Information and Communication Theory Group
Delft University of Technology
http://ict.ewi.tudelft.nl/

P.O. Box 5013
2600 GA Delft
The Netherlands

September 25, 2009

EWI-ICT TR 2009–001

Modeling Semantic Similarities in Multiple Maps

Laurens van der Maaten
ICT Group, Delft University of Technology

Geoffrey Hinton
Department of Computer Science, University of Toronto

Abstract

Models that represent words as points in a semantic space are subject to fundamental limitations of metric spaces. These limitations prevent semantic space models from faithfully representing, for example, the pairwise similarities between word meanings as revealed by word association data. In particular, semantic space models cannot faithfully represent intransitive pairwise similarities or the similarities of words that have multiple meanings. In this paper, we present a model that alleviates the limitations of semantic space models by constructing a collection of maps that represent complementary structure in the similarity data. Our model is a variant of a similarity choice model known as Stochastic Neighbor Embedding that constructs multiple maps and allows each object to occur as a point in several different maps. We apply the model to a set of word association data, demonstrating that it can successfully represent intransitive semantic relations as well as words with multiple meanings, and that it outperforms traditional semantic space models in the prediction of word associations. We compare the model to alternative representations of semantic structure, such as topic models and semantic networks.

Modeling Semantic Similarities in Multiple Maps

Laurens van der Maaten
ICT Group, Delft University of Technology

Geoffrey Hinton
Department of Computer Science, University of Toronto

1 Introduction

Multidimensional scaling [55] is a well-known computational model that represents similar objects, for instance, words that exhibit a certain semantic similarity, by nearby points in a metric spatial representation. Models that use the spatial proximity of two points to represent the similarity between two objects are often referred to as second-order isomorphic models [11], because they do not aim to model each object by a representation that has similar properties as the object (i.e., to implement a first-order isomophism as in, e.g., the recognition-by-components theorem in vision [1]), but instead, they aim to model the similarity between two objects by the proximity of their two representations. When modeling semantic similarities, such second-order isomorphic models are often referred to as semantic space models [49, 44, 26, 22].

Over the last decade, research on multidimensional scaling models has mainly focused on the development of more sophisticated similarity measurements between objects by means of, for example, geodesic or diffusion distances [21, 54] or by computing Euclidean distances in a (possibly infinite-dimensional) feature space through the "kernel trick" [46, 19]. However, these approaches do not address fundamental limitations of multidimensional scaling models that are due to the characteristics of metric spaces. A metric space is a space in which the following four *metric axioms* hold: (1) non-negativity of distances, (2) identity of indiscernibles, (3) symmetry of distances, and (4) the triangle inequality. If we denote the distance between object A and object B by $d(A, B)$, the four metric axioms may be denoted by

$$d(A, B) \geq 0, \tag{1}$$

$$d(A, B) = 0 \text{ iff } A = B, \tag{2}$$

$$d(A, B) = d(B, A), \tag{3}$$

$$d(A, C) \leq d(A, B) + d(B, C). \tag{4}$$

The metric axioms give rise to the following three limitations of metric spaces in terms of the similarities they can represent: (1) the triangle inequality that holds in metric spaces induces transitivity of similarities, (2) in a metric space, the number of points that can have the same point as their nearest neighbor is limited[1], and (3) in a metric space, similarities are symmetric. As a result of these limitations, multidimensional scaling cannot, for instance, faithfully model word association data that does not obey the metric axioms. We discuss the three limitations of metric spaces and their consequences for modeling semantic similarities in more detail below.

The first limitation of metric spaces is due to the triangle inequality, which basically states that if point A is close to point B and B is close to point C, A has to be close to C as well. In practice, this

[1] This is not the only limitation on the neighborhood relations of points in a metric space. For instance, the maximum number of equidistant points in a metric space is limited as well.

constraint may well be violated by the implicit structure of word associations. Consider, for instance, the word *tie*, which has a semantic relation with words such as *suit* and *tuxedo*. In a low-dimensional metric map of the input objects, these three words need to be close to each other. However, the word *tie* is ambiguous: it also has a semantic relation with words such as *rope* and *knot*, and should therefore be close to these words as well. As a result, the words *suit* and *rope* will be modeled fairly close together in a semantic space even though the words exhibit very little similarity other than their association with *tie*. In semantic networks such as Wordnet [12], this problem is circumvented by differentiating between the different senses of a word, but often it is very difficult to specify the sense of a word. The word *monarchy*, for example, would not typically be viewed as ambiguous but it certainly has nuances of meaning. Associations to this word can relate it to other forms of government or to the pageantry that surrounds royalty.

The second limitation of a metric space is that only a limited number of points can have the same point as their nearest neighbor. For instance, in a two-dimensional space, at most five points can have the same point as their nearest neighbor (by arranging them in a pentagon that is centered on the point). As a result, a (low-dimensional) metric semantic space cannot faithfully model the large number of similarities of 'central' objects with other objects. This is problematic because word meanings are characterized by a high 'centrality', i.e., by the presence of words that are similar to a large portion of the other words [56]. The high centrality of semantics can be understood from the properties of semantic networks, which are scale-free networks that are characterized by a high clustering coefficient [52]. This implies that semantic data is bound to contain central concepts. For instance, large numbers of mammals have a closer semantic relation with the word *mammal* than with each other, and as a result, these mammals would like to have the word *mammal* as their nearest neighbor in a spatial model. It is impossible to achieve this in a low-dimensional metric semantic space because only a limited number of points can have the same nearest neighbor.

The third limitation of metric spaces is that the (dis)similarities in these spaces are symmetric, whereas the (dis)similarities between entities in the world are often non-symmetric. Tversky illustrated this problem with a famous example on the similarity between China and North Korea [56]: "People typically have the intuition that North Korea is more similar to China than China is to North Korea". The presumed reason for the asymmetry in human similarity judgements is that a person's representation of China typically contains a large number of features, of which only some features are shared with North Korea, whereas the representation of North Korea involves a small number of features, most of which are shared with China.

The three limitations of (low-dimensional) metric spaces discussed above lead Tversky to argue against multidimensional scaling as a model for semantic similarities [56], since its fundamental limitations make it unsuitable for faithfully modeling these similarities. Instead, he advocates the representation of concepts in terms of sets of features, in which the similarity between concept A and concept B depends on the relative amount of features that concept A shares with B. Under such a representation, the asymmetric similarity between, for instance, China and North Korea can appropriately be modeled.

In this paper, we present an alternative solution to the three limitations of semantic space models by introducing a new variant of multidimensional scaling. Our model has two main differences from traditional models for multidimensional scaling. Firstly, it converts the pairwise similarities between objects into conditional probabilities[2] and tries to model the objects in the (semantic) space in such a way that they give rise to a similar set of conditional probabilities. Hence, the second-order isomorphism is between two conditional probability distributions. Secondly, instead of a using a single metric map to represent the conditional probabilities, our model employs a collection of maps that together represent the conditional probabilities. Each object has a corresponding point in every map in the collection, and each of these points has a "mixing proportion" that indicates its "importance" in each map. The mixing proportions of all the points corresponding to the same object sum to one. The conditional probability that represents the similarity of two objects is then modeled by a sum over all maps of the conditional probability that arises from each map. The conditional probability in a map depends on both the mixing proportions of the two points in the map and on their proximity. If two points are close together in a map

[2] Note that for certain types of data, such as word association data, the data already takes the form of such conditional probabilities.

in which both points have a high mixing proportion, the multiple map model will assign them a fairly high joint probability, even if these points are very far apart in some of the other maps. Note that using, for instance, two two-dimensional maps in this "disjunctive" way is very different from the standard "conjunctive" approach of using a four-dimensional map and then treating the first two dimensions as one map and the last two dimensions as another map [41]. In the conjunctive approach, a pair of points needs to be close together in *all* of the two-dimensional projections in order to represent high similarity between the corresponding objects. In the disjunctive approach, by contrast, high similarity in one map cannot be vetoed by low similarity in another map.

The proposed multiple map model is capable of faithfully representing nonmetric similarities between objects, such as the strength of associations between words. For example, the word *tie* can be close to *tuxedo* but far from *knot* in one map, and close to *knot* but far from *tuxedo* in another map. This captures its similarity to both *tuxedo* and *knot* without forcing *tuxedo* to be close to *knot*. In the paper, we show that it is possible to learn multiple map representations from observed co-occurence or association data. We present visualizations of the multiple maps that show how the model is able to, e.g., identify different senses of words and modeling them in different maps. Also, we show that the use of multiple maps if beneficial in generalization tasks: a multiple map model is better at predicting word association than a single map model.

The remainder of the paper consists of five main parts. First, we review an asymmetric variant of the similarity choice model called 'Stochastic Neighbor Embedding' or 'SNE' [16] that forms the basis of our multiple map model. Second, we present the multiple map model, which can be viewed as a generalization of SNE. Third, we present the results of experiments in which we visualize a large dataset of word association data. The results demonstrate the potential of multiple maps for semantic representation and illustrate how they deal with intransitive similarities and objects with high centrality. Fourth, we present experiments in which we show the benefits of multiple map models for generalization: we show that a multiple map model is better at predicting unseen word associations than a single map model, even when the single map model is allowed to use exponentially more space. Fifth, we conclude the paper with a discussion of the similarities and differences of the multiple map model with (i) other semantic space models, (ii) semantic networks, and (iii) topic models [14].

