Probabilistic Multilevel Clustering via Composite Transportation Distance

Nhat Ho[⋆,†] **Viet Huynh**[⋆,‡] **Dinh Phung**[‡] **Michael I. Jordan**[†]
[†] University of California at Berkeley, Berkeley, USA [‡]Monash University, Australia

Abstract

We propose a novel probabilistic approach to multilevel clustering problems based on composite transportation distance, which is a variant of transportation distance where the underlying metric is Kullback-Leibler divergence. Our method involves solving a joint optimization problem over spaces of probability measures to simultaneously discover grouping structures within groups and among groups. By exploiting the connection of our method to the problem of finding composite transportation barycenters, we develop fast and efficient optimization algorithms even for potentially large-scale multilevel datasets. Finally, we present experimental results with both synthetic and real data to demonstrate the efficiency and scalability of the proposed approach.

1 Introduction

Clustering is a classic and fundamental problem in machine learning. Popular clustering methods such as K-means and mixture models have been the workhorses of exploratory data analysis. However, the underlying model for such methods is a simple flat partition or a mixture model, which do not capture multilevel structures (e.g., words are grouped into documents, documents are grouped into corpora) that arise in many applications in the physical, biological or cognitive sciences. The clustering of multilevel structured data calls for novel methodologies beyond classical clustering.

One natural approach for capturing multilevel structures is to use a hierarchy in which data are clustered locally into groups, and those groups are partitioned in a "global clustering." Attempts to develop algorithms of this kind can be roughly classified into two categories. The first category makes use of probabilistic models, often based on Dirichlet process priors. Examples in this vein include the Hierarchical Dirichlet Process (HDP) [23], Nested Dirichlet Process (NDP) [19], Multilevel Clustering with Context (MC^2) [15], and Multilevel Clustering Hierarchical Dirichlet Process (MLC-HDP) [26]. Despite the flexibility and solid statistical foundation of these models, they have seen limited application to large-scale datasets, given concerns about the computational scaling of the sampling-based algorithms that are generally used for inference under these models.

A second category of multilevel methods is based on tools from optimal transport theory, where algorithms such as Wasserstein barycenters provide scalable computation [4, 5]. These methods trace their origins to a seminal paper by Pollard [18] which established a connection between the K-means algorithm and the problem of determining a discrete probability measure that is close in Wasserstein distance [24] to the empirical measure of the data. Based on this connection, it is possible to use Wasserstein distance to develop a combined local/global multilevel clustering method [7].

The specific multilevel clustering method proposed in [7] has, however, its limitations. Most notably, as that method uses K-means as a building block, it is only applicable to continuous data. When being used to cluster discrete data, it yields poor results. In this work, we make use of a novel form of transportation distance, which is termed as *composite transportation distance* [16], to overcome this limitation, and to provide a more general multilevel clustering method. The salient feature of composite transportation distance is that it utilizes Kullback-Leibler (KL) divergence as the underlying metric of optimal transportation distance, in contrast to the standard Euclidean metric that has been used in optimal transportation approaches to

[⋆] Nhat Ho and Viet Huynh contributed equally to this work and are in alphabetical order.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

clustering to date.

In order to motivate our use of composite transportation distance, we start with a one-level structure data in which the data are generated from a finite mixture model, e.g., a mixture of (multivariate) Gaussian distributions or multinomial distributions. Unlike traditional estimators such as maximum likelihood estimation (MLE), the high-level idea of using composite transportation distance is to determine optimal parameters to minimize the KL cost of moving the likelihood from one cluster to another cluster. Intuitively, with such a distance, we can employ the underlying geometric structure of parameters to perform efficient clustering with the data. Another advantage of composite transportation distance is its flexibility to generalize to multilevel structure data. More precisely, by representing each group in a multilevel clustering problem by an unknown mixture model (local clustering), we can determine the optimal parameters, which can be represented as local (probability) measures, of each group via optimization problems based on composite transportation distance. Then, in order to determine global clustering among these groups, we perform a composite transportation barycenter problem over the local measures to obtain a global measure over the space of mixture models, which serves as a partition of these groups. As a result, our final method, which we refer to as *multilevel composite transportation* (MCT), involves solving a joint optimal transport optimization problem with respect to both a local clustering and a global clustering based on the cost matrix encoding KL divergence among atoms. The solution strategy involves using the fast computation method of Wasserstein barycenters combined with coordinate descent.

In summary, our main contributions are the following: (i) A new optimization formulation for clustering based on a variety of multilevel data types, including both continuous and discrete observations, based on composite transportation distance; (ii) We provide a highly scalable solution strategy for this optimization formulation; (iii) Although our approach avoids the use of the Dirichlet process as a building block, the approach has much of the flexibility the hierarchical Dirichlet process in its ability to share atoms among local clusterings. We thus are able to borrow strength among clusters, which improves statistical efficiency under certain applications, e.g., image annotation in computer vision.

The paper is organized as follows. Section 2 provides preliminary background on composite transportation distance and composite transportation barycenters. Section 3 formulates the multilevel composite transportation optimization model, while Section 4 presents simulation studies with both synthetic and real data. Finally, we conclude the paper with a discussion in Section 5. Technical details of proofs and algorithm development are provided in the Supplementary Material.

Although our model focuses on the finite mixture case, one can add a regularization term to control the complexity of the model (aka the number of clusters) similar to DP-means [12] or use the (Poisson) prior as a regularization [14].

