



Posterior sampling of word embeddings

Downloaded from: <https://research.chalmers.se>, 2026-10-21 00:32 UTC

Citation for the original published paper (version of record):

Yrjänäinen, V., Boström, I., Jonasson, J. et al (2026). Posterior sampling of word embeddings. *Statistics and Computing*, 36(6). <http://dx.doi.org/10.1007/s11222-026-10996-5>

N.B. When citing this work, cite the original published paper.



Posterior sampling of word embeddings

Väinö Yrjänäinen¹ · Isac Boström² · Johan Jonasson² · Måns Magnusson¹

Received: 22 December 2025 / Accepted: 29 September 2026
© The Author(s) 2026

Abstract

We study Bayesian inference for probabilistic word embedding models, including skip-gram with negative sampling (SGNS) and continuous bag-of-words (CBoW). These models are widely used for inference, but their posteriors are non-identifiable under general linear transformations, which prevents meaningful Bayesian uncertainty quantification. We provide a simple and principled identifiability constraint that yields a well-defined posterior without altering the data model. Building on this, we develop scalable, fully Bayesian posterior sampling algorithms for SGNS and CBoW using Pólya–Gamma augmentation. Our block Gibbs samplers yield exact samples from the posterior and scale well to large corpora due to their straightforward parallelisation. We discuss the efficient implementation of the Gibbs sampler, and provide a fast TensorFlow reference implementation. In a combination of simulations and empirical experiments, we demonstrate that the sampler is both accurate and computationally efficient. It matches the accuracy of NUTS-HMC at a fraction of the computational cost and avoids the variance underestimation of mean-field variational inference. On large real datasets, we show that posterior means yield improved predictive performance, while posterior credible intervals provide meaningful uncertainty quantification for semantic similarity.

Keywords Probabilistic modelling · Markov chain Monte Carlo · Bayesian statistics · Natural language processing · Gibbs sampling

1 Introduction

Word embeddings are vector representations of words that capture statistical relationships within textual data. Traditional word embedding methods, such as word2vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014), provide point estimates for each word, representing them as fixed vectors

in a high-dimensional space. In recent years, word embeddings have been increasingly applied to inference tasks in the social sciences to investigate a wide range of linguistic and cultural aspects (Garg et al. 2018; Kozłowski et al. 2019; Rodman 2020; Durrheim et al. 2023; Abramson et al. 2024; Aceves and Evans 2024; Ma et al. 2025). However, standard word embedding methods (e.g., word2vec and GloVe) do not quantify the uncertainty associated with the embeddings, which is essential for reliable inference and interpretation in scientific settings.

Probabilistic word embeddings (Rudolph et al. 2016; Bamler and Mandt 2017) extend traditional word embeddings and formulate word embeddings as a full probabilistic model. Nevertheless, quantifying uncertainty in these models poses significant challenges due to large datasets, non-identifiability, and large parameter spaces commonly involved in practical applications. Proposed by Bamler and Mandt (2017), mean-field variational inference (MFVI) is today the only Bayesian posterior estimation method of word embeddings. Although computationally efficient, it is well known that MFVI underestimates posterior variance (Wang and Blei 2018), limiting its use in scientific settings. Fur-

Väinö Yrjänäinen and Isac Boström contributed equally to this work.

✉ Väinö Yrjänäinen
vaino.yrjanainen@statistik.uu.se

Isac Boström
isac.bostrom@chalmers.se

Johan Jonasson
jonasson@chalmers.se

Måns Magnusson
mans.magnusson@statistik.uu.se

¹ Department of Statistics, Uppsala University, Kyrkogårdsgatan 10, 75120 Uppsala, Sweden

² Department of Mathematical Sciences, Chalmers University of Technology, Chalmers Tvärgata 3, 412 58 Göteborg, Sweden

thermore, No U-Turn Sampler Hamiltonian Monte Carlo (NUTS-HMC) can be used for embeddings but is usually infeasible due to the high computational cost. Thus, while there has been work on word embedding uncertainty, accurate Bayesian uncertainty quantification for embeddings remains elusive.