2 Stochastic Neighbor Embedding

Stochastic Neighbor Embedding and its recent extensions [16, 9, 57, 60] are popular multidimensional scaling techniques that are often used in machine learning to learn low-dimensional data representations from high-dimensional vectorial data in such a way, that the small pairwise distances between objects (i.e., the local structure of the high-dimensional data) are preserved as well as possible. Stochastic Neighbor Embedding can be viewed as an asymmetric variant of the similarity choice model [25, 48, 51, 33].

The input of SNE consists of a collection of N conditional probability distributions P_i that represent the pairwise similarities between the N input objects. These conditional probability distributions have entries $p_{j|i}$ that represent, e.g., the probability that object j is associated with object i, or the relative number of co-occurrences of object j with object i. For instance, word association data is obtained by measuring how often the word *cup* (object j) comes to mind after the word *coffee* (object i) was presented as stimulus. In learning settings in which the input objects are points in a high-dimensional space (as is the case in many machine learning applications [2]), the probabilities may be computed by using a procedure that is based on Shepard's universal law of generalization [50, 16]. However, we do not consider such learning settings in this paper.

SNE models the input similarities by giving each of the N objects a location $\mathbf{y}_i$ in a metric map. The aim of SNE is to do this in such a way that the pairwise similarities $p_{j|i}$ are modeled as well as possible in the map. In order to evaluate the pairwise similarities of objects in the map, we define N conditional probability distributions Q_i with respect to the coordinates $\mathbf{y}_i$ of the objects in the low-dimensional map. The conditional probability distribution Q_i has entries $q_{j|i}$, which represent the probability of point i picking point j as its neighbor in the map, thereby measuring the pairwise similarity between the points

$\mathbf{y}_i$ and $\mathbf{y}_j$. In SNE, the pairwise similarity $q_{j|i}$ is defined[3] to be proportional to the density of point y_j under a Gaussian distribution that is centered on point $\mathbf{y}_i$

$$q_{j|i} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{i \neq k} \exp\left(-\|\mathbf{y}_i - \mathbf{y}_k\|^2\right)}, \tag{5}$$

This definition of the similarity $q_{j|i}$ is basically an implemention of Shepard's universal law of generalization [50]: we assume that the probability that object j is associated with object i decays exponentially with the squared Euclidean distance between the corresponding points in the space. Alternative motivations for computing $q_{j|i}$ as exponentially decaying with the squared Euclidean distance between the map points are based on, among others, the maximum-entropy characteristic of the Gaussian distribution [10] and on the theory of reproducible kernel Hilbert spaces [46]. The latter theory states that the values $p_{j|i}$ are measurements of inproducts between objects in an infinite-dimensional feature space, and is covered in detail in [19].

If the map $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ appropriately models the pairwise similarities $p_{j|i}$ of the input objects, all of the conditional probability distributions P_i and Q_i are equal. In SNE, such a setting of the map $\mathbf{Y}$ is identified by minimizing the sum of the natural differences between the conditional probability distributions P_i and Q_i, the Kullback-Leibler divergences, with respect to the coordinates $\mathbf{y}_i$ of the datapoints in the map. Mathematically, SNE thus minimizes the cost function[4]

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}. \tag{6}$$

The asymmetric nature of the Kullback-Leibler divergence leads SNE to focus on appropriately modeling the large pairwise similarities $p_{j|i}$ between the input objects. In other words, similar input objects really need to be close together in the low-dimensional map in order to minimize the cost function C. In this respect, SNE differs from traditional classical scaling [55], which as a result of the use of a squared error criterion mainly focuses on modeling dissimilar input objects as far apart in the map.

The minimization of the SNE cost function is typically performed using a simple gradient descent method[5], the details of which are described in [16]. The gradient of the cost function with respect to the low-dimensional map coordinates $\mathbf{y}_i$ is given by

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(\mathbf{y}_i - \mathbf{y}_j). \tag{7}$$

The gradient thus defines a collection of $N(N-1)$ springs between the points in the map, in which the stiffness of the spring is given by the discrepancy $p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j}$ between the similarities $p_{j|i}$ and $q_{j|i}$. The length and direction of each spring is given by $(\mathbf{y}_i - \mathbf{y}_j)$. Minimizing the cost function in Equation 6 can thus be seen as the minimization of the potential energy in a system of springs, though the stiffnesses of the springs change on each iteration.

Stochastic Neighbor Embedding and its variants have been shown to outperform existing techniques for multidimensional scaling [57] such as Sammon mapping [44], Isomap [54], and diffusion maps [21]. Another important advantage of SNE over other approaches to multidimensional scaling is its use of conditional probability distributions, which makes it a very natural technique to model asymmetric data such as word associations. These two advantages of SNE, as well as its relation to Shepard's universal law of generalization, have led us to select it as the basis for our multiple map model. However, we note that in principle it is possible to implement the idea of using multiple maps in other multidimensional scaling techniques as well.

[3] Throughout the paper, we do not consider self-similarities. We simply assume that $p_{i|i} = q_{i|i} = 0$.

[4] Note that since the distributions P_i are fixed, minimizing the sum of Kullback-Leibler divergences $\sum_i KL(P_i||Q_i)$ is identical to minimizing the sum of the cross-entropies of P_i and Q_i.

[5] We should note that other optimization procedures have been proposed for SNE, for example, using trust region methods [31] or simulated annealing [16].

3 Multiple map SNE

The probabilistic nature of SNE allows for a natural extension to a multiple map version [9], which is a desirable property that traditional multidimensional scaling techniques do not have. As we will show, the limitations of metric spaces can be circumvented by the use of multiple maps. We propose a multiple map version of SNE that constructs a collection of M maps, all of which contain N points (one for each of the N input objects). In each map with index m, a point with index i has a so-called mixing proportion $\pi_i^{(m)}$ that measures the "importance" of point i in map m. Because of the probabilistic interpretation of our model, we require the mixing proportions to be positive, and we define the sum of the mixing proportions of a single point over all maps to be 1. In other words, we constrain the mixing proportions $\pi_i^{(m)}$ to make sure that $\pi_i^{(m)} \geq 0, \forall i, m$ and $\sum_m \pi_i^{(m)} = 1, \forall i$. We redefine the conditional probability distribution $q_{j|i}$, which represents the similarity between the objects with index i and j under the model, as the weighted sum of the pairwise similarities between the points corresponding to input objects i and j over all M maps. Mathematically, we redefine $q_{j|i}$ in the multiple map SNE model as

$$q_{j|i} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} \exp\left(-\|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right)}{\sum_{m'} \sum_{i \neq k} \pi_i^{(m')} \pi_k^{(m')} \exp\left(\|\mathbf{y}_i^{(m')} - \mathbf{y}_k^{(m')}\|^2\right)}. \tag{8}$$

The cost function of the multiple map version of SNE is still given by Equation 6, however, it is now optimized with respect to the $N \times M$ low-dimensional map points $\mathbf{y}_i^{(m)}$ and with respect to the $N \times M$ mixing proportions $\pi_i^{(m)}$.

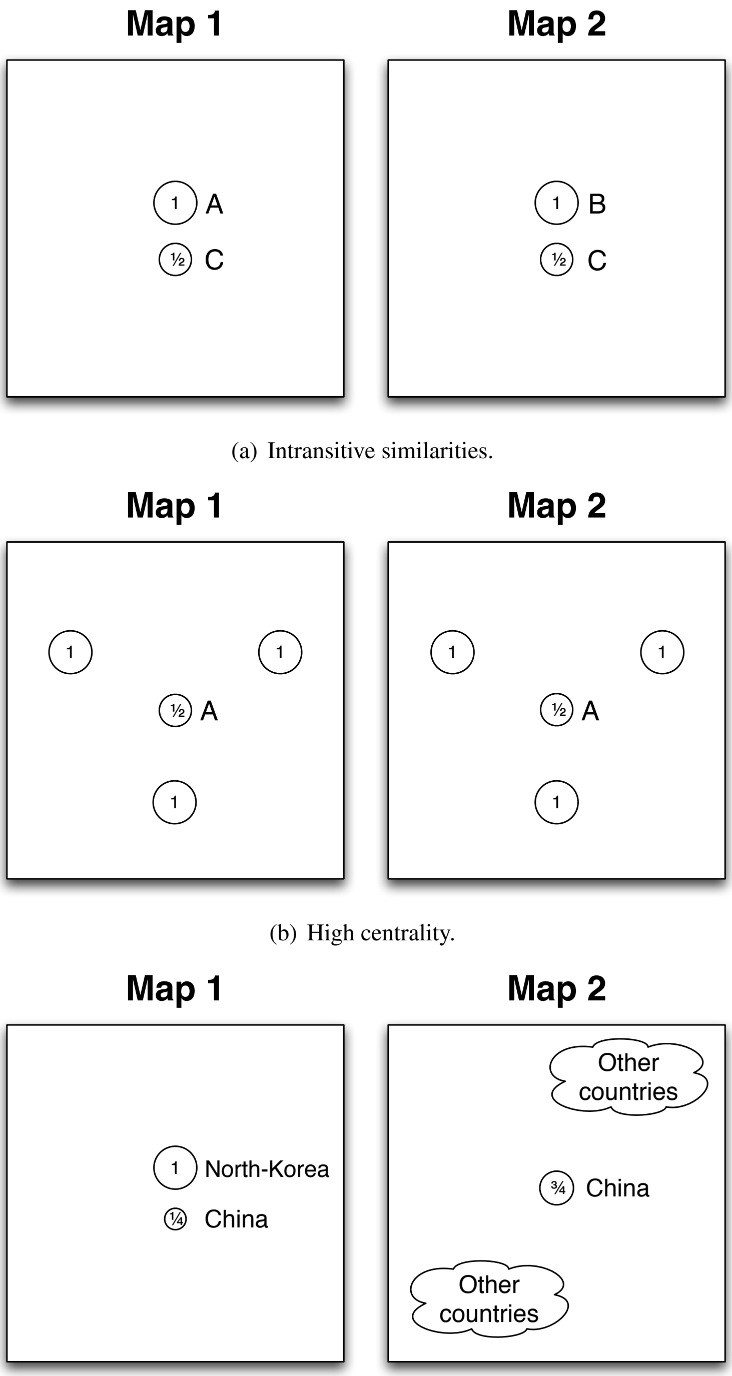
Because the mixing proportions $\pi_i^{(m)}$ for a single point i should sum to 1 over all maps, direct optimization of the cost function C with respect to the parameters $\pi_i^{(m)}$ is tedious. To circumvent this problem, we represent the mixing proportions $\pi_i^{(m)}$ in terms of "mixing weights" using an idea that is similar to that of softmax units, which are commonly used in neural networks [5]. The mixing proportions $\pi_i^{(m)}$ are represented in terms of the mixing weights $w_i^{(m)}$ as follows

