\mathbf{x})$. The expression of $p_{\mathbf{x}}$ in this parametrization is

$$p_{\mathbf{x}'}(x') = p_{\mathbf{x}}(\pi^{-1}(x')).$$

Let $f : \mathcal{M} \to \mathbb{R}$ be a real function on $\mathcal{M}$. We define the expectation of f under $\mathbf{x}$ as

$$\mathbb{E}[f(\mathbf{x})] = \int f(x) p_{\mathbf{x}}(x) \eta_g(dx)$$

We have to be careful when computing $\mathbb{E}[f(\mathbf{x})]$ so that we do it independently of the parametrization. We have to use the fact that

$$\int f(x) p_{\mathbf{x}}(x) \eta_g(dx) = \int f(\pi^{-1}(x')) p_{\mathbf{x}'}(x') \sqrt{|G(x')|} dx',$$

where $G(x')$ is the Fisher matrix at x' in the parametrization π. Hence,

$$\mathbb{E}[f(\mathbf{x})] = \int f(\pi^{-1}(x')) \rho_{\mathbf{x}'}(x') dx'.$$

where $\rho_{\mathbf{x}'}(x') = p_{\mathbf{x}'}(x') \sqrt{|G(x')|} = p_{\mathbf{x}}(\pi^{-1}(x')) \sqrt{|G(x')|}$ is the expression of $p_{\mathbf{x}}$ in the parametrization for integration purposes, that is, its expression with respect to the Lebesgue measure dx' instead of $d\eta_g$.

We note that $\rho_{\mathbf{x}'}$ depends on the chart used whereas $p_{\mathbf{x}}$ is intrinsic to the manifold.

2.2. Formalizing Remoteness and Closeness

Intuitively, the objective of this section is to create a probability distribution over pairs of probability distributions that assigns higher probability to those pairs of probability distributions which are "close".

We assume that we measure how distant two points are in $\mathcal{M}$ by means of a *remoteness function* $r : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$, such that $r(x, y) \geq 0$ for each $x, y \in \mathcal{M}$. Note that r does not need to be transitive, symmetric, or reflexive.

As can be seen in Appendix A, r induces a total order $\leq_r$ in $\mathcal{M} \times \mathcal{M}$. We say that two remoteness functions r, s are *order-equivalent* if $\leq_r = \leq_s$.

Proposition 1. *Let $\gamma, \beta \in \mathbb{R}, \gamma, \beta > 0$. Then, $\gamma \cdot r + \beta$ is order-equivalent to r.*

Proof. $a \leq_r b$ iff $r(a) \leq r(b)$ iff $\gamma \cdot r(a) \leq \gamma \cdot r(b)$ iff $\gamma \cdot r(a) + \beta \leq \gamma \cdot r(b) + \beta$ iff $a \leq_{\gamma \cdot r + \beta} b$. $\square$

We say that a probability density function $p : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ *implements* a remoteness function r if $\geq_p = \leq_r$. This is equivalent to stating that for each $x, y, z, t \in \mathcal{M}$ we have that $p(x, y) \geq p(z, t)$ iff $r(x, y) \leq r(z, t)$. That is, a density function implements a remoteness function r if it assigns higher probability density to those pairs of points which are closer according to r.

Once we have clarified what it means for a probability to implement a remoteness function, we introduce a specific way of creating probabilities to that.

Definition 2. *Let $f_r : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ be $f_r(x, y) = \exp(-r(x, y))$. If $Z_r = \int_{\mathcal{M}} \int_{\mathcal{M}} f_r d\eta_g d\eta_g$ is finite, we define the density function*

$$p_r(x, y) = \frac{f_r(x, y)}{Z_r} = \frac{\exp(-r(x, y))}{Z_r}. \tag{7}$$

We refer to the corresponding probability distribution as a closeness distribution.

