SIGNAL DECOMPOSITION USING MULTISCALE ADMIXTURE MODELS

Matus Telgarsky and John Lafferty

Computer Science Department, and
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213 USA

ABSTRACT

Admixture models are "mixtures of mixtures" that decompose an object into multiple latent components, with the component proportions varying stochastically across objects. Recent work in machine learning has successfully developed admixture models for text, and work in population genetics has developed such models to analyze complex groups of individuals having mixed ancestry. We introduce a family of graphical admixture models for decomposing a signal into multiple components based on a wavelet representation of the signal. Two models are developed, one using a fixed segmentation of the signal, another using recursive dyadic partitioning. Variational algorithms are derived for inferring mixture proportions and estimating parameters.

Index Terms— Graphical model, wavelets, variational inference, recursive dyadic partitioning, unsupervised signal segmentation and labeling.

1. INTRODUCTION

Many types of data can be interpreted as being naturally constructed from several different components. A single text article, for example, may be composed on a few different themes or topics; an organism within a large population may have a genotype that reflects an ancestry from multiple subpopulations. The underlying components are latent, but might be revealed from either expert knowledge or a statistical analysis of large amounts of data. Recent research in machine learning has developed powerful new tools for automating such analysis.

The latent Dirichlet allocation model [1] was proposed by Blei et al. for decomposing text documents into latent "topics," which are the high level themes in a large document collection. This basic technique is serving as a foundation for new tools for automated document analysis [2, 3]; similar models have been independently developed for population genetics [4]. Such hierarchical probabilistic models have been generalized to other kinds of data as well, notably to natural images that have been presegmented into chunks [5, 6, 7]. These models provide a method for dimensionality reduction for large collections of unstructured data, which can be useful for information retrieval, collaborative filtering, classification, and topic-directed browsing.

In the latent Dirichlet allocation model, the mixing proportions are randomly drawn for each instance, while the mixture components, or topics, are shared across documents. The words of each document are assumed to be independently drawn from the resulting mixture of multinomials. Thus, the model is exchangeable; if the words in a document are jumbled up, the model's predictions remain the same. For many types of data however, order matters. In financial time series, natural images, or acoustic signals, the ordering of the observations carries crucial information, which an exchangeable model ignores. Scramble the words in a document, and a human reader can still easily discern the main themes; scramble the pixels in an image or the notes in a musical piece, and the result will appear unintelligible.

In this paper we integrate latent topic models with graphical models of signals based on wavelet representations, combining the efficiency of multiscale methods with the flexibility of Markov random fields. Seminal work in this direction includes [8, 9, 10]. In the work of Crouse et al. [10], for instance, hidden Markov trees are used to model the wavelet coefficients for signals drawn from different classes. However, the classes are assumed to be known in labeled training data. While this approach can be used to develop segmentation procedures to label composite signals, it cannot be used to automatically *discover* the underlying signal components. Motivated by latent topic models of text that can discover the semantic themes in a document collection, our goal is to develop admixture models for signals that uncover the primary building blocks from which those signals are composed.

2. LATENT TOPIC MODELS FOR SIGNALS

The latent Dirichlet allocation model [1] is a simple, but elegant and effective graphical model used to decompose text

Research supported in part by NSF grants IIS-0312814 and IIS-0427206, and by a grant from Google.

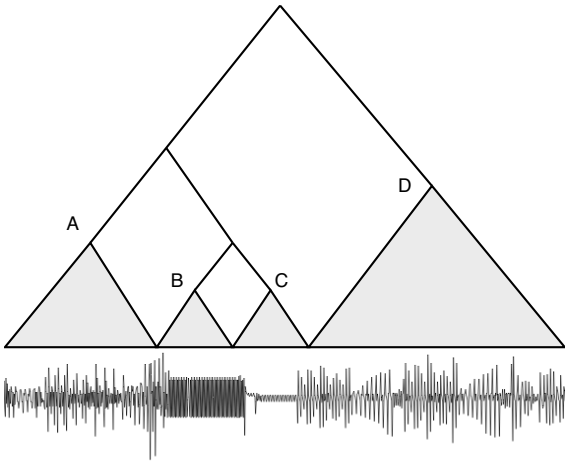

Fig. 1. Excerpt from the Chopin étude in C, encoded as note differences, extracted from MIDI. Multiple "themes" or "textures" are apparent. The modeling approach is to posit that different segments, under dyadic partitioning, are generated from wavelet coefficient models associated with a corresponding latent variable at each leaf. The different models constitute the "themes" or "topics" learned across a large corpus of signals.

corpora into semantic "topics." The model is generative, and synthesizes a document by first drawing a weighting over the topics, then independently sampling a topic for each word position, and finally sampling each word from the multinomial corresponding to its assigned topic. While the words are conditionally independent given the mixing proportions, topically related words are highly correlated.

