Mixtures of Multinomials

Jason D. M. Rennie
jrennie@gmail.com

September 1, 2005*

Abstract

We consider two different types of multinomial mixtures, (1) a word-level mixture, and (2) a document-level mixture. We show that the word-level mixture is, in fact, no different than a regular multinomial. However, the document-level mixture can produce distributions that a regular multinomial cannot. We point out that Latent Dirichlet Allocation contains a word-level mixture that can be collapsed into a single multinomial. Finally, we discuss a natural parameter variant of the word-level mixture.

1 Mixtures of Multinomials

The multinomial model over term frequencies is defined as

$$P(\vec{x}|\vec{\mu}) = \frac{l!}{\prod_i x_i!} \prod_{i=1}^{n} \mu_i^{x_i}, \tag{1}$$

where x_i is the number of times word i occurs, μ_i is the mean parameter for word i (i.e. μ_i is the expected rate of occurrence), and $l = \sum_i x_i$ is the document length. Note that the multinomial is conditioned on document length.

We can view the multinomial as the following generative process. We have a hopper filled with balls. Each ball represents one of the n words and the hopper and balls are designed so that the chance of pulling a ball labeled word i is μ_i. Each hopper draw is a "word" event. Balls are replaced before the next draw. We tally word counts, but no not keep track of order information. The resulting set of counts is a "document" event.

1.1 Word-Level Mixtures

For a word-level mixture, we use two hoppers, α and β, each designed to generate word events according to its parameter vector. For each "word" event, we draw one ball from each hopper, but only count one of the balls. With probability λ,

*Updated September 29, 2005

we count the ball drawn from the α hopper, otherwise we count the ball drawn from the β hopper. The distribution specified by this model is

$$Q(\vec{x}|\lambda, \vec{\alpha}, \vec{\beta}) = \frac{l!}{\prod_i x_i!} \prod_{i=1}^{n} (\lambda\alpha_i + (1-\lambda)\beta_i)^{x_i}. \tag{2}$$

Clearly, a multinomial with parameter vector $\lambda\vec{\alpha} + (1-\lambda)\vec{\beta}$ specifies the same distribution.

1.2 Document-Level Mixtures

For a document-level mixture, we again use two hoppers, α and β. However, this time we randomly select one hopper for the entire set of word-level draws. With probability λ, we select α. Thus, the distribution specified by this document-level mixture is simply the convex combination of two multinomials:

$$R(\vec{x}|\lambda, \vec{\alpha}, \vec{\beta}) = \lambda P(\vec{x}|\vec{\alpha}) + (1-\lambda) P(\vec{x}|\vec{\beta}). \tag{3}$$

We cannot write this as a single multinomial. To show this, we consider the following parameter settings, $\lambda = 0.5$, $\vec{\alpha} = (1,0)$, and $\vec{\beta} = (0,1)$. For a length 2 document, we get the following distribution:

$$R((0,2)|\lambda, \vec{\alpha}, \vec{\beta}) = 0.5 \quad R((2,0)|\lambda, \vec{\alpha}, \vec{\beta}) = 0.5 \quad R((1,1)|\lambda, \vec{\alpha}, \vec{\beta}) = 0 \tag{4}$$

No multinomial can produce such a distribution.

1.3 Discussion

The word-level mixture is used in Latent Dirichlet Allocation (LDA) [1]. Blei et al. write out the LDA model[1] as

$$p(\vec{w}) = \int_\theta \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n;\beta) p(z_n|\theta) \right) p(\theta;\alpha) d\theta, \tag{5}$$

where θ is the vector of topic weights and β is the matrix of multinomial parameters (we assume that each column stores parameters for a multinomial). The inner mixture ($\sum_{z_n}$) is a word-level mixture, so we can re-write the mixture as a single multinomial,

$$p(\vec{w}) = \int_\theta \left(\prod_{n=1}^{N} p(w_n|\beta\theta) \right) p(\theta;\alpha) d\theta. \tag{6}$$

Note that $\beta\theta$ is a matrix-vector multiplication resulting in a multinomial parameter vector. Thanks to David Blei for noting that this observation was previously made by [2]. Buntine and Jakulin use the term “admixture” to describe the word-level mixture. We note that Ueda and Saito’s Parametric Mixture Model is a word-level mixture [4].

[1] Note that these scalar and parameter names do not correspond to the names used elsewhere in this document.