In this paper, we present the following contributions:

1. We study the identifiability of word embeddings in the context of large-scale Bayesian inference and prove that current embeddings are not identified. We present a conceptually easy and computationally feasible way to identify the skip-gram with negative samples (SGNS) and the Continuous Bag-of-Words (CBoW) models, and mathematically prove that this approach identifies the models.
2. We propose a computationally efficient and scalable Gibbs sampler for both the SGNS and CBoW models. Further, we show that the Gibbs sampler matches the gold-standard NUTS-HMC in accuracy while being faster, and is more accurate than the established word embedding estimation approach, mean-field variational inference.
3. We show that our proposed Gibbs sampler is more accurate than MFVI and as accurate as HMC, but more computationally efficient than HMC for large models and data, a common setting in real-world applications.

2 Probabilistic word embeddings

Probabilistic word embeddings (Rudolph et al. 2016; Bamler and Mandt 2017; Rudolph and Blei 2017) extend the standard word embedding models of (Mikolov et al. 2013) to a probabilistic model.

Let $(w_i)_{i=1}^{N_w}$ be a dataset, a sequence of words, where each word w_i belongs to the *vocabulary* $W = \{1, \dots, V\}$, $V \in \mathbb{N}$, the set of all relevant words in the corpus.¹ For each word $w \in W$ we associate two embedding vectors: A *target vector* $\rho_w \in \mathbb{R}^K$ and a *context vector* $\alpha_w \in \mathbb{R}^K$, where K denotes the dimensionality of the embedding. We organize the embedding vectors into matrices such that $\rho \in \mathbb{R}^{V \times K}$ is the matrix of target vectors and $\alpha \in \mathbb{R}^{V \times K}$ the matrix of

context vectors

$$\rho = \begin{bmatrix} (\rho_1)^\top \\ (\rho_2)^\top \\ \vdots \\ (\rho_V)^\top \end{bmatrix} \quad \text{and} \quad \alpha = \begin{bmatrix} (\alpha_1)^\top \\ (\alpha_2)^\top \\ \vdots \\ (\alpha_V)^\top \end{bmatrix}. \quad (1)$$

The complete set of parameters $\theta \in \mathbb{R}^{2V \times K}$ is then defined as

$$\theta = \begin{bmatrix} \rho \\ \alpha \end{bmatrix}. \quad (2)$$

For SGNS, the data is preprocessed to a list of word co-occurrences $\mathcal{D} = ((w_1, \dots, w_N), (v_1, \dots, v_N), (x_1, \dots, x_N)) \in W^N \times W^N \times \{0, 1\}^N$. The SGNS likelihood, up to a proportionality constant, models co-occurrences in the data via the dot products of ρ and α

$$\log p(\mathcal{D} | \theta) = \sum_{i=1}^N \left(\underbrace{\sum_{v \in C_i^+} \log \sigma(\rho_{w_i}^\top \alpha_v)}_{\text{positive samples}} + \underbrace{\sum_{v \in C_i^-} \log(1 - \sigma(\rho_{w_i}^\top \alpha_v))}_{\text{negative samples}} \right) \quad (3)$$

where C_i^+ is the context window, or the set of positive samples, for word w_i at index i in the data, and $\sigma(x) = (1 + e^{-x})^{-1}$ is the logistic function. Specifically, the context window is the set of words within the distance $M \in \mathbb{N}$ from the centre word, i.e.

$$C_i^+ = \{w_{i-M}, \dots, w_{i-1}, w_{i+1}, \dots, w_{i+M}\}. \quad (4)$$

For each positive sample, $n_s \in \mathbb{N}$ negative samples forming C_i^- are simulated randomly from the empirical distribution $\hat{F}$ of all words in W , sometimes with slight adjustments (Mikolov et al. 2013). Here $x_i = 1$ indicates that (w_i, v_i) is a positive sample and $x_i = 0$ a negative one. It can further be computationally and conceptually beneficial to formulate the likelihood as a sum over word pairs, i.e.

$$\log p(\mathcal{D} | \theta) = \sum_{(w,v)} \left(n_{wv}^+ \log \sigma(\rho_w^\top \alpha_v) + n_{wv}^- \log(1 - \sigma(\rho_w^\top \alpha_v)) \right), \quad (5)$$

where n_{wv}^+ and n_{wv}^- denotes the number of times that $(w, v) \in W \times W$ appear in the positive and negative samples.