In more detail, suppose that $X = \{X_i\}_{i=1}^N$ is a document, thought of as a "bag of words," with $X_i \in \mathbb{V}$ for some fixed vocabulary. The LDA model generates X according to the process

$$\begin{aligned} \theta &\sim \text{Dirichlet}(\alpha) \\ Z_i \,|\, \theta &\sim \text{Mult}(\theta), \text{ for each } i \\ X_i \,|\, Z_i &\sim \text{Mult}(\beta_{Z_i}), \text{ for each } i \end{aligned}$$

The document likelihood is thus given by the computationally intractable integral

$$p(X \,|\, \alpha, \beta) = \int p(\theta \,|\, \alpha) \prod_{i=1}^{N} \sum_a \theta_a \, p(X_i \,|\, \beta_a) \, d\theta$$

We will modify this model by assuming X is a signal represented in a wavelet basis, and that there is a latent labeled dyadic segmentation of X, with each leaf generated from a mixture model associated with the corresponding label. We first describe our approach at an intuitive level, and specify it in more detail in the following section.

To generate a one-dimensional signal of length 2^n, we first generate a tree over the signal using recursive dyadic partitioning; a prior over tree depth can be incorporated to favor trees that are not too detailed. For each leaf $\ell \in \text{leaves}(T)$ in this segmentation tree, a topic Z_ℓ is sampled from the topic mixture vector θ. The portion of the signal spanned by node ℓ is generated by sampling the wavelet coefficients for this portion of the signal, according to an appropriate model. A simple, yet still effective choice is to sample the coefficients from a mixture of Gaussians associated with topic Z_ℓ; more sophisticated models based on hidden Markov trees [10] are alternatives. Thus, the model takes the following generative form:

$$\begin{aligned} T &\sim \text{Tree}(\delta) \\ \theta &\sim \text{Dirichlet}(\alpha) \\ Z_\ell \,|\, \theta &\sim \text{Mult}(\theta), \text{ for each leaf } \ell \\ X_\ell \,|\, Z_\ell &\sim \text{Mixture}\left(\mu^{(Z_\ell)}, \sigma^{(Z_\ell)}\right), \text{ for each leaf } \ell \end{aligned}$$

Here $\text{Tree}(\delta)$ denotes a prior model over dyadic partitions; for example $p(T) \propto \exp(-\delta|T|)$. The use of dyadic partitions and variational approximations enables dynamic programming algorithms that efficiently sum over all labelings and segmentations of the signal, as described below.

3. VARIATIONAL INFERENCE FOR TWO MODELS

Using the paradigm of seeding more complex models with simpler models [11], we first train Model 1 using fixed segmentations, and use it to initialize Model 2, which sums over all dyadic partitions.

3.1. Model 1

Our simplest model begins with a single segmentation of each signal into segments of fixed length (δ); the signal segments correspond to words in an LDA model. Let X_ℓ denote signal segment ℓ and let $\widehat{X}_{\ell i}$ denote the wavelet coefficients for the segment (other orthogonal function representations can be used as well, including overcomplete bases). A simple model of the coefficients is independent sampling under a Gaussian mixture:

$$p(X_\ell \,|\, a) \;=\; \prod_{i \in \ell} \sum_k \lambda_k^{(a)} p\left(X_{\ell i} \,|\, \mu_k^{(a)}, \sigma_k^{(a)}\right)$$

The mixture models may vary with the level in the multiscale wavelet decomposition, or they may be tied according to a hidden Markov tree. The data likelihood is given by

$$p(X \,|\, \alpha, \mu, \sigma) = \int p(\theta \,|\, \alpha) \prod_\ell \sum_a \theta_a \, p(X_\ell \,|\, \mu^{(a)}, \sigma^{(a)}) \, d\theta$$

The mean-field variational algorithm for Model 1 parallels the algorithm for LDA [1]. The variational approximation is factored according to

$$q_X(\theta, \{Z_\ell\}) = q(\theta \,|\, \gamma) \prod_\ell \varphi_{\ell, Z_\ell}$$

where γ are Dirichlet parameters and $\varphi_{\ell,a}$ are multinomial parameters, representing the estimated posterior probability that segment ℓ is generated from topic a. Using a coordinate ascent algorithm, the variational lower bound on the signal likelihood is maximized by iterating the following steps until convergence:

$$\begin{aligned}\gamma_a &= \alpha_a + \sum_\ell \varphi_{\ell,a} \\ \varphi_{\ell,a} &\propto p(X_\ell \,|\, \mu^{(a)}, \sigma^{(a)}) \, \exp(\Psi(\gamma_a))\end{aligned}$$

Here Ψ denotes the digamma function. The estimated mixture weights and labels for each segment are then given by

$$\mathbb{E}_q(\theta_a) = \frac{\gamma_a}{\sum_b \gamma_b}, \quad \widehat{a}(\ell) = \arg\max_a \; \varphi_{\ell,a}$$

To estimate the model parameters, the variational parameters $\phi_{\ell a}$ are used in a variational E-step, updating the sufficient statistics of the leaf topic models.