$$\pi_i^{(m)} = \frac{e^{-w_i^{(m)}}}{\sum_{m'} e^{-w_i^{(m')}}}. \tag{9}$$

By defining the mixing proportions in this way, they are guaranteed to be positive and to sum up to 1. As a result, the minimization of the cost function can be performed with respect to the unconstrained mixture weights $w_i^{(m)}$. This significantly simplifies the optimization of the cost function (which is still given by Equation 6) using gradient descent.

The gradients that are necessary to perform the minimization of the cost function are given in Appendix A. In our experiments, we used a simple gradient descent method that employs an additional momentum term to stabilize the gradient search. In other words, the gradient at each iteration is added to an exponentially decaying sum of the gradients at previous iterations in order to determine the changes in the parameters at each iteration of the gradient search. The momentum term is employed in order to speed up the gradient search without creating the oscillations that are caused by simply increasing the step size. Moreover, we employ an approach called "early exaggeration" [57]: in the early stages of the optimization, we multiply the conditional probabilities $p_{j|i}$ by, say, 4. As a result, the $p_{j|i}$'s are much too large to be appropriately modeled by their corresponding $q_{j|i}$'s (which still sum up to 1). This encourages the optimization to model the largest $p_{j|i}$'s by relatively large $q_{j|i}$'s, thereby creating tight widely separated clusters in the maps that facilitate the identification of an appropriate global organization of the maps.

The multiple map model presented above is capable of circumventing the limitations of multidimensional scaling that caused Tversky to reject it as a model for semantic representation. In particular, the multiple map model has three main advantages over traditional multidimensional scaling models: (1) it can represent intransitive similarities, (2) it can represent data with high centrality even in low-dimensional semantic spaces, and (3) it can represent asymmetric similarities. We discuss the three advantages of the multiple map model separately below.



(a) Intransitive similarities.

(b) High centrality.

(c) Asymmetric similarities.

Figure 1: Illustration of how multiple map SNE can model intransitive similarities, data with high centrality, and asymmetric similarities.

1) Intransitive similarities. Consider our introductory example with the word *tie*, which is semantically similar to *tuxedo* and to *knot*. The word *tie* should be modeled close to *tuxedo* and *knot*, but the words *tuxedo* and *knot* should not be modeled close to each other. In contrast to single map multidimensional scaling techniques, multiple map SNE can appropriately model this example as follows.

Assume we have three datapoints A, B, and C that are embedded into two maps (see Figure 1(a)). Multiple map SNE can give point A a mixing proportion of 1 in the first map, point B a mixing proportion of 1 in the second map, and point C a mixing proportion of $\frac{1}{2}$ in both maps, and it can give all three points have the same spatial location in both maps. Then, the pairwise similarity between point A and C is equal to $1 \times \frac{1}{2} = \frac{1}{2}$, and the pairwise similarity between point B and C is also equal to $\frac{1}{2}$. However, the pairwise similarity between point A and B is 0, because the points A and B have no mixing proportion in each others maps. Hence, the representation constructed by multiple map SNE does not satisfy the triangle inequality, as a result of which it can model intransitive semantic similarities such as our example with *tie*, *tuxedo*, and *knot*.

Each time we want to add an object that violates the triangle inequality to an existing map, we need to put one copy of the object in the existing map and another copy in a different map. This way of using multiple maps to violate the triangle inequality is not the same as simply using different maps to capture different topics. If the triangle inequality is violated by three words with the same topic, it is necessary to put two of them in at least two maps. This leads to many small clusters in each of the multiple maps and clusters within one map that are not adjacent may have little in common. It would have been nice if each map captured a different topic, but the cost function does not significantly favor such arrangements of the words over the maps.

2) High centrality. In a metric space, only a limited number of points can have the same point as their nearest neighbor, as a result of which it is not possible to model the large number of similarities of 'central' objects with other objects appropriately in a low-dimensional metric map. Data with high centrality can be modeled appropriately by multiple map SNE even if the dimensionality of the maps is low, essentially, because multiple maps provide much more space than a single map. We illustrate the capability of multiple map SNE to model data with high centrality by an example.

Assume we have six objects that all have the same 'central' object A as their most similar object. In a single map, only five of the objects can be modeled in such a way that they have the low-dimensional model of object A as their nearest neighbor. In contrast, when two maps are available, the data can be modeled in such a way that the low-dimensional models of the all six objects have the model of object A as their nearest neighbor, and that these objects are to embedded in two-dimensional map(s). For instance, this can be achieved by giving A a mixing proportion of $\frac{1}{2}$ in both maps, modeling the first three objects close to the model of object A in the first map with mixing proportion 1, and modeling the remaining three objects close to the model of object A in the second map with mixing proportion 1. This example is illustrated in Figure 1(b).

Multiple map SNE can thus successfully model 'central' objects, such as the *mammal* in our introductory example, even in low-dimensional semantic spaces. Clearly, the number of points that can have the same point as their nearest neighbor in multiple map SNE depends on the number of maps and on the dimensionality of these maps.

3) Asymmetric similarities. A metric low-dimensional map constructed by a single map multidimensional scaling technique (such as SNE) cannot appropriately model asymmetric similarities, such as the similarity between China and North Korea in Tversky's famous example. In contrast, asymmetric similarities between objects can be modeled in multiple maps. We illustrate this capability using by modeling the similarity between China and North-Korea in multiple maps.

Assume (1) that we have two maps, (2) that North Korea has a mixing proportion of 1 in the first map and a mixing proportion of 0 in the second map, and (3) that China has a mixing proportion of $\frac{1}{4}$ in the first map and a mixing proportion of $\frac{3}{4}$ in the second map. In addition, assume (4) that North Korea and China are mapped close to each other in map 1, and (5) that China is modeled close to other countries in map 2. This example is illustrated in Figure 1(c). In the example, North Korea is modeled as very similar to China, whereas China is much less similar to North Korea, because it shares a large number of features with other countries as well. The actual similarity between China to North Korea under the model depends on the locations and mixing proportions of the other countries in both maps, i.e., on the amount of features that China shares with North Korea, relative to the amount of features that China shares with other countries. Nevertheless, the representation constructed by multiple map SNE successfully models the asymmetric similarity between China and North Korea.

4 Visualization Experiments

In this section, we present experiments in which we visualize the maps that were learned by the multiple map model from a large dataset of word associations. Our experiments reveal that the multiple map model can successfully model the nonmetric structure of the word associations, for instance, by modeling different senses of the same word in different maps.

4.1 Experimental setup

We performed experiments in which we used the multiple map model to visualize the Florida State University word association dataset [32]. The dataset contains association data for $10,617$ words, $5,019$ of which were used as input stimuli. The dataset contains a semantic similarity value for each pair of words that was computed as follows. Human subjects were given a specific word and asked to name associated words. After normalization, a condition probability $p_{j|i}$ is obtained that measures the probability that a human subject produces word j in response to word i. The conditional probabilities $p_{j|i}$ are a natural way to provide input to the multiple map model.

The word association dataset has all three characteristics that make it difficult to visualize the data using traditional multidimensional scaling approaches. First, it contains numerous examples of intransitive semantic relations, such as our introductory example with *tie*, *tuxedo*, and *rope*. Second, it contains a number of fairly 'general' words that have semantic relations with many other words. The high centrality of the Florida State University word association dataset is reflected in the high clustering coefficient of the data [52]. The most central word in the data is the word *field*, which has a semantic similarity to 33 other words in the data. Third, the word association data contains numerous examples of asymmetric similarities. For instance, the word *scissors* has a similarity of 0.879 with the word *cut*, whereas the similarity of *cut* with *scissors* is only 0.034.

