MODEL COMPARISON FOR SEMANTIC GROUPING

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a probabilistic framework for quantifying the semantic similarity between two groups of embeddings. We formulate this as a model comparison task in which we contrast a generative model that encodes similarity between the two groups versus one that does not. We illustrate how this framework can be used for the Semantic Text Similarity (STS) task using clear assumptions about how the embeddings of words are generated. We apply information criteria based model comparison to overcome the shortcomings of Bayesian model comparison, whilst still penalising model complexity. We achieve competitive results by applying the proposed framework with an appropriate choice of likelihood on the STS datasets.

1 INTRODUCTION

The problem of semantic textual similarity (STS), measuring how closely the meaning of one piece of text corresponds to that of another, has been studied in the hope of improving performance across various problems in Natural Language Processing (NLP), including information retrieval (Zheng & Callan, 2015). Recent progress in the generation of word embeddings has allowed the encoding of words using distributed vector representations, which capture semantic information through their location in the learned embedding space. However, it remains a challenge to adapt these representations of individual words to express semantic similarity between words groups, like documents, sentences, and other textual formats.

Recent methods for STS rely on additive composition of word vectors (Arora et al., 2016; Blacoe & Lapata, 2012; Mitchell & Lapata, 2008; 2010) or complex deep learning architectures (Kiros et al., 2015), which summarise a sentence through a single embedding. Resulting sentence vectors are compared using cosine similarity, a choice which is well supported by empirical results, but which is not backed by a solid theoretical foundation. It has been suggested that this should be addressed by analysing the geometry of the embedding space (Almahairi et al., 2015; Hill et al., 2016; Wieting & Gimpel, 2017). However, this approach has so far not led to significant insight.

We approach the problem of STS from a novel direction. The compositional nature of distributed representations demonstrated in Mikolov et al. (2013) and Pennington et al. (2014) indicate the presence of semantic groups in the representation space of word embeddings; an idea which has been further explored in Athiwaratkun & Wilson (2017). Under this assumption, the task of semantic similarity can be formulated as the following question: "Are two arbitrary subgroups of words partitions of the same semantic group?" Figure 1 demonstrates this concept.

In this work we propose a method that brings a paradigm shift from estimating sentence embeddings to a model comparison framework in the task of semantic text similarity. We employ a generative model for sentences similar to Arora et al. (2016), but our goal differs from theirs in that our similarity is based on the aforementioned model comparison test, whilst theirs is based on the inner product of sentence embeddings derived as maximum likelihood estimators.

Assuming a probability distribution of word embeddings over the unit hypersphere, we achieve results comparable to Arora et al. (2016) on the STS dataset in $\mathcal{O}(nd)$ time on average – compared to the $\mathcal{O}(nd^2)$ average complexity of their method (where n is the number of words in a sentence, and d is the embedding size). A further advantage of our framework is that it naturally lends itself to online and unsupervised settings. We define an online setting as one in which we have access to only a single query pair at a time.

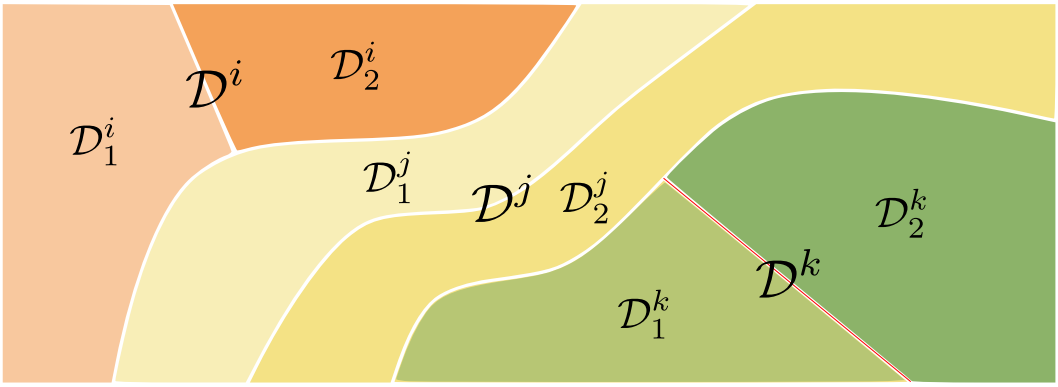

Figure 1: This paper models the generation of embeddings of words in a sentence through a latent variable (the semantic groups). Above we visualize the semantic group space of embeddings (with $\mathcal{D}^i, \mathcal{D}^j, \mathcal{D}^k$ representing different semantic groups). This enables us to perform a model comparison test to assess similarity - are the two sentences part of the same semantic group like $\mathcal{D}_1^i$ and $\mathcal{D}_2^i$ or are they from different semantic groups like $\mathcal{D}_1^i$ and $\mathcal{D}_2^k$?

2 Methodology and Background

In this work we formulate the task of semantic similarity between two arbitrary groups of objects as a model comparison problem. Taking inspiration from Ghahramani & Heller (2006) and Marshall et al. (2006) we propose the generative models for groups (e.g. sentences) $\mathcal{D}_1, \mathcal{D}_2$ seen in Figure 2. We adopt and adapt this idea from the Bayesian model comparison setting to an Information Criterion based test.