3.2. Model 2

In Model 2 we incorporate a segmentation model, using recursive dyadic partitioning. We present the technique for one-dimensional signals, but it extends easily to two dimensions. Let $p(T)$ be a distribution over all dyadic trees over X. We require that $p(T)$ is additive on the leaves. The likelihood now sums over all segmentations:

$$p(X \,|\, \delta, \alpha, \mu, \sigma) = \sum_T p(T \,|\, \delta) \int p(\theta \,|\, \alpha) \prod_{\ell \in \text{leaves}(T)} \sum_a \theta_a \, p\left(X_\ell \,|\, \mu^{(a)}, \sigma^{(a)}\right) d\theta$$

Variational inference in this case requires dynamic programming. The variational approximation takes the factored form

$$q_X(T, \theta, Z) = q(T \,|\, \eta) \, q(\theta \,|\, \gamma) \, q(Z \,|\, \varphi)$$

We take $q(\theta \,|\, \gamma)$ to be Dirichlet with parameter γ, and take $q(Z \,|\, \varphi) = \prod_\ell \varphi_{\ell,Z_\ell}$ to be a product of multinomials, as before. The variational distribution over trees yields a segmentation model for the signal; we take this to be a branching process

$$q(T \,|\, \eta) = \prod_{\ell \in \text{leaves}(T)} \eta_\ell \prod_{n \in \text{internal}(T)} (1 - \eta_n)$$

Here the Bernoulli parameter η_n indicates whether or not node n is split; at the lowest level we require $\eta_n = 1$. The variational parameters are now γ, η, and φ. The structure of the signal emerges in the tree model $q(T \,|\, \eta)$ and the segment topic probabilities $\{\varphi_{\ell a}\}$. The variational lower bound on the log-likelihood is

$$\begin{aligned}\ell(X) \geq\; & \mathbb{E}_q \log p(\theta) + \mathbb{E}_q \log p(T) \\ & + \sum_\ell q(\ell \in \text{leaves}(T)) \sum_a \varphi_{\ell a} \left(\mathbb{E}_q \log \theta_a + \log p(X_\ell \,|\, a)\right) \\ & + H(q(\theta)) + H(q(T)) + \mathbb{E}_q H(q(Z \,|\, T))\end{aligned}$$

with the probability $q(\ell \in \text{leaves}(T))$ calculated as

$$q(\ell \in \text{leaves}(T)) = \eta_\ell \prod_{n \in \text{anc}(\ell)} (1 - \eta_n)$$

The entropy of the variational distribution over trees is

$$H(q(T)) = \sum_\ell H(\eta_\ell) \prod_{n \in \text{anc}(\ell)} (1 - \eta_n)$$

and the (average) entropy $\mathbb{E}_q(H(q(Z \,|\, T)))$ is

$$\mathbb{E}_q H(q(Z \,|\, T)) = -\sum_\ell q(\ell \in \text{leaves}(T)) \sum_a \varphi_{\ell a} \log \varphi_{\ell a}$$

3.2.1. Iterative updates for variational parameters

We adopt a coordinate ascent algorithm where φ and γ are estimated holding η fixed, and then η is estimated holding φ and γ fixed. The updates for γ and φ are

$$\begin{aligned}\gamma_a &= \alpha_a + \sum_\ell q(\ell \in \text{leaves}(T)) \varphi_{\ell,a} \\ \varphi_{\ell,a} &\propto p(X_\ell \,|\, \mu^{(a)}, \sigma^{(a)}) \, \exp(\Psi(\gamma_a))\end{aligned}$$