2 Composite transportation distance

Throughout this paper, we let Θ be a bounded subset of $\mathbb{R}^d$ for a given dimension $d \geq 1$. Additionally, $\{f(x|\theta),\ \theta \in \Theta\}$ is a given exponential family of distributions with natural parameter θ:

$$f(x|\theta) := h(x) \exp\left(\langle T(x), \theta\rangle - A(\theta)\right),$$

where $A(\theta)$ is the log-partition function which is convex. We define P_θ to be the probability distribution whose density function is $f(x|\theta)$. Given a fixed number of K components, we denote a finite mixture distribution as follows:

$$P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K} := \sum_{k=1}^{K} \omega_k P_{\theta_k}, \tag{1}$$

where $\boldsymbol{\omega}_K = (\omega_1, \ldots, \omega_K) \in \Delta^K$, which is a probability simplex in $K-1$ dimensions, and $\boldsymbol{\Theta}_K = \{\theta_k\}_{k=1}^K \in \Theta^K$ are the weights and atoms. Then, the probability density function of mixture model can be expressed

$$p_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}(\boldsymbol{x}) := \sum_{k=1}^{K} \omega_k f(x|\theta_k).$$

We also use $Q_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}$ to denote a finite mixture of at most K components to avoid potential notational clutter.

2.1 Composite transportation distance

For any two finite mixture probability distributions $P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}$ and $P_{\boldsymbol{\omega}'_{K'}, \boldsymbol{\Theta}'_{K'}}$ and any two given numbers K and K', we define the composite transportation distance between $P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}$ and $P_{\boldsymbol{\omega}'_{K'}, \boldsymbol{\Theta}'_{K'}}$ as follows

$$\widehat{W}(P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}, P_{\boldsymbol{\omega}'_{K'}, \boldsymbol{\Theta}'_{K'}}) := \inf_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\omega}_K, \boldsymbol{\omega}'_{K'})} \langle \boldsymbol{\pi}, \boldsymbol{M} \rangle, \tag{2}$$

where the cost matrix $\boldsymbol{M} = (M_{ij})$ satisfies $M_{ij} = \mathrm{KL}(f(x|\theta_i), f(x|\theta'_j))$ for $1 \leq i \leq K$ and $1 \leq j \leq K'$. Here, $\langle ., . \rangle$ denotes the dot product (or Frobenius inner product) of two matrices and $\Pi(\boldsymbol{\omega}_K, \boldsymbol{\omega}'_{K'})$ is the set of all probability measures (or equivalently transportation plans) $\boldsymbol{\pi}$ on $[0,1]^{K \times K'}$ that have marginals $\boldsymbol{\omega}_K$ and $\boldsymbol{\omega}'_{K'}$ respectively.

Detailed form of cost matrix Since $f(x|\theta)$ is an exponential family, we can compute $\mathrm{KL}\left(f(x|\theta), f(x|\theta')\right)$ in closed form as follows [27, 11, Ch.8]:

$$\mathrm{KL}\left(f(x|\theta'), f(x|\theta)\right) = D_A\left(\theta, \theta'\right),$$

where $D_A\left(\cdot,\cdot\right)$ is the Bregman divergence associated with log-partition function $A\left(\cdot\right)$ of f, i.e.,

$$D_A\left(\theta, \theta'\right) = A\left(\theta\right) - A\left(\theta'\right) - \left\langle \nabla A\left(\theta'\right), \left(\theta - \theta'\right)\right\rangle.$$

Therefore, the cost matrix $\boldsymbol{M}$ has an explicit form

$$M_{ij} = A\left(\theta_i\right) - A\left(\theta'_j\right) - \left\langle \nabla A\left(\theta'_j\right), \left(\theta_i - \theta'_j\right)\right\rangle \quad (3)$$

for $1 \leq i \leq K$ and $1 \leq j \leq K'$.

Composite transportation distance on the space of finite mixtures of finite mixtures We can recursively define finite mixtures of finite mixtures, and define a suitable version of composite transportation distance on this abstract space. In particular, consider a collection of N finite mixture probability distributions with at most K components $\left\{P_{\boldsymbol{\omega}^i_K, \boldsymbol{\Theta}^i_K}\right\}_{i=1}^N$ and a collection of $\bar{N}$ finite mixture probability distributions with at most $\bar{K}$ components $\left\{P_{\overline{\boldsymbol{\omega}}^i_{\bar{K}}, \overline{\boldsymbol{\Theta}}^i_{\bar{K}}}\right\}_{i=1}^{\bar{N}}$. We define two finite mixtures of these distributions as follows

$$\mathcal{P} = \sum_{i=1}^N \tau_i P_{\boldsymbol{\omega}^i_K, \boldsymbol{\Theta}^i_K},\ \mathcal{Q} = \sum_{i=1}^{\bar{N}} \overline{\tau}_i P_{\overline{\boldsymbol{\omega}}^i_{\bar{K}}, \overline{\boldsymbol{\Theta}}^i_{\bar{K}}},$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_N) \in \Delta^N$ and $\overline{\boldsymbol{\tau}} = (\overline{\tau}_1, \ldots, \overline{\tau}_{\bar{N}}) \in \Delta^{\bar{N}}$. Then, the composite transportation distance between $\mathcal{P}$ and $\mathcal{Q}$ is

$$\widehat{W}(\mathcal{P}, \mathcal{Q}) := \inf_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\tau}, \overline{\boldsymbol{\tau}})} \left\langle \boldsymbol{\pi}, \overline{\boldsymbol{M}}\right\rangle,$$

where the cost matrix $\overline{\boldsymbol{M}} = \left\{\overline{M}_{ij}\right\}$ is defined as

$$\overline{M}_{ij} = \widehat{W}(P_{\boldsymbol{\omega}^i_K, \boldsymbol{\Theta}^i_K}, P_{\overline{\boldsymbol{\omega}}^j_{\bar{K}}, \overline{\boldsymbol{\Theta}}^j_{\bar{K}}}),$$

for $1 \leq i \leq N$ and $1 \leq j \leq \bar{N}$. Note that, in a slight notational abuse, $\widehat{W}(.,.)$ is used for both the finite mixtures and finite mixtures of finite mixtures.