2 A Different Parameterization

So far, we have used the mean parameterization of the multinomial. I.e. the (mean) parameters, $\vec{\mu}$, are proportional to the expected multinomial model outcome,

$$E_{P(\vec{x}|\vec{\mu})}\left[\vec{x}\right] = l\vec{\mu}, \tag{7}$$

where l is the fixed document length. Although this parameterization is intuitively pleasing, it is an awkward representation. The mean parameter vector is constrained to be non-negative and sum to one; parameter values cannot be changed individually; and, the number of dimensions of the parameter space is one less than the length of the parameter vector. See Collins et al. for a discussion of learning natural parameter mixtures for any exponential model [3].

2.1 Natural Parameters

Here, we consider an alternate, "natural" parameterization. A model is in the exponential family if it can be written as

$$\log P(\vec{x}|\vec{\theta}) = \log P_0(\vec{x}) + \vec{x}^T\vec{\theta} - G(\vec{\theta}), \tag{8}$$

where $\vec{\theta}$ are the natural parameters of the model. Using this parameterization, the multinomial distribution is

$$P(\vec{x}|\vec{\theta}) = \frac{l!}{\prod_i x_i!}\prod_{i=1}^{n}\left(\frac{\exp(\theta_i)}{\sum_{i'}\exp(\theta_{i'})}\right)^{x_i}, \tag{9}$$

where $P_0(\vec{x}) = \frac{l!}{\prod_i x_i!}$ and $G(\vec{\theta}) = l\log\left[\sum_{i'}\exp(\theta_{i'})\right]$. Note that if there are N mean parameters, we only need $N-1$ natural parameters. To make translation between the two parameterizations easier, we assume $\theta_N = 0$. Given the natural parameters for a multinomial model, the mean parameters are $\mu_i = \frac{\exp(\theta_i)}{\sum_{i'}\exp(\theta_{i'})}$.

2.2 Mixtures

Here we define what we call a natural parameter (multinomial) mixture. Like the word-level mixture we discussed earlier, we mix at the parameter level. The difference is that we mix natural parameters rather than mean parameters. Let $\vec{\theta}$ and $\vec{\phi}$ be the natural parameters for two multnomials. Then, our natural parameter mixture is simply a multinomial with natural parameters equal to a convex combination of the two original paramter vectors,

$$P(\vec{x}|\lambda\vec{\theta} + (1-\lambda)\vec{\phi}) \propto \prod_{i=1}^{n}\left(\frac{\exp(\lambda\theta_i + (1-\lambda)\phi_i)}{\sum_{i'}\exp(\lambda\theta_{i'} + (1-\lambda)\phi_{i'})}\right)^{x_i}, \tag{10}$$

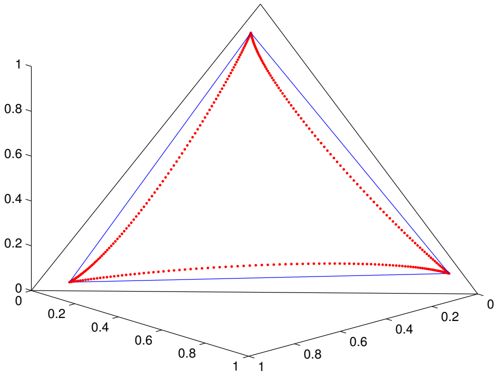

Figure 1: Shown are the boundaries of regions enclosing all possible parameter mixtures of the points $(.9, .07, .03)$, $(.07, .03, .9)$, and $(.03, .9, .07)$, viewed in the mean parameter space. The blue lines bound the region of possible mean parameter mixtures (described as word-level mixtures earlier in this document). The red dots bound the region of possible natural parameter mixtures. The black lines bound the three-parameter simplex.

where $0 \leq \lambda \leq 1$ is the mixture parameter. Note that this mixture does not have the same interpretation as the "word-level" mixture we discussed earlier. Like the word-level mixture, the natural parameter mixture can be written as a single multinomial and can be viewed as a modification of the word-level probabilities. However, it cannot be viewed as a convex combination of word-level probabilities; instead, it is a geometric average of the word-level probabilities. Figure 1 gives a pictoral view of mean and natural parameter mixtures. Viewed in mean-parameter space, the boundaries of word-level mixtures form a triangle. The boundaries for natural parameter mixtures look like a bowed triangle with curved sides bent slightly inward; also, note the skew due to the placement of the corners. Note that the natural parameter mixture is related to the Dirichlet distribution.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14*, 2002.

[2] W. Buntine and A. Jakulin. Applying discrete pca in data analysis. In *Twentieth Conference on Uncertainty in Artificial Intelligence*, 2004.

[3] M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems 14*, 2002.

[4] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In S. T. S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 721–728. MIT Press, 2003.