Markov Mixed Membership Models

Aonan Zhang
John Paisley

Department of Electrical Engineering, Columbia University, New York, NY, USA

Abstract

We present a Markov mixed membership model (Markov M3) for grouped data that learns a fully connected graph structure among mixing components. A key feature of Markov M3 is that it interprets the mixed membership assignment as a Markov random walk over this graph of nodes. This is in contrast to tree-structured models in which the assignment is done according to a tree structure on the mixing components. The Markov structure results in a simple parametric model that can learn a complex dependency structure between nodes, while still maintaining full conjugacy for closed-form stochastic variational inference. Empirical results demonstrate that Markov M3 performs well compared with tree structured topic models, and can learn meaningful dependency structure between topics.

1. Introduction

Mixed membership modeling is a statistical framework for modeling grouped data where each group is represented as a unique mixture over a shared structure (Airoldi et al., 2014). A wide range of data fall within the scope of mixed membership models, including documents (Blei et al., 2003), images (Li & Perona, 2005), and the genome (Pritchard et al., 2000). Exchangeability assumptions can be relaxed to extend mixed membership models to link data (Airoldi et al., 2008), heterogeneous data (Chang & Blei, 2009; Wang & Blei, 2011) and matrix factorization (Mackey et al., 2010). In this paper, we focus on the case where each group’s data is assumed i.i.d. given its mixed membership mixing measure.

For discrete grouped data, the most basic mixed membership model is latent Dirichlet allocation (LDA) (Blei et al., 2003), which assumes a finite set of discrete distributions, and models each group as a mixture over these distributions using a Dirichlet prior. The simple “flat” Dirichlet prior assumes no structure among the atoms, and so the model can overfit as the number of components increases beyond a certain number. To capture finer resolution without overfitting, structure has been introduced to the atom relationships, for example by modeling pairwise correlations (Blei & Lafferty, 2007) or tree structures (Blei et al., 2010).

Among these models, the tree-structured model is especially interesting for the structure it can learn (Blei et al., 2010; Li et al., 2012; Kim et al., 2012; Ahmed et al., 2013; Paisley et al., 2015). Because the components are given a strict parent/child relationship, tree models can discover components of different granularities having top-down dependencies. In topic modeling, this is natural since topics can be more or less specific and the children of one topic can further specify the more general content of the parent topic that unites them.

To consider two Bayesian nonparametric instances, the nested Chinese restaurant process (nCRP) (Blei et al., 2010) and nested hierarchical Dirichlet process (nHDP) (Paisley et al., 2015) are two tree-structured models that select distributions on paths from a root node (see Figure 1). For example, the nCRP selects the atoms for a group by following a path from root to leaf node; the nHDP generalizes this by selecting a subtree of atoms for each group. Still, in both models it is assumed that there is a clear tree-structured relationship among nodes. In this paper, we explore a related modeling framework that allows for a more flexible range of node connections by assuming a Markov structure among the nodes.

To this end, we present the Markov mixed membership model (Markov M3) for grouped data that learns a fully connected graph structure among mixing components. With respect to tree-structured models, our proposed Markov model is straightforward in that, rather than imposing a tree-structured transition rule between nodes, we model the nodes as a fully connected graph with a first-order Markov rule for transitioning between nodes (see Figure 1). We therefore refer to this as a graph-based mixed membership model. In the context of topic models,
Markov Mixed Membership Models

An advantage of our proposed framework is that it avoids the combinatorial issues encountered during inference by models such as the nCRP (Wang & Blei, 2009), and does not require complicated pruning procedures such as required by the nHDP (Paisley et al., 2015). In contrast, Markov M3 is a fully conjugate-exponential family model that allows for closed-form updates that doesn’t require the complicated procedures of the nCRP or nHDP. As with those two models, our model is easily learned with stochastic variational inference allowing for processing of large data sets (Hoffman et al., 2013).

In Section 2 we present Markov M3 and show the nCRP to be the nonparametric limiting case of the model. We present a closed-form stochastic variational inference algorithm for the model in Section 3. In Section 4 we present experiments on two large grouped discrete data sets.