Another variant of the probabilistic word2vec is the Continuous Bag-of-Words (CBoW) model. There, the effect of the context $C^\pm$ is aggregated to a sum of the context vectors

$$\alpha_{C_i^+} = \alpha_{C_i^-} = \sum_{v \in C_i^+} \alpha_v, \quad (6)$$

¹ Throughout the article, we use an implicit mapping from the set of words $W = \{\text{dog}, \text{cat}, \dots\}$ onto indices $\{1, \dots, V\}$, e.g. $\{\text{dog} \mapsto 1, \text{cat} \mapsto 2, \text{cow} \mapsto 3\}$. Hence, the word vector ρ_v for word $v \in W$ is just the v th column of the matrix ρ . The numeric indices have no meaning in word embeddings, and thus no information is lost.

where C_i^+ and C_i^- are the context windows for positive and negative samples, defined as Eq. (4). In CBoW model, the context is usually the same for both positive and negative samples, and the negative sampling is done for the target word instead. The data is then preprocessed into center words and their co-occurring context windows $\mathcal{D} = ((w_1, \dots, w_N), (C_1, \dots, C_N), (x_1, \dots, x_N)) \in W^N \times W^{N \times M} \times \{0, 1\}^N$. This results in a likelihood

$$\log p(\mathcal{D} | \theta) = \sum_{i=1}^N \left(\underbrace{x_i \log \sigma(\rho_{w_i}^\top \alpha_{C_i})}_{\text{positive samples}} + \underbrace{(1 - x_i) \log(1 - \sigma(\rho_{w_i}^\top \alpha_{C_i}))}_{\text{negative samples}} \right). \quad (7)$$

where the w_i 's are sampled from the empirical distribution of words $\hat{F}$.

For both Bayesian SGNS and CBoW, a common prior for the embedding vectors is a spherical Gaussian prior

$$\begin{aligned} \rho_w &\sim \mathcal{N}(\mathbf{0}, \lambda^{-1} \mathbf{I}), \quad w \in W, \\ \alpha_w &\sim \mathcal{N}(\mathbf{0}, \lambda^{-1} \mathbf{I}), \quad w \in W \end{aligned} \quad (8)$$

where $\lambda \in \mathbb{R}_+$ is a hyperparameter. Other priors used for embedding vectors, and common in practical applications, include dynamic priors (Rudolph and Blei 2017; Bamler and Mandt 2017), grouped priors (Rudolph et al. 2017), and informative priors focused on interpretability (Bodell et al. 2019), and graph priors (Yrjänäinen and Magnusson 2022), most of which are Gaussian.

3 Identification of the probabilistic SGNS model

The SGNS likelihood in Eq. 3 only consists of dot products between the target vectors ρ_w and the context vectors α_v . Because dot products are preserved under invertible linear transformations, we can apply the following transformations for any invertible matrix $\mathbf{P}$ in the general linear group of degree K , denoted $\text{GL}(K)$, i.e., $\mathbf{P} \in \text{GL}(K)$:

$$\rho'_w = \mathbf{P} \rho_w, \quad \alpha'_v = \mathbf{P}^{-\top} \alpha_v. \quad (9)$$

Under these transformations the dot product is invariant, $(\rho'_w)^\top \alpha'_v = (\rho_w)^\top \alpha_v$, and the likelihood remains unchanged. As a result, there are $K \times K$ degrees of freedom corresponding to these transformations that do not affect the likelihood. This makes the SGNS model non-identifiable in the sense that the mapping from the parameter space onto likelihoods is not one-to-one (Bickel and Doksum 2015).