Note that p_r is defined intrinsically. Following the explanation in the previous section, let π be a chart of $\mathcal{M}$ defined almost everywhere. The representation of this pdf in the parametrization $x', y' = (\pi(x), \pi(y))$ is simply

$$p_{r'}(x', y') = p_r(\pi^{-1}(x'), \pi^{-1}(y')) \tag{8}$$

and its representation for integration purposes is

$$\rho_{r'}(x', y') = p_r(\pi^{-1}(x'), \pi^{-1}(y')) \sqrt{|G(x')|} \sqrt{|G(y')|} \tag{9}$$

Proposition 2. *It it exists, p_r implements r.*

Proof. The exponential is a monotonous function and the minus sign in the exponent is used to revert the order. $\square$

Proposition 3. *If r is measurable and $\mathcal{M}$ has finite volume, then Z_r is finite, and hence p_r implements r.*

Proof. Note that since $r(x, y) \geq 0$, we have that $f_r(x, y) \leq 1$, and hence f_r is bounded. Furthermore, f_r is measurable since it is a composition of measurable functions. Now, since any bounded measurable function in a finite volume space is integrable, Z_r is finite. $\square$

Obviously, once we have established a closeness distribution p_r we can define its marginal and conditional distributions in the usual way. We note $p_r(x)$ (resp. $p_r(y)$) as the marginal over x (resp. y). We note $p_r(x|y)$ (resp. $p_r(y|x)$) as the conditional density of x given y (resp. y given x).

3. KL-Closeness Distributions for Multinomials

In this section we study closeness distributions on M_n (the family of multinomial distributions of dimension n, or the family of finite discrete distributions over $n + 1$ atoms). To do that, first we need to establish the remoteness function. It is well known that there is an isometry between M_n and the positive orthant of the n dimensional sphere (S_n) of radius 2 (see Section 7.4.2. in [17]). This isometry allows us to compute the volume of the manifold as the area of the sphere of radius 2 on the positive orthant.

Proposition 4. *The volume of M_n is* $Vol(M_n) = \frac{\pi^{\frac{n+1}{2}}}{\Gamma(\frac{n+1}{2})}$.

Proof. The area of a sphere S_n of radius r is $A_{n,r} = \frac{2\pi^{\frac{n+1}{2}} r^n}{\Gamma(\frac{n+1}{2})}$. Taking $r = 2$, $A_{n,2} = \frac{\pi^{\frac{n+1}{2}} 2^{n+1}}{\Gamma(\frac{n+1}{2})}$. Now, there are 2^{n+1} orthants, so the positive orthant amounts for $\frac{1}{2^{n+1}}$ of that area, as stated. $\square$

Figure 3 shows that the volume of the space of multinomial distributions over $n + 1$ atoms reaches its maximum at $n = 7$. The main takeover of Proposition 4 is that the volume of M_n is finite, because this allows us to prove the following result:

Proposition 5. *For any measurable remoteness function r on M_n there is a closeness distribution p_r implementing it.*

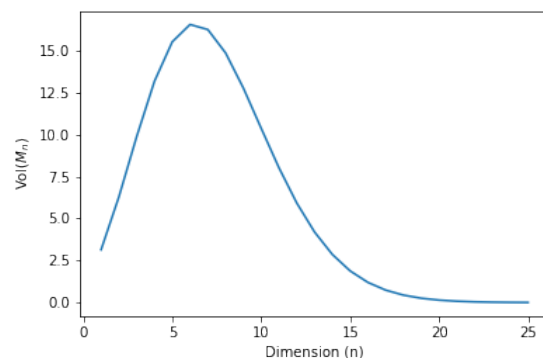**Proof.** Directly from Proposition 3 and the fact that M_n has finite volume. $\square$

Figure 3. Volume of the family of multinomial distributions as dimension increases

A reasonable choice of remoteness function for a statistical manifold is the Kullback-Leibler (KL) divergence. The next section analyzes the closeness distributions that implement KL in $\mathcal{M}_n$.

3.1. Closeness Distributions for KL as Remoteness Function

Let $\theta \in \mathcal{M}_n$. Thus, θ is a discrete distribution over $n + 1$ atoms. We write θ_i to represent $p(x = i|\theta)$. Note that each θ_i is independent of the parametrization an thus it is an intrinsic quantity of the distribution.

Let $\theta, \mu \in \mathcal{M}_n$. The KL divergence between θ and μ is

$$D(\mu, \theta) = \sum_{i=1}^{n+1} \mu_i \log \frac{\mu_i}{\theta_i}.$$

We want to study the closeness distributions that implement KL in $\mathcal{M}_n$. The detailed derivation of these results can be found in Appendix B. The closeness pdf according to Equation (7) is

$$p_D(\mu, \theta) = \frac{1}{Z_D} \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i}$$

The marginal for μ is

$$p_D(\mu) = \frac{1}{Z_D} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i} B(\mu + \frac{1}{2})$$

where $B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$ is the multivariate Beta function.