2.1 Bayesian Model Comparison

The Bayes factor, traditionally used in Bayesian model comparison under the graphical models in Figure 2 is defined as

$$\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = \log \frac{p(\mathcal{D}_1, \mathcal{D}_2|\mathcal{M}_1)}{p(\mathcal{D}_1, \mathcal{D}_2|\mathcal{M}_2)} = \log \frac{p(\mathcal{D}_1, \mathcal{D}_2|\mathcal{M}_1)}{p(\mathcal{D}_1|\mathcal{M}_2)p(\mathcal{D}_2|\mathcal{M}_2)}. \tag{1}$$

To obtain the evidence $p(\mathcal{D}|\mathcal{M}_i)$ the parameters of the model must be marginalised (where $\oplus$ means concatenation):

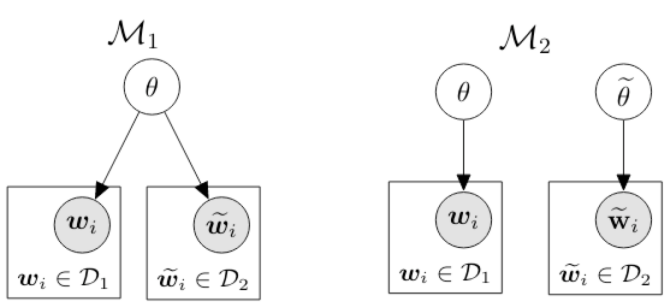

Figure 2: On the left, $\mathcal{M}_1$ assumes that both datasets are generated i.i.d. from the same parametric distribution. On the right, $\mathcal{M}_2$ assumes that the datasets are generated i.i.d. from distinct parametric distributions.

$$p(\mathcal{D}_1, \mathcal{D}_2|\mathcal{M}_1) = \int p(\mathcal{D}_1, \mathcal{D}_2|\theta)p(\theta)d\theta = \int \prod_{\boldsymbol{w}_k \in \mathcal{D}_1 \oplus \mathcal{D}_2} p(\boldsymbol{w}_k|\theta)p(\theta)d\theta,$$
$$p(\mathcal{D}_i|\mathcal{M}_2) = \int \prod_{\boldsymbol{w}_k \in \mathcal{D}_i} p(\boldsymbol{w}_k|\theta)p(\theta)d\theta.$$

Computing the semantic similarity score of the two groups $\mathcal{D}_1, \mathcal{D}_2$ under the Bayesian framework requires selecting a reasonable model likelihood $p(\boldsymbol{w}_k|\theta)$, prior $p(\theta)$, and computing the marginal evidence specified above. Computing the evidence can be computationally intensive and usually requires approximation.

We have extended the method Ghahramani & Heller (2006) to compare two groups. The method by Marshall et al. (2006) applies the same score as in Equation 1 in the context of merging datasets whilst we apply it in the setting of information retrieval and semantic text similarity.

We differ from both methods in that we choose to work with information criteria based model comparison as opposed to using the Bayes factor. The Bayes factor is very sensitive to the choice of prior and can result in estimates that heavily underfit the data (especially under a vague prior), having the tendency to always select the simpler model; which is argued further by M. S. Bartlett (1957) and Akaike et al. (1981). This is handled in Ghahramani & Heller (2006) by using the empirical Bayes procedure, a shortcoming of which is the issue of double counting (Berger, 2000) and thus being prone to over-fitting. We address this issue by exploring information theoretic criteria to design a log-ratio-like test that is prior free and robust to overfitting.

2.2 Information Criteria

We seek to compare models $\mathcal{M}_1$ and $\mathcal{M}_2$ using information criteria (IC) to assess the goodness of fit of each model. There are multiple information criteria used for model selection, each with different settings to which they are better suited. The information criteria which we will be working with have the general form

$$\mathrm{IC}(\mathcal{D}, \mathcal{M}) = -\left(\alpha\hat{\mathcal{L}}_{\mathcal{M}} + \Omega\left(\mathcal{D}, \mathcal{M}\right)\right),$$

where $\hat{\mathcal{L}}_{\mathcal{M}} = \mathcal{L}(\hat{\theta}|\mathcal{D}, \mathcal{M})$ is the maximised value of the log likelihood function for model $\mathcal{M}$, α is a scalar derived for each IC, and $\Omega\left(\mathcal{D}, \mathcal{M}\right)$ represents a model complexity penalty term which is model and IC specific. Using this general formulation for the involved information criteria yields the following similarity score:

$$\begin{aligned}\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) &= -\,\mathrm{IC}(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_1) + \mathrm{IC}(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_2),\\ \mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) &= \alpha\left(\hat{\mathcal{L}}(\hat{\theta}_{1,2}|\mathcal{M}_1) - (\hat{\mathcal{L}}(\hat{\theta}_1|\mathcal{M}_2) + \hat{\mathcal{L}}(\hat{\theta}_2|\mathcal{M}_2))\right)\\ &\quad - \Omega\left(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_1\right) + \Omega\left(\{\mathcal{D}_1, \mathcal{D}_2\}, \mathcal{M}_2\right).\end{aligned}$$

Examples of these criteria can be put into two broad classes. Bayesian Information Criterion (BIC) is an example of an IC that approximates the model evidence directly, as defined in Schwarz et al. (1978). Empirically it has been shown that the BIC is likely to underfit the data, especially when the number of samples is small (Dziak et al., 2012). Sentences contain a relatively small number of words (samples), thus we focus on the second class - Information Theoretic Criteria.