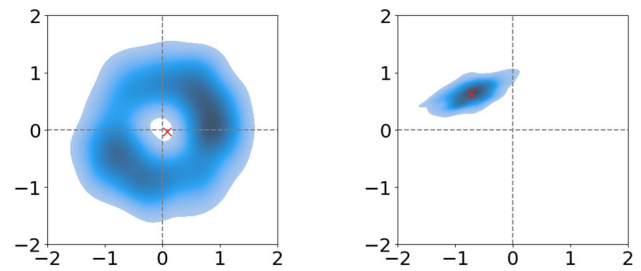


Fig. 1 The marginal distribution of $\rho_{1,1}$ and $\rho_{1,2}$ displays a circular shell-like symmetry (left), which is eliminated by fixing α_I to a MAP estimate (right). The posterior mean is marked with a red cross. $N = 2 \cdot 10^4$, $V = 10$, $K = 2$. Samples generated using Stan HMC

Similar symmetries of the likelihood can be found in other models that are defined in terms of products of parameters. Notably, some factor models also have $\text{GL}(K)$ symmetries in their likelihood, pertaining to the factor loading matrix $\mathbf{\Lambda}$ and factor correlation matrix $\mathbf{\Phi}$. These symmetries can be eliminated via a combination of fixing values to zeros, nonzeros and rank restrictions to the parameter matrices (Peeters 2012). However, as the invariance in factor models pertains to the covariance $\mathbf{\Lambda} \mathbf{\Phi} \mathbf{\Lambda}^\top$ which is quadratic in $\mathbf{\Lambda}$, in contrast to the potentially nonsymmetric, bilinear $\alpha \rho^\top$ in word embeddings, the factor restrictions posed on $\mathbf{\Lambda}$ and $\mathbf{\Phi}$ cannot be directly applied to ρ and α . Furthermore, due to the complex effects of parameter restrictions on the posterior, Bayesian approaches often instead use post-hoc rotation alignment such as Procrustes rotations (Aßmann et al. 2016), or a combination of Varimax and label normalisation (Papastamoulis and Ntzoufras 2022). However, these post-hoc rotations do not solve the full $\text{GL}(K)$ invariance, resulting in inflated posterior variance in eg. the commonly used cosine distance (see Appendix A for details).

In the context of word embeddings, Mu et al. (2019) propose adding L_2 regularisation to identify the likelihood. However, their proof only shows that the MAP estimate is unique up to a rotation. The likelihood still retains *orthogonal* symmetries $p(\theta | x) = p(\theta \mathbf{O} | x)$, where $\mathbf{O}^\top \mathbf{O} = \mathbf{I}$. The posterior distribution exhibits a *spherical-shell*-like structure, where even a very concentrated posterior will be distributed on the shell of a hypersphere, as illustrated in Fig. 1. Intuitively, one might view this as any orientation of the embeddings is valid.

We propose a simple solution: fixing K^2 elements of α to a constant value, which eliminates the $\text{GL}(K)$ symmetries. Specifically, we consider θ for which the K first context vectors, α_I , where $I = \{1, \dots, K\}$, form a fixed and predetermined invertible matrix $\mathbf{M} \in \mathbb{R}^{K \times K}$, while the rest of the context vectors form an unrestricted matrix $\alpha_J \in \mathbb{R}^{(V-K) \times K}$, $J = \{K+1, \dots, V\}$. This adds a similar, though not equivalent restriction on the parameters as some Bayesian approaches to factor modeling (Geweke and

Zhou 1996; Aßmann et al. 2016). In Proposition 1, we show that the restricted model can represent any likelihood that the unrestricted model can, and in Proposition 2, we show that the likelihood of the restricted model is identifiable.

Proposition 1 *Given any embedding $\hat{\alpha}$, $\hat{\rho}$, and any invertible matrix $\mathbf{M} \in \mathbb{R}^{K \times K}$, and either the SGNS or the CBoW likelihood, the following embedding*

$$\rho = \hat{\rho}(\hat{\alpha}_I^{-1}\mathbf{M})^{-\top}, \quad \alpha = \begin{bmatrix} \mathbf{M} \\ \hat{\alpha}_J \hat{\alpha}_I^{-1} \mathbf{M} \end{bmatrix} \quad (10)$$

has the same likelihood

$$p(\mathcal{D} \mid \rho, \alpha) = p(\mathcal{D} \mid \hat{\rho}, \hat{\alpha}) \quad (11)$$

for any dataset $\mathcal{D}$.