Now holding φ and γ fixed, and assuming for simplicity that $p(T)$ is uniform, the updates for η_ℓ can be shown to satisfy

$$\begin{aligned}\log\left(\frac{\eta_\ell}{1 - \eta_\ell}\right) = & -\Lambda_{\text{left}(\ell)} - \Lambda_{\text{right}(\ell)} \\ & + \sum_a \varphi_{\ell a} \left(\mathbb{E}_q \log \theta_a + \log p(X_\ell \,|\, a) - \log \varphi_{\ell a}\right)\end{aligned}$$

where Λ_n is computed recursively as

$$\begin{aligned}\Lambda_n &= (1 - \eta_n) \left(\Lambda_{\text{left}(n)} + \Lambda_{\text{right}(n)}\right) \\ &+ \eta_n \sum_a \varphi_{\ell a} \left(\mathbb{E}_q \log \theta_a + \log p(X_\ell \,|\, a) - \log \varphi_{\ell a}\right)\end{aligned}$$

After variational inference is carried out, the estimated labels and segmentation is given by

$$(\widehat{T}, \{\widehat{a}(\ell)\}) = \arg\max_{T,a} \; q(T) \prod_{\ell \in \text{leaves}(T)} \varphi_{\ell,a}$$

where the $\arg\max$ is computed using dynamic programming.

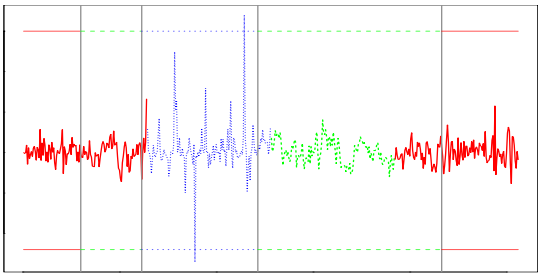

Fig. 2. Synthetic signal with true segmentation demarcated by vertical lines, and true labeling given by horizontal lines. Varying line types of signal indicate labeling as inferred by Model 1, with four segments spanning the five component signal.

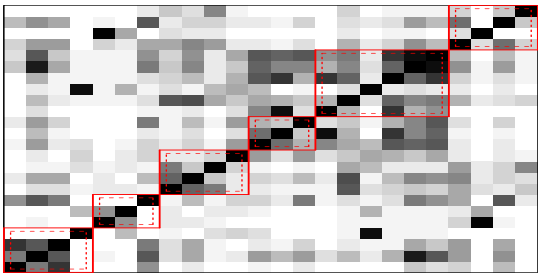

Fig. 3. Hellinger distance between topic weights of 24 music pieces, filtered through a sigmoid; the boxes delineate each composer, from left to right: Bach, Chopin, Debussy, Haydn, Mozart, Scriabin. The strongest inter-composer correlations are between Mozart and Haydn, with some correlation between Bach's Goldberg variations and Mozart's early works, both of which are also the most thematically cohesive works, as inferred by their dark regions.

4. EXPERIMENTAL RESULTS

Testing data came from two sources: synthetic autoregressive signals (some in the same form used in [10]) and key-normalized MIDI files. The synthetic data provides grounds to flex the segmentation muscles, whereas the music data is to show the models' ability to correlate examples.

The simplest synthetic data case, that of a signal composed of fixed length samples of different signal types concatenated together, generally achieves zero classification error when analyzed using Model 1, even in difficult situations where multiple signals of the same form are present. More realistically, the component lengths were sampled from a Poisson process, and Model 1 must still follow fixed boundaries. Figure 4 illustrates such an example. Here there are four signal types ("topics") whereas only three are used in the learned model. The model generalizes and uses one topic to cover two similar topics (not shown), with another learned topic having high variance to cover situations where the segmentation window occupies different topics evenly. Even with four topics this situation persists. Model 2 handles these cases elegantly, although it is much slower to train.

Future work includes further development of the models, including the use of hidden Markov trees to model wavelet coefficients, and application to both music data and multi-modal forms, such as pictures and text. All source code is planned for Internet release.

5. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, January 2003.

[2] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smith, "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. 2004, pp. 487–494, AUAI Press.

[3] David M. Blei and John D. Lafferty, "Dynamic topic models," in *Proceedings of the Twenty-third International Conference on Machine Learning (ICML 2006)*, A. Moore and W. Cohen, Eds. IMLS/ICML, 2006.

[4] J. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, pp. 945–959, June 2000.

[5] D. Blei and M. Jordan, "Modeling annotated data," in *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2003, pp. 127–134, ACM Press.

[6] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *IEEE Computer Vision and Pattern Recognition*, 2005.

[7] J. Sivic, B. Rusell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *International Conference on Computer Vision (ICCV 2005)*, October 2005.

[8] M. Luettgen, W. Karl, A. Willsky, and R. Tenney, "Multiscale representations of Markov random fields," *IEEE Trans. Signal Proc.*, vol. 41, no. 12, pp. 3377–3395, 1993.

[9] C. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Proc.*, vol. 3, no. 2, pp. 162–177, 1994.

[10] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, 1998.

[11] Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.