2.2 Learning finite mixtures with composite transportation distance

In this section, we assume that $X_1, \ldots, X_n$ are i.i.d. samples from the mixture density $p_{\boldsymbol{\omega}^0_{k_0}, \boldsymbol{\Theta}^0_{k_0}}(x) = \sum_{i=1}^{k_0} \omega_i^0 f(x|\theta_i^0)$, where $k_0 < \infty$ is the true number of components. Since k_0 is generally unknown, we fit this model by a mixture of K distributions where $K \geq k_0$.

Note that the order of parameters is reversed in KL and Bregman divergence.

2.2.1 Inference with composite transportation distance

Denote $P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ as an empirical measure with respect to samples $X_1, \ldots, X_n$. To facilitate the discussion, we define the following composite transportation distance between an empirical measure P_n and the mixture probability distribution $P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}$

$$\widehat{W}(P_n, P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}) := \inf_{\boldsymbol{\pi} \in \Pi\left(\frac{1}{n}\mathbf{1}_n, \boldsymbol{\omega}_K\right)} \left\langle \boldsymbol{\pi}, \boldsymbol{M}\right\rangle, \quad (4)$$

where $\boldsymbol{M} = (M_{ij}) \in \mathbb{R}^{n \times K}$ is a cost matrix defined as $M_{ij} := -\log f(X_i|\theta_j)$ for $1 \leq i \leq n, 1 \leq j \leq K$. Furthermore, $\Pi\left(\cdot,\cdot\right)$ is the set of transportation plans between $\mathbf{1}_n/n$ and $\boldsymbol{\omega}_K$.

To estimate the true weights $\boldsymbol{\omega}^0_{k_0}$ and true components θ_i^0 as $1 \leq i \leq k_0$, we perform an optimization with transportation distance $\widehat{W}$ as follows:

$$(\widehat{\boldsymbol{\omega}}_{n,K}, \widehat{\boldsymbol{\Theta}}_{n,K}) = \arg\min_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K} \widehat{W}(P_n, P_{\boldsymbol{\omega}_K, \boldsymbol{\Theta}_K}). \quad (5)$$

The estimator $(\widehat{\boldsymbol{\omega}}_{n,K}, \widehat{\boldsymbol{\Theta}}_{n,K})$ is usually referred to as the Minimum Kantorovitch estimator [1].

Algorithm 1 Composite Transportation Distance with Mixtures

Input: Data $D = \{X_i\}_{i=1}^n$; the number of clusters K the regularized hyper-parameter $\lambda > 0$.
Output: Optimal weight-atoms $\{\omega_j, \theta_j\}_{j=1}^K$
Initialize weights $\{\omega_j\}_{j=1}^K$ and atoms $\{\theta_j\}_{j=1}^K$.
while not converged **do**
1. Update weights ω_j:
for $j = 1$ **to** K **do**
Compute transportation plan π_{ij} as

$$\pi_{ij} = \left(f(X_i|\theta_j)\right)^{1/\lambda} \Big/ \left(n\sum_{k=1}^K \left(f(X_i|\theta_j)\right)^{1/\lambda}\right)$$

for $1 \leq i \leq n$
Update weight $\omega_j = \sum_{i=1}^n \pi_{ij}$.
end for
2. Update atoms θ_j:
for $j = 1$ **to** J **do**
Update atoms θ_j as solution of equation $\nabla A\left(\theta_j\right) = \sum_{i=1}^n \frac{\pi_{ij}}{\omega_j} T(X_i)$.
end for
end while

2.2.2 Regularized composite transportation distance

As is the case with the traditional optimal transportation distance, the composite transportation distance $\widehat{W}$ does not have a favorable computational

complexity. Therefore, we consider an entropic regularizer to speed up its computation [4]. More precisely, we consider the following regularized version of $\widehat{W}(P_n, P_{\boldsymbol{\omega}_K,\boldsymbol{\Theta}_K})$:

$$\inf_{\boldsymbol{\pi}\in\Pi\left(\frac{1}{n}\mathbf{1}_n,\boldsymbol{\omega}_K\right)} \langle \boldsymbol{\pi}, \boldsymbol{M}\rangle - \lambda\mathbb{H}(\boldsymbol{\pi}),$$

where $\lambda > 0$ is a penalization term and $\mathbb{H}(\boldsymbol{\pi}) := -\sum_{i,j}\pi_{ij}\log\pi_{ij}$ is an entropy of $\boldsymbol{\pi} \in \Pi(\mathbf{1}_n/n, \boldsymbol{\omega}_K)$. Equipped with this regularization, we have a regularized version of the optimal estimator in (5):

$$\min_{\boldsymbol{\omega}_K,\boldsymbol{\Theta}_K} \inf_{\boldsymbol{\pi}\in\Pi\left(\frac{1}{n}\mathbf{1}_n,\boldsymbol{\omega}_K\right)} \langle \boldsymbol{\pi}, \boldsymbol{M}\rangle - \lambda\mathbb{H}(\boldsymbol{\pi}). \quad (6)$$

We summarize the algorithm for determining local solutions of the above objective function in Algorithm 1. The details for how to obtain the updates of weight and atoms in Algorithm 1 are deferred to the Supplementary Material. Given the formulation of Algorithm 1, we have the following result regarding its convergence to a local optimum.

