Model Comparison For Semantic Grouping Supplementary Material

Francisco Vargas [1] **Kamen Brestnichki** [1] **Nils Hammerla** [1]

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. Reparametrisation of the vMF Distribution

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

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

This reparametrisation simplifies the calculation of partial derivatives. 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 given by applying the map from the cartesian MLE to the polars.

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

where both the derivation and approximation for the MLE estimates are derived in (Banerjee et al., 2005). Let $\mathcal{D} = \{\boldsymbol{\phi}_i\}_{i=1}^n$ be the dataset. The log likelihood is then

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

[1]Babylon Health. Correspondence to: Francisco Vargas <vargfran@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

C. Partial Derivative Calulations (vMF Likelihood)

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

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

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 $\boldsymbol{w}(\boldsymbol{\phi})$

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

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 $\partial^2\cos\theta_k/\partial\theta_k^2$ or $\partial^2\sin\theta_k/\partial\theta_k^2$, 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 (where $v = \frac{d}{2} - 1$)

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

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.

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

Evaluated at the MLE by definition $\kappa^{-1}\frac{\partial\mathcal{L}(\boldsymbol{\theta},\kappa|\mathcal{D})}{\partial\theta_k}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\kappa=\hat{\kappa}} = 0$

Assuming $l < k$ (Hessian is symmetric)

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

where the sum starts from $max(k,l) = k$ because all terms below that are zero. Then at the MLE $\cot\theta_l\frac{\partial\mathcal{L}(\boldsymbol{\theta},\kappa|\mathcal{D})}{\partial\theta_k}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\kappa=\hat{\kappa}} = 0.$

D. Partial Derivatives Calculation (Gaussian Likelihood)

The partial derivatives for the diagonal Gaussian likelihood are (we take derivatives with respect to precision $\lambda_k^2 = 1/\sigma_k^2$)

$$\begin{aligned}\frac{\partial}{\partial\mu_k}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\boldsymbol{w}) &= \sum_{i=1}^{n}\lambda_k^2\left(x_k^{(i)} - \mu_k\right), \\ \frac{\partial^2}{\partial\mu_k^2}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\boldsymbol{w}) &= -n\lambda_k^2, \\ \frac{\partial}{\partial\lambda_k^2}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\boldsymbol{w}) &= \frac{n}{2\lambda_k^2} - \frac{1}{2}\sum_{i=1}^{n}\left(x_k^{(i)} - \mu_k\right)^2, \\ \frac{\partial^2}{\partial(\lambda_k^2)^2}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\boldsymbol{w}) &= -\frac{n}{2\lambda_k^4}, \\ \frac{\partial^2}{\partial\lambda_k^2\partial\mu_k}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\boldsymbol{w}) &= \sum_{i=1}^{n}\left(x_k^{(i)} - \mu_k\right).\end{aligned}$$

Evaluating at the MLE we get

$$\frac{\partial^2}{\partial \mu_k^2}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\mathcal{D})\Big|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}},\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = -n\hat{\lambda}_k^2 = -\frac{n}{\hat{\sigma}_k^2},$$
$$\frac{\partial^2}{\partial (\lambda_k^2)^2}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\mathcal{D})\Big|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}},\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = -\frac{n}{2\hat{\lambda}_k^4} = -\frac{n\hat{\sigma}_k^4}{2},$$
$$\frac{\partial^2}{\partial \lambda_k^2 \partial \mu_k}\mathcal{L}(\boldsymbol{\mu},\boldsymbol{\lambda}|\mathcal{D})\Big|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}},\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} = \sum_{i=1}^{n}\left(x_k^{(i)} - \hat{\mu}_k\right) = 0.$$

Substituting these derivatives into the definition of $\hat{\boldsymbol{\mathcal{I}}}$ and $\hat{\boldsymbol{J}}$ we get

$$\hat{\boldsymbol{\mathcal{I}}}_{\mu_k,\mu_k} = \hat{\lambda}_k^2, \quad \hat{\boldsymbol{\mathcal{I}}}_{\lambda_k^2,\lambda_k^2} = -4\hat{\lambda_k}^{-4} + \frac{(\hat{\boldsymbol{\mu}}_4)_k}{4}$$
$$\hat{\boldsymbol{J}}_{\mu_k,\mu_k} = -\hat{\lambda}_k^2, \quad \hat{\boldsymbol{J}}_{\lambda_k^2,\lambda_k^2} = -\frac{1}{2}\hat{\lambda}_k^{-4}$$