2. Model Description

Mixed membership models are applicable to data sets where the observations are grouped, i.e., where viewing the data on the instance-level results in subsets of the data. For example, each document in a set can be represented as a group of words. Mixed membership models assume that the groups share a data generating (global) structure, but with different distributions to account for group-level (local) differences. For example, topic models share the same set of distributions on words, but mix over them differently for each group.

A common assumption made by mixed membership models is that the data is exchangeable within and across groups. In this case, where there exists a random mixed membership measure (i.e., De-finitte’s measure) that results in i.i.d. generation of the data. LDA is an example of an exchangeable mixed membership model both within and across groups, whereas dynamic topic models assume partial exchangeability since the order of documents matters (Blei & Lafferty, 2006). We present a fully exchangeable model in which each group mixes on paths selected from a fully connected graph according to a Markov random walk.

2.1. Markov mixed membership models

We define the generative process for the Markov mixed membership model. The following procedure is also summarized in Algorithm 1. Let \( w_d \) be the set of data for group \( d \). We model this data as an i.i.d. set drawn from a mixture \( G_d \) with mixing distribution \( \nu_d \sim \text{GEM}(\gamma_0) \) and a group-specific sequence of atoms \( (\hat{\beta}_1, \hat{\beta}_2, \ldots) \). We can draw from the GEM stick-breaking distribution by sampling,

\[
 u_{di} \sim \text{Beta}(1, \gamma_0), \quad \nu_{di} = u_{di} \prod_{j=1}^{i-1} (1 - u_{dj}). \tag{1}
\]

We use the mixture \( G_d = \sum_{i=1}^{\infty} \nu_{di} \hat{\beta}_i \) to generate group \( w_d = (w_{d1}, \ldots, w_{dn}) \) by sampling

\[
 \ell_{dn} \sim \text{Discrete}(\nu_d), \quad w_{dn} | \ell_{dn} \sim f(\hat{\beta}_{\ell_{dn}}). \tag{2}
\]
The distribution $f$ is problem-specific. In this paper we take $f$ to be a discrete distribution and $\beta$ a $V$-dimensional probability vector. We assume there are $K$ atoms, where $\beta_i \in \beta = \{\beta_1, \ldots, \beta_K\}$, and define the prior

$$\beta_k \sim \text{Dir}(\beta_0 1_V).$$

In the nCRP (Blei et al., 2010), the sequence of $\hat{\beta}_i$ is selected by following a path from the root node to the leaf node of a tree. Instead, we assume a first order Markov structure for this sequence. We construct a Markov transition distribution on $\beta$ by drawing

$$\theta_k \sim \text{Dir}(\alpha_0 / K, \ldots, \alpha_0 / K)$$

for each $k$. The distribution on the sequence $\hat{\beta}_d$ is then $P(\hat{\beta}_1 = \beta_1 | \hat{\beta}_{i-1} = \beta_{i-1}, \theta) = \theta_{j, \hat{j}}$. The variable $z_d$ shown in Algorithm 1 and used for inference indexes this sequence: $z_{di} = j'$ if $\hat{\beta}_i = \beta_j'$. We assume the same Dirichlet prior on the initial state distribution $\pi$ as well.

We observe that if we were to set $\ell_{dn} = n$, we would assign $w_{dn}$ to atom $\hat{\beta}_n$, and the result would be a standard hidden Markov model (HMM). What differentiates our model, and reintroduces group-level exchangeability, is that each word chooses which of the selected states it belongs to i.i.d. $\nu_d$. The analogy to the HMM can be pursued by thinking of the words of a document first being partitioned into ordered sets according to $\nu_d$, and then drawing each group according to an HMM. We will see how this way of considering the model leads to variational inference that builds on inference for the HMM (Beal, 2003).

### 2.2. Relationship to tree-structured models

As Figure 1 illustrates, the major difference between graph-based and tree-based mixed membership models is the dependence structure between nodes. Where models such as the nCRP impose a strict parent/child hierarchy, Markov M3 in a sense captures the potential for each node to be the parent of all others (which simply results from a shift of perspective about Markov chains). The limited modeling ability of the nCRP is primarily due to the rigid single path allowed per group. In topic modeling, this forces each selected topic to be a strict subset of those previously selected. This assumption was relaxed by recent tree-structured alternatives (Kim et al., 2012; Ahmed et al., 2013; Paisley et al., 2015). For example, as Figure 1 indicates, the nHDP allows for multiple paths per document, so two general topics can be combined in a single document by allowing words to select paths in different directions.

