LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Gero Walter

A Technical Note on the Dirichlet-Multinomial Model

Technical Report Number 131, 2012
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A Technical Note on the Dirichlet-Multinomial Model

The Dirichlet Distribution as the Canonically Constructed Conjugate Prior

Gero Walter
Department of Statistics
Ludwig-Maximilians-Universität München
gero.walter@stat.uni-muenchen.de

October 4, 2012

Abstract

This short note contains an explicit proof of the Dirichlet distribution being the conjugate prior to the Multinomial sample distribution as resulting from the general construction method described, e.g., in Bernardo and Smith (2000). The well-known Dirichlet-Multinomial model is thus shown to fit into the framework of canonical conjugate analysis (Bernardo and Smith 2000, Prop. 5.6, p. 273), where the update step for the prior parameters to their posterior counterparts has an especially simple structure. This structure is used, e.g., in the Imprecise Dirichlet Model (IDM) by Walley (1996), a simple yet powerful model for imprecise Bayesian inference using sets of Dirichlet priors to model vague prior knowledge, and furthermore in other imprecise probability models for inference in exponential families where sets of priors are considered.

1 Conjugate Priors, Canonical Construction

Before turning to the proof in the next section, we introduce the relevant concepts, and give some background on the use of canonically constructed conjugate priors in imprecise Bayesian inference.

Conjugate priors are an important tool in Bayesian statistics. A prior is called *conjugate* (to a certain sample distribution, or likelihood, e.g., the

Normal, or the Binomial, distribution) if it leads, by updating via Bayes' Rule, to a posterior that is from the same parametric class as the prior. Thus, the posterior remains easily tractable, making inferences based on it straightforward, and the update step is fully described by the change of parameter values in the conjugate class of parametric distributions.

There are general results regarding the construction of conjugate priors to a given sample distribution for several general classes of distributions. Here, we consider the *canonical exponential family* [1, Def. 4.12, p. 202] class of sample distributions, which covers a wide range of sample distributions relevant to statistical practice.[1]

A sample distribution is said to belong to the *canonical exponential family* if its density or mass function satisfies the decomposition[2]

$$f(x|\theta)dx \propto \exp\big\{\langle\psi, \tau^*(x)\rangle - \mathbf{b}(\psi)\big\}dx\,. \tag{1}$$

Here, $\psi \in \Psi \subset \mathbb{R}^q$, with $q \in \mathbb{N}_{>0}$, is the so-called *canonical* parameter of the distribution, being a transformation of the (possibly vectorial) parameter $\theta \in \Theta$ commonly used. $\mathbf{b}(\psi)$ is a scalar function of ψ (or, in turn, of θ), while $\tau^*(x)$ is a function of a single sample x with $\tau^*(x) \in \mathcal{T} \subset \mathbb{R}^q$.[3]

The *canonical conjugate prior* on ψ can be constructed as [1, p. 272]

$$p(\psi|n^{(0)}, y^{(0)})d\psi \propto \exp\Big\{n^{(0)}\Big[\langle y^{(0)}, \psi\rangle - \mathbf{b}(\psi)\Big]\Big\}d\psi\,, \tag{2}$$

with $n^{(0)}$ and $y^{(0)}$ the parameters by which a certain prior is specified.[4] The domain of $y^{(0)}$ is $\mathcal{Y}$, the interior of the convex hull of $\mathcal{T}$; the scalar $n^{(0)}$ must take strictly positive values for the prior to be proper (integrable to 1).

When updating a prior (2) via Bayes' Rule with an i.i.d. sample of size n, the posterior parameters $y^{(n)}$ and $n^{(n)}$ are calculated as

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)}+n}\cdot y^{(0)} + \frac{n}{n^{(0)}+n}\cdot\frac{\tau(\mathbf{x})}{n}\,, \qquad n^{(n)} = n^{(0)} + n\,. \tag{3}$$

We may thus denote the posterior $p(\psi|\mathbf{x}, n^{(0)}, y^{(0)})$ by $p(\psi|n^{(n)}, y^{(n)})$.