The conditional for θ given μ:

$$p_D(\theta \mid \mu) = \frac{\prod_{i=1}^{n+1} \theta_i{}^{\mu_i}}{B(\mu + \frac{1}{2})} \tag{10}$$

Equation (10) is very similar to the expression of a Dirichlet distribution. In fact, the expression of $p_D(\theta \mid \mu)$ for integration purposes in the expectation parameterization, namely $\rho_D(\theta \mid \mu)$, is that of a Dirichlet distribution:

$$\rho_D(\theta \mid \mu) = Dirichlet(\theta; \mu + \frac{1}{2}) \tag{11}$$

Equation (11) deserves some attention. We have defined the joint density $p_D(\mu, \theta)$ so that pairs of distributions (μ, θ) that are close in terms of KL divergence are assigned a higher probability than pairs of distributions (μ^*, θ^*) which are further away in terms of KL. Hence, the conditional $p_D(\theta \mid \mu)$ assigns a larger probability to those distributions θ which are close in terms of KL to μ. This means that whenever we have a probabilistic model which encodes two multinomial distributions θ and μ, and we are interested in introducing that θ should be close to μ, we can introduce the assumption that $\theta \sim Dirichlet(\mu + \frac{1}{2})$.

Interesting as it is for modeling purposes, the use of Equation (11) however does not allow the modeler to convey information regarding the strength of the link. That is, θ's in the KL-surrounding of μ will be more probable, but there is no way to establish how much more probable. We know by Proposition 1 that for any remoteness function r, we can select $\gamma, \beta > 0$, and $\gamma \cdot r + \beta$ is order-equivalent to r. We can take advantage of that fact and use γ to encode the strength of the probabilistic link between θ and μ. If instead of using the KL (D) as the remoteness function, we opt for $\gamma \cdot D$, following a parallel development to the one above we will find that

$$\rho_{\gamma \cdot D}(\theta \mid \mu) = Dirichlet(\theta; \gamma\mu + \frac{1}{2}). \tag{12}$$

Now, Equation (12) allows the modeler to fix a large value of γ to encode that it is extremely unlikely that θ separates from μ, or a value of γ close to 0 to encode that the link between θ and μ is highly loose. Furthermore it is important to realize that Equation (12) allows us to interpret any already existing model which incorporates Dirichlet (or Beta) distributions with the only requirement that each of its concentration parameters is larger than $\frac{1}{2}$. Say we have a model in which $\theta \sim Dirichlet(\alpha)$. Then, defining μ by coordinates as $\mu_i = \frac{\alpha_i - \frac{1}{2}}{-\frac{n+1}{2} + \sum_{i=1}^{n+1} \alpha_i}$, we can interpret the model as imposing θ to be close to μ with intensity $\gamma = \frac{\alpha_1 - \frac{1}{2}}{\mu_1}$. Note that, extending this interpretation a bit to the extreme, since the strength of the link reduces as $\gamma \to 0$, a "free" Dirichlet will have all of its weights set to $\frac{1}{2}$. This coincides with the classical prior suggested by Jeffreys [18,19] for this very same problem. This is reasonable since Jeffreys' prior was constructed to be independent of the parametrization, that is, to be intrinsic to the manifold, similarly to what we are doing.

3.2. Visualizing the Distributions

In the previous section we have seen an expression for $p_{\gamma \cdot D}(\theta \mid \mu)$. Since the KL divergence is not symmetric, we have that $p_{\gamma \cdot D}(\mu \mid \theta)$ is different from $p_{\gamma \cdot D}(\theta \mid \mu)$. Unfortunately, we have not been able to provide a closed form expression for $p_{\gamma \cdot D}(\mu \mid \theta)$. However, it is possible to compute it numerically in order to compare both conditionals.