2.2.1 Information Theoretic Criteria

The Information Theoretic Criteria (ITC) are a family of model selection criteria. The task they address is evaluating the expected quality of an estimated model specified by $\mathcal{L}(\hat{\theta}|\boldsymbol{w})$ when it is used to generate unseen data from the true distribution $G(\boldsymbol{w})$, as defined in Konishi & Kitagawa (2008b).

These family of criteria perform this evaluation by using the Kullback-Leibler (KL) divergence between the true model $G(\boldsymbol{w})$ and the fitted model $\mathcal{L}(\hat{\theta}|\boldsymbol{w})$, with the aim of selecting the model (from a given set of models) that minimizes the quantity

$$\begin{aligned} D_{KL}\left(G(\boldsymbol{w})\middle|\middle|p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})\right) &= \mathbb{E}_G\left[\ln\frac{G(\boldsymbol{w})}{p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})}\right] \\ &= -H_G(\boldsymbol{w}) - \mathbb{E}_G\left[\ln p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})\right]. \end{aligned}$$

The entropy of the true model $H_G(\boldsymbol{w})$ will remain constant across different likelihoods. Thus, the quantity of interest in the definition of the information criterion under consideration is given by the expected log likelihood under the true model $\mathbb{E}_G[\ln p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})]$. The goal then, is to find a good estimator for this quantity. One such estimator is given by the normalized maximum likelihood:

$$\mathbb{E}_{\hat{G}}\left[\ln p(\boldsymbol{w}|\hat{\boldsymbol{\theta}})\right] = \frac{1}{n}\sum_{i=1}^{n}\ln p(\boldsymbol{w}_i|\hat{\boldsymbol{\theta}}),$$

where $\hat{G}$ represents the empirical distribution. This estimator introduces a bias that varies with respect to the dimension of the model's parameter vector $\boldsymbol{\theta}$ and requires a correction in order to carry out a fair comparison of information criteria between models. In Takeuchi (1976) a model specific correction is derived resulting in the following information criterion, called Takeuchi Information Criterion (TIC):

$$\mathrm{TIC}(\mathcal{D},\mathcal{M}) = -2\left(\hat{\mathcal{L}}_{\mathcal{M}} - \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}\right)\right),$$

where

$$\hat{\boldsymbol{J}} = \frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w}_i)\big|_{\hat{\boldsymbol{\theta}}}, \qquad \hat{\boldsymbol{\mathcal{I}}} = \frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w}_i)\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\top}(\boldsymbol{\theta}|\boldsymbol{w}_i)\big|_{\hat{\boldsymbol{\theta}}}.$$

Then, under the TIC criterion we have the following similarity (derivation in Appendix A):

$$\begin{aligned} \mathbf{sim}(\mathcal{D}_1,\mathcal{D}_2) = 2\Big(&\mathcal{L}(\hat{\theta}_{1,2}|\mathcal{M}_1) - \mathcal{L}(\hat{\theta}_1|\mathcal{M}_2) - \mathcal{L}(\hat{\theta}_2|\mathcal{M}_2) \\ &- \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_{1,2}\hat{\boldsymbol{J}}_{1,2}^{-1}\right) + \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_1\hat{\boldsymbol{J}}_1^{-1}\right) + \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_2\hat{\boldsymbol{J}}_2^{-1}\right)\Big). \end{aligned}$$

To make this entire procedure concrete, we provide Algorithm 1.

For the case where we assume our model has the same parametric form as the true model and as $n \to \infty$, the equality $\hat{\boldsymbol{\mathcal{I}}} = \boldsymbol{J}$ holds resulting in a penalty of $\mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}\right) = \mathrm{tr}(\boldsymbol{I}_k) = k$ where k is the number of model parameters, as shown in Konishi & Kitagawa (2008a). This results in the Akaike Information Criterion (AIC) by Akaike (1974):

$$\mathrm{AIC}(\mathcal{D},\mathcal{M}) = -2(\hat{\mathcal{L}}_{\mathcal{M}} - k).$$

The AIC simplification of TIC relies on several assumptions that only hold true in the big data limit and rely on assuming our model $\mathcal{M}$ has the same parametric form as the true model (Konishi & Kitagawa (2008a)). In general TIC is a more robust approximation, especially in the application we study where we consider the datasets to be sentences and thus the number of samples is small.

