Controlling Sparseness in Multinomial Models

Juan K. Lin
Department of Statistics
Rutgers University
Piscataway, NJ 08854
jklin@stat.rutgers.edu

1 Introduction

There has been much recent interest in multinomial mixture models which have appeared under the names Latent Class Models [1], Aggregate Markov Models [2], Non-negative Matrix Factorization (NMF)[3], and probabilistic LSA (pLSA)[4]. When viewed from a matrix decomposition perspective, these models discover sparse descriptions of data as mixtures of sparse basis functions. Many of the recent applications of these models have been in the fields of natural language processing and information retrieval. We formulate variations of this multinomial mixture model in the setting of source-destination traffic analysis. In analogy with controlling the covariance structure of Gaussians from diagonal covariance, to graphical Gaussian models, Factor Analysis, to full covariance structure Gaussians, we present multinomial mixture models containing structure with various complexities.

2 Multinomial Traffic Models

The starting point of our analysis is traffic data consisting of n_{ij} counts of traffic from source $X = i$ to destination $X' = j$. We assume that all sources are destinations, and destinations sources. In this section, we describe sparse multinomial mixture models of traffic flow.

2.1 Stochastic Square Root Model

We begin with a stochastic square root traffic model. Let T_{ki} be the empirical conditional probability of traffic transition from source i to destination k. We assume that the traffic flow data was generated by two time step iterations of an underlying traffic transition matrix H. Maximum likelihood estimation of H is equivalent to minimization of the following Kullback-Leibler divergence:

$$\mathcal{D}(T_{ki} \parallel \sum_j H_{kj} H_{ji}).$$

In Vardi and Lee[5], the numerous applications of a non-negative probabilistic inverse is described. Drawing on the connection with linear algebra, the stochastic square root model in essence describes a non-negative probabilistic square root. Even though H contains the same number of degrees of freedom as the empirical transition matrix T, the minimization of the Kullback-Leibler divergence imposes sparseness constraints on H. Since the latent Markov process has two time steps to describe the observed empirical traffic, the stochastic square root transition matrix tends to have significantly fewer positive elements. In essence, the observed traffic data is described by a sparse underlying traffic model. Analytic and empirical results will appear in a more extended paper.

2.2 Highway Traffic Model

We continue with descriptions of latent variable traffic models with varying structural complexities [6]. Discrete latent variables H and H' are introduced which characterize the underlying entrance hubs and exit hubs on the highway. We assume that all entrances are exits, and vice versa. Our model of traffic flow consists of onramp traffic from sources to highway entrances, highway traffic from entrances to exits, and offramp traffic from highway exits to destinations. The model assigns a probability of going from source i to destination j of:

$$p(i, j) = \sum_{k,l} \alpha_{ik} \beta_{kl} \gamma_{jl},$$

where $\alpha_{ik} = P(X = i|H = k)$, $\beta_{kl} = P(H = k, H' = l)$, and $\gamma_{jl} = P(X' = j|H' = l)$. In words, α_{ik} is the fraction of traffic at entrance k from source i, β_{kl} is the probability of going from entrance k to exit l on the highway, and γ_{jl} is the fraction of traffic at exit l that proceed to destination j. The double sum in the expression is over all highway entrances and exits. Note that the traffic model is probabilistic, and in general allows for more than one highway route from source to destination. We further impose a constraint equating the onramp and offramp traffic distributions:

$$\gamma_{jl} = \alpha_{jl}.$$

Thus the fraction of traffic at exit l which continue to destination j is equal to the fraction of traffic at entrance l which originate from j. The model parameters are specified by $\alpha(x|h) = P(x|h)$ and $\beta(h,h') = P(h,h')$, which specify respectively the onramp/offramp traffic distribution, and highway traffic between the entrances and exits. Let the total amount of observed traffic be $N = \sum_{i,j} n_{ij}$, and let $\tilde{p}_{ij} = n_{ij}/N$ be the observed empirical joint distribution $\tilde{p}(x=i, x'=j)$. The log-likelihood function is given by

$$\mathcal{L} = N \sum_{x,x'} \tilde{p}(x,x') \log[\sum_{h,h'} \alpha(x|h)\beta(h,h')\alpha(x'|h')].$$

Maximizing the likelihood of the observed source-destination traffic counts is equivalent to minimizing the following Kullback-Leibler divergence:

$$\mathcal{D}(\tilde{p}(x,x') \parallel \sum_{h,h'} \alpha(x|h)\beta(h,h')\alpha(x'|h')).$$

2.3 Hub Traffic Models

A traffic model with the same structure as pLSA/NMF [1]-[4] seeks to minimize the KL-divergence:

$$\mathcal{D}(\tilde{p}(x,x') \parallel \sum_{h} \eta(x|h)\kappa(h)\theta(x|h)).$$

Here $\eta(x|h)$ and $\theta(x|h)$ represent onramp and offramp traffic distributions to latent hubs, while $\kappa(h)$ describes the probability of traversing through each latent hub. We refer to this as the "hub" traffic model. A simpler "symmetric hub" model can be obtained by constraining the hub model to have identically distributed onramp and offramp distributions.