Figure 4 shows a comparison of $p_{\gamma \cdot D}(\mu \mid \theta)$ and $p_{\gamma \cdot D}(\theta \mid \mu)$. According to what is suggested in [20], for a proper interpretation of the densities we show its density function, which is intrinsic to the manifold, instead of its expression in the parametrization, as is commonly done. Note that from Equation (12), the value of $p_{\gamma \cdot D}(\theta \mid \mu)$ is 0 at $\theta = 0$ and $\theta = 1$. In Figure 4, we can see that this is not the case for $p_{\gamma \cdot D}(\mu \mid \theta)$ neither at $\mu = 0$ nor at $\mu = 1$. In fact we see that $p_{\gamma \cdot D}(\theta \mid \mu)$ always starts below $p_{\gamma \cdot D}(\mu \mid \theta)$ at $\theta = 0$ (resp. $\mu = 0$). Then, as θ (resp. μ) grows, it is always the case that $p_{\gamma \cdot D}(\theta \mid \mu)$ goes over $p_{\gamma \cdot D}(\mu \mid \theta)$, to end decreasing again below it when θ (resp μ) approaches to 1.
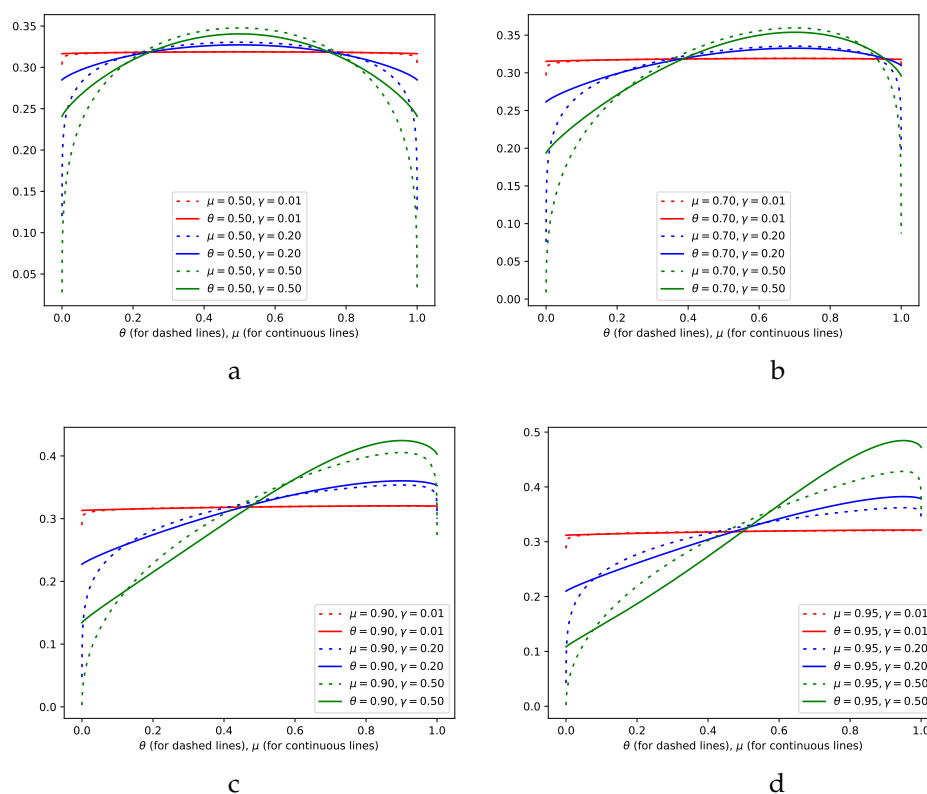
Figure 4. Comparison of $p_{\gamma \cdot D}(\theta \mid \mu)$ and $p_{\gamma \cdot D}(\mu \mid \theta)$. In (**a**) $p_{\gamma \cdot D}(\theta \mid \mu = 0.5)$ and $p_{\gamma \cdot D}(\mu \mid \theta = 0.5)$. In (**b**) $p_{\gamma \cdot D}(\theta \mid \mu = 0.7)$ and $p_{\gamma \cdot D}(\mu \mid \theta = 0.7)$. In (**c**) $p_{\gamma \cdot D}(\theta \mid \mu = 0.9)$ and $p_{\gamma \cdot D}(\mu \mid \theta = 0.9)$. In (**d**) $p_{\gamma \cdot D}(\theta \mid \mu = 0.95)$ and $p_{\gamma \cdot D}(\mu \mid \theta = 0.95)$.