As seen in Figure 1, Markov M3 in one sense returns to the one path per group structure of the nCRP, but allows for exploration of the entire space like the nHDP. This provides another remedy to the rigidity of the nCRP. It also offers modeling capabilities not found in the nHDP, since in that model, the presence of two subtrees is not causally linked according to the prior—the presence of one subtree says nothing about the presence of another. With Markov M3, there is a causal connection between two atoms according to the Markovian generative structure (which we observe is not symmetric). Markov M3 is again like the nCRP in that, once it selects its path of atoms, it mixes on them using a probability vector drawn from a stick-breaking distribution.

The Markov mixed membership model can be viewed as a possible parametric version of the nested Chinese restaurant process, in that we can show that the nCRP is the limiting process of the model in Algorithm 1 as $K$ goes to infinity. To roughly sketch this, consider the marginal probability measure $G_m^K = \sum_{k=1}^{K} \theta_{mk} \delta_{\beta_k}$ constructed for the $m$th node. Ishwaran & Zarepour (2002) proved that $\lim_{K \rightarrow \infty} G_m^K = G_m \sim \text{DP}(\alpha_0 \mu)$. In the limit $K \rightarrow \infty$, $\theta_{mk} = 0$ with probability one, while $\sum_k \theta_{mk} = 1$. In this case, the nonzero probability can be shown to be on a disjoint set of atoms for each $G_m$, with probability one, despite the fact that they share atoms in the finite approximation $G_m^K$. Practically speaking, this means that a state transition sequence sampled from the infinite limit of Markov M3 will never return to the same node twice, and so the model can equivalently be thought of as selecting a path in a tree. We formally state this in the following Proposition.

**Proposition 1** As $K$ goes to infinity, the Markov mixed membership model recovers the underlying mixing measure of the nested Chinese restaurant process.

### 2.3. Related work

Graph-based mixed membership models have been applied to grouped data in other settings. For example, mixed membership models have been applied to graph data, where instances are linked as a graph, and stochastic block models (Airoldi et al., 2008) have been proposed to model the links between instances through clustering. Mixed membership models have also been applied to collaborative filtering (Mackey et al., 2010; Wang & Blei, 2011) and link prediction (Chang & Blei, 2009). Exploring exchangeable structures in graphs has also received recent theoretical interest (Orbanz & Roy, 2015).

We also observe that mixed membership models can be applied to the more traditional use of hidden Markov models for sequential data. For example Paul (2014) considers a similarly named process for the fundamentally different problem of nonexchangeable sequence modeling.
3. Scalable Variational Inference

We derive a variational inference algorithm for learning an approximate posterior of all model variables shown in Algorithm 1. We can factorize the joint distribution of our model as

\[
p(w, \beta, \theta, z, u, \ell, \pi) = p(\beta)p(\theta)p(\pi) \times \prod_d p(w_d|\beta, z_d, \ell_d)p(\pi|\ell_d)p(u_d)p(\ell_d,u_d). \tag{5}
\]

We apply mean-field variational inference to approximate the posterior \(p(\beta, \theta, z, u, \ell|w)\) by defining a factorized \(q\) distribution on these variables and locally maximizing the variational objective function

\[
\mathcal{L} = E_q[\ln p(w, \beta, \theta, z, u, \ell)] - E_q[\ln q]
\]

using coordinate ascent, which approximately minimizes the KL-divergence between the true posterior and \(q\). We restrict \(q\) to the following factorized form

\[
q(\beta, \theta, \pi, z, u, \ell) = q(\beta)q(\theta)q(\pi)q(z)q(u)q(\ell), \tag{6}
\]

which we further factorize as

\[
q(\beta)q(\theta)q(\pi) = q(\pi|\alpha, \pi) \prod_{k=1}^K q(\beta_k|\lambda_k)q(\theta_k|\alpha_k),
\]

\[
q(z)q(u) = \prod_d q(z_d|\varphi_d)\prod_i q(u_d|a_{di}, b_{di}),
\]

\[
q(\ell) = \prod_d \prod_n q(\ell_{dn}|\phi_{dn}). \tag{7}
\]