Algorithm 1 Calculate $\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2)$

1: **input**: Sentences $\mathcal{D}_1 := \{\boldsymbol{w}_k\}_{k=1}^m, \mathcal{D}_2 = \{\widetilde{\boldsymbol{w}}_k\}_{k=1}^l$
2: Compute the maximum likelihood estimates under the 2 models :
$\hat{\boldsymbol{\theta}}_{1,2} := \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}_1 \oplus \mathcal{D}_2)$
$\hat{\boldsymbol{\theta}}_1 := \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}_1)$
$\hat{\boldsymbol{\theta}}_2 := \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}_2)$
3: Calculate the grad vector and Hessian for each model:
$\nabla_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w})$, $\nabla_{\boldsymbol{\theta}_2} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w})$, $\nabla_{\boldsymbol{\theta}_{1,2}} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w})$
$\nabla^2_{\boldsymbol{\theta}_1} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w})$, $\nabla^2_{\boldsymbol{\theta}_2} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w})$, $\nabla^2_{\boldsymbol{\theta}_{1,2}} \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{w})$
4: $\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) := \alpha \left(\hat{\mathcal{L}}(\hat{\theta}_{1,2}|\mathcal{M}_1) - (\hat{\mathcal{L}}(\hat{\theta}_1|\mathcal{M}_2) + \hat{\mathcal{L}}(\hat{\theta}_2|\mathcal{M}_2)) \right)$
$- \operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_{1,2}\hat{\boldsymbol{J}}_{1,2}^{-1}\right) + \left(\operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_1\hat{\boldsymbol{J}}_1^{-1}\right) + \operatorname{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_2\hat{\boldsymbol{J}}_2^{-1}\right)\right)$
5: **output:** $\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2)$

2.3 Von Mises-Fisher Likelihood

Word embeddings are d dimensional vectors of real numbers, that are traditionally learned by optimizing a dot product between target words and context vectors (Mikolov et al., 2013). Due to this training setup, cosine similarity is used to measure the semantic similarity of words in various information retrieval tasks. Thus, we want to use a suitable distribution induced by the cosine similarity measure. We model our embeddings as vectors lying on the surface of the $d-1$ dimensional unit hyper-sphere $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and i.i.d according to a von Mises-Fisher (vMF) likelihood (Fisher et al. (1993)):

$$p(\boldsymbol{w}|\boldsymbol{\mu}, \kappa) = \frac{1}{Z(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{w}\right),$$
$$p(\boldsymbol{w}|\boldsymbol{\mu}, \kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)} \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{w}\right),$$

where $\boldsymbol{\mu}$ is the mean direction vector and κ is the concentration parameter, with supports $||\boldsymbol{\mu}|| = ||\boldsymbol{w}|| = 1,\ \kappa \geq 0$. The term $I_\nu(\kappa)$ Corresponds to a modified Bessel function of the first kind with order ν.

We reparametrise the random variable to polar hyperspherical $\boldsymbol{w}(\boldsymbol{\phi})$ ($\boldsymbol{\phi} = (\phi_1, ..., \phi_{d-1})^\top$) as adopted in Mabdia (1975):

$$p(\boldsymbol{\phi}|\boldsymbol{\theta}, \kappa) = \left(\frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)}\right) \left|\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{\phi}}\right| \exp\left(\kappa \boldsymbol{\mu}(\boldsymbol{\theta})^\top \boldsymbol{w}(\boldsymbol{\phi})\right),$$

where

$$w_i(\boldsymbol{\phi}) = ((1-\delta_{id})\cos\phi_i + \delta_{id}) \prod_{k=1}^{i-1} \sin\phi_k\ , \quad \mu_i(\boldsymbol{\theta}) = ((1-\delta_{id})\cos\theta_i + \delta_{id}) \prod_{k=1}^{i-1} \sin\theta_k,$$
$$\left|\frac{\partial \boldsymbol{w}}{\partial \boldsymbol{\phi}}\right| = \prod_{k=1}^{d-2} (\sin\phi_k)^{d-k-1}.$$

This reparametrisation simplifies the calculation of the Hessian and the empirical Fisher information matrix. The maxima of the likelihood remains unchanged since $|\partial \boldsymbol{w}/\partial \boldsymbol{\phi}|$ does not depend on $\boldsymbol{\theta}$ thus the MLE estimate in the hyper-shperical coordinates parametrisation is just given by applying the map from the cartesian MLE to the polars.

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^n \boldsymbol{w}_i}{||\sum_{i=1}^n \boldsymbol{w}_i||}, \quad \hat{\boldsymbol{\theta}} = \boldsymbol{\mu}^{-1}(\hat{\boldsymbol{\mu}}),$$
$$A_d(\kappa) = \frac{I_{d/2}}{I_{d/2-1}}, \quad \bar{R} = \frac{||\sum_{i=1}^n \boldsymbol{w}_i||}{n},$$
$$\hat{\kappa} = A_d^{-1}(\bar{R}) \approx \frac{\bar{R}(d - \bar{R}^2)}{1 - \bar{R}^2}.$$

where both the derivation and approximation for the MLE estimates are derived in Banerjee et al. (2005).