4. Reinterpreting the Beta-Binomial Model

We are now ready to go back to the rodents example provided in the introduction. The main idea we would like this hierarchical model to capture is that the θ_i's are somewhat similar. We do this by introducing a new random variable μ to which we would like each θ_i to be close to (see Figure 5). Furthermore, we introduce another variable γ that controls how tightly coupled the θ_i's are to μ. Now, μ represents a proportion, and priors for proportions have been well studied, including the "Bayes-Laplace rule" [21] which recommends $Beta(1,1)$, the Haldane prior [22] which is an improper prior $\lim_{\alpha \to 0^+} Beta(\alpha, \alpha)$, and the Jeffreys' prior [18,19] $Beta(\frac{1}{2}, \frac{1}{2})$. Following the arguments in the previous section, here we stick with the Jeffreys' prior. A more difficult problem is the selection of the prior for γ, where we still do not have a well founded choice. Note that taking a look at Equation (12), γ's role acts similarly (although not exactly equal) to an equivalent sample size. Thus, the

prior over γ could be thought as a prior over the equivalent sample size with which μ will be incorporated as prior into the determination of each of the θ_i's. In case the size of each sample (n_i) is large, there will not be much difference between a hierarchical model and modeling of each of the 71 experiments as independent experiments. So, it makes sense for the prior over γ to concentrate on a relatively small equivalent sample sizes. Following this line of thought we propose γ to follow a *Gamma*($\alpha = 1, \beta = 0.1$).
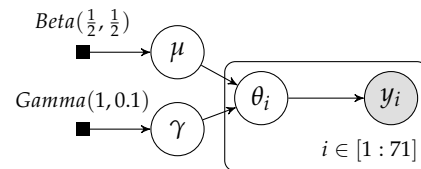

Figure 5. Reinterpreted hierarchical graphical model for the rodents example.

To summarize, the hierarchical model we obtain based on closeness probability distributions is:

$$\mu \sim Beta(\frac{1}{2}, \frac{1}{2}) \tag{13}$$

$$\gamma \sim Gamma(1, 0.1) \tag{14}$$

$$\theta_i \sim Beta(\gamma\mu + \frac{1}{2}, \gamma(1 - \mu) + \frac{1}{2}) \qquad \forall i \in [1:71] \tag{15}$$

$$y_i \sim Binomial(n_i, \theta_i) \qquad \forall i \in [1:71]. \tag{16}$$

Figure 6 shows that the posteriors generated by both models are similar, and put the parameter μ (the pooled average) between 0.08 and 0.15, and the parameter γ (the intensity of the link between μ and each of the θ_i's) between 5 and 25. Furthermore, the model is relatively insensitive to the parameters of the prior for γ as long as they do create a sparse prior. Thus, we see that selecting the prior as $\Gamma(1, 0.5)$ creates a prior that is too concentrated on the low values of γ (that is, it imposes a relatively mild closeness link between μ and each of the θ_i's). This changes the estimation a lot. However, $\Gamma(1, 0.01)$ creates a posterior similar to that of $\Gamma(1, 0.1)$, despite being more spread.
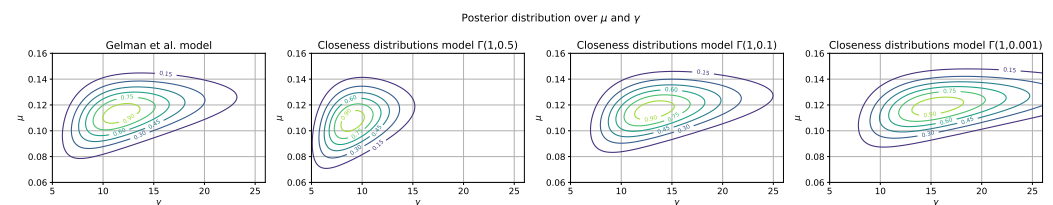

Figure 6. Comparison of posteriors between a closeness distribution model and that proposed by Gelman et al. in [2].