Finally, computing the model complexity penalty we get

$$\operatorname{tr}(\hat{\boldsymbol{\mathcal{I}}}\hat{\boldsymbol{J}}^{-1}) = \frac{1}{2}\left(d + \sum_{i=1}^{d}\frac{(\hat{\boldsymbol{\mu}}_4)_i}{\hat{\sigma}_i^4}\right) = \frac{d}{2} + \sum_{i=1}^{d}\frac{\hat{\kappa}_i}{2}.$$

E. Bayes Factor and Bayesian Information Criterion

We first define the Bayes Factor for the Gaussian likelihood with parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1})$. We assume a Wishart prior

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\kappa_0\boldsymbol{\Lambda})^{-1}\right)\mathcal{W}_i(\boldsymbol{\Lambda}|\nu_0, \boldsymbol{T}_0^{-1}),$$

which yields the following Normal-Wishart posterior (Murphy, 2007)

$$p(\boldsymbol{\mu},\boldsymbol{\Lambda}|\mathcal{D}) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_n, (\kappa_n\boldsymbol{\Lambda})^{-1}\right)\mathcal{W}_i(\boldsymbol{\Lambda}|\nu_n, \boldsymbol{T}_n^{-1}).$$

The posterior parameters are

$$\begin{aligned}
\nu_n &= \nu_0 + n,\\
\kappa_n &= \kappa_0 + n,\\
\boldsymbol{\mu}_n &= \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\mathbf{w}}}{\kappa_n},\\
\boldsymbol{S} &= \sum_{k=1}^{n}(\boldsymbol{w}_k - \bar{\boldsymbol{w}})(\boldsymbol{w}_k - \bar{\boldsymbol{w}})^\top,\\
\boldsymbol{T}_n &= \boldsymbol{S} + \boldsymbol{T}_0 + \frac{n\kappa_0}{2\kappa_n}(\bar{\mathbf{w}} - \boldsymbol{\mu}_0)(\bar{\mathbf{w}} - \boldsymbol{\mu}_0)^\top.
\end{aligned}$$

The evidence (Murphy, 2007) is

$$p(\mathcal{D}) = \frac{1}{\pi^{nd/2}}\left(\frac{\kappa_0}{\kappa_n}\right)^{\frac{d}{2}}\frac{|\boldsymbol{T}_0|^{\frac{\nu_0}{2}}}{|\boldsymbol{T}_n|^{\frac{\nu_n}{2}}}\frac{\Gamma_d(\nu_n/2)}{\Gamma_d(\nu_0/2)}.$$

Table 1. Spearman correlations using GloVe embeddings and Gaussian likelihood. The AIC and BIC use a diagonal covariance matrix, while the Bayes Factor uses a full covariance matrix.

	AIC	Bayes Factor	BIC
STS-12	**0.6031**	0.4592	0.5009
STS-13 (-SMT)	**0.6132**	0.5687	0.6011
STS-14	**0.6445**	0.5829	0.5926
STS-15	**0.7171**	0.6627	0.6625
STS-16	**0.7346**	0.5389	0.5826

Using $p(\mathcal{D}_1,\mathcal{D}_2|\mathcal{M}_1) = p(\mathcal{D}_1 \oplus \mathcal{D}_2|\mathcal{M}_1)$ we compute the Bayes factor for $\mathcal{M}_1, \mathcal{M}_2$ in closed form

$$\mathbf{sim}(\mathcal{D}_1,\mathcal{D}_2) = \left(\frac{\kappa_n\kappa_m}{\kappa_0\kappa_l}\right)^{\frac{d}{2}}\frac{|\boldsymbol{T}_n|^{\frac{\nu_n}{2}}|\boldsymbol{T}_m|^{\frac{\nu_m}{2}}}{|\boldsymbol{T}_l|^{\frac{\nu_l}{2}}|\boldsymbol{T}_0|^{\frac{\nu_0}{2}}}\frac{\Gamma_d\left(\frac{\nu_l}{2}\right)\Gamma_d\left(\frac{\nu_l}{2}\right)}{\Gamma_d\left(\frac{\nu_n}{2}\right)\Gamma_d\left(\frac{\nu_m}{2}\right)}. \tag{1}$$

where $|\mathcal{D}_1| = n,\ |\mathcal{S}_2| = m$ and $|\mathcal{D}_1 \oplus \mathcal{D}_2| = m + n = l$.