2.4 Analysis

Representing the model parameters α and β as matrices, the highway traffic model seeks an approximation of the empirical traffic distribution $\tilde{p}$ by minimizing

$$\mathcal{D}(\tilde{p} \parallel \alpha\beta\alpha^t).$$

In comparison, the hub traffic model seeks to minimize

$$\mathcal{D}(\tilde{p} \parallel AB).$$

The highway model assumes more structure in the traffic data, and is a constrained version of the hub model. In particular, a highway model can always be represented as a hub model by equating corresponding terms in $(\alpha\beta)(\alpha^t) = (A)(B)$, effectively folding in the highway traffic between entrances and exits into the onramp traffic distribution specified by A. This comes at the cost of reduced sparseness of the onramp traffic distribution, and an increase in complexity of the hub model. Without equating onramp to offramp traffic in the highway model, the highway traffic has extra degrees of freedom since we can always write $\alpha\beta\gamma = (\alpha\beta)(I)(\gamma)$. Here the onramp traffic incorporates the highway traffic, and now there is no cross-traffic between entrances and exits. By equating onramp to offramp traffic, these degrees of freedom are removed in the highway traffic term β. The highway and hub traffic models differ in complexity, sparseness and structure. In the following section, the highway and hub traffic models will be compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores, as well as predictive test-set log-likelihoods.

3 Numerical Experiments

We analyzed simulated internet traffic data based on an undirected connectivity graph between Autonomous Systems (AS). The connectivity graph consists of AS paths in BGP routing tables collected by the server *route-views.oregon-ix.net*. This data is the basis of the power-law analysis in [7], and is publicly available at *topology.eecs.umich.edu/data.html*. After trimming out nodes with no edges, we are left with an undirected binary AS connectivity graph with 13233 interconnected AS nodes.

We compared the highway traffic model to the hub traffic model normalizing for the complexity differences of the two models. With k latent hub states in the hub model, and n sources/destinations, the hub model has $[2k(n-1)+k-1]$ parameters. In contrast, the highway model with the same number k or entrances/exits contains only $[k(n-1)+k^2-1]$ parameters. We compared the two models using the Akaike Information Criterion (AIC), Bayesian Informatin Criterion (BIC), and predictive test-set log-likelihoods. The simulated traffic data was constructed as follows. For the training set, we performed 100000 *single* random walk steps on the connectivity graph. The 100000 source nodes were sampled in accordance with the stationary distribution of the random walk on the connectivity graph. Traffic from source nodes are assumed to follow each of the edge paths with equal probability. Since multinomial mixture models can be prone to over-fitting problems, we added a single pseudo-count traffic for each edge in the connectivity graph. If traffic from a source to a destination is not observed in the test set, but appears in the training set, the traffic models may assign zero probability to the test set likelihood. Early stopping will effectively stop parameter updates if an update assigns zero probability to a traffic path that appears in the test set. An additional inverse annealing (heating) is used in [4] to smooth multinomial parameters and prevent sparseness. However, this directly comes at the cost of sparseness of representation. For the test set, 20000 single random walk steps were simulated.

In Table 1 the AIC and BIC scores for the highway and hub models are tabulated for a number of different k values. For each model and each k, 10 EM runs with random parameter initializations are performed. Scores for the best respective runs are reported in the table. The highway traffic model has significantly better (lower) AIC and BIC scores than the hub traffic model.

values $\times 10^6$	k=26	k=51	k=100	k=197
Highway AIC	**4.90**	**5.41**	**6.59**	**9.07**
Hub AIC	5.56	6.72	9.16	14.15
Highway BIC	**8.34**	**12.2**	**19.9**	**35.0**
Hub BIC	12.4	20.2	35.5	66.1

Table 1. AIC and BIC scores for the highway and hub models.

We also compared predictive test-set log-likelihoods for a highway model and hub model with comparable degrees of freedom. Comparing the $k = 51$ highway model with 677432 parameters with the $k = 26$ hub model with 688089 parameters, the best test set log-likelihoods ($\times 10^5$) were -2.64 and -2.74 for the highway and hub models respectively. Comparing the $k = 100$ highway model with 1333199 parameters with the $k = 51$ hub model with 1349714 parameters, the best test set log-likelihood($\times 10^5$) were -2.55 and -2.63 respectively. Finally, the $k = 197$ highway model with 2645512 parameters and the $k = 100$ hub model with 2646499 parameters had best test set log-likelihoods($\times 10^5$) of -2.46 and -2.54 respectively. In all three comparisons, the highway model had slightly fewer parameters, but significantly higher (less negative) predictive test set log-likelihoods.

4 Discussion

The symmetric hub model, highway model, and hub model are constructed with varying complexity and associated sparseness structure in the multinomial mixture. The highway traffic model extracts community structure based on source-destination traffic information, but in addition captures the aggregate "highway" traffic between the communities. This important distinction extends the highway traffic analysis beyond clustering and allows it to extract out underlying backbone traffic structure from traffic data matrix.

5 References

[1] Everitt, B. (1984). An Introduction to Latent Variable Models. London: Chapman & Hall.

[2] Saul, L. and Pereira, F. (1997) Aggregate and mixed-order Markov models for statistical language processing. In *Proceedings of the second conference on empirical methods in natural language processing*, 81-89.

[3] Lee, D. and Seung, S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(675), 788-791.

[4] Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177-196.

[5] Vardi, Y and Lee, D. (1993). From Image Deblurring to Optimal Investments: Maximum Likelihood Solutions for Positive Linear Inverse Problems, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 55, No. 3. (1993), pp. 569-612.

[6] Lin, J.K. (2006). Traffic Models for Community-based Ranking and Navigation, to appear, Lecture Notes in Computer Science: Workshop on Internet and Network Economics, Springer-Verlag.

[7] Qien, C., Chang, H., Govindan, R., Jamin, S., Shenker, S. and Willinger, W. (1992) The Origin of Power Laws in Internet Topologies Revisited, *Proc. of IEEE Infocom, 2002*