5. Hierarchical Dirichlet Multinomial Model

Recently [9–11], Azzimonti et al. have proposed a hierarchical Dirichlet multinomial model to estimate conditional probability tables (CPTs) in Bayesian networks. Given two discrete finite random variables X (over domain $\mathcal{X}$) and Y over domain ($\mathcal{Y}$) which are part of a Bayesian network, and such that Y is the only parent of X in the network, the CPT for X is responsible of storing $p(X|Y)$. The usual CPT model (the so called Multinomial-Dirichlet) adheres to *parameter independence* and stores $|\mathcal{Y}|$ different independent Dirichlet distributions over each of the $\theta_{X|y}$. Instead, Azzimonti et al. propose the hierarchical Multinomial-Dirichlet model, where "the parameters of different conditional distributions belonging to the same CPT are drawn from a common higher-level distribution". Their model can be summarized equationally as

$$\alpha \sim s \cdot Dirichlet(\alpha_0)$$
$$\theta_{X|y} \sim Dirichlet(\alpha) \qquad\qquad \forall y \in \mathcal{Y}$$
$$X_y \sim Categorical(\theta_{X|y}) \qquad\qquad \forall y \in \mathcal{Y}$$

and graphically as shown in Figure 7.

The fact that the Dirichlet distribution is the conditional of a closeness distribution allows us to think about this model as a generalization of the model presented for the rat example. Thus, the Hierarchical Dirichlet Multinomial model can be understood as introducing the assumption that there is a probability distribution with parameter μ, that is close in terms of its KL divergence to each of the $y \in \mathcal{Y}$ different distributions, each of them parameterized by $\theta_{X|y}$. Thus, in equational terms, we have that the model can be rewritten as

$$\mu \sim Dirichlet(\frac{1}{2}, \ldots, \frac{1}{2}) \tag{17}$$

$$\gamma \sim Gamma(1, 0.1) \tag{18}$$

$$\theta_{X|y} \sim Dirichlet(\gamma\mu + \frac{1}{2}) \qquad\qquad \forall y \in \mathcal{Y} \tag{19}$$

$$X_y \sim Categorical(\theta_{X|y}) \qquad\qquad \forall y \in \mathcal{Y} \tag{20}$$

and depicted as shown in Figure 8. Note that γ in our reinterpreted model plays a role quite similar to the one that s played on Azzimonti's model. To maintain the parallel with the model developed for the rodents example, here we have also assumed a $Gamma(1, 0.1)$ as prior over γ, instead of the punctual distribution assumed in [10], but we could easily mimic their approach and specify a single value for γ.
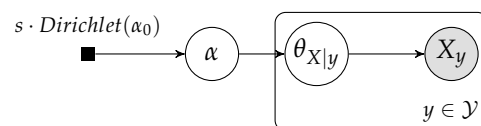
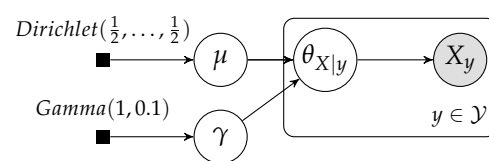**Figure 7.** PGM for the hierarchical Dirichlet Multinomial model proposed in [10].

Figure 8. Reinterpreted PGM for the hierarchical Dirichlet Multinomial model.

Note that we are not claiming that we are improving the Hierarchical Dirichlet Multinomial model, we are just reinterpreting it in a way that is easier to understand conceptually.

6. Conclusions and Future Work

We have introduced the idea and the formalization remoteness functions and closeness distributions in Section 2. We have proven that any remoteness function induces a closeness distribution, provided that the volume of the space of distributions is finite. We have particularized closeness distributions for multinomials in Section 3, taking the remoteness function as the KL-divergence. By analyzing two examples, we have shown that closeness distributions can be a useful tool to the probabilistic model builder. We have seen that they can provide additional rationale and geometric intuitions for some commonly used hierarchical models.

In the Crowd4SDG and Humane-AI-net projects, these mathematical tools could prove useful in the understanding and development of consensus models for citizen science, improving the ones presented in [23]. Our plan is to study this in future work.

In this paper we have concentrated on discrete closeness distributions. The study of continuous closeness distributions remains for future work.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Total Order Induced by a Function

Definition A1. *Let Z be a set and $f : Z \to \mathbb{R}$ a function. The binary relation $\leq_f$ (a subset of $Z \times Z$) is defined as*

$$a \leq_f b \text{ iff } f(a) \leq f(b) \tag{A1}$$

Proposition A1. $\leq_f$ *is a total (or lineal) order in Z.*

Proof. Reflexivity, transitivity, antisimmetry, and totality are inherited from the fact that $\leq$ is a total order in Z. □

Appendix B. Detailed Derivation of the KL Based Closeness Distributions for Multinomials

Closeness Distributions for KL as Remoteness Function

Let $\theta \in \mathcal{M}_n$. Thus, θ is a discrete distribution over $n + 1$ atoms. We write θ_i to represent $p(x = i|\theta)$. Note that each θ_i is independent of the parametrization an thus it is an intrinsic quantity of the distribution.