Theorem 1. *The Algorithm 1 monotonically decreases the objective function (6) of the regularized composite transportation distance for finite mixtures.*

2.3 Composite transportation barycenter for mixtures of exponential families

In this section, we consider a problem of finding composite transportation barycenters for a collection of mixtures of exponential family. For $J \geq 1$, let $\left\{P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}\right\}_{j=1}^J$ be a collection of J mixtures of exponential families as described in (1), and let $\{a_j\}_{j=1}^J \in \Delta^J$ be weights associated with these mixtures. The transportation barycenter of these probability measures is a mixture of exponential family with at most L components, and is defined as an optimal solution of the following problem:

$$\operatorname*{argmin}_{\mathbf{w}_L,\boldsymbol{\Psi}_L} \sum_{j=1}^J a_j \widehat{W}\left(Q_{\mathbf{w}_L,\boldsymbol{\Psi}_L}, P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}\right), \quad (7)$$

where $\mathbf{w}_L = \{w_l\}_{l=1}^L \in \Delta^L$ and $\boldsymbol{\Psi}_L = \{\psi_l\}_{l=1}^L \in \Theta^L$ are unknown weights and parameters that we need to optimize. Recall that, to avoid notational clutter, we use $Q_{\mathbf{w}_L,\boldsymbol{\Psi}_L}$ to denote a finite mixture with at most L components. Since $P^j_{\boldsymbol{\omega}^j_{K_j},\boldsymbol{\Theta}^j_{K_j}}$ and $Q_{\mathbf{w}_L,\boldsymbol{\Psi}_L}$ are mixtures of exponential families, Eq. (7) can be rewritten as

$$\operatorname*{argmin}_{\mathbf{w}_L,\boldsymbol{\Psi}_L} \sum_{j=1}^J a_j \min_{\boldsymbol{\pi}^j\in\Pi\left(\boldsymbol{\omega}^j_{K_j},\mathbf{w}_L\right)} \left\langle \boldsymbol{\pi}^j, \boldsymbol{M}^j\right\rangle,$$

where the cost matrices $\boldsymbol{M}^j = (M^j_{uv})$ satisfy $M^j_{uv} = \mathrm{KL}(f(x|\theta^j_u), f(x|\psi_v))$ for $1 \leq j \leq J$, which has the closed form defined in Eq. (3) since f is from an exponential family of distributions.

2.3.1 Regularized composite transportation barycenter

We incorporate regularizers in the composite transportation barycenter. In particular, we write the objective function to be minimized as

$$\operatorname*{argmin}_{\mathbf{w}_L,\boldsymbol{\Psi}_L} \sum_{j=1}^J a_j \min_{\boldsymbol{\pi}^j\in\Pi\left(\boldsymbol{\omega}^j_{K_j},\mathbf{w}_L\right)} \left\langle \boldsymbol{\pi}^j, \boldsymbol{M}^j\right\rangle - \lambda\mathbb{H}\left(\boldsymbol{\pi}^j\right). \quad (8)$$

We call this objective function the *regularized composite transportation barycenter*. Due to space constraints, we present the detailed algorithm for determining local solutions of this objective function in the Supplementary Material.

3 Probabilistic clustering with multilevel structural data

Assume that we have J groups of independent data, $X_{j,i}$, where $1 \leq j \leq J$ and $1 \leq i \leq n_j$; i.e., the data are presented in a two-level grouping structure. Our goal is to find simultaneously the local clustering for each data group and the global clustering across groups.

3.1 Multilevel composite transportation (MCT)

To facilitate the discussion, for each $1 \leq j \leq J$, we denote the empirical measure associated with group j as

$$P^j_{n_j} := \frac{1}{n_j}\sum_{i=1}^{n_j}\delta_{X_{j,i}}.$$

Additionally, we assume that the number of local and global clusters are bounded. In particular, we allow local group j to have at most K_j clusters, which can be represented as a mixture of exponential families $P^j_{\boldsymbol{\omega}_j,\Theta^j_{K_j}}$, while we have at most $\boldsymbol{C}$ global clusters among J given groups. Here, each global cluster can be represented as a finite mixture distribution $Q^m_{\mathbf{w}^m_L,\boldsymbol{\Psi}^m_L}$ with at most L clusters, where $\mathbf{w}^m_L = (w^m_1, \ldots, w^m_L)$ and $\boldsymbol{\Psi}^m_L = (\psi^m_1, \ldots, \psi^m_L)$ are global weights and atoms for $1 \leq m \leq \boldsymbol{C}$, respectively.

3.1.1 Local clustering and global clustering

With the local clustering, we perform composite transportation distance optimization for group j, which can be expressed as in (5). More precisely, this step can be

viewed as finding optimal local weights $\boldsymbol{\omega}^j_{K_j}$ and local atoms $\Theta^j_{K_j}$ to minimize the composite transportation distance $\widehat{W}(P^j_n, P_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}})$ for all $1 \leq j \leq J$. Regarding the global clustering with J given groups, we can treat the finite mixture probability distribution $P_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ of each group as observations in the space of distributions over probability distributions. Thus we achieve a clustering of these distributions by means of an optimization with the following composite transportation distance on the space of finite mixtures of finite mixtures:

$$\inf_{\boldsymbol{\mathcal{Q}}} \widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right),$$

where we denote $\boldsymbol{P} := \frac{1}{J}\sum_{j=1}^{J} \delta_{P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}}$ and $\boldsymbol{\mathcal{Q}} := \sum_{m=1}^{\boldsymbol{C}} b_m \delta_{Q^m_{\mathbf{w}^m_L, \boldsymbol{\Psi}^m_L}}$.