We select all \(q\) distributions to be in the same family as the prior defined in Algorithm 1. In the following, we use coordinate ascent to optimize the variational objective, with respect to the variational parameters. We first discuss the update of local variables \(z_d, u_d, \ell_d\), which are only dependent on a single instance. Then we discuss the global variables \(\beta, \theta, \pi\) that depends on multiple instances. Though we have proposed a parametric model in the size of the graph, \(K\), we have defined the Markov chain for each group to be infinite in length. For inference, we introduce a truncation of this stick-breaking construction to level \(T\), as is typically done (Blei & Jordan, 2006). We note that truncation-free methods are a possible remedy (Wang & Blei, 2012).

3.1. Local variables

The most complex part of inference is in learning the Markov sequence \(z_d\) that selects atoms for group \(d\). We can derive the explicit form of its posterior from

\[
q(z_d) \propto \exp \left( \sum_i \mathbb{E}[\ln p(z_{di}|z_{di-1}, \theta)] \right)
\]

\[
+ \sum_i \sum_n \phi_{dn}(i) \mathbb{E}[\ln p(w_d|\ell_d,z_{di},\beta)], \tag{8}
\]

In Eq. (8) we can break \(q(z_d)\) into single potentials and pairwise potentials as indicated. We can then solve using the forward-backward algorithm to infer the posterior joint marginals. In fact, this is exactly the procedure for learning the state transitions of an HMM with the important difference that the emission at step \(i\) is not predefined, but instead the soft clustering of data in \(w_d\) induced by the variational parameters of \(\ell_d\), which is changing with each iteration. A simple modification can be made to the forward-backward algorithm that accounts for the new emission process (please see the supplemental material).

Let \(\chi^f_{di}(k)\) and \(\chi^b_{di}(k)\) be the outputs of the forward and backward algorithms, respectively. As with the HMM, we will use the quantities

\[
\varphi_{di}(k) \propto \chi^f_{di}(k)\chi^b_{di}(k) \tag{9}
\]

\[
\xi_{di}(k, k') \propto \chi^f_{di}(k)\chi^b_{di+1}(k'), \tag{10}
\]

where we define

\[
\kappa_{di}(k, k') = \exp \{\mathbb{E}[\ln \theta_{k,k'}] + \sum_n \phi_{dn}(k') \mathbb{E}[\ln \beta_{k',w_d}]\}.
\]

Normalization is over \(k\) for \(\varphi\) and the pair \((k, k')\) for \(\xi\). The multinomial variational parameter \(\phi_{dn}\) corresponds to the data allocation variable \(\ell_{dn}\), found by calculating

\[
\phi_{dn}(i) \propto \exp \left( \mathbb{E}[\ln u_{di}] + \sum_k \varphi_{di}(k) \mathbb{E}[\ln \beta_{k,w_d}] \right). \tag{11}
\]

We observe that the last term sums over atom assignments for the \(i\)th value in the Markov sequence. The first expectation is from the stick-break construction

\[
\mathbb{E}[\ln u_{di}] = \mathbb{E}[\ln u_{di}] + \sum_{j=1}^{i-1} \mathbb{E}[\ln(1-u_{dj})] \tag{12}
\]

The final local variables are the stick-breaking proportions \(u_d\). Given the allocation distributions \(q(\ell_{dn})\), the update of the beta \(q\) distribution of \(u_{di}\) is

\[
a_{di} = 1 + \sum_n \phi_{dn}(i), \quad b_{di} = \gamma_0 + \sum_{n,i' > i} \phi_{dn}(i'). \tag{13}
\]

We can iterate several times updating the local variables for each document before moving on to the global variables.