The log likelihood is then (where $\mathcal{D} = \{\boldsymbol{\phi}_i\}_{i=1}^n$ is the dataset)

$$\mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) = \kappa \boldsymbol{w}(\boldsymbol{\phi})^T \boldsymbol{\mu}(\boldsymbol{\theta}) - \log Z(\kappa) + \log \left| \frac{\partial \boldsymbol{w}}{\partial \boldsymbol{\phi}} \right|,$$
$$\mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n)) = \sum_{i=1}^n \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}_i).$$

The first and second order partial derivatives of the vMF log likelihood are (derived in Appendix B):

$$\frac{\partial}{\partial \theta_k} \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) = \kappa \sum_{j=k}^d w_j(\boldsymbol{\phi}) \mu_j(\boldsymbol{\theta})((1 - \delta_{kj}) \cot \theta_k - \delta_{kj} \tan \theta_k),$$
$$\frac{\partial^2}{\partial \theta_k^2} \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) = -\kappa \sum_{j=k}^d w_j(\boldsymbol{\phi}) \mu_j(\boldsymbol{\theta}),$$
$$\frac{\partial}{\partial \kappa} \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) = \boldsymbol{w}(\boldsymbol{\phi})^\top \boldsymbol{\mu}(\boldsymbol{\theta}) - \frac{I_{\frac{d}{2}}(\kappa)}{I_{\frac{d}{2}-1}(\kappa)},$$
$$\frac{\partial^2}{\partial \kappa^2} \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) = \frac{I_{\frac{d}{2}}(\kappa)(I_{\frac{d}{2}-2}(\kappa) + I_{\frac{d}{2}}(\kappa)) - I_{\frac{d}{2}-1}(\kappa)(I_{\frac{d}{2}-1}(\kappa) + I_{\frac{d}{2}+1}(\kappa))}{2 I_{\frac{d}{2}-1}(\kappa)^2}.$$

We prove (in Appendix B) that the mixed derivatives of the vMF log likelihood are a constant (with respect to $\boldsymbol{\phi}$) times $\partial \mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) / \partial \theta_k$. Thus, evaluated at the MLE, these entries are zero.

$$\left. \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n)}{\partial \kappa \partial \theta_k} \right|_{\kappa = \hat{\kappa}, \theta_k = \hat{\theta_k}} = \hat{\kappa}^{-1} \left. \frac{\partial \mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n)}{\partial \theta_k} \right|_{\theta_k = \hat{\theta_k}} = 0.$$

Assuming $l < k$ (Hessian is symmetric and diagonal has different form)

$$\left. \frac{\partial \mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n)}{\partial \theta_k \partial \theta_l} \right|_{\theta_l = \hat{\theta}_l, \theta_k = \hat{\theta_k}} = \cot \hat{\theta}_l \left. \frac{\partial \mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n)}{\partial \theta_k} \right|_{\theta_k = \hat{\theta_k}} = 0.$$

Thus, $\hat{\boldsymbol{J}}$ is a diagonal matrix, with diagonal $\boldsymbol{j} = (\hat{J}_{11}, ..., \hat{J}_{dd})^\top$:

$$\hat{\boldsymbol{J}} = \text{diag}(\boldsymbol{j}),$$
$$\hat{\boldsymbol{J}}^{-1} = \text{diag}\left((\hat{J}_{11}^{-1}, ..., \hat{J}_{dd}^{-1})^\top \right),$$
$$\text{tr}(\boldsymbol{\mathcal{I}} \hat{\boldsymbol{J}}^{-1}) = \sum_{i=1}^d \hat{J}_{ii}^{-1} \hat{\mathcal{I}}_{ii} = \hat{J}_{11}^{-1} \left(\frac{\partial}{\partial \kappa} \mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n) \right)^2 + \sum_{i=2}^d \hat{J}_{ii}^{-1} \left(\frac{\partial}{\partial \theta_{i-1}} \mathcal{L}(\boldsymbol{\theta}, \kappa|\{\boldsymbol{\phi}_i\}_{i=1}^n) \right)^2.$$

This quantity only requires $\mathcal{O}(nd)$ operations to compute and thus does not increase the asymptotic complexity of the algorithm.

The closed form of the similarity measure for two sentences $\mathcal{D}_1, \mathcal{D}_2$ of length m and l respectively under this model is then

$$\begin{aligned}\mathbf{sim}(\mathcal{D}_1, \mathcal{D}_2) = 2\Big(&(m+l)\hat{\kappa}_{1,2}\bar{R}_{1,2} - m\hat{\kappa}_1\bar{R}_1 - l\hat{\kappa}_2\bar{R}_2 \\ &- (m+l)\log Z(\hat{\kappa}_{1,2}) + m\log Z(\hat{\kappa}_1) + l\log Z(\hat{\kappa}_2) \\ &- \mathrm{tr}(\boldsymbol{\mathcal{I}}_{1,2}\hat{\boldsymbol{J}_{1,2}}^{-1}) - \mathrm{tr}(\boldsymbol{\mathcal{I}}_1\hat{\boldsymbol{J}_1}^{-1}) - \mathrm{tr}(\boldsymbol{\mathcal{I}}_2\hat{\boldsymbol{J}_2}^{-1})\Big),\end{aligned}$$

where the Jacobian terms (from the reparametrisation) cancel out. The subscripts indicate the sentence, with $1, 2$ meaning the concatenation of the two sentences.