3.1.2 MCT formulation

Since the finite mixture probability distributions $P_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ in each group are unobserved, we determine them by minimizing the objective cost functions in the local clustering and global clustering simultaneously. In particular, we consider the following objective function:

$$\inf_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}, \boldsymbol{\mathcal{Q}}} \sum_{j=1}^{J} \widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}\right) + \zeta \widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right), \quad (9)$$

where $\zeta > 0$ serves as a penalization term between the global cluster and local cluster. We call this problem *Multilevel Composite Transportation (MCT)*.

3.2 Regularized version of MCT

To obtain a favorable computation profile with MCT, we consider a regularized version of the composite transportation distances in both the local and global structures. To simplify the discussion, we denote $\boldsymbol{\pi}^j \in \Pi(\frac{1}{n_j}\mathbf{1}_{n_j}, \boldsymbol{\omega}^j_{K_j})$ as *local transportation plans* between $P^j_{n_j}$ and $P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ for all $\leq j \leq J$. Thus, the following formulation holds

$$\widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}\right) = \inf_{\boldsymbol{\pi}^j \in \Pi(\frac{1}{n_j}\mathbf{1}_{n_j}, \boldsymbol{\omega}^j_{K_j})} \left\langle \boldsymbol{\pi}^j, \boldsymbol{M}^j \right\rangle,$$

where $\boldsymbol{M}^j$ is the cost matrix between $P^j_{n_j}$ and $P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ that is defined as

$$[M^j]_{uv} = -\log f(X_{j,u}|\theta^j_v),$$

for $1 \leq u \leq n_j$ and $1 \leq v \leq K_j$. Therefore, we can consider the regularized version of composite transportation distance at each group j as follows:

$$\widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}\right) - \lambda_l \mathbb{H}(\boldsymbol{\pi}^j), \quad (10)$$

where $\lambda_l > 0$ is a penalization term for each group. Regarding the global structure, according to the definition of composite transportation distance for probability measure of measures, we have

$$\widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right) = \inf_{\boldsymbol{a} \in \Pi(\frac{1}{J}\mathbf{1}_J, \boldsymbol{b})} \sum_{j,m} a_{jm} \widehat{W}(P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L, \boldsymbol{\Psi}^m_L}),$$

where $\boldsymbol{a} = (a_{jm})$ in the above infimum is a *global transportation plan* between $\boldsymbol{P}$ and $\boldsymbol{\mathcal{Q}}$. Here, we can further rewrite $\widehat{W}(P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L, \Psi^m_L})$ as

$$\widehat{W}(P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}, Q^m_{\mathbf{w}^m_L, \Psi^m_L}) = \inf_{\boldsymbol{\tau}^{j,m} \in \Pi(\boldsymbol{\omega}^j_{K_j}, \mathbf{w}^m_L)} \left\langle \boldsymbol{\tau}^{j,m}, \boldsymbol{\gamma}^{j,m} \right\rangle,$$

where the cost matrix $\boldsymbol{\gamma}^{j,m}$ is defined as KL divergence between two exponential family atoms in Eq. (3):

$$\gamma^{j,m}_{k,l} := A\left(\theta^j_k\right) - A\left(\psi^m_l\right) - \left\langle \nabla A\left(\psi^m_l\right), \left(\theta^j_k - \psi^m_l\right) \right\rangle.$$

To facilitate the discussion later, we denote $\boldsymbol{\tau}^{j,m}$ the *partial global transportation plan* between $P_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}$ and $Q^m_{\mathbf{w}^m_L, \Psi^m_L}$. Therefore, we can regularize the composite transportation distance with global structure as

$$\widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right) - \lambda_a \mathbb{H}(\boldsymbol{a}) - \lambda_g \sum_{j=1}^{J}\sum_{m=1}^{\boldsymbol{C}} \mathbb{H}(\boldsymbol{\tau}^{j,m}), \quad (11)$$

where λ_a corresponds to a penalization term for global structure while λ_g represent a penalization term for a partial global transportation plan. Combining the results from Eqs. (10) and (11), we obtain the overall objective function of MCT:

$$\inf_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}, \boldsymbol{\mathcal{Q}}} \sum_{j=1}^{J} \widehat{W}\left(P^j_{n_j}, P^j_{\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}}\right) + \zeta \widehat{W}\left(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}\right) - R\left(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{a}\right), \quad (12)$$

where $R\left(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{a}\right) := \lambda_l \sum_{j=1}^{J} \mathbb{H}(\boldsymbol{\pi}^j) + \zeta[\lambda_a \mathbb{H}(\boldsymbol{a}) + \lambda_g \sum_{j=1}^{J}\sum_{m=1}^{\boldsymbol{C}} \mathbb{H}(\boldsymbol{\tau}^{j,m})]$ is a combination of all regularized terms for the local and global clustering. We call this objective function *regularized MCT*.