The BIC is defined as

$$\mathrm{BIC}(\mathcal{D},\mathcal{M}) = -2\mathcal{L}(\hat{\boldsymbol{\theta}}|\mathcal{D},\mathcal{M}) + k\log n \approx -p(\mathcal{D}|\mathcal{M}),$$

and acts as a direct approximation to the model evidence (Schwarz et al., 1978). Thus, the similarity under the BIC is

$$\begin{aligned}\mathbf{sim}(\mathcal{D}_1,\mathcal{D}_2) = 2\big(&\mathcal{L}(\hat{\boldsymbol{\theta}}_{1,2}|\mathcal{M}_1) - \mathcal{L}(\hat{\boldsymbol{\theta}}_1|\mathcal{M}_2) - \mathcal{L}(\hat{\boldsymbol{\theta}}_2|\mathcal{M}_2)\big)\\ &- k\log\frac{n+m}{nm},\end{aligned} \tag{2}$$

where n, m are defined as above.

Equations 1 and 2 represents our similarity score under a Gaussian likelihood, for the Bayes Factor and BIC respectively.

Table E compares BIC and the Bayes Factor using a Gaussian likelihood to the approach presented in the paper. We see that while these approaches are competitive on STS-13, STS-14 and STS-15, they both give severely worse results on STS-12 and STS-16, with more than 0.08 difference. This motivates our choice to do a penalised likelihood ratio test instead of doing full Bayesian inference of the evidences.

F. TIC Robustness

In this section, we compare the TIC and AIC on the two likelihoods described in the main text. Table F presents the results of that comparison. As we can see, with each word embedding the Gaussian AIC correction outperforms the TIC correction on average. Looking at Figure 1, it becomes apparent why — the more parameters a Gaussian has, the more dependent its correction is on the number

Table 2. Comparison of Spearman correlations on the STS datasets between the TIC and AIC corrections for the diagonal covariance Gaussian and vMF likelihood functions.

Embedding	Method	STS12	STS13	STS14	STS15	STS16	Average
FastText	vMF+TIC	**0.5219**	**0.5147**	**0.5719**	**0.6456**	**0.6347**	**0.5762**
	vMF+AIC	0.5154	**0.5107**	**0.5697**	**0.6425**	0.**6330**	**0.5726**
	Diag+TIC	0.5882	**0.6585**	0.6678	0.7205	0.7060	0.6632
	Diag+AIC	**0.6193**	0.6335	**0.6721**	**0.7328**	**0.7518**	**0.6764**
GloVe	vMF+TIC	**0.5421**	**0.5598**	**0.5736**	**0.6474**	0.6168	**0.5859**
	vMF+AIC	0.5331	0.5465	0.5653	**0.6434**	**0.6331**	0.5802
	Diag+TIC	0.5773	**0.6467**	**0.6559**	**0.7141**	0.7019	**0.6536**
	Diag+AIC	**0.6031**	0.6132	0.6445	**0.7171**	**0.7346**	**0.6564**
Word2Vec GN	vMF+TIC	**0.5665**	**0.5735**	**0.6062**	**0.6681**	**0.6510**	**0.6115**
	vMF+AIC	0.5519	**0.5770**	**0.6055**	**0.6715**	**0.6560**	**0.6094**
	Diag+TIC	0.5673	0.6234	0.6460	0.6942	0.6559	0.6363
	Diag+AIC	**0.5957**	**0.6358**	**0.6614**	**0.7213**	**0.7187**	**0.6618**

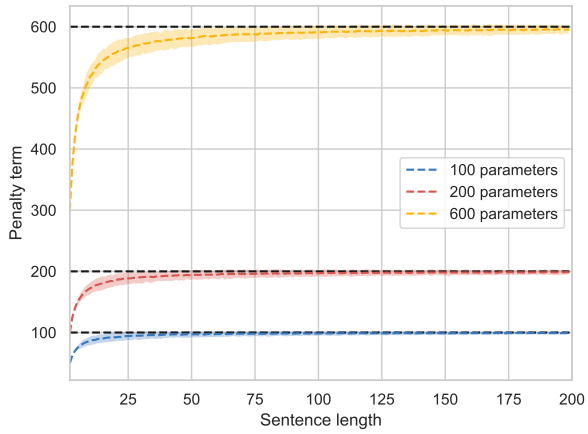

Figure 1. TIC Correction penalty for varying sample sizes, samples generated from standardised normal $\mathcal{N}(\mathbf{0}, \mathbb{I}_d)$.