It is the weighted average structure for $y^{(n)}$ in (3) which is the key to a simple calculus for imprecise Bayesian inference based on sets of conjugate

[1] The class contains, e.g., the Normal, Multinomial, Poisson, Exponential, and Gamma sample models.

[2] $\langle\cdot,\cdot\rangle$ denotes the scalar product.

[3] For an i.i.d. sample $\mathbf{x}$ of size n, (1) can be modified by replacing $\tau^*(x)$ with $\tau(\mathbf{x}) = \sum_{i=1}^n \tau^*(x_i)$, and $\mathbf{b}(\psi)$ with $n\mathbf{b}(\psi)$, respectively.

[4] In our notation, $^{(0)}$ denotes prior parameters; $^{(n)}$ denotes posterior parameters.

priors. In imprecise Bayesian inference, sets of priors can be considered in order to model partial or very weak prior knowledge. A convex set of priors, often called *prior credal set*, is updated element by element to obtain the *posterior credal set*. This procedure can be rigorously justified in the theoretical framework developed by [4], where it is equivalent to applying the *Generalized Bayes' Rule* [4, §6.4] in a statistical context. When a set of priors is defined as (the convex hull of) parametric priors $p(\psi|n^{(0)}, y^{(0)})$ where $(n^{(0)}, y^{(0)})$ varies in a set $\Pi\!\Pi^{(0)}$, the corresponding set of posteriors is given by (the convex hull of) parametric posteriors $p(\psi|n^{(n)}, y^{(n)})$, whose parameters $(n^{(n)}, y^{(n)})$ are obtained by (3). For fixed $n^{(0)}$, $y^{(n)}$ is linear in $y^{(0)}$, and thus extreme points of $\Pi\!\Pi^{(0)} = n^{(0)} \times \mathcal{Y}^{(0)}$ are updated to the extreme points of the posterior parameter set $\Pi\!\Pi^{(n)}$, making also an imprecise inference calculus tractable.

The IDM [5] is based on this very calculus in case of the Dirichlet-Multinomial model, and [2] have first described it for the canonical exponential family setting in general as presented above. It was further generalized by the definition of so-called LUCK-models in [7, 6] to apply also to other settings.

2 The Proof

In the remainder, we will construct the conjugate prior to the Multinomial sampling model according to (2), and subsequently show that the resulting prior is indeed a Dirichlet. The result shown here can, without proof, be found in [2, Table 1], which tabulates priors constructed for a number of sample models.[5]

For the construction, we will represent the Multinomial distribution as a multivariate Bernoulli. A multivariate Bernoulli is equivalent to a Multinomial distribution with sample size 1, and thus i.i.d. repetitions of a multivariate Bernoulli distribution lead to the Multinomial distribution.

The density of the multivariate Bernoulli with categories $j = 0, 1, \dots, k$, where for the observation vector $\mathbf{x}$ with elements x_j holds that $\mathbf{x} \in \{0,1\}^k \cap \big\{\mathbf{x} : \sum_{j=1}^k x_j \in \{0,1\}\big\}$, and for the parameter vector $\boldsymbol{\theta}$ with elements θ_j holds $\boldsymbol{\theta} \in (0,1)^k \cap \{\boldsymbol{\theta} : 0 < \sum_{j=1}^k \theta_j < 1\}$ (so $\theta_0 := 1 - \sum_{j=1}^k \theta_j$), is as

[5] However, for the Multinomial sampling model, the table contains a sign error for $\mathbf{b}(\boldsymbol{\psi})$.

follows:

$$\begin{aligned}p(\mathbf{x}\,|\,\boldsymbol{\theta}) &= \left(\prod_{j=1}^{k}\theta_j^{x_j}\right)\left(1-\sum_{j=1}^{k}\theta_j\right)^{1-\sum_{j=1}^{k}x_j} = \theta_0\prod_{j=1}^{k}\left(\frac{\theta_j}{\theta_0}\right)^{x_j}\\ &= \exp\left\{\sum_{j=1}^{k}x_j\ln\left(\frac{\theta_j}{\theta_0}\right)-\big(-\ln(\theta_0)\big)\right\}.\end{aligned}$$