Proof in Appendix A.

Proposition 2 *Given any invertible matrix $\mathbf{M} \in \mathbb{R}^{K \times K}$, and either the SGNS or the CBoW likelihood, any two embeddings $\alpha = \begin{bmatrix} \mathbf{M} \\ \alpha_J \end{bmatrix}$, ρ and $\alpha' = \begin{bmatrix} \mathbf{M} \\ \alpha'_J \end{bmatrix}$, ρ' , for which $\alpha \neq \alpha'$ or $\rho \neq \rho'$, the likelihood*

$$p(\mathcal{D} \mid \rho, \alpha) \neq p(\mathcal{D} \mid \hat{\rho}, \hat{\alpha}) \quad (12)$$

for some dataset $\mathcal{D}$.

Proof in Appendix A.

Without loss of generality, we can choose any set of K context or word vectors to fix to have the same result, given that they form a full-rank matrix. The identifiability correction enables a direct comparison of the parameter values. This allows us to calculate $\hat{R}$ values (Gelman and Rubin 1992) to assess convergence to the stationary distribution and effective sample sizes (Gelman et al. 2013, p. 286) for specific parameters. It also makes it possible to calculate a posterior mean θ , which would otherwise just be zero, as $p(\theta \mid \mathcal{D}) = p(-\theta \mid \mathcal{D})$ for any θ and $\mathcal{D}$ without the fix. To retain cosine similarities of the MAP estimate, which is identified under Gaussian priors, we further propose to fix the K context vectors to their MAP estimates. We empirically verified that choosing different sets of anchor words leads to nearly identical posteriors for cosine similarities, see Appendix B.

3.1 Comparing identified and non-identified embeddings

Rotation-invariant features are commonly used to analyze word embeddings. These include Euclidean distances between the words (column distances) and cosine similarity (Garg et al. 2018; Eisenstein 2018; Kozłowski et al. 2019).

The co-occurrence probability of any word pair $(w, v) \in W \times W$ is calculated as

$$P(w \wedge v \mid \theta) = \sigma(\alpha_w^\top \rho_v). \quad (13)$$

Using this, we define the distance between two embeddings as the root mean squared error (RMSE) of the co-occurrence probabilities over all word pairs

$$\text{RMSE}_{co}(\theta, \theta') = \frac{1}{V} \sqrt{\sum_{v, w \in W} (P(w \wedge v \mid \theta) - P(w \wedge v \mid \theta'))^2} \quad (14)$$

Notably, RMSE_{co} is invariant under invertible transformations $\mathbf{P}$. When $\theta = \theta'$ then $\text{RMSE}_{co}(\theta, \theta') = 0$, but when $\text{RMSE}_{co}(\theta, \theta') = 0$, $\theta' = \begin{bmatrix} \rho \mathbf{P} \\ \alpha \mathbf{P}^{-1} \end{bmatrix}$ for some transformation $\mathbf{P} \in \text{GL}(K)$. Hence, we can compare both unidentified and identified embeddings. We then use it as a similarity metric between two embeddings. This is also the basis of our convergence analysis when true parameters are present. We interpret $\text{RMSE}_{co}(\hat{\theta}, \theta_{true}) = 0$ as the estimate $\hat{\theta}$ having converged to the true parameter θ_{true} .

4 A Gibbs sampler for word embeddings

For SGNS the posterior distribution of ρ_w , $w \in W$ is, conditionally on α , independent of the other $\rho_{w'}$, $w' \in W$, $w' \neq w$. This holds similarly for α_w as well

$$\begin{aligned} p(\rho \mid \alpha, \mathcal{D}) &\propto \prod_{w \in W} p(\rho_w \mid \alpha, \mathcal{D}) \\ p(\alpha \mid \rho, \mathcal{D}) &\propto \prod_{v \in W} p(\alpha_v \mid \rho, \mathcal{D}) \end{aligned} \quad (15)$$

Moreover, conditionally on α (or similarly on ρ), the likelihood for ρ is identical to logistic regression when treating α as covariates,

$$p(\mathcal{D} \mid \rho_w, \alpha) \propto \prod_{v \in W} \sigma(\alpha_v^\top \rho_w)^{n_{wv}^+} \sigma(-\alpha_v^\top \rho_w)^{n_{wv}^-}. \quad (16)$$