In our experiments, we construct 40 maps in which we embed the $5,019$ words that were used as input stimuli in the collection of the data (i.e., the $5,019$ words for which we have the conditional probabilities $p_{j|i}$). The dimensionality of each map is set to 2. We trained the model using $2,000$ iterations of gradient descent, in which we employed an additional momentum term. We set the momentum term to 0.5 during the first 250 iterations, and to 0.8 afterwards. For the learning rate, we employed an adaptive learning rate scheme that is commonly applied in the training of neural networks [18]. The adaptive scheme aims to speed up the optimization by using a different (variable) learning rate for each parameter in the model. The scheme iteratively increases the learning rate for parameters for which the sign of the gradient is stable, whereas it rapidly decreases the learning rate when the sign of the gradient changes on successive weight updates. In our experiments, we set the initial value of the learning rate to 0.1. We also used early exaggeration with a factor of 4 during the first 50 iterations of the gradient descent.

In preliminary experiments, we found the approach to be fairy robust under changes in the various optimization parameters. Simpler optimization approaches in which the adaptive learning rate scheme and early exaggeration are not employed are capable of producing good results as well, but they are slower and the maps are generally slightly less good.

The transformation of high-dimensional data into two dimensions that happens in visualization often leads to a problem that is the result of the exponential volume difference between the high-dimensional and the two dimensional space. This problem is sometimes referred to as the crowding problem, and it can be alleviated by using a heavy-tailed distribution to measure the similarities between points in a map [57]. Therefore, in the visualization experiments, we redefine the similarities under the model $q_{j|i}$

in such a way that the similarity of two points in a map is measured using a Cauchy distribution

$$q_{j|i} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} \left(1 + \|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right)^{-1}}{\sum_{m'} \sum_{i \neq k} \pi_i^{(m')} \pi_k^{(m')} \left(1 + \|\mathbf{y}_i^{(m')} - \mathbf{y}_k^{(m')}\|^2\right)^{-1}}. \tag{10}$$

This slightly changes the gradients of the cost function, but apart from that, it just like the multiple maps model proposed in the previous section.

We visualize the 40 word maps by showing them in an annotated scatter plot, in which the size of a dot represents the mixing proportion of a word in a specific map. To prevent the visualizations from being too cluttered, datapoints with a mixing proportion below 0.1 were removed from the visualization. To increase the legibility of the plots, the annotations in the scatter plot were manually aligned to reduce the overlaps between annotations, while ensuring that word labels are still near their corresponding point in the map.

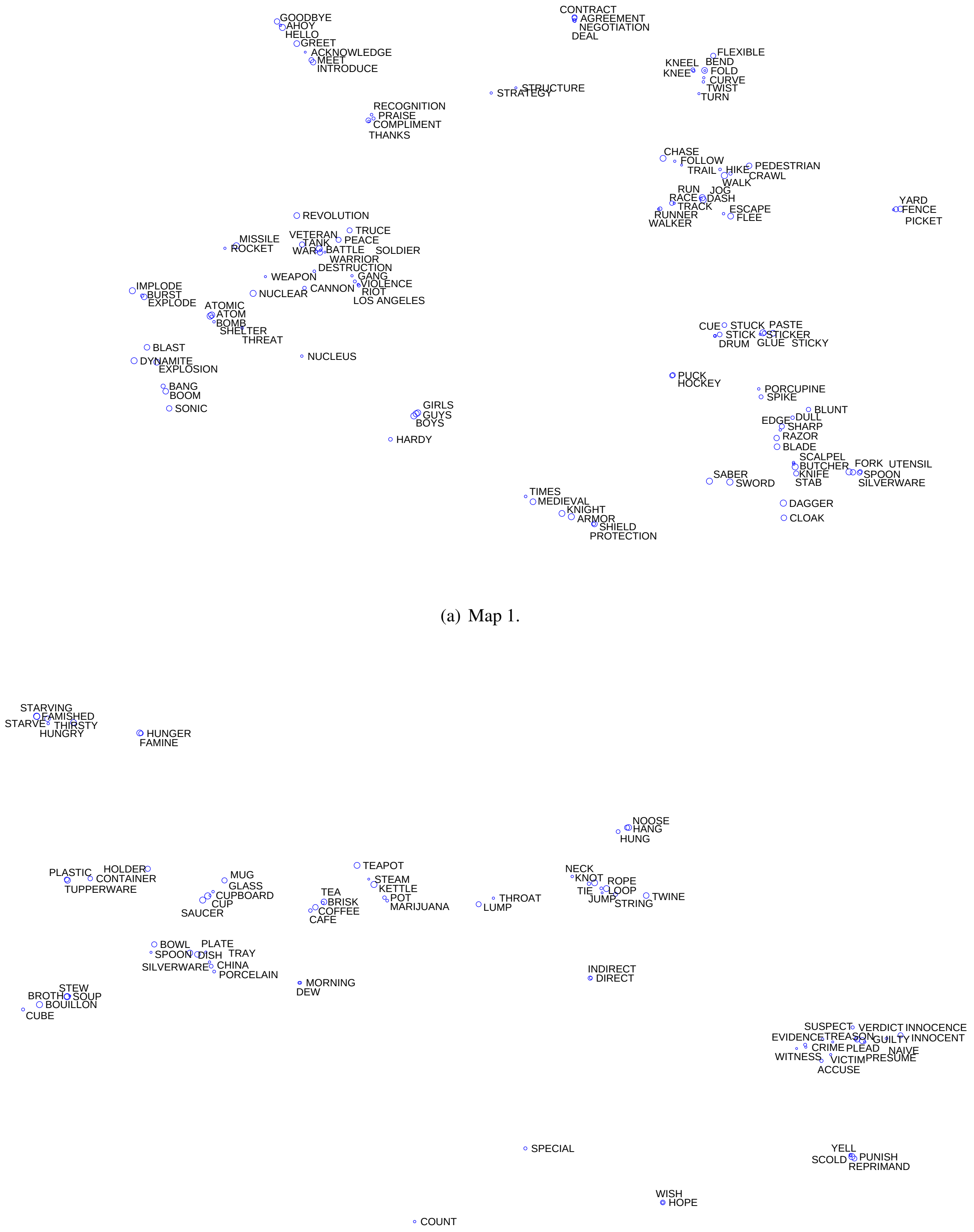
4.2 Results

Figure 2 shows 6 of the 40 maps that were constructed the multiple map model. The results reveal that the maps retain the similarity structure of the association data fairly well[6]. Because the data contains too many topics, a single map does not generally visualize a single topic. Instead, most maps reveal two or three main topics, as well as some very small local structures. For instance, map 2(d) visualizes the topics *sports* and *clothing*, and it shows small local structures that are related to, e.g., the Statue of Liberty: *monument - statue - liberty -freedom*.

The results reveal how a multiple map model circumvents the limitations of low-dimensional spaces. In particular, the model successfully models intransitive associations of words. For instance, the semantic relation of the word *tie* with words such as *suit*, *tuxedo*, and *prom* is modeled in map 2(a), whereas in map 2(d), the semantic relation of the word *tie* with *rope* and *knot* is modeled. In addition, map 2(e) reveals the semantic relation of *tie* with words such as *ribbon* and *bow*. As a second example, the semantic relation of the word *cheerleader* with various kinds of sports is modeled in map 2(d), whereas map 2(f) reveals the association of the word *cheerleader* with words such as *gorgeous*, *beauty*, and *sexy*. A third example is the word *monarchy*, which is modeled close to words that are related to royalty such as *king*, *queen*, *crown*, and *royal* in map 2(c). In map 2(f), the word *monarchy* is modeled close to other governmental forms such as *oligarchy*, *anarchy*, *democracy*, and *republic*.

The results also reveal how the multiple map model represents asymmetric pairwise similarities. For instance, map 2(c) reveals that the word *dynasty* is more often associated with the word *China* than the other way around. In map 2(c), the representations of both words are close to one another, however, the word *China* has a much smaller mixing proportion than *dynasty* in map 2(c). As a result, the denominator of Equation 8 is much higher for the word *China* than for the word *dynasty*, which implies that the word *dynasty* is closer to the word *China* under the model than the other way around.