3.3 Algorithm for regularized MCT

We now describe our detailed strategy for obtaining a locally optimal solution of *regularized MCT*. In particular, our algorithm consists of two key steps: local clustering updates and global clustering updates. For simplicity of presentation, we assume that at step t of our algorithm, we have the following updated values of our parameters: $\boldsymbol{\omega}^j_{K_j}, \boldsymbol{\Theta}^j_{K_j}, \mathbf{w}^m_L, \boldsymbol{\Psi}^m_L$, for $1 \leq j \leq J$ and $1 \leq m \leq \boldsymbol{C}$.

Algorithm 2 Probabilistic Multilevel Clustering

Input: Data $D = \{X_{j,i}\}_{j=1,i=1}^{J,n_j}$; the number of local clusters K_j and global clusters $\boldsymbol{C}$; the number of components in each global cluster L; the penalization term ζ; the regularized hyper-parameters λ_l, λ_g.

Output: local and global parameters $\boldsymbol{\omega}_{K_j}^j, \boldsymbol{\Theta}_{K_j}^j, \mathbf{w}_L^m, \boldsymbol{\Psi}_L^m$ for $1 \le j \le J$ and $1 \le m \le \boldsymbol{C}$.

Initialize these local and global parameters.

while not converged **do**

1. Update local parameters:.

for $j = 1$ **to** J **do**

Update $\boldsymbol{\omega}_{K_j}^j$, $\boldsymbol{\Theta}_{K_j}^j$ as optimal solutions of (13).

end for

2. Update global parameters:

for $m = 1$ **to** $\boldsymbol{C}$ **do**

Update $\mathbf{w}_L^m, \boldsymbol{\Psi}_L^m$ as optimal solutions of (14).

end for

end while

Local clustering updates To obtain updates for local weights $\boldsymbol{\omega}_{K_j}^j$ and local atoms $\boldsymbol{\Theta}_{K_j}^j$, we solve the following combined regularized composite transportation barycenter problem:

$$\inf_{\boldsymbol{\omega}_{K_j}^j, \boldsymbol{\Theta}_{K_j}^j} \widehat{W}\left(P_{n_j}^j, P_{\boldsymbol{\omega}_{K_j}^j, \boldsymbol{\Theta}_{K_j}^j}^j\right) - \lambda_l \mathbb{H}(\boldsymbol{\pi}^j) \tag{13}$$

$$+ \sum_{m=1}^{\boldsymbol{C}} a_{jm} \widehat{W}(P_{\boldsymbol{\omega}_{K_j}^j, \boldsymbol{\Theta}_{K_j}^j}^j, Q_{\mathbf{w}_L^m, \Psi_L^m}^m) - \zeta\lambda_g \sum_{m=1}^{\boldsymbol{C}} \mathbb{H}(\boldsymbol{\tau}^{j,m}),$$

where $\boldsymbol{\pi}^j$ is the local transportation plan between $P_{n_j}^j$ and $P_{\boldsymbol{\omega}_{K_j}^j, \boldsymbol{\Theta}_{K_j}^j}^j$ at step t while $\boldsymbol{a}$ and $\boldsymbol{\tau}^{j,m}$ are respectively the global transportation plan and partial global transportation plans at this step. The idea of obtaining local solution of the above objective function is identical to that of (8); therefore, we defer the detailed presentation of this algorithm to the Supplementary Material.

Global clustering updates In order to update the global weights $\mathbf{w}_L^m$ and global atom parameters $\boldsymbol{\Psi}_L^m$, we consider the following optimization problem:

$$\inf_{\boldsymbol{\mathcal{Q}}} \widehat{W}(\boldsymbol{P}, \boldsymbol{\mathcal{Q}}) - \zeta[\lambda_a \mathbb{H}(\boldsymbol{a}) + \lambda_g \sum_{j=1}^{J} \sum_{m=1}^{\boldsymbol{C}} \mathbb{H}(\boldsymbol{\tau}^{j,m})]. \tag{14}$$

The algorithm for obtaining the local solutions of this objective is based on bacycenter computation algorithms in [5] for updating barycenter weights $\mathbf{w}_L^m$ and the partial global transportation plan $\boldsymbol{\tau}^{j,m}$. The natural parameters of global atoms of the barycenters are weighted averages of local atoms from all group j:

$$\psi_v = \frac{\sum_{j=1}^{J} \sum_{u=1}^{L} \pi_{uv}^j \theta_u^j}{\sum_{j=1}^{J} \sum_{u=1}^{L} \pi_{uv}^j}.$$

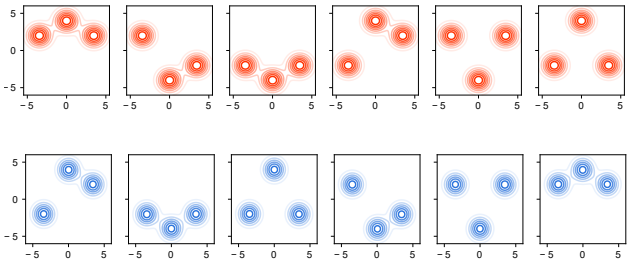

Figure 1: Synthetic multilevel Gaussian data (orange at top) and inferred clusters (blue at bottom).

The detailed derivation of this algorithm is deferred to the Supplementary Material. In summary, the main steps of updating the local and global clustering updates are summarized in Algorithm 2. We have the following result guaranteeing the local convergence of this algorithm.

Theorem 2. *Algorithm 2 monotonically decreases the objective function of regularized MCT* (12) *until local convergence.*

4 Experimental studies

We first evaluate the model via simulation studies, then demonstrate its applications on text and image modeling using two real-world datasets.

4.1 Simulated data

We evaluate the effectiveness of our proposed clustering algorithm by considering two types (discrete and continuous) of synthetic data generated from multilevel processes as follows.