Like the logistic likelihood, the SGNS likelihood can thus be represented as a Polya-Gamma mixture of exponentiated polynomials of ρ and α

$$\begin{aligned} p(\mathcal{D} \mid \rho_w, \alpha) &\propto \prod_{v \in W} \sigma(\alpha_v^\top \rho_w)^{n_{wv}^+} \sigma(-\alpha_v^\top \rho_w)^{n_{wv}^-} \\ &= \prod_{v \in W} \exp(\kappa_{wv} \alpha_v^\top \rho_w) \mathbb{E}_{\omega_{wv}} [\exp(-\omega_{wv}/2 (\alpha_v^\top \rho_w)^2)] \end{aligned} \quad (17)$$

where $\kappa_{wv} = n_{wv}^+/2 - n_{wv}^-/2$ and $\omega_{wv} \sim \text{Polya-Gamma}$

$(n_{wv}^+ + n_{wv}^-, 0)$.

Given the additional latent variables ω_v , the conditional distributions of ρ_w and α_v become Gaussian. Moreover, the conditional distribution of ω is also in the Polya-Gamma family (Polson et al. 2013), so they can be efficiently sampled. Since the conditional distributions of all variables are known, a Gibbs sampler can be formulated, something observed in logistic matrix factorization (Klami 2015; Liu et al. 2018). The resulting Gibbs sampler is presented in Algorithm 1, and a detailed derivation of the SGNS Gibbs sampler is provided in Appendix C. 1.

Similar to SGNS, the CBoW posterior can also be represented as a Polya-Gamma mixture of Gaussian kernels

$$p(\mathcal{D} \mid \rho, \alpha) \propto \prod_{i=1}^N \sigma(\alpha_{C_i}^\top \rho_{w_i})^{x_i} \sigma(-\alpha_{C_i}^\top \rho_{w_i})^{(1-x_i)} \\ = \prod_{i=1}^N \exp(\kappa_i \alpha_{C_i}^\top \rho_{w_i}) \mathbb{E}_{\omega_i} [\exp(-\omega_i \cdot (\alpha_{C_i}^\top \rho_{w_i})^2 / 2)] \quad (18)$$

where $\kappa_i = x_i - 1/2$. As in SGNS, all conditional distributions are straight-forward to sample from: α and ρ have Gaussian distributions conditional on the other parameters, while the ω_i 's are Polya-Gamma distributed. The full CBOW Gibbs sampler is derived in Appendix C. 2 and presented in Algorithm 2.

Algorithm 1 SGNS Gibbs sampler

Require:

- 1: Vocabulary W and vocabulary size $V = |W|$.
- 2: Number of observations $N \in \mathbb{N}$
- 3: Dataset $\mathcal{D} = (\mathcal{D}_w, \mathcal{D}_v, \mathcal{D}_x) = ((w_1, \dots, w_N), (v_1, \dots, v_N), (x_1, \dots, x_N)) \in W^N \times W^N \times \{0, 1\}^N$.
- 4: Embedding dimensionality $K \in \mathbb{N}$.
- 5: Number of iterations $L \in \mathbb{N}$.

Ensure: $(\rho_{w,l}, \alpha_{w,l})_{l=1}^L$ samples.

- 6: **for** $w \in W$ **do**
- 7: Sample $\rho_{w,0} \sim \mathcal{N}(\mu_{\rho_w}, \Sigma_{\rho_w})$
- 8: Sample $\alpha_{w,0} \sim \mathcal{N}(\mu_{\alpha_w}, \Sigma_{\alpha_w})$
- 9: **end for**
- 10: **for** $l \in \{1, \dots, L\}$ **do**
- 11: **for** $w \in W$ **do**
- 12: Let $I_w = \{i : w_i \in \mathcal{D}_w, w_i = w\}$
- 13: $A \leftarrow (\dots, \alpha_{v_i}, \dots), i \in I_w$
- 14: $\kappa \leftarrow (\dots, (x_i - \frac{1}{2}), \dots), i \in I_w$
- 15: Sample $\omega \sim \text{PG}(1, A\rho_{w,l-1})$
- 16: $S_{\rho_w} \leftarrow (A^\top \text{diag}(\omega)A + \Sigma_{\rho_w}^{-1})^{-1}$
- 17: $m_{\rho_w} \leftarrow S_{\rho_w}(A^\top \kappa + \Sigma_{\rho_w}^{-1} \mu_{\rho_w})$
- 18: Sample $\rho_{w,l} \sim \mathcal{N}(m_{\rho_w}, S_{\rho_w})$
- 19: **end for**
- 20: **for** $w \in W$ **do**
- 21: Sample $\alpha_{w,l}$ similarly to $\rho_{w,l}$.
- 22: **end for**
- 23: **end for**