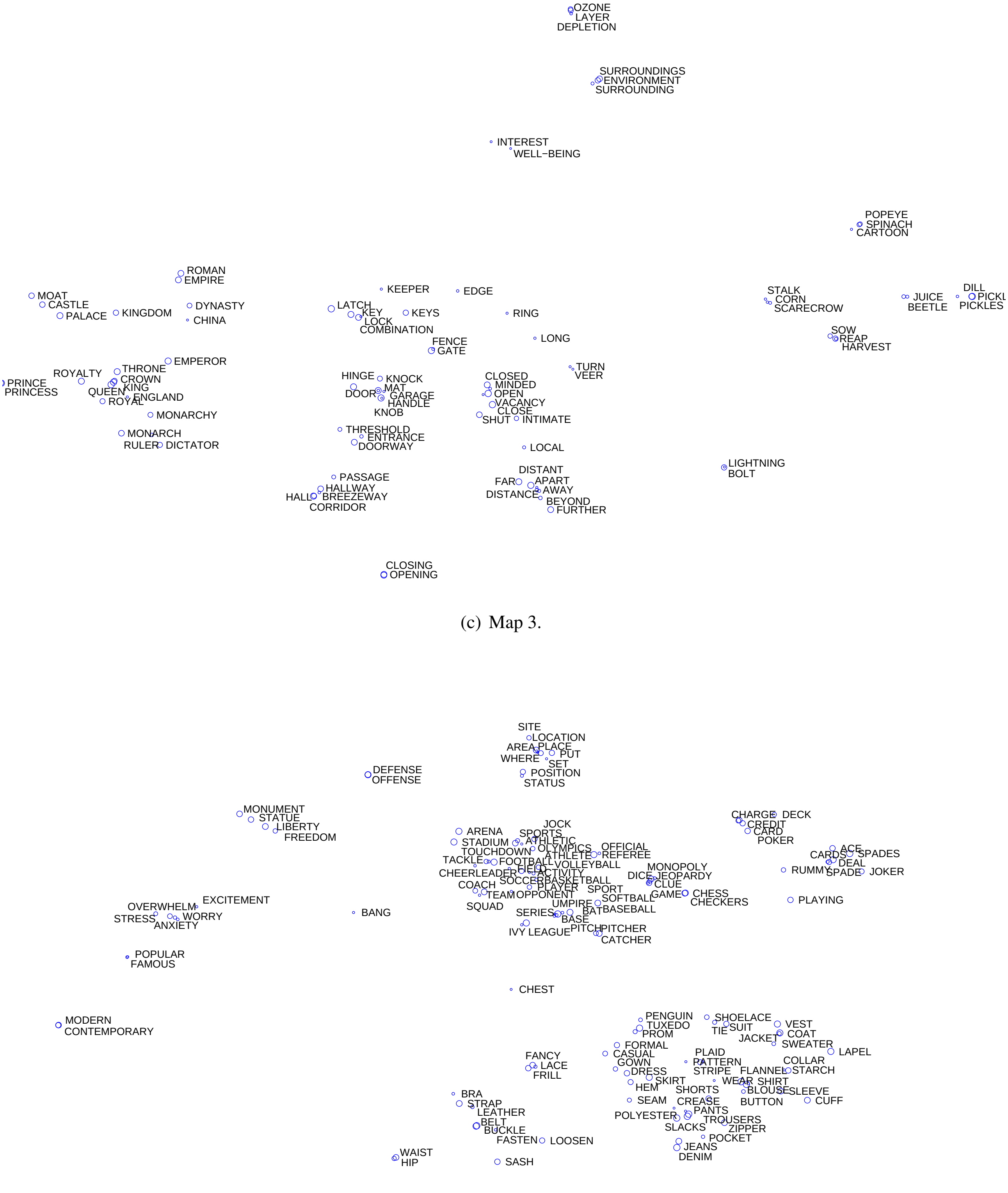
[6] Please note that the word association data does not exactly capture semantic similarity. For instance, in map 2(f), the word *beauty* is shown next to the word *beast*, revealing the word association that results from a famous Disney movie.



(a) Map 1.

(b) Map 2.

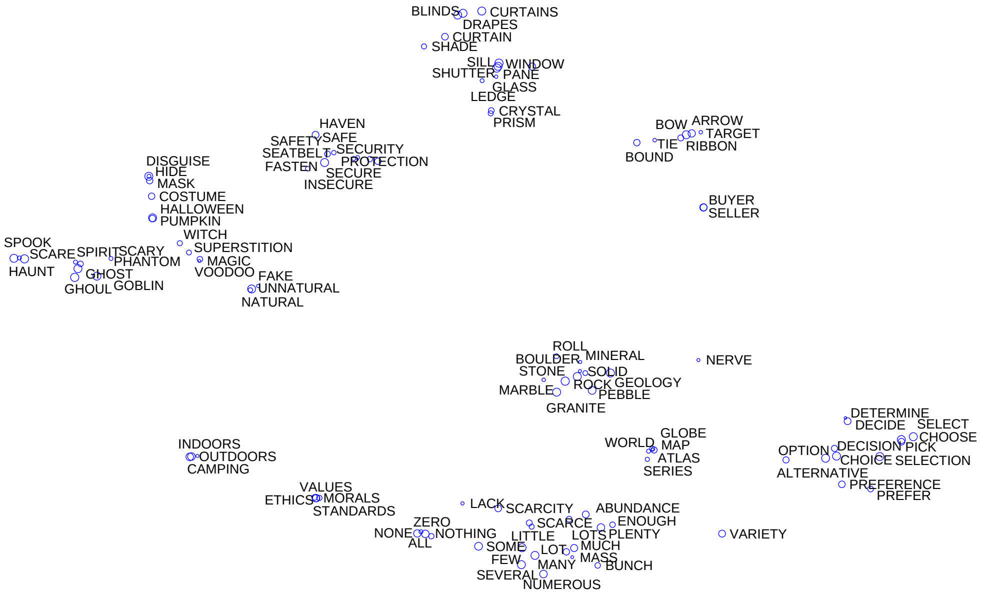
Figure 2: Results of the multiple map model on the word association dataset (a-b). Because of space limitations, we only show 6 of the original 40 maps.



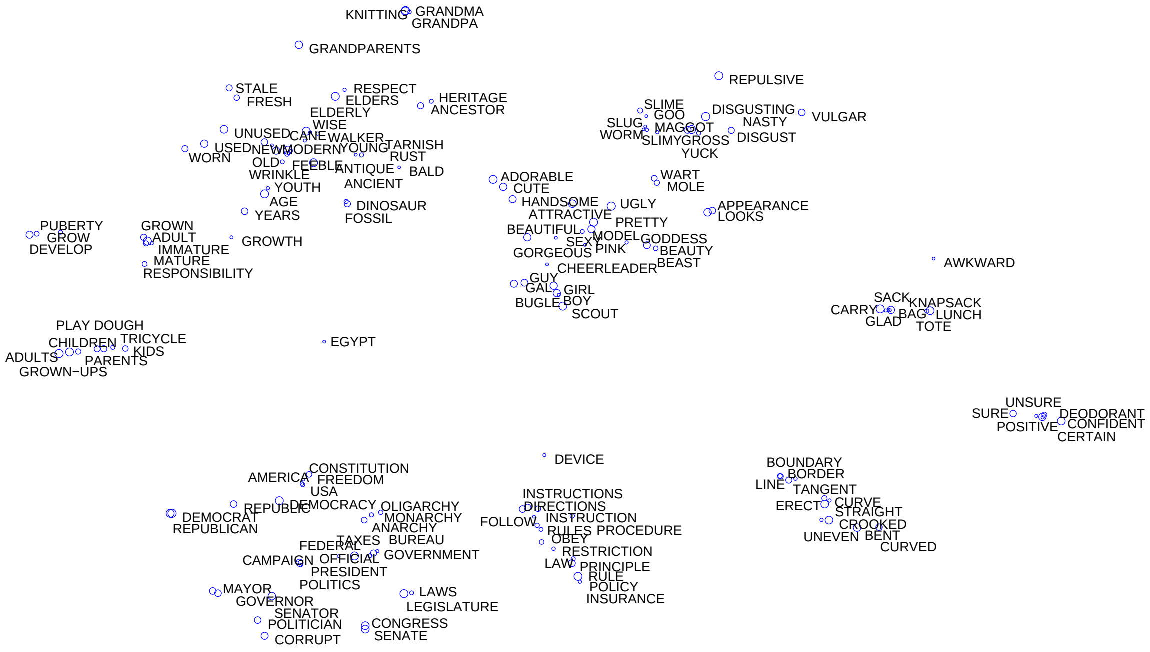
(c) Map 3.

(d) Map 4.

Figure 2: Results of the multiple map model on the word association dataset (c-d). Because of space limitations, we only show 6 of the original 40 maps.



(e) Map 5.



(f) Map 6.

Figure 2: Results of the multiple map model on the word association datasets (e-f). Because of space limitations, we only show 6 of the original 40 maps.

5 Generalization Experiments

In the previous section, we have shown how multiple map models can exploit the availability of multiple maps to, for example, model intransitive similarities between semantic concepts. However, these visualization results in themselves do not provide sufficient evidence to suggest their use as a computational model for semantic representation. In particular, one may argue that many of Tversky's arguments against multidimensional scaling can readily be resolved by increasing the dimensionality of the semantic space, or that these arguments are of limited relevance in prediction tasks in which context is available [14]. Such objections are frequently supported by the successful application of vector space models in information retrieval [22, 36, 59, 42].

In settings in which contextual clues are not available, such as in word association prediction, these objections against Tversky's arguments break down. Indeed, it is possible to model word associations well in very high-dimensional single map models, but this good performance is likely to be the result of *overfitting*. By allowing the semantic space to have a very large number of dimensions, it is possible to model given word associations appropriately even though the model does not form a good representation for the underlying process that generates the word associations. Such an overfitted single map model will perform poorly in generalization tasks, for instance, in a task in which the model has to predict held-out word associations. In order to investigate the performance of single and multiple map models in such a generalization task, we performed word association prediction experiments with both types of models.

5.1 Experimental setup

In the generalization experiments, we randomly divided the Florida State University word association data into three parts: 80% of the associations is used as training data, 10% is used as validation data, and 10% of the associations is used as test data. To train the model on the training data while ignoring the validation and test data, we need to redefine the cost function in such a way that it measures the error only over the association pairs that are part of the training data. Hence, instead of minimizing Equation 6 we minimize

$$C^{(train)} = \sum_i \sum_{j \neq i} \delta_{ij}^{(train)} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{11}$$

where $\delta_{ij}^{(train)}$ is an indicator variable that is 1 if the association between i and j is part of the training data, and 0 otherwise. Note that in the random selection of the training associations, we make sure that $\delta_{ij}^{(train)} = \delta_{ji}^{(train)}$. Also note that this redefinition of the pairwise similarities under the model in the training phase (slightly) changes the gradient of the cost function. In the same way that we redefined the cost function $C^{(train)}$ for the training data, we can also define the cost function for the test data $C^{(test)}$ and the cost function for the validation data $C^{(valid)}$. These functions measure the prediction error of the model.