3.2. Global variables

The global variables include the Markov transition probabilities and the atoms. The update to the variational parameters of the initial state and transition probabilities, \(\pi\) and \(\theta\), are identical to the HMM,

\[
\alpha_{\pi,k} = \frac{\alpha_0}{K} + \sum_d \varphi_{d1}(k), \tag{14}
\]

\[
\alpha_{k,k'} = \frac{\alpha_0}{K} + \sum_d \sum_i \xi_{di}(k, k'). \tag{15}
\]
calculated as in Eq. (14)–(16). Then the stochastic updates the coordinate ascent updates restricted to $C$. Given the subset $C$, the decaying learning rate set local calculations to a small subset $C$. The right-most summation calculates the probability that word $w_{dn}$ and topic $\beta_k$ are both assigned to the same point in the Markov sequence $(z_{d1}, z_{d2}, \ldots)$, which is required for word $w_{dn}$ to belong to component $\beta_k$.

3.3. Stochastic variational inference

Since Markov M3 is a conjugate-exponential family model, it is immediately amenable to stochastic variational inference (SVI) (Hoffman et al., 2013). For models such as tree-based models and the proposed graph-based model, such large data extensions can help in learning the greater level of structure defined by the model prior. As with other mixed membership models, we can exploit the fact that the variational objective function factorizes as

$$L = -E_q[\ln q] + E_q[\ln p(\beta, \theta, \pi)] + \sum_d E_q[\ln p(w_d, z_d, u_d, \ell_d|\beta, \theta, \pi)].$$

Using SVI, we stochastically optimize $L$ by restricting the local calculations to a small subset $C_t$ of the $D$ groups at iteration $t$. Given the subset $C_t$, SVI proceeds by (1) optimizing all local variables indexed by $C_t$, (2) forming the global updates restricted to $C_t$, and (3) averaging these updates with the current values. Let $\hat{\alpha}_{\pi, k}$, $\hat{\alpha}_{k, k'}$, and $\hat{\lambda}_{kv}$ be the coordinate ascent updates restricted to $C_t$. These are calculated as in Eq. (14)–(16). Then the stochastic updates to the true values are

$$\alpha_{\pi, k}'(t+1) = (1 - \rho_t)\alpha_{\pi, k}'(t) + \rho_t(D/C_t)\hat{\alpha}_{\pi, k},$$

$$\alpha_{k, k'}'(t+1) = (1 - \rho_t)\alpha_{k, k'}'(t) + \rho_t(D/C_t)\hat{\alpha}_{k, k'},$$

$$\lambda_{kv}'(t+1) = (1 - \rho_t)\lambda_{kv}'(t) + \rho_t(D/C_t)\hat{\lambda}_{kv}.$$ (18)

The decaying learning rate $\rho_t$ must satisfy $\sum_{t=1}^{\infty} \rho_t = \infty$, $\sum_{t=1}^{\infty} \rho_t^2 < \infty$ to ensure convergence (Bottou, 1998). We set $\rho_t = (\tau_0 + t)^{-\kappa}$, where $\tau_0 > 0$ and $0.5 < \kappa \leq 1$.

4. Experiments

Our experiments with Markov M3 focus on grouped discrete data problems. We first consider topic modeling on small and large scale problems. We then show qualitative results on a music tagging problem, where the union of quantized song features and user tags provides the discrete grouping on a song level.

4.1. Document modeling

**Batch comparisons.** We first apply our model to three datasets easily learned with batch inference: *Huffington Post*, *Science* and *NIPS* papers. The statistics from each data set is shown in Table 1. We split each data set into a training set and a test set, also shown in Table 1. For the testing set we split each document into a 90/10 split and learned local parameters on 90% of words in the document and predicted 10% for prediction given the inferred topic proportions. We present the quantitative performance using perplexity, which can be calculated as

$$\text{perplexity} = \exp \left( - \frac{\sum_{n \in w_{TS}} \log p(w_n|w_{TR})}{|w_{TS}|} \right),$$

where $w_{TR}, w_{TS}$ represent training and test words in the test set respectively.

We compare performance of our model with LDA, the correlated topic model (CTM) (Blei & Lafferty, 2007) and the nested HDP (nHDP) (Paisley et al., 2015). The nHDP was shown to give better predictive ability than the nCRP, and so we do not compare with that algorithm. We have also noted that when $K$ goes to infinity, the Markov M3 recovers the nCRP mixed membership model and so performance of our model tends to the nCRP.

