FITTING A HIERARCHICAL LOGISTIC NORMAL DISTRIBUTION

JONATHAN HUANG AND TOMASZ MALISIEWICZ

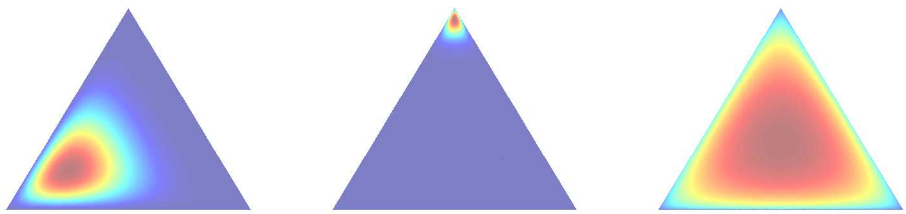

FIGURE 1. Dirichlet Distributions for various parameter settings on a 2-simplex. Red corresponds to high probability density and blue corresponds to low probability density.

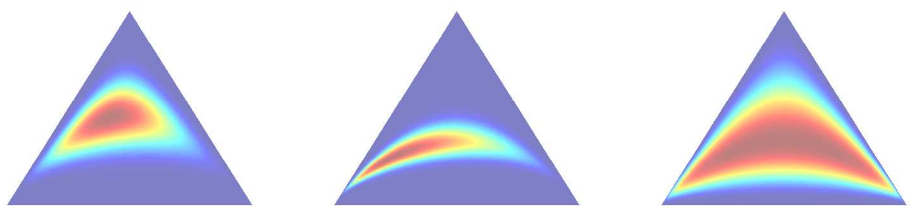

FIGURE 2. Logistic Normal Distributions for various parameter settings on a 2-simplex. Note that unlike the Dirichlet, its level sets can bound nonconvex regions.

The *Logistic-Normal* distribution [AS80] is a distribution over a simplex which forms a richer class of distributions than Dirichlets and better captures inter-component correlations. The process of drawing a k-dimensional Logistic-Normal random variable u is as follows:

(1) Draw $v \sim N(\mu, \Sigma)$ where $N(\mu, \Sigma)$ is a $k-1$ dimensional Normal distribution.
(2) Define $v_k = 0$.
(3) Let
$$\theta = \frac{\exp v}{\sum_{j=1}^{k} \exp v_j}$$
(This is the projection of $\exp(v)$ to the simplex)

The probability density for θ can be explicitly written as

$$p(\theta; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|} \left(\prod_{j=1}^{k} \theta_j \right)^{-1} \exp\left[-\frac{1}{2} \{\log(\theta/\theta_k) - \mu\} \Sigma^{-1} \{\log(\theta/\theta_k) - \mu\} \right]$$

[1]

We present the method for fitting the Hierarchical Logistic-Normal (HLN) distribution given by Hoff [Hof03]. The HLN distribution can be described by the following generative process.

(1) Draw $v_j \sim N(\mu, \Sigma)$ where $N(\mu, \Sigma)$ is a $k-1$ dimensional Normal distribution.
(2) Define $v_{jk} = 0$.
(3) Let
$$\theta_j = \frac{\exp v}{\sum_{j=1}^{k} \exp v_j}$$
(4) For $i = 1, \ldots, n$, draw $z_{ji} \sim \text{Multinomial}(\theta)$

Notice that if the v_j are known, then finding the maximum likelihood estimates of μ and Σ is easy. Since they are unknown, the strategy will be instead to alternate between estimating $v_1, \ldots v_m$ for each document, and estimating μ and Σ using EM. Let $\hat{\mathbf{p}}(z)$ be the empirical distribution function (normalized histogram) of the topic assignments in a document. The conditional likelihood of $\mathbf{v}$ given $\mathbf{z} = (z_1, \ldots, z_n)$ for a given document can be written down using Bayes rule:

$$\begin{aligned} P(\mathbf{v}|\mathbf{z}, \mu, \Sigma) &\propto P(\mathbf{z}|\mathbf{v})P(\mathbf{v}|\mu, \Sigma) \\ &= \frac{\exp\left(\sum_{i=1}^{k-1} v_i n \hat{\mathbf{p}}_i\right)}{\left(1 + \sum_{j=1}^{k-1} \exp v_j\right)^n} \exp\left(-\frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu)\right) \end{aligned}$$