Algorithm 2 CBoW Gibbs sampler

Require:

- 1: Vocabulary W and vocabulary size $V = |W|$.
- 2: Number of observations $N \in \mathbb{N}$ and window size $M \in \mathbb{N}$.
- 3: Dataset $\mathcal{D} = (\mathcal{D}_w, \mathcal{D}_C, \mathcal{D}_x) = ((w_1, \dots, w_N), (C_1, \dots, C_N), (x_1, \dots, x_N)) \in W^N \times W^{N \times M} \times \{0, 1\}^N$.
- 4: Embedding dimensionality $K \in \mathbb{N}$.
- 5: Number of iterations $L \in \mathbb{N}$.

Ensure: $(\rho_{w,l}, \alpha_{w,l})_{l=1}^L$ samples.

- 6: **for** $w \in W$ **do** ▷ Initialize from prior
- 7: Sample $\rho_{w,0} \sim \mathcal{N}(\mu_{\rho_w}, \Sigma_{\rho_w})$
- 8: Sample $\alpha_{w,0} \sim \mathcal{N}(\mu_{\alpha_w}, \Sigma_{\alpha_w})$
- 9: **end for**
- 10: **for** $l \in \{1, \dots, L\}$ **do**
- 11: **for** $w \in W$ **do**
- 12: Let $I_w = \{i : w_i \in \mathcal{D}_w, w_i = w\}$
- 13: $Q \leftarrow (\dots, \alpha_{C_i}, \dots), i \in I_w$
- 14: $\kappa \leftarrow (\dots, (x_i - \frac{1}{2}), \dots), i \in I_w$
- 15: Sample $\omega \sim \text{PG}(1, Q\rho_{w,l})$
- 16: $S_{\rho_w} \leftarrow (Q^\top \text{diag}(\omega)Q + \Sigma_{\rho_w}^{-1})^{-1}$
- 17: $m_{\rho_w} \leftarrow S_{\rho_w}(Q^\top \kappa + \Sigma_{\rho_w}^{-1} \mu_{\rho_w})$
- 18: Sample $\rho_{w,l} \sim \mathcal{N}(m_{\rho_w}, S_{\rho_w})$
- 19: **end for**
- 20: **for** $w \in W$ **do**
- 21: Let $I_v = \{i : C_i \in \mathcal{D}_C, w \in C_i\}$
- 22: $R \leftarrow (\dots, \rho_{w_i}, \dots), i \in I_v$
- 23: $\kappa \leftarrow (\dots, (x_i - \frac{1}{2}), \dots), i \in I_v$
- 24: $\delta \leftarrow (\dots, \rho_i^\top (\alpha_{C,i} - \alpha_{w,l-1}), \dots), i \in I_v$
- 25: Sample $\omega \sim \text{PG}(1, R\alpha_{w,l-1})$
- 26: $S_{\alpha_w} \leftarrow (R^\top \text{diag}(\omega)R + \Sigma_{\alpha_w}^{-1})^{-1}$
- 27: $m_{\alpha_w} \leftarrow S_{\alpha_w}(R^\top (\kappa - \delta \circ \omega) + \Sigma_{\alpha_w}^{-1} \mu_{\alpha_w})$
- 28: Sample $\alpha_{w,l} \sim \mathcal{N}(m_{\alpha_w}, S_{\alpha_w})$
- 29: **end for**
- 30: **end for**