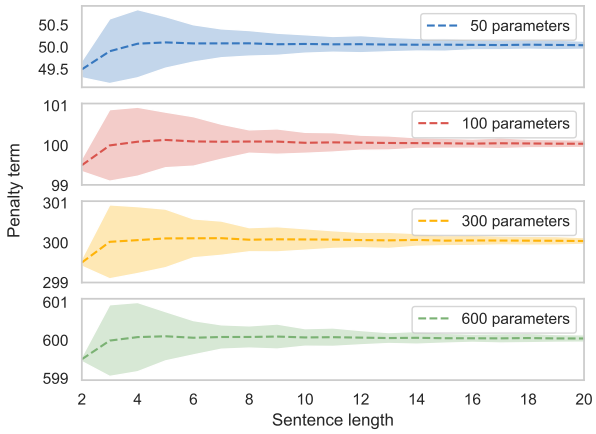

Figure 2. TIC Correction penalty for varying sample sizes, samples generated from Uniform distribution on the unit hypersphere $U(\mathbb{S}_{d-1})$.

of words in the sentence. This is reminiscent of the linear scaling with number of words in the BIC penalty discussed in Appendix E, which was shown to perform badly. On the other hand, looking at Figure 2, we see that the TIC for the vMF distribution has very low variance, and is generally not dependent on the number of word embeddings in the word group. This gives intuition why the AIC and TIC for the vMF give very similar results.

G. Running Times

In our case $d = 300$, and n ranges from 2 to 30. If we ignore word embedding loading times, on the entire STS dataset (STS12 to STS16), the SIF+PCA algorithm takes 1.02 seconds for the calculation of the PC components and 1.92 seconds for the cosine-based similarity. Our entire diagonal Gaussian AIC method takes 1.97 seconds, which is about 33 percent faster.

H. Spherical Gaussian Results

We compute the average AIC of the two Gaussians using FastText embeddings. The diagonal covariance Gaussian has an average AIC of -4738, while the spherical has an average AIC of -3945. The results of each similarity measure induced by the two likelihoods are shown in Table H. As we can see, the diagonal Gaussian covariance likelihood

Table 3. Comparison of Spearman correlations on the STS datasets between the AIC corrections for the diagonal covariance Gaussian and spherical covariance Gaussian likelihood functions.

Embedding	Method	STS12	STS13	STS14	STS15	STS16	Average
FastText	Spherical+AIC	0.5817	0.5912	0.6199	0.6866	0.7076	0.6312
	Diag+AIC	**0.6193**	**0.6335**	**0.6721**	**0.7328**	**0.7518**	**0.6764**
GloVe	Spherical+AIC	0.5639	0.5718	0.5890	0.6667	0.6732	0.6073
	Diag+AIC	**0.6031**	**0.6132**	**0.6445**	**0.7171**	**0.7346**	**0.6564**
Word2Vec GN	Spherical+AIC	0.5844	0.6131	0.6308	0.6864	0.6975	0.6368
	Diag+AIC	**0.5957**	**0.6358**	**0.6614**	**0.7213**	**0.7187**	**0.6618**

Table 4. Pairs for which the non-parametric test rejected the null hypothesis (i.e. the results were statistically significant).

Embedding	Method 1	Method 2
FastText	SIF	WMD
FastText	SIF+PCA	WMD
GloVe	Diag+AIC	MWV
GloVe	SIF+PCA	MWV

produces a better similarity measure.

I. Significance Analysis

We perform the same triplet sampling procedure as in (Zhelezniak et al., 2019) to generate estimates of Spearman correlations. Using the bootstrap estimated correlation coefficients we then carry out a non parametric two sample test (Gretton et al., 2012) in order to determine if the samples are drawn from different distributions.

Out of all possible pairs in [Diag+AIC, SIF, SIF+PCA, MWV, WMD] we list the pairs for which the significance analysis rejected the null hypothesis in Table H. We can see that the only significant differences are between MWV/WMD and the rest of the methods, thus showing we are competitive to the methods from (Arora et al., 2016). The code to reproduce these results can be found at https://github.com/Babylonpartners/MCSG.

References

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

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

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

Murphy, K. P. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007.

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

Zhelezniak, V., Savkov, A., Shen, A., Moramarco, F., Flann, J., and Hammerla, N. Y. Don't settle for average, go for the max: Fuzzy sets and max-pooled word vectors. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkxXg2C5FX.