During the training stage, we minimize $C^{(train)}$ using the same gradient descent method described earlier, but now we also employ early stopping [30, 6]. That is, we stop the minimization of the training error $C^{(train)}$ when the prediction error on the validation data, $C^{(valid)}$, starts to increase. In this way, we prevent the model from overfitting on the training data. As in the visualization experiments, we used a momentum term of 0.5 during the first 250 iterations, and of 0.8 afterwards. Again, we use early exaggeration with a factor of 4 during the first 250 iterations, and we used an adaptive learning rate scheme [18] with an initial learning rate of $1,000$. We assess the quality of the trained models by measuring the error in the predicted word associations on the held-out test data, $C^{(test)}$.

In the generalization experiments, the dimensionality of the maps is typically much larger than 2. As a result, the multiple map model is not hampered by the crowding problem, which forced us to use heavy-tailed similarity measurements in the visualization experiments. In the generalization experiments, therefore, we use the definition of pairwise similarities that is based on Shepard's universal law of generalization (see Equation 8), i.e., the similarity between two points in a map is exponentially decaying with their squared Euclidean distance in the map.

The multiple map model has two parameters which need to be set by the user: the number of maps M and the dimensionality of the maps D. The number of degrees of freedom in the multiple map model is equal to $NMD + N(M-1)$, as a result of which a model with $M = 1$ and $D = 70$ has roughly the same number of degrees of freedom as model with $M = 2$ and $D = 35$. In experiments in which we compared a single map model with multiple map models, therefore, we make sure that the product of the number of maps M and the map dimensionality D is fixed. In order to determine the value of the product MD for our experiments, we performed experiments with the single map model for a large number of values of D (ranging from 2 to 150), and we determined the number of dimensions D^* for which the generalization error of the single maps is minimized. In the experiments with the multiple map model, we set the values of M and D in such a way that the product MD is roughly equal to D^*. Hence, we gave the single map model a major advantage: the single map model is allowed to select its optimal number of degrees of freedom, and the comparison with the multiple map models is performed using that number of degrees of freedom.

5.2 Results

The results of the experiments in which we ran the single map model for a wide range of values of D (ranging from 2 to 150 with steps of 2 dimensions) revealed the single map model is best at word association prediction when D is set to 70 dimensions. The training, test, and validation errors of the 70-dimensional single map model ($M = 1$) are presented in Table 1.

In Table 1, we also present the prediction errors of the training, validation, and test data obtained after learning multiple map models ($M > 1$). Note that the parameters were set in such a way as to make sure that the product MD is roughly equal to 70. The training errors presented in the table measure how well the model represents the word associations that were used for training. The validation errors measure the prediction error on the held-out validation set that was used for early stopping. The test errors presented in the table measure the generalization performance of the trained models, i.e., the ability of the model to predict word associations[7]. The best performance across the models on each of the datasets is typeset in boldface.

The results reveal that, even though the single map model is best at representing the presented word association data (i.e., it has the lowest training error), a model with two maps performs better at the generalization task than the single map model (i.e., it has the lowest test error). The multiple map model with 2 maps of 35 dimensions achieves an error of 0.5195, whereas a single map model with 70 dimensions achieves an error of 0.5305. These results are quite remarkable, given that (1) the multiple map model has less space to model the word associations because volume decreases *exponentially* with dimensionality and (2) the selection of the model size was performed in such a way as to favor the single map model.

To demonstrate the advantage of having multiple maps, we also performed experiments with single map models with 35 dimensions. The results of these experiments reveal that such a model performs significantly worse than a model with two maps of 35 dimensions: the test error of the single map model is 0.5916, whereas the test error of the multiple map model is only 0.5195.

Although the results of the experiments clearly demonstrate the benefits of having multiple maps, they also reveal that for more than two maps, the exponential loss of space in the model hampers the performance of the multiple map model. Still, a model with three maps of 22 dimensions still performs on par with a single map model with 70 dimensions[8].

6 Discussion

In the previous two sections, we presented the results of visualization and generalization experiments that reveal the merits of multiple map SNE over single map multidimensional scaling techniques such

[7] Note that the prediction errors cannot be directly compared across datasets, because they are not normalized: the errors on the training set are always higher because the training set contains more associations.

[8] Again, we would like to emphasize that a 70-dimensional space is much larger than a 22-dimensional space, as the volume of a space decreases exponentially with it dimensionality.

Number of maps	*Number of dimensions*	*Training error*	*Validation error*	*Test error*
1	35	3.8937	0.5559	0.5916
1	70	**3.8289**	0.5048	0.5315
2	35	3.9579	**0.5032**	**0.5195**
3	22	4.0826	0.5164	0.5338
4	17	4.2255	0.5328	0.5461
5	14	4.2894	0.5367	0.5545

Table 1: Generalization errors for single and multiple map models.

as SNE. The ability of multiple map SNE to faithfully model nonmetric similarity data may have applications in, for instance, the visualization of stimulus-response pairs resulting from psychological or behavioral experiments, or the visualization of species co-occurence counts that are gathered by biologists [45]. However, in this section we will focus on the potential of multiple maps SNE as a computational model for semantic representation. We showed that the model may overcome many of the problems of semantic space models [49, 56, 22]. Below, we compare the theoretical properties of multiple map SNE with those of three alternative computational models for semantic representation: (1) semantic space models, (2) semantic networks, and (3) topic models.

1) Semantic space models. Semantic space models are similar to multiple map SNE in that they represent semantic concepts as points in a space in such a way, that similar concepts are represented close together in the space. In other words, semantic space models are based on the idea of implementing a second-order isomorphism between the representation space and the concepts in the world [11], which means that words with similar semantics should have a similar representation in the space.

Traditionally, multidimensional scaling models have been the most popular semantic space models [55, 49, 44], but these models are hampered by the limitations of metric spaces that we discussed in the introduction. For classical scaling [55], it is possible to partially overcome these problems by exploiting the structure from the eigenvectors that correspond to the negative eigenvalues of the Gram matrix when modeling non-metric similarities, as these eigenvectors may contain structural information on the metricity violations in the pairwise dissimilarity matrix [24, 23]. However, such an approach is limited for two main reasons. First, it is hard to interpret the map that corresponds to the negative part of the eigenspectrum: the map that corresponds to the positive part of the eigenspectrum is a metric approximation to the similarities, and the "negative map" is constructed in such a way as to correct the errors in the "positive map". Second, in contrast to our multiple model, approaches that employ the negative part of the eigenspectrum can only construct two metric maps[9]: a positive map and a negative map.

An alternative approach to address the limitations of metric spaces is by using an extended two-way Euclidean model with common and specific dimensions [58]. In such a model, the dissimilarity d_{ij} between two object representations $\mathbf{y}_i$ and $\mathbf{y}_j$ is extended with a variable s_i that measure the specifity of the object with index i. The model performs standard multidimensional scaling, where $d_{ij} = \sqrt{\|\mathbf{y}_i - \mathbf{y}_j\|^2 + s_i + s_j}$. The main advantage of this model over the traditional scaling models is that it can represent data with high centrality successfully. However, the model is not capable of representing intransitive semantic similarities faithfully.

Another alternative semantic space model, called "trajectory mapping", aims to address the limitations of low-dimensional metric spaces by constructing an object representation consists of a set of paths (the so-called "trajectories") through the objects that connects all objects that have a certain common feature [37]. The main drawback of this model is that it requires the "extrapolation" of features. How this extrapolation should be performed for semantic features remains unclear.

More recently, the Latent Semantic Analysis (LSA) model has gained popularity. LSA is a model

[9] We should note it is possible to consider these two maps as a hyperbolic space in which distance measures can be defined that do not obey the triangle inequality [34].

for semantic representation that was originally designed for use in information retrieval systems [22]. It computes a low-rank approximation of a word association or word co-occurence matrix by means of singular value decomposition (SVD). The most important output of LSA is formed by the k principal left-singular vectors of the low-rank approximation, where the importance of the singular vectors is determined by their corresponding singular values. The left-singular vectors provide a spatial representation for the words in the data in an orthogonal basis spanned by k vectors, hence, they represent words as points in the k-dimensional metric space $\mathbb{R}^k$. Semantic similarity in this space is typically measured using the cosine distance between word vectors, as a result of which semantic similarities under the LSA model obey all metric axioms. Therefore, LSA [22] and its probabilistic counterpart pLSA [17] are not fundamentally different from other semantic representation models that rely on second-order isomorphic representations. LSA is thus subject to all of the objections against multidimensional scaling that were formulated by Tversky[10] [56], as a result of which multiple map SNE has important advantages over (probabilistic) LSA. In contrast to (probabilistic) LSA, multiple map SNE can successfully model intransitive similarities and asymmetric similarities between semantic objects.