Continuous data We start with six clusters of data, each of which is a mixture of three Gaussian components. Figure 1 depicts the ground truth of the six mixtures we generate the data from. We uniformly generated 100 groups of data, each group belonging to one of the six aforementioned clusters. Once the cluster index of a data group was defined, we generated 500 data points from the corresponding mixture of Gaussians.

Discrete data Data was generated from five clusters of 25-dimensional bar topics, each of which is a mixture of four bar topics out of total ten topics as shown in Figure 2 (second row). Each cluster shares two topics with any other cluster. We then generated 500 groups of data, each group belonging to one of the five aforementioned clusters. Once the cluster index of a data group is defined, we generate 100 data points from the mixture of bar topics of that cluster.

Datasets	#groups(J)	#dim	#points(n_j)	#clusters($\boldsymbol{C}$)
Continuous data	100	2	500	6
Discrete data	500	25	100	5

(a) Statistics of synthetic datasets

Datasets	#groups(J)	#dim	#clusters($\boldsymbol{C}$)
LabelMe	1,800	30	8
NUS-WIDE	1,040	238	13

(b) Statistics of real-world datasets

Table 1: Summarization of synthetic and realworld datasets

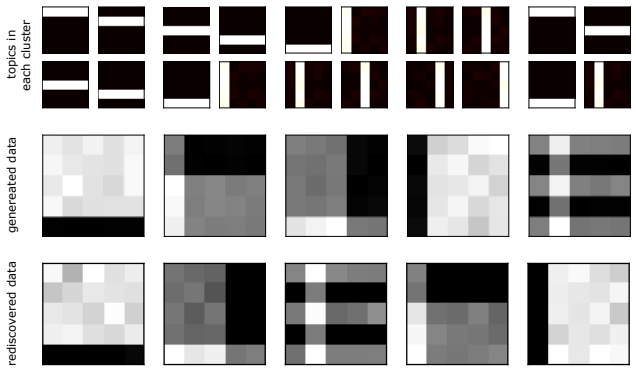

Figure 2: Synthetic multilevel bar topic data (two top rows) and rediscovered output (bottom row)

Clustering results We ran the proposed method with *synthetic continuous data* using the following local and global penalization hyper-parameters: λ_l and λ_g are set equal to 1.3 and 10, respectively. We model each atom in the (local and global) mixture models as an isotropic multivariate Gaussian. As shown in the bottom row of Figure 1, the model is able to rediscover the clustering structure in the generated dataset. Comparing with the top row of the figure, there is permutation in the order of discovered mixture models due to the label switching. Similarly, we use Categorical distribution to model each atom in the mixture models of the proposed model. Each observation X_{ji} now is a one-hot vector. We simulated from ten topics including five horizontal and five vertical bars. The top row of Figure 2 depicts a collection of four bar topics that data of a cluster may be generated from; i.e., the first cluster contains data simulated from a mixture of four horizontal bar topics. In the middle row, we depict the histogram plot of all data generated from each cluster while the bottom row shows the plot of clusters discovered by our proposed model. There is only a slight difference in the plot between ground truth and the inferred mixture of bar topics. These results demonstrate the effectiveness and flexibility of our algorithms in learning both continuous and discrete data.

We also measured the NMI (Normalized Mutual Information) between the ground truth labels and learned groups and obtained around 0.98.

Methods	NMI	ARI	AMI
K-means	0.37	0.282	0.365
SVB-MC2	0.315	0.206	0.273
W-means	0.423	0.35	0.416
MCT	**0.485**	**0.412**	**0.477**

Table 2: Clustering performance on LabelMe (continuous) dataset.

Methods	NMI	ARI	AMI
K-means	0.35	0.093	0.22
SVB-MC2	0.295	0.139	0.249
W-means	0.356	0.089	0.203
MCT	**0.423**	**0.255**	**0.39**

Table 3: Clustering performance on NUS-WIDE (discrete) dataset.

4.2 Real-world data

We now demonstrate our proposed model on two real-world datasets: the LabelMe dataset [20, 17] with continuous observations and the NUS-WIDE [2] with discrete observations. Statistics for these datasets are presented in Table 1b.

LabelMe dataset This consists of 2,688 annotated images which are classified into eight scene categories including *tall buildings, inside city, street, highway, coast, open country, mountain, and forest* [20]. Each image contains multiple annotated regions. Each region, which is annotated by users, represents an object in the image. We remove the images containing less than four annotated regions and obtained totally 1,800 images. We then extract GIST features [13], a visual descriptor to represent perceptual dimensions and oriented spatial structures of a scene, for each region in an image. We use PCA to reduce the number of dimensions to 30.

NUS-WIDE dataset We used a subset of the original NUS-WIDE dataset [2] which contains images of

Sample images and annotated regions can be found at http://people.csail.mit.edu/torralba/code/spatialenvelope/

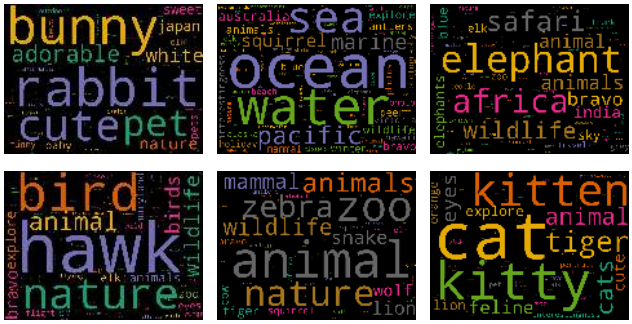

Figure 3: Tag cloud of six clusters discovered by the proposed model with the NUS-WIDE dataset.