3 EXPERIMENTS

We assess our methods performance on the Semantic Text Similarity (STS) datasets (Agirre et al., 2012; 2013; 2014; 2015; 2016). The objective of these tasks is to predict the similarity between two given sentences, validated against human scores. In our experiments, we utilise the pre-trained, commonly used GloVe word embeddings (Pennington et al., 2014). We remove stop words and punctuation using NTLK (Bird et al., 2009) from the dataset. This is done for two reasons - firstly, these bring little semantic meaning to a sentence. Secondly, the vMF distribution is unimodal and thus highly susceptible to stop words that form their own mode. For our method, we centre the distribution of word embeddings and normalize them to be of length 1, so that they can be modelled by the vMF distribution. Additionally, after the stop word removal, some of the sentences are left with a single word, making the MLE of the κ parameter ill-defined. We overcome this issue by padding each sentence with an arbitrary embedding of a stop word or punctuation symbol (i.e. '.' or '?').

We compare our method against mean word vector (MWV), word mover distance (WMD) (Kusner et al., 2015) [1], smooth inverse frequency (SIF), and SIF with principal component removal as defined in Arora et al. (2016) [2]. We re-ran these models under our experimental setup, to ensure a fair comparison. The reported results in this paper are slightly higher than the ones observed in the original works, due to stop word removal. The metric used is the average Spearman correlation score over each dataset, weighted by the number of sentences. The choice of Spearman is given by its non-parametric nature (assumes no distribution over the scores), as well as measuring any monotonic relationship between the two compared quantities.

3.1 ONLINE EXPERIMENT

	Our method	WMD	MWV	SIF
STS-12	0.5944	0.5516	**0.5990**	0.5980
STS-13 (-SMT)	**0.6676**	0.5007	0.6371	0.6623
STS-14	**0.6614**	0.5811	0.6459	0.6551
STS-15	**0.7244**	0.6704	0.7008	0.7042
STS-16	**0.6877**	0.6246	0.6663	0.6612

Table 1: Spearman correlations on the STS datasets using GloVe vectors, when pre-processing of the datasets are disallowed.

In this experiment, we do not allow pre-processing of the sentence-level dataset to simulate the conditions seen during an online information-retrieval scenario. The results are presented in Table

[1] https://github.com/mkusner/wmd

[2] https://github.com/PrincetonML/SIF

3.1. We are able to out-perform the standard weighting induced by mean word vector, as well as the smooth inverse frequency weighting technique.

3.2 OFFLINE EXPERIMENT

	Our method	SIF + PCA
STS-12	**0.5944**	0.5871
STS-13 (-SMT)	0.6676	**0.7029**
STS-14	0.6614	**0.6846**
STS-15	0.7244	**0.7288**
STS-16	0.6877	0.6875

Table 2: Spearman correlations on the STS datasets using GloVe vectors with pre-processing allowed.

For this experiment, we compare against the SIF weightings, augmented with the additional pre-processing technique seen in Arora et al. (2016). The results are shown in Table 3.2. We remain competitive with their results on the last two datasets, acheive better performance on STS-12, but have an overall lower average.

4 DISCUSSION

We've presented a new approach to similarity measurement, that achieves competitive performance to standard methods. Our method requires a set of clear choices - model, likelihood and information criterion. From that, a comparison framework is naturally derived, which supplies us with a similarity measure.

In this study, we found that the von Mises-Fisher distribution naturally lends itself to representing word embeddings, if their magnitude is disregarded and a unimodal distribution over individual sentences is assumed. Both of these assumptions can be relaxed, but further work is needed to derive the resulting distribution's properties. Adapting the Takeuchi Information Criterion to reduce the resulting model-comparison bias, we perform a statistically justified model-comparison induced similarity score. This framework is suitable for a variety of modelling scenarios, due to the modelling freedom in specifying the generative process. The graphical model we employ is adaptable to encode structural dependencies beyond the i.i.d. data-generating process we have assumed throughout this study.

ACKNOWLEDGMENTS

REFERENCES

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 385–393. Association for Computational Linguistics, 2012.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pp. 32–43, 2013.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 81–91, 2014.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. Semeval-2015 task 2:

Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 252–263, 2015.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 497–511, 2016.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Hirotugu Akaike et al. Likelihood of a model and information criteria. *Journal of econometrics*, 16 (1):3–14, 1981.

Amjad Almahairi, Kyle Kastner, Kyunghyun Cho, and Aaron Courville. Learning distributed representations from reviews for collaborative filtering. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 147–154. ACM, 2015.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations, 2017*, 2016.

Ben Athiwaratkun and Andrew Gordon Wilson. Multimodal word distributions. *arXiv preprint arXiv:1704.08424*, 2017.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep): 1345–1382, 2005.

James O. Berger. Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association*, 95(452):1269–1276, 2000. ISSN 01621459. URL `http://www.jstor.org/stable/2669768`.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

William Blacoe and Mirella Lapata. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 546–556. Association for Computational Linguistics, 2012.

John J Dziak, Donna L Coffman, Stephanie T Lanza, and Runze Li. Sensitivity and specificity of information criteria. *The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University*, 16(30):140, 2012.

Nicholas I Fisher, Toby Lewis, and Brian JJ Embleton. *Statistical analysis of spherical data*. Cambridge university press, 1993.

Zoubin Ghahramani and Katherine A Heller. Bayesian sets. In *Advances in neural information processing systems*, pp. 435–442, 2006.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*, 2016.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015. URL `http://arxiv.org/abs/1506.06726`.

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008a.

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008b.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/kusnerb15.html.