Other important semantic space models are models based on distributed representations that are typically employed in connectionist models of semantic representation [29, 20, 35, 38]. Distributed representations are fairly similar to multiple map SNE in that they allow an object to be represented by multiple points. However, an important problem with distributed representations is that automatically extracting a distributed semantic representation from text involves significant computational challenges, such as deciding how many senses each word should have and when those senses are being used. Until now, these problems have been alleviated by constructing the networks based on data that consists of labeled pairs of words and their meanings [39]. In contrast, multiple map SNE automatically learns a semantic representation from word associations (that can, in turn, be automatically extracted from text corpora [26]) and infers from the data how many senses each word has. The only restriction is that the number of senses for a single word cannot exceed the predefined number of maps, but it is unlikely that this restriction is violated if a sensible number of maps is used. This property of multiple map SNE gives it an important advantage over connectionist models for semantic representation.

2) Semantic networks. Semantic associative networks provide an intuitive way to model semantic similarities, and they provide simple solutions to problems such as word prediction, word disambiguation, and gist extraction [8, 7]. A semantic network consists of nodes that represent the words, and edges that represent the semantic similarities between the two words that the edges connect. When a word is observed, the node that corresponds to this word is activated. The resulting activation spreads through the semantic network, thereby activating nodes that are nearby in terms of the diffusion distance through the network. The strength of the activations in the nodes represents the semantic similarity of their corresponding words with the observed word.

Activations in undirected semantic networks can readily be represented in a distributed semantic representation [15, 47], and as a result, an undirected semantic network can be converted into a semantic space model using a bijective mapping. The semantic space corresponding to an undirected semantic network typically has a very high dimensionality, as a result of which the model has no problems with representing 'central' concepts. However, undirected semantic networks cannot represent asymmetric or intransitive semantic relations, because they obey the symmetry axiom and the triangle inequality, respectively. The former problem can be overcome by defining semantic networks as directed graphs, in which the weight of an edge from A to B may be different from the weight of the edge between B and A, causing similarities in the network to become asymmetric. However, this does not resolve problems with intransitive similarities. If node A has a strong connection to node B, and node B has a strong connection to node C, activation from node A will spread to node C, which makes A and C semantically related under the model. Multiple map SNE thus has significant advantages over models based on semantic networks, in particular, because it can infer the different senses of a word from the data. In contrast, semantic networks require manual specification of the senses[11].

[10] We are not the first authors to note the limitations of Latent Semantic Analysis. See for a more extensive coverage of the limitations of LSA, e.g., [14].

[11] Note that specifying senses may not be as trivial as it seems, as we explained in the introduction for the two senses of the

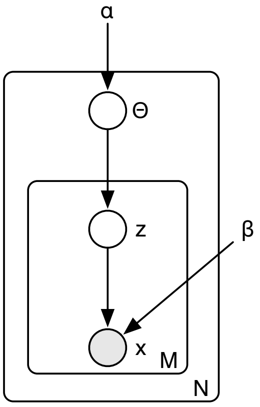


Figure 3: Generative process of Latent Dirichlet Allocation.

3) Topic models. Recently, a family of (Bayesian) latent variable models that originates from information retrieval has been proposed as computational models for semantic representation [14]. The most important examples of such models are the so-called *topic models*. Recently proposed topic models include Latent Dirichlet Allocation [4], the author model [27], the author-topic model [40], and the author-topic-recipient model [28]. Because of the popularity of Latent Dirichlet Allocation (LDA), we will focus on that model here. However, our discussion also holds for other topic models.

LDA was originally developed to model large text corpora. The key idea of LDA is that each word x has a topic z (picked from k topics) that is drawn from a topic distribution θ that is specific for a document. The graphical model of LDA is shown in Figure 3. The corresponding underlying generative process is given by

- For each of the N documents in the corpus:
  - Choose a topic distribution $\theta \sim Dirichlet(v)$
  - For each of the M words in the document:
    - Choose a topic $z \sim Multinomial(\theta)$
    - Choose a word $x \sim Multinomial(\beta_z)$

The latent variables in LDA are formed by: (i) k multinomial distributions z over all words and (ii) a distribution θ over these multinomial distributions[12]. The k multinomial distributions z can be viewed upon as topics, and each topic has its own multinomial distribution over words. The variable k is a parameter that sets the number of topics that is employed in the semantic representation. It may either be set by the user, or it may be learned from the data using non-parametric Bayesian techniques [3, 53].

Under a topic model, two words can be viewed as semantically similar if they both have a high probability under at least one of the k topics [14]. This provides topic models with the same desirable properties that multiple map SNE has. In particular, a topic model is capable of modeling intransitive semantic similarities in different topics. Analogous to our example with *tie*, *tuxedo*, and *knot*, in LDA, *tie* and *tuxedo* could be given a high probability in one topic and *tie* and *knot* could be given high probability in another topic, which would not make *tuxedo* similar to *knot* under the model. In the same way, LDA can model central objects by giving them a high probability in a large number of topics, which automatically gives rise to asymmetric similarities. The only requirement is that (as in multiple map SNE) sufficient topics are available to model the required centrality. The topics in LDA can be thought of as an equivalent for the maps in multiple map SNE.

The main difference between topic models and multiple map SNE is that, in contrast to LDA, multiple map SNE can (1) be trained directly on association or co-occurrence data and (2) capture subtle

word *monarchy*.

[12] The distribution over the multinomial distributions over all words is parametrized by means of a Dirichlet distribution, which is the conjugate prior of the multinomial distribution, see, e.g., [13].

semantic structure in the spatial structure of the maps. The first capability may be relevant depending on the input data that is available. The merits of the second capability are illustrated, for instance, in the 'sports' cluster in Figure 2(d), where the subtle semantic difference between physical sports such as *football*, *baseball*, and *volleybal*, and mental sports such as *chess*, *checkers*, and *poker* is captured in the spatial structure of the cluster (from left to right). In addition, multiple map SNE has the advantage that it can model small semantic structures that are not closely related to other semantic structures, such as the *Popeye* - *spinach* - *cartoon* cluster in Figure 2(c), without resorting to the construction of a new map or topic.

A minor disadvantage of multiple map SNE is that it is not tailored to clustering the concepts in the data, i.e., the concepts that have a high mixing proportion in a specific map do not necessarily all correspond to the same topic. This behavior of the model is due to the structure of the objective function: due to the asymmetry of the Kullback-Leibler divergence, the objective function does not severely penalize cases in which dissimilar objects (low $p_{j|i}$) both have a high mixing proportion in the same map (high q_{ij}). In cases in which it is desirable that maps only model a single topic, it is better to minimize the sum of the inverse Kullback-Leibler divergences $\sum_i KL(Q_i||P_i)$ instead of the sum of the "normal" divergences $\sum_i KL(P_i||Q_i)$. However, one should note that this may have a negative influence on the spatial layout of the maps, since in terms of spatial layout of the maps, the inverse Kullback-Leibler divergence will focus on modeling dissimilar objects far apart (i.e., focus on global data structure) instead of on modeling similar objects close together. We also note that clusters or topics do not necessarily need to be "clean" in order for a model to perform well. In fact, product of expert models outperform standard mixture topic models such as LDA, even though they do not learn very precise topics [43].

7 Conclusions

We presented a variant of multidimensional scaling that is capable of representing input objects in a collection of maps. The model alleviates the fundamental limitations of traditional multidimensional scaling techniques that are due to the metric axioms that hold in semantic space models. We presented results of visualization and generalization experiments on a dataset of word association data, revealing that the multiple map model is capable of accurately representing and predicting central concepts, as well as asymmetric and intransitive semantic relations. We compared the characteristics of the multiple map model with those of other computational models for semantic representation, and argued that the multiple map model has important advantages over popular semantic space models such as Latent Semantic Analysis. In particular, the multiple map model has characteristics that are similar to those of topic models that were recently proposed as computational models for semantic representation.

8 Acknowledgements

Part of this work was performed when Laurens van der Maaten was at Tilburg University. Laurens van der Maaten acknowledges support by NWO-CATCH, project RICH (grant 640.002.401), and support by the EU-FP7 Network of Excellence on Social Signal Processing (SSPNet). Geoffrey Hinton is a fellow of CIFAR and is supported by NSERC and CFI, as well as by gifts from Google and Microsoft.

A Gradients of the multiple map model

The multiple map SNE model minimizes the sum of Kullback-Leibler divergences between the pairwise similarities $p_{j|i}$ (where $p_{j|i} \geq 0$ and $\sum_j p_{j|i} = 1$) and the pairwise similarities in the multiple map with respect to the coordinates in the map $\mathbf{y}_i^{(m)}$ and the mixing proportions $\pi_i^{(m)}$ (which are in turn defined as a function of the mixture weights $w_i^{(m)}$). Mathematically, the cost function is given by

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_{j \neq i} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}, \tag{12}$$

where $q_{j|i}$ is the weighted sum of pairwise similarities between i and j over all low-dimensional maps

$$q_{j|i} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} \exp\left(-\|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right)}{\sum_{m'} \sum_{k \neq i} \pi_k^{(m')} \pi_l^{(m')} \exp\left(-\|\mathbf{y}_i^{(m')} - \mathbf{y}_k^{(m')}\|^2\right)}. \tag{13}$$

In the equation, $\pi_i^{(m)}$ is the mixing proportion of point i in map m, which is defined in terms of the mixing weights $w_i^{(m)}$ as

$$\pi_i^{(m)} = \frac{e^{-w_i^{(m)}}}{\sum_{m'} e^{-w_i^{(m')}}}. \tag{14}$$

The gradients of the cost function C with respect to the parameters of the model are given below.