The conditional log-likelihood and its derivatives are straightforward (but not fun) to derive:

$$\log P(\mathbf{v}|\mathbf{z}, \mu, \Sigma) = \sum_{i=1}^{k-1} v_i n \hat{\mathbf{p}}_i - n \log\left(1 + \sum_{j=1}^{k-1} \exp v_j\right) - \frac{1}{2}(v-\mu)^T \Sigma^{-1}(v-\mu) + C$$

$$\frac{\partial \log P(\mathbf{v}|\mathbf{z}, \mu, \Sigma)}{\partial \mathbf{v}} = n\left(\hat{\mathbf{p}} - \frac{\exp \mathbf{v}}{1 + \sum_{j=1}^{k-1} \exp v_j}\right) - \Sigma^{-1}(v-\mu)$$

$$\begin{aligned} \frac{\partial^2 \log P(\mathbf{v}|\mathbf{z}, \mu, \Sigma)}{\partial v_i \partial v_j} &= -\Sigma_{ij}^{-1} - n\left[\delta\{i=j\} \frac{\exp v_j}{1 + \sum_{l=1}^{k-1} \exp v_l}\right. \\ &\left. - \left(\frac{\exp v_i}{1 + \sum_{l=1}^{k-1} \exp v_l}\right)\left(\frac{\exp v_j}{1 + \sum_{l=1}^{k-1} \exp x_l}\right)\right] \end{aligned}$$

By maximizing the conditional log-likelihood, the conditional mode of $\mathbf{v}$ can be found. [2]

Let $\hat{\mu}$ be the conditional mode of $\mathbf{v}$ and $\hat{I}$ be the Fisher Information matrix (negative Hessian) evaluated at $\hat{\mu}$. Then asymptotically,

$$f(\mathbf{v}|\mathbf{z}, \mu, \Sigma) \approx \mathcal{N}(\mathbf{v}|\hat{\mu}, \hat{I}^{-1})$$

---

[1] Since θ is actually a k-dimensional vector, we concatenate a zero to the end of μ and pad Σ and Σ^{-1} on the right and bottom by a column and row of zeros respectively.

[2] In practice, we find that (Polak-Ribiere) Conjugate Gradient tends to be more dependable than the Newton-Raphson method in high dimensions. We used Carl Rasmussen's Conjugate Gradient Matlab code for this.

To estimate the Logistic Normal parameters μ and Σ, we iterate between computing conditional modes, and updating μ, Σ. The algorithm is as follows

(1) Initialize μ_0, Σ_0.
(2) Until convergence,
   (a) For each document $j \in \{1, \ldots, m\}$, estimate $\hat{\mu}_j$ and $\hat{I}_j$ with respect to current model parameters μ_l and Σ_l.
   (b) Update μ, Σ:

$$\mu_{l+1} = \frac{1}{m} \sum_{j=1}^{m} \hat{\mu}_j$$

$$\Sigma_{l+1} = \frac{1}{m} \sum_{j=1}^{m} \left[(\hat{\mu}_j - \mu_{l+1})(\hat{\mu}_j - \mu_{l+1})^T + \hat{I}_j^{-1} \right]$$

References

[AS80] J Aitchison and S.M. Shen, *Logistic-normal distributions: Some properties and uses*, Biometrika **67** (1980).

[Hof03] Peter Hoff, *Nonparametric modelling of hierarchically exchangeable data*, Tech. report, Department of Statistics, University of Washington, 2003.

E-mail address: jch1@cs.cmu.edu

E-mail address: tomasz@cmu.edu