The perplexity results using a different number of topic are shown in Figure 2. For the nonparametric nHDP we truncate its posterior topic number to 175, which is higher than the maximum number of topics used by the parametric models. On all datasets, the Markov M3 consistently performs better than other parametric models. The reason is that Markov M3 can more flexibly model all pairwise topic dependencies, while LDA only considers there to be a slight negative correlation among topics, and CTM considers pairwise correlation without dependency information. We observe that the best performance for Markov M3 is better than nHDP, which gives evidence that a graph struc-
Stochastic learning. For a large-scale problem, we train our model using stochastic variational inference on the *New York Times* dataset, which contains 1.8 million documents, and compare its predictive performance on a held-out test set with online LDA (Hoffman et al., 2010a). For both models we use the same topic initialization. We also use a learning rate of \((10 + t)^{-0.75}\) for both models, and a batch size of \(|C_t| = 500\). For Markov M3, we truncate the path length to 15 and set \(\gamma_0 = 1\).

In Figure 3, we show the predictive performance throughout the learning process of both models, considering various number of topics. We see that the stochastic version of Markov M3 performs better than online LDA, which is consistent with the results on smaller scale problems.

Markov M3 learns a transition distribution over all topics. In Fig. 5, we depict the most probable transitions from several topics within the graph. We limit connections and directions to those above a threshold for clarity and use size to roughly indicate probability. As can be seen, most transition probabilities between topics are very low, thus are not displayed on the graph. Among all the topics, a few of them are general (e.g., topic 67 with top words 'say', 'life', 'man'), but most topics are more specific. Topics naturally form small cliques, where the transition probability within a cluster is significantly higher than transitions across clusters. There are also connections between topics in different
but similar domains. For example, there is a high transition probability between a “film” topic (18) to a “music” topic (46).

We further illustrate the Markov property by focusing on the document level. Since each document learns a distribution on paths through the graph, we show the most probable path found using the Viterbi algorithm. We show the paths selected for three documents in Figure 4, where the arrow indicates path direction and the size of the node indicates the proportion for that topic. In general, Markov M3 tends to visit earlier topics with larger proportions, as is encouraged by the stick-breaking prior. The topics selected in these examples are clearly interpretable, and the sequence captures information about topics relations.

**Sensitivity analysis.** We empirically analyze the effect of the truncation level of the Markov chain and the stick-breaking concentration parameter in our model. The truncation level indicates the number of topics we allow to each document. When the truncation level is too small, each document can only explore very limited number topics. As Figure 6 shows, the model is less accurate in this case.

However, when the length of Markov random walk become too large, performance also decreases. This is because the update to the transition distributions $\theta_k$ treat the entire sequence of each document equally. If there are many empty topics in the Markov sequence, as indicated by the $q$ distributions on $u_{dt}$, the information in the $q$ distributions of the transition matrix can become less informative. In this sense, truncation of the model is important and should be set so that most available topics are used by a document. The concentration parameter $\gamma$ defines the smoothness for the stick-breaking proportions. Figure 6 shows that choosing a smooth prior can help improve performance.
we learn a distribution on paths, and we only show the most probable path using Viterbi. For this testing problem, since we don’t have any tags, we feed the quantized audio features into our model and to learn their Markov transitions over factors. We then use the tagging portion of the selected atoms to represent the path selected. From Figure 7, we find that Markov M3 pulls out the correct, but noisy tags (marked as red) with high probability, and also discovers other possibly relevant tags.

5. Conclusions and Future Work

We proposed a Markov mixed membership model (Markov M3) that explores a fully connected graph structure among components. Markov M3 provides a new way of performing mixed membership modeling with structured distributions, and an alternative to similar tree-based models. We showed how Markov M3 gives a new perspective on the nCRP by showing nCRP to be a limiting case of Markov M3. We showed the effectiveness of this modeling framework on small and large datasets for discrete grouped data such as documents and quantized music. In future work, we are interested in exploring an Bayesian nonparametric extension of Markov M3 that still allows its components to be fully connected. Another direction is to develop extensions of Markov M3 to problems with nonexchangeable data.

References


Wang, C. and Blei, D. Collaborative topic modeling for recommending scientific articles. In Knowledge Discovery and Data Mining, 2011.