13 kinds of animals comprising 2,054 images in training subset. Each image is annotated with several tags out of 1,000 tags. We filtered out images with less than three tags and obtained 1,040 images with the remaining number tags of 238. After preprocessing, we have a dataset with 1,040 groups; each data point in a group is a one-hot vector of 238 dimensions representing a tag word annotated for that group (image).

Baseline methods We *quantitatively* compare our proposed method to baseline approaches discussed in [7], including *K-means*, *W-means*, and *SVB-MC2 without context* [9]. We use three popular metrics: NMI (Normalized Mutual Information) [21, 16.3], ARI (Adjusted Rand Index) [8], and AMI (Adjusted Mutual Information) [25] to evaluate the clustering performance.

Experimental results We conducted experiments on the LabelMe dataset with the number of local atoms set equal to $K = 5$, the number of global atoms set to $L = 15$, and the number of clusters set to $\boldsymbol{C} = 8$. We ran 10-fold cross-validation to choose the best hyper-parameters for penalized terms which are $\lambda_l = 3$ and $\lambda_g = 3$. As shown in Table 2, our proposed method is superior to the baseline methods in terms of clustering performance.

We also compared clustering performance using the discrete real-world dataset NUS-WIDE. We chose $K = 2$, $L = 4$, and $\boldsymbol{C} = 13$ with the 10-fold cross-validation hyper-parameters $\lambda_l = 1$ and $\lambda_g = 1.6$. Results are presented in Table 3. Since baseline methods are applicable only to continuous data, we have normalized the discrete data for each image and then applied the baseline methods to cluster the dataset. The results show that the clustering performance of K-means and W-means is inferior to that of our proposed model which directly models discrete data. Moreover, the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) of these model show that their clustering outcomes are not robust.

To illustrate the *qualitative* results of the proposed model, we selectively choose six out of thirteen clusters discovered and computes the proportion of tags presented in each cluster. Figure 3 depicts tag-clouds of these clusters. Each tag-cloud consistently manifests the cluster content. For example, the top-left tag-cloud denote the cluster of rabbits, which is one of the ground-truth subsets of images.

5 Discussion

We have proposed a probabilistic model that uses a novel composite transportation distance to cluster data with potentially complexed hierarchical multilevel structures. The proposed model is able to handle both discrete and continuous observations. Experiments on simulated and real-world data have shown that our approach outperforms competing methods that also target multilevel clustering tasks. Our developed model is based on the exponential family assumption with data distribution and thereby applies naturally to other data types; e.g., a mixture of Poisson distributions [10]. Finally, there are several possible directions for extensions from our work. First, it is of interest to extend our approach to richer settings of hierarchical data similar to those considered in MC^2 [15]; e.g., when group-level context is available in the data. Second, our method requires knowledge of the upper bounds with the numbers of clusters both in local and global clustering. It is of practical importance to develop methods that are able to estimate these cardinalities efficiently. Third, regarding computational scalability, we can leverage the recent development of stochastic computation [3] and distributed/parallel computation [22, 6] of Wasserstein barycenter into our algorithm development which allows us to scale up our learning problem for millions of data groups.

including squirrel, cow, cat, zebra, tiger, lion, elephant, whales, rabbit, snake, antlers, hawk and wolf

References

[1] F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006.

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval*, page 48. ACM, 2009.

[3] S. Claici, E. Chien, and J. Solomon. Stochastic wasserstein barycenters. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 998–1007, 2018.

[4] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[5] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[6] P. Dvurechenskii, D. Dvinskikh, A. Gasnikov, C. Uribe, and A. Nedich. Decentralize and randomize: Faster algorithm for wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 10783–10793, 2018.

[7] N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Q. Phung. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1501–1509, 2017.

[8] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[9] V. Huynh, D. Q. Phung, S. Venkatesh, X. Nguyen, M. D. Hoffman, and H. H. Bui. Scalable nonparametric Bayesian multilevel clustering. In *UAI*, 2016.

[10] R. R. Jayasekare, R. Gill, and K. Lee. Modeling discrete stock price changes using a mixture of Poisson distributions. *Journal of the Korean Statistical Society*, 45(3):409–421, 2016.

[11] M. I. Jordan. An introduction to probabilistic graphical models. 2003.

[12] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[14] J. W. Miller and M. T. Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

[15] V. Nguyen, D. Phung, X. Nguyen, S. Venkatesh, and H. Bui. Bayesian nonparametric multilevel clustering with group-level contexts. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

[16] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013.

[17] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[18] D. Pollard. Quantization and the method of K-means. *IEEE Transactions on Information Theory*, 28:199–205, 1982.

[19] A. Rodriguez, D. B. . Dunson, and A. E. Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103:1131–1144, 2008.

[20] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[21] H. Schütze, C. D. Manning, and P. Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, 2008.

[22] M. Staib, S. Claici, J. Solomon, and S. Jegelka. Parallel streaming Wasserstein barycenters. In *Advances in Neural Information Processing Systems 31*, 2017.

[23] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.

[24] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

[25] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[26] D. F. Wulsin, S. T. Jensen, and B. Litt. Nonparametric multi-level clustering of human epilepsy seizures. *Annals of Applied Statistics*, 10:667–689, 2016.

[27] J. Zhang, Y. Song, G. Chen, and C. Zhang. Online evolutionary exponential family mixture. In *IJCAI*, pages 1610–1615, 2009.