Let $\theta, \mu \in \mathcal{M}_n$. The KL divergence between θ and μ is

$$D(\mu, \theta) = \sum_{i=1}^{n+1} \mu_i \log \frac{\mu_i}{\theta_i}.$$

The closeness pdf according to Equation (7) is

$$\begin{aligned} p_D(\mu, \theta) &= \frac{1}{Z_D} \exp(-D(\mu, \theta)) \\ &= \frac{1}{Z_D} \exp\left(-\sum_{i=1}^{n+1} \mu_i \log \frac{\mu_i}{\theta_i}\right) \\ &= \frac{1}{Z_D} \prod_{i=1}^{n+1} \theta_i^{\mu_i} \prod_{i=1}^{n+1} \mu_i^{-\mu_i}. \end{aligned}$$

Now, it is possible to assess the marginal for μ

$$
\begin{aligned}
p_D(\mu) &= \int_\theta p_D(\mu, \theta) \eta_g(d\theta) \\
&= \int_\theta \frac{1}{Z_D} \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i} \eta_g(d\theta) \\
&= \frac{1}{Z_D} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i} \int_\theta \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \eta_g(d\theta)
\end{aligned}
\tag{A2}
$$

where we recall that η_g is the measure induced by the Fisher metric and it is not connected to μ. To continue, we need to compute $\int_\theta \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \eta_g(d\theta)$ as an intrinsic quantity of the manifold, that is, invariant to changes in parametrization. We are integrating $f(\theta) = \prod_{i=1}^{n+1} \theta_i{}^{\mu_i}$. We can parameterize the manifold using θ itself (the expectation parameters). In this parameterization, the integral can be written as

$$
\begin{aligned}
\int_\theta \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \eta_g(d\theta) &= \int_\theta \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \sqrt{|\, G(\theta)\, |} \, d\theta \\
&= \int_\theta \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \prod_{i=1}^{n+1} \theta_i{}^{-\frac{1}{2}} d\theta \\
&= \int_\theta \prod_{i=1}^{n+1} \theta_i{}^{\mu_i - \frac{1}{2}} d\theta \\
&= B\!\left(\mu + \frac{1}{2}\right),
\end{aligned}
\tag{A3}
$$

where the last equality comes from identifying it as a Dirichlet integral of type 1 (see 15-08 in [24]), and $B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$ is the multivariate Beta function. Combining Equation (A2) with Equation (A3) we get

$$
p_D(\mu) = \frac{1}{Z_D} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i} B\!\left(\mu + \frac{1}{2}\right).
\tag{A4}
$$

From here, we can compute the conditional for θ given μ:

$$
\begin{aligned}
p_D(\theta \mid \mu) &= \frac{p_D(\mu, \theta)}{p_D(\mu)} \\
&= \frac{\frac{1}{Z_D} \prod_{i=1}^{n+1} \theta_i{}^{\mu_i} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i}}{\frac{1}{Z_D} \prod_{i=1}^{n+1} \mu_i{}^{-\mu_i} B\!\left(\mu + \frac{1}{2}\right)} \\
&= \frac{\prod_{i=1}^{n+1} \theta_i{}^{\mu_i}}{B\!\left(\mu + \frac{1}{2}\right)}.
\end{aligned}
\tag{A5}
$$

Equation (A5) is very similar to the expression of a Dirichlet distribution. In fact, the expression of $\rho_D(\theta \mid \mu)$ in the expectation parameterization is that of a Dirichlet distribution:

$$
\begin{aligned}
\rho_D(\theta \mid \mu) &= p_D(\theta \mid \mu) \sqrt{|\, G(\theta)\, |} \\
&= \frac{\prod_{i=1}^{n+1} \theta_i{}^{\mu_i}}{B\!\left(\mu + \frac{1}{2}\right)} \prod_{i=1}^{n+1} \theta_i{}^{-\frac{1}{2}} \\
&= \frac{\prod_{i=1}^{n+1} \theta_i{}^{\mu_i - \frac{1}{2}}}{B\!\left(\mu + \frac{1}{2}\right)} \\
&= Dirichlet\!\left(\theta; \mu + \frac{1}{2}\right).
\end{aligned}
\tag{A6}
$$