With $\boldsymbol{\psi}$ and $\mathbf{b}(\boldsymbol{\psi})$ derived from the sample model as

$$\psi_j = \ln\left(\frac{\theta_j}{\theta_0}\right),\ j = 1,\ldots,k \qquad \text{and} \qquad \mathbf{b}(\boldsymbol{\psi}) = -\ln(\theta_0),$$

the conjugate prior is at first constructed as a density over $\boldsymbol{\psi}$ and then transformed to a density over $\boldsymbol{\theta}$:

$$p(\boldsymbol{\psi}\,|\,n,\mathbf{y})\,d\boldsymbol{\psi} \propto \exp\left\{n\left[\sum_{j=1}^{k}y_j\ln\left(\frac{\theta_j}{\theta_0}\right)-\big(-\ln(\theta_0)\big)\right]\right\}d\boldsymbol{\psi}$$

Written as a density over $\boldsymbol{\theta}$, we have

$$p(\boldsymbol{\theta}\,|\,n,\mathbf{y})\,d\boldsymbol{\theta} \propto \exp\left\{n\left[\sum_{j=1}^{k}y_j\ln\left(\frac{\theta_j}{\theta_0}\right)-\big(-\ln(\theta_0)\big)\right]\right\}\cdot\left|\det\left(\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}\right)\right|\,d\boldsymbol{\theta}\,,$$

with the the elements of the Jacobian matrix $\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}$ being

$$\begin{aligned}\frac{d\psi_i}{d\theta_i} &= \frac{1}{d\theta_i}\ln\left(\frac{\theta_i}{1-\sum_{j=1}^{k}\theta_j}\right) = \frac{1-\sum_{j=1}^{k}\theta_j}{\theta_i}\cdot\frac{1-\sum_{j=1}^{k}\theta_j+\theta_i}{(1-\sum_{j=1}^{k}\theta_j)^2} = \frac{\theta_0+\theta_i}{\theta_0\theta_i}\\ \frac{d\psi_h}{d\theta_i} &= \frac{1}{d\theta_i}\ln\left(\frac{\theta_h}{1-\sum_{j=1}^{k}\theta_j}\right) = \frac{1-\sum_{j=1}^{k}\theta_j}{\theta_h}\cdot\frac{\theta_h}{(1-\sum_{j=1}^{k}\theta_j)^2} = \frac{1}{\theta_0}\,, \quad i \neq h\end{aligned}$$

Thus,

$$\det\left(\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}\right) = \det\begin{pmatrix}\frac{\theta_0+\theta_1}{\theta_0\theta_1} & \frac{1}{\theta_0} & \cdots & \frac{1}{\theta_0}\\ \frac{1}{\theta_0} & \frac{\theta_0+\theta_2}{\theta_0\theta_2} & \ddots & \vdots\\ \vdots & \ddots & \ddots & \frac{1}{\theta_0}\\ \frac{1}{\theta_0} & \cdots & \frac{1}{\theta_0} & \frac{\theta_0+\theta_k}{\theta_0\theta_k}\end{pmatrix}$$

$$\begin{aligned}
&= \left(\frac{1}{\theta_0}\right)^k \det \begin{pmatrix} \frac{\theta_0}{\theta_1}+1 & 1 & \dots & 1 \\ 1 & \frac{\theta_0}{\theta_2}+1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & \frac{\theta_0}{\theta_k}+1 \end{pmatrix} \\
&\overset{*}{=} \left(\frac{1}{\theta_0}\right)^k \prod_{j=1}^k \frac{\theta_0}{\theta_j} \cdot \left(1 + \begin{pmatrix}1 & \dots & 1\end{pmatrix} \begin{pmatrix} \frac{\theta_1}{\theta_0} & & 0 \\ & \ddots & \\ 0 & & \frac{\theta_k}{\theta_0}\end{pmatrix} \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right) \\
&= \prod_{j=1}^k \frac{1}{\theta_j} \cdot \left(1 + \sum_{i=1}^k \frac{\theta_i}{\theta_0}\right) = \left(\prod_{j=1}^k \frac{1}{\theta_j}\right) \frac{\theta_0 + \sum_{i=1}^k \theta_i}{\theta_0} \\
&= \left(\prod_{j=1}^k \frac{1}{\theta_j}\right) \frac{1}{\theta_0} = \prod_{j=0}^k \frac{1}{\theta_j}\,,
\end{aligned}$$