Maurice S M. S. Bartlett. A comment on d. v. lindley's statistical paradox. *Biometrika*, 44(3/4): 533–534, 1957.

KV Mabdia. Distribution theory for the von mises-fisher distribution and ite application. *Statistical Distributions for Scientific Work*, 1:113–30, 1975.

Phil Marshall, Nutan Rajguru, and Anže Slosar. Bayesian evidence as a tool for comparing datasets. *Physical Review D*, 73(6):067302, 2006.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pp. 236–244, 2008.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Kei Takeuchi. The distribution of information statistics and the criterion of goodness of fit of models. *Mathematical Science*, 153:12–18, 1976.

John Wieting and Kevin Gimpel. Revisiting recurrent networks for paraphrastic sentence embeddings. *arXiv preprint arXiv:1705.00364*, 2017.

Guoqing Zheng and Jamie Callan. Learning to reweight terms with distributed representations. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pp. 575–584, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3621-5. doi: 10.1145/2766462.2767700. URL http://doi.acm.org/10.1145/2766462.2767700.

A DERIVATION OF TIC FOR $\mathcal{M}_2$

The MLE for $p(\boldsymbol{\phi}_{1,2}|\boldsymbol{\theta}, \mathcal{M}_2)$ can be derived by simply estimating the separate MLE solutions for $p(\boldsymbol{\phi}_1|\boldsymbol{\theta}_1, \mathcal{M}_2)$ and $p(\boldsymbol{\phi}_2|\boldsymbol{\theta}_2, \mathcal{M}_2)$. What is not as obvious is that the penalty term follows the estimation pattern.

Gradient vectors for $\mathcal{M}_2$ are given by (where $\oplus$ is concatenation)

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\phi}_{1,2}) = \nabla_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1|\boldsymbol{\phi}_1) \oplus \nabla_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_2|\boldsymbol{\phi}_2),$$

and Hessian results in a block diagonal matrix

$$\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\phi}_{1,2}) = \begin{bmatrix} \nabla^2_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1|\boldsymbol{\phi}_1) & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_2|\boldsymbol{\phi}_2) \end{bmatrix},$$

with inverse:

$$\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\phi}_{1,2})^{-1} = \begin{bmatrix} \nabla^2_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1|\boldsymbol{\phi}_1)^{-1} & \mathbf{0} \\ \mathbf{0} & \nabla^2_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_2|\boldsymbol{\phi}_2)^{-1} \end{bmatrix}.$$

Computing $\mathrm{tr}(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}^{-1}})$ then yields

$$\begin{aligned}
\mathrm{tr}(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}^{-1}}) &= \mathrm{tr}\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\phi}_{1,2})\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\phi}_{1,2})^\top\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{\phi}_{1,2})^{-1}\right) \\
&= \mathrm{tr}\left(\begin{bmatrix} \hat{\boldsymbol{\mathcal{I}}}_{11}\nabla^2_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1|\boldsymbol{\phi}_1)^{-1} & \hat{\boldsymbol{\mathcal{I}}}_{12}\nabla^2_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_2|\boldsymbol{\phi}_2)^{-1} \\ \hat{\boldsymbol{\mathcal{I}}}_{12}\nabla^2_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1|\boldsymbol{\phi}_1)^{-1} & \hat{\boldsymbol{\mathcal{I}}}_{22}\nabla^2_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_2|\boldsymbol{\phi}_2)^{-1} \end{bmatrix}\right) \\
&= \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_{11}\nabla^2_{\boldsymbol{\theta}_1}\mathcal{L}(\boldsymbol{\theta}_1|\boldsymbol{\phi}_1)^{-1}\right) + \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_{22}\nabla^2_{\boldsymbol{\theta}_2}\mathcal{L}(\boldsymbol{\theta}_2|\boldsymbol{\phi}_2)^{-1}\right) \\
&= \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_1\hat{\boldsymbol{J}_1}^{-1}\right) + \mathrm{tr}\left(\hat{\boldsymbol{\mathcal{I}}}_2\hat{\boldsymbol{J}_2}^{-1}\right).
\end{aligned}$$

B PARTIAL DERIVATIVE CALULATIONS

We first show the following result, which is useful for the full derivation. For $k \leq j$

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\mu_j(\boldsymbol{\theta}) &= \frac{\partial}{\partial\theta_k}\left((1-\delta_{kd})\cos\theta_k + \delta_{kd}\right)\prod_{i=1}^{j-1}\sin\theta_i \\
&= \left((1-\delta_{kj})\frac{\cos\theta_k}{\sin\theta_k} - \delta_{kj}\frac{\sin\theta_k}{\cos\theta_k}\right)\left((1-\delta_{kd})\cos\theta_k + \delta_{kd}\right)\prod_{i=1}^{i-1}\sin\theta_i \\
&= \left((1-\delta_{kj})\cot\theta_k - \delta_{kj}\tan\theta_k\right)\mu_j(\boldsymbol{\theta}),
\end{aligned}$$

where the second line comes from the fact that $\sin\theta_k$ (or $\cos\theta_k$) gets transformed into a $\cos\theta_k$ (or $-\sin\theta_k$), and thus we can revert to the original definition of μ_j by multiplying with a $\cot\theta_k$ (or $-\tan\theta_k$). If $k > j$, this derivative is 0. Thus, for a single data point $\boldsymbol{w}(\boldsymbol{\phi})$