To ease up the notation of the gradients somewhat, we denote the denominators of the the definition of $q_{j|i}$ in Equation 8 by Z_i, i.e., we define

$$Z_i = \sum_{m'} \sum_{k \neq i} \pi_i^{(m')} \pi_k^{(m')} \exp\left(-\|\mathbf{y}_i^{(m')} - \mathbf{y}_k^{(m')}\|^2\right). \tag{15}$$

The gradient of the cost function with respect to the low-dimensional map points $\mathbf{y}_i^{(m)}$ is given by

$$\frac{\delta C}{\delta y_i^{(m)}} = 2 \sum_j \left(\frac{\delta C}{\delta d_{ij}^{(m)}} + \frac{\delta C}{\delta d_{ji}^{(m)}} \right) \left(\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)} \right), \tag{16}$$

where $\frac{\delta C}{\delta d_{ij}^{(m)}}$ denotes the gradient of the cost function C with respect to the squared pairwise distance between point $\mathbf{y}_i$ and $\mathbf{y}_j$ in map m (i.e., $d_{ij}^{(m)} = \|\mathbf{y}_i - \mathbf{y}_j\|^2$). This gradient is given by

$$\frac{\delta C}{\delta d_{ij}^{(m)}} = \frac{\pi_i^{(m)} \pi_j^{(m)} \exp\left(-\|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right)}{q_{j|i} Z_i} (p_{j|i} - q_{j|i}). \tag{17}$$

The gradient of the cost function with respect to the mixing proportions $\pi_i^{(m)}$ is given by

$$\frac{\delta C}{\delta w_i^{(m)}} = \pi_i^{(m)} \left(\left(\sum_{m'} \pi_i^{(m')} \frac{\delta C}{\delta \pi_i^{(m')}} \right) - \frac{\delta C}{\delta \pi_i^{(m)}} \right), \tag{18}$$

where the gradient of the cost function with respect to the mixing proportions $\pi_i^{(m)}$ is given by

$$\frac{\delta C}{\delta \pi_i^{(m)}} = \sum_j \left(\frac{1}{q_{j|i} Z_i} (q_{j|i} - p_{j|i}) + \frac{1}{q_{i|j} Z_j} (q_{i|j} - p_{i|j}) \right) \pi_j^{(m)} \exp\left(-\|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right). \tag{19}$$

References

[1] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[2] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006.

[3] D.M. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, pages 17–24, Cambridge, MA, 2004. The MIT Press.

[4] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F. Fogelman-Soulie and J. Hérault, editors, *Neurocomputing: Algorithms, Architectures, and Applications*. Springer-Verlag, New York, NY, 1989.

[6] R. Caruana, S. Lawrence, and L. Giles. Overtting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, volume 13, pages 402–408, Cambridge, MA, 2001. The MIT Press.

[7] A.M. Collins and E.F. Loftus. A spreading activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

[8] A.M. Collins and M.R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, 8:240–247, 1969.

[9] J.A. Cook, I. Sutskever, A. Mnih, and G.E. Hinton. Visualizing similarity data with a mixture of maps. *JMLR Workshop and Conference Proceedings*, 2:67–74, 2007.

[10] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 1991.

[11] S. Edelman and S. Duvdevani-Bar. Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 1997.

[12] C. Fellbaum. *Wordnet: An Electronic Lexical Database*. Bradford Books, Cambridge, MA, USA, 1998.

[13] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, NY, 1995.

[14] T.L. Griffiths, M. Steyvers, and J.L. Tenenbaum. Topics in semantic representation. *Psychological Review*, 114(2):211–244, 2007.

[15] G.E. Hinton. Implementing semantic networks in parallel hardware. In G.E. Hinton and J.A. Anderson, editors, *Parallel Models of Associative Memory*. Erlbaum, Hillsdale, NJ, 1981.

[16] G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840, Cambridge, MA, 2003. The MIT Press.

[17] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the* 22^{th} *Annual International SIGIR Conference*, pages 50–57, New York, NY, 1999. ACM Press.

[18] R.A. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1:295–307, 1988.

[19] F. Jäkel, B. Schölkopf, and F.A. Wichmann. Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, 52(2):297–303, 2008.

[20] A.H. Kawamoto. Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, 32:474–516, 1993.

[21] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

[22] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[23] J. Laub, J. Macke, K.-R. Müller, and F.A. Wichmann. Inducing metric violations in human similarity judgements. In *Advances in Neural Information Processing Systems*, volume 19, pages 777–784, Cambridge, MA, 2007. The MIT Press.

[24] J. Laub and K.-R. Müller. Feature discovery in non-metric pairwise data. *Journal of Machine Learning Research*, 5:801–818, 2004.

[25] R.D. Luce. Detection and recognition. In R.D. Luce, R.R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, pages 103–190, New York, NY, 1963. Wiley.

[26] K. Lund, C. Burgess, and R.A. Atchley. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the* 17^{th} *Annual Conference of the Cognitive Science Society*, pages 660–665, Mahwah, NJ, 1995. Erlbaum.

[27] A. McCallum. Multi-label text classication with a mixture model trained by em. In *AAAI Workshop on Text Learning*, New York, NY, 1999. ACM.

[28] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email. Technical Report UM-CS-2004-096, Department of Computer Science, University of Massachusetts, Amherst, MA, 2004.

[29] J.L. McClelland and D.E. Rumelhart. An interactive activiation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, 88(5):375–407, 1981.

[30] N. Morgan and H. Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 630–637, San Mateo, CA, 2009. Morgan Kaufman Publishers Inc.

[31] K. Nam, H. Je, and S. Choi. Fast Stochastic Neighbor Embedding: A trust-region algorithm. In *Proceedings of the IEEE International Joint Conference on Neural Networks 2004*, volume 1, pages 123–128, New York, NY, 2004. ACM Press.

[32] D.L. Nelson, C.L. McEvoy, and T.A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms, 1998.

[33] R. Nosofsky. Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, 38(5):415–432, 1985.

[34] E. Pekalska and R.P.W. Duin. *The dissimilarity representation for pattern recognition: Foundations and Applications*. World Scientific, Singapore, 2005.

[35] D.C. Plaut. Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, 12:765–805, 1997.

[36] B. Rehder, M.E. Schreiner, M.B. Wolfe, D. Laham, T.K. Landauer, and W. Kintsch. Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25:337–354, 1998.

[37] W. Richards and J.J. Koenderink. Trajectory mapping: a new nonmetric scaling technique. *Perception*, 24(11):1315–1331, 1995.

[38] J.M. Rodd, M.G. Gaskell, and W.D. Marslen-Wilson. Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, 28:89–104, 2004.

[39] T. Rogers and J. McClelland. *Semantic cognition: A parallel distributed processing approach*. The MIT Press, Cambridge, MA, 2004.

[40] M. Rosen-Zvi, T. Grifths, and M. Steyversand P. Smyth. The author-topic model for authors and documents. In *Proceedings of the* 20^{th} *Conference on Uncertainty in Articial Intelligence*, Arlington, VA, 2004. AUAI Press.

[41] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[42] R.R. Salakhutdinov and G.E. Hinton. Semantic hashing. In *Proceedings of the SIGIR 2007 Workshop on Information Retrieval and Applications of Graphical Models*, pages 52–63, New York, NY, 2007. ACM Press.

[43] R.R. Salakhutdinov and G.E. Hinton. Replicated softmax: an undirected topic model. In *Advances in Neural Information Processing Systems*, volume 22, Cambridge, MA, 2010. The MIT Press.

[44] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5):401–409, 1969.

[45] S. Schmidtlein, P. Zimmermann, R. Schüpferling, and C. Weiss. Mapping the floristic continuum: Ordination space position estimated from imaging spectroscopy. *Journal of Vegetation Science*, 18:131–140, 2007.

[46] B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, Cambridge, MA, 2002.

[47] L. Shastri and V. Ajjanagadde. From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16(3):417–494, 1993.

[48] R.N. Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22:325–345, 1957.

[49] R.N. Shepard. Cognitive psychology: A review of the book by U. Neisser. *American Journal of Psychology*, 81:285–289, 1968.

[50] R.N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[51] J.E.K. Smith. Models of identification. In R. Nickerson, editor, *Attention and Performance VIII*, Hillsdale, NJ, 1980. Erlbaum.

[52] M. Steyvers and J.B. Tenenbaum. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78, 2005.

[53] Y. Teh, M.I. Jordan, M. Beal, and D.M. Blei. Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, volume 17, pages 1385–1392, Cambridge, MA, 2004. The MIT Press.

[54] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[55] W.S. Torgerson. Multidimensional scaling I: Theory and method. *Psychometrika*, 17:401–419, 1952.

[56] A. Tversky and J.W. Hutchinson. Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(11):3–22, 1986.

[57] L.J.P. van der Maaten and G.E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2431–2456, 2008.

[58] S. Winsberg and J. Douglas Carrol. A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. *Psychometrika*, 54(2):217–229, 1989.

[59] M.B. Wolfe, M.E. Schreiner, B. Rehder, D. Laham, P.W. Foltz, W. Kintsch, and T. Landauer. Learning from text: Matching readers and text by latent semantic analysis. *Discourse Processes*, 25:309–336, 1998.

[60] Z. Yang, I. King, E. Oja, and Z. Xu. Heavy-tailed symmetric stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, volume 22, Cambridge, MA, 2010. The MIT Press.