References

1. Van de Schoot, R.; Depaoli, S.; King, R.; Kramer, B.; Märtens, K.; Tadesse, M.G.; Vannucci, M.; Gelman, A.; Veen, D.; Willemsen, J.; et al. Bayesian statistics and modelling. *Nat. Rev. Methods Prim.* **2021**, *1*, 1. [CrossRef]
2. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall/CRC: London, UK, 2013.
3. Allenby, G.M.; Rossi, P.E.; McCulloch, R. Hierarchical Bayes Models: A Practitioners Guide. *SSRN Electron. J.* **2005**. [CrossRef]
4. Lee, S.Y.; Lei, B.; Mallick, B. Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS ONE* **2020**, *15*, e0236860. [CrossRef] [PubMed]
5. Lee, S.Y.; Mallick, B.K. Bayesian Hierarchical Modeling: Application Towards Production Results in the Eagle Ford Shale of South Texas. *Sankhya B* **2021**. [CrossRef]
6. Tarone, R.E. The Use of Historical Control Information in Testing for a Trend in Proportions. *Biometrics* **1982**, *38*, 215–220. [CrossRef]
7. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
8. Obermeyer, F.; Bingham, E.; Jankowiak, M.; Pradhan, N.; Chiu, J.; Rush, A.; Goodman, N. Tensor variable elimination for plated factor graphs. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 4871–4880.
9. Azzimonti, L.; Corani, G.; Zaffalon, M. Hierarchical Multinomial-Dirichlet Model for the Estimation of Conditional Probability Tables. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; pp. 739–744. [CrossRef]
10. Azzimonti, L.; Corani, G.; Zaffalon, M. Hierarchical estimation of parameters in Bayesian networks. *Comput. Stat. Data Anal.* **2019**, *137*, 67–91. [CrossRef]
11. Azzimonti, L.; Corani, G.; Scutari, M. Structure Learning from Related Data Sets with a Hierarchical Bayesian Score. In Proceedings of the International Conference on Probabilistic Graphical Models, PMLR, Aalborg, Denmark, 23–25 September 2020; pp. 5–16. ISSN: 2640-3498.
12. Amari, S.I. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 194.
13. Dudley, R.M. *Real Analysis and Probability*, 2nd ed.; Cambridge Studies in Advanced Mathematics; Cambridge University Press: Cambridge, UK, 2002. [CrossRef]
14. Jost, J. *Riemannian Geometry and Geometric Analysis*; Springer: Berlin/Heidelberg, Germany, 2011. [CrossRef]
15. Pennec, X. *Probabilities and Statistics on Riemannian Manifolds: A Geometric Approach*; Technical Report RR-5093; INRIA: Rocquencourt, France, 2004.
16. Amann, H.; Escher, J. *Analysis III*; Birkhäuser: Basel, Switzerland, 2009. [CrossRef]
17. Kass, R.E.; Vos, P.W. *Geometrical Foundations of Assimptotic Inference*; Wiley-Interscience: Hoboken, NJ, USA, 1997.
18. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1946**, *186*, 453–461. [CrossRef]
19. Jeffreys, H. *The Theory of Probability*; Oxford University Press: Oxford, UK, 1998.
20. Cerquides, J. Parametrization invariant interpretation of priors and posteriors. *arXiv* **2021**, arXiv:2105.08304.
21. Laplace, P.S.m.d. *Essai Philosophique sur les Probabilités*; Courcier: Le Mesnil-Saint-Denis, France, 1814.
22. Haldane, J.B.S. A note on inverse probability. *Math. Proc. Camb. Philos. Soc.* **1932**, *28*, 55–61. [CrossRef]
23. Cerquides, J.; Mülâyim, M.O.; Hernández-González, J.; Ravi Shankar, A.; Fernandez-Marquez, J.L. A Conceptual Probabilistic Framework for Annotation Aggregation of Citizen Science Data. *Mathematics* **2021**, *9*, 875. [CrossRef]
24. Jeffreys, H.; Swirles Jeffreys, B. *Methods of Mathematical Physics*; Cambridge University Press: Cambridge, UK, 1950.