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\mathcal{L}(\boldsymbol{\theta}, \kappa|\boldsymbol{\phi}) &= \frac{\partial}{\partial\theta_k}\kappa\boldsymbol{w}(\boldsymbol{\phi})^\top\boldsymbol{\mu}(\boldsymbol{\theta}) - \frac{\partial}{\partial\theta_k}\log Z(\kappa) \\
&= \kappa\boldsymbol{w}(\boldsymbol{\phi})^\top\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_k} \\
&= \kappa\sum_{j=k}^{d} w_j(\boldsymbol{\phi})\mu_j(\boldsymbol{\theta})\left((1-\delta_{kj})\cot\theta_k - \delta_{kj}\tan\theta_k\right),
\end{aligned}$$

where the sum starts from k, as for $j < k$, the derivative is zero.

The derivative with respect to κ is derived as follows:

$$\begin{aligned}\frac{\partial}{\partial\kappa}\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi}) &= \frac{\partial}{\partial\kappa}\kappa\boldsymbol{w}(\boldsymbol{\phi})^\top\boldsymbol{\mu}(\boldsymbol{\theta}) - \frac{\partial}{\partial\kappa}\log Z(\kappa)\\ &= \boldsymbol{w}(\boldsymbol{\phi})^\top\boldsymbol{\mu}(\boldsymbol{\theta}) - \frac{I_{\frac{d}{2}}(\kappa)}{I_{\frac{d}{2}-1}(\kappa)},\end{aligned}$$

where the derivative of the second term is a known result.

We next focus on second order derivatives:

$$\frac{\partial^2}{\partial\theta_k^2}\mu_j(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_k^2}\left((1-\delta_{id})\cos\theta_i + \delta_{id}\right)\prod_{i=1}^{j-1}\sin\theta_i.$$

Unless this derivative is zero, we notice that we take the derivative $\frac{\partial^2\cos\phi_k}{\partial\phi_k^2}$ or $\frac{\partial^2\sin\phi_k}{\partial\phi_k^2}$, both of which result in the negative of the original function. Thus

$$\frac{\partial^2}{\partial\theta_k^2}\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi}) = -\kappa\sum_{j=k}^{d} w_j(\boldsymbol{\phi})\mu_j(\boldsymbol{\theta}).$$

The below result is given

$$\begin{aligned}\frac{\partial^2}{\partial\kappa^2}\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi}) &= \frac{\partial}{\partial\kappa}\left(-\frac{I_{\frac{d}{2}}(\kappa)}{I_{\frac{d}{2}-1}(\kappa)}\right)\\ &= \frac{I_{\frac{d}{2}}(\kappa)(I_{\frac{d}{2}-2}(\kappa)+I_{\frac{d}{2}}(\kappa)) - I_{\frac{d}{2}-1}(\kappa)(I_{\frac{d}{2}-1}(\kappa)+I_{\frac{d}{2}+1}(\kappa))}{2I_{\frac{d}{2}-1}(\kappa)^2}.\end{aligned}$$

Next, we show that the second order mixed derivatives are a constant (with respect to $\boldsymbol{\phi}$) times $\partial\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi})/\partial\theta_k$, i.e.

$$\begin{aligned}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi})}{\partial\kappa\partial\theta_k} &= \frac{\partial^2}{\partial\kappa\partial\theta_k}\kappa\boldsymbol{w}(\boldsymbol{\phi})^\top\boldsymbol{\mu}(\boldsymbol{\theta}) - \frac{\partial}{\partial\kappa\partial\theta_k}\log Z(\kappa)\\ &= \boldsymbol{w}(\boldsymbol{\phi})^\top\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_k}\\ &= \kappa^{-1}\frac{\partial\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi})}{\partial\theta_k},\end{aligned}$$

and assuming $l < k$

$$\begin{aligned}\frac{\partial\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi})}{\partial\theta_k\partial\theta_l} &= \kappa\boldsymbol{w}(\boldsymbol{\phi})^\top\frac{\partial\boldsymbol{\mu}(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_l}\\ &= \kappa\sum_{j=\max(k,l)}^{d} w_j(\boldsymbol{\phi})\mu_j(\boldsymbol{\theta})((1-\delta_{kj})\cot\theta_k - \delta_{kj}\tan\theta_k)((1-\delta_{lj})\cot\theta_l - \delta_{lj}\tan\theta_l)\\ &= \kappa\sum_{j=k}^{d} w_j(\boldsymbol{\phi})\mu_j(\boldsymbol{\theta})\cot\theta_l((1-\delta_{kj})\cot\theta_k - \delta_{kj}\tan\theta_k)\\ &= \cot\theta_l\kappa\sum_{j=l}^{d} w_j(\boldsymbol{\phi})\mu_j(\boldsymbol{\theta})((1-\delta_{kj})\cot\theta_k - \delta_{kj}\tan\theta_k)\\ &= \cot\theta_l\frac{\partial\mathcal{L}(\boldsymbol{\theta},\kappa|\boldsymbol{\phi})}{\partial\theta_k}.\end{aligned}$$