where equality * holds by the theorem [3, Theorem A 16 (x), Appendix A3, p. 494], stating that

$$\det(A + a\,a^\top) = \det(A)(1 + a^\top A^{-1} a) \qquad \text{if} \qquad \det(A) \neq 0$$

for all appropriately sized matrices A and column vectors a.

With $\det\left(\frac{d\boldsymbol{\psi}}{d\boldsymbol{\theta}}\right) = \prod_{j=0}^k \frac{1}{\theta_j}$, we get

$$\begin{aligned}
p(\boldsymbol{\theta} \,|\, n, \mathbf{y}) &\propto \exp\left\{ n \left[\sum_{j=1}^k y_j \ln\left(\frac{\theta_j}{\theta_0}\right) - \big(- \ln(\theta_0)\big) \right] \right\} \cdot \left| \prod_{j=0}^k \frac{1}{\theta_j} \right| \\
&= \exp\left\{ n \left[\sum_{j=1}^k y_j \Big(\ln(\theta_j) - \ln(\theta_0) \Big) + \ln(\theta_0) \right] - \sum_{j=0}^k \ln(\theta_j) \right\} \\
&= \exp\left\{ n \left[\sum_{j=1}^k y_j \ln(\theta_j) + \ln(\theta_0) \underbrace{\Big(1 - \sum_{j=1}^k y_j \Big)}_{=:y_0} \right] - \sum_{j=0}^k \ln(\theta_j) \right\} \\
&= \exp\left\{ n \left[\sum_{j=0}^k y_j \ln(\theta_j) \right] - \sum_{j=0}^k \ln(\theta_j) \right\} \\
&= \exp\left\{ \sum_{j=0}^k (n\, y_j - 1) \ln(\theta_j) \right\} = \exp\left\{ \sum_{j=0}^k \ln\left(\theta_j^{n\, y_j - 1} \right) \right\}
\end{aligned}$$

$$= \prod_{j=0}^{k} \theta_j^{n\, y_j - 1} \,,$$

which is the core of a Dirichlet density $\mathrm{Dir}(n, \mathbf{y})$ over $\boldsymbol{\theta}$. Therefore, the Dirichlet distribution is the canonically constructed conjugate prior to the multivariate Bernoulli. Since i.i.d. repetitions do not interfere with conjugacy, the Dirichlet distribution is the canonically constructed conjugate prior also to the Multinomial distribution.

Acknowledgements

The author would like to thank Erik Quaeghebeur for helpful discussions.

References

[1] J.-M. Bernardo and A. Smith. *Bayesian Theory.* Wiley, Chichester, 2000.

[2] E. Quaeghebeur and G. de Cooman. Imprecise probability models for inference in exponential families. In F. Cozman, R. Nau, and T. Seidenfeld, editors, *ISIPTA '05. Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, pages 287–296, 2005.

[3] C.R. Rao, H. Toutenburg, Shalabh, and C. Heumann. *Linear Models and Generalizations.* Springer Series in Statistics. Springer, Berlin, extended edition, 2008. Least squares and alternatives, With contributions by Michael Schomaker.

[4] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[5] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society, Series B*, 58:3–34, 1996.

[6] G. Walter and T. Augustin. Imprecision and prior-data conflict in generalized Bayesian inference. *Journal of Statistical Theory and Practice*, 3:255–271, 2009.

[7] G. Walter, T. Augustin, and A. Peters. Linear regression analysis under sets of conjugate priors. In G. de Cooman, J. Vejnarova, and M. Zaffalon, editors, *ISIPTA '07: Proceedings of the Fifh International Symposium on Imprecise Probability: Theories and Applications*, pages 445–455, 2007.