4.1 Efficient implementation

Some design choices critically affect the efficiency of the Gibbs sampler. First, the operations in the algorithms are presented word-for-word, whereas in practice they are mostly independent and can be executed in parallel in large batches (see Appendix C for details). In fact, the only parameters that are *not* conditionally independent given the other blocks are the α_v 's in Algorithm 2. Still, the most computationally expensive calculations, the product $R^\top \text{diag}(\omega)R$ and the inversion and factorization of $S_{\alpha_w}^{-1}$ in Algorithm 2 only depend on ω and ρ , not the other α_v 's. In applications, informative priors are often applied on ρ , and even then, most ρ_w 's are conditionally independent of each other.

Furthermore, both Algorithm 1 and 2 ostensibly rely on matrix inversion to obtain $m_{(\cdot)}$ for both the ρ 's and α 's. However, these inversions are only used to calculate the product with two vectors per iteration. Thus, sampling Gaussians via the Cholesky decomposition of $S_{(\cdot)}^{-1}$ is substantially more computationally efficient. The mean and the random offset can be obtained via a triangular solve given the Cholesky factor of $S_{(\cdot)}^{-1}$.

In the SGNS model, multiple co-occurrences of a word pair can be aggregated to one observation with a count n_{wv} . This speeds up both the Gibbs sampler and the Stan reference implementations. The aggregation is explained in detail in Appendix C.1.3.

For CBoW specifically, there is a slight trade-off between efficiency in terms of raw samples and the effective sample size (ESS). If the sampling is split into B -sized batches so that within each batch, the α_v 's are conditionally independent (i.e. when the words do not share context windows), it is possible to avoid factorising larger $(B \cdot K) \times (B \cdot K)$ sized matrices and instead factorise B matrices of size $K \times K$. This can be far more computationally efficient, especially on a GPU. On the other hand, organising the sampling in conditionally dependent blocks is beneficial for the efficiency of a Gibbs sampler in terms of the ESS (Roberts and Sahu 1997).

4.2 Computational complexity

For SGNS, aggregated over unique word pairs, the dominant cost per Gibbs iteration is the calculation of the precision matrix $S_{(\cdot)}^{-1}$, which scales as $O(N_u K^2)$ and its Cholesky decomposition, $O(V K^3)$. Here N_u denotes the number of unique co-occurring word pairs, and by definition $N_u \leq V^2$. Thus, the total asymptotic complexity for SGNS is $O(N_u K^2 + V K^3)$.

For CBoW the complexity per iteration is $O(N K^2 + V K^3)$, with N appearing directly, since the observations cannot be aggregated in the same way as in SGNS. Thus, for small N , SGNS and CBoW have a similar scaling, but as N increases, SGNS is bounded by $O(V^2 K^2 + V K^3)$ whereas CBoW continues to scale linearly with N .

For NUTS-HMC, the evaluation of the posterior gradient has a complexity of $O(N_u K + V K)$ and $O(N K)$, respectively. In addition to that, the appropriate number of leapfrog steps in NUTS-HMC increases with the model dimensionality. This depends on the posterior, but is usually at least $O(d^{1/4})$ where d is the number of parameters (Neal et al. 2011). All in all, this makes for a complexity of $O(N_u K^{5/4} V^{1/4})$ or more. Since K is essentially a constant, rarely exceeding 300, this is slower than the Gibbs sampler: $O(N_u)$ vs $O(N_u V^{1/4})$. See Appendix D for a detailed breakdown.

5 Experiments

In the experiments, we study the Gibbs samplers presented in Sec. 4, and compare them with the current state-of-the-art for estimation, namely MAP estimation, MFVI, and HMC sampling. We use HMC implemented in Stan² and the MFVI

algorithm presented by Bamler and Mandt (2017), using the Stan autodiff implementation of MFVI. As another baseline we also ran Metropolis Adjusted Langevin Sampling (MALA), whose results are presented in Appendix B.1. It generated very low effective sample sizes and was thus omitted from the comparisons. All code for the experiments is provided on GitHub³.

To study convergence and uncertainty estimates, we use a simulated dataset with known parameters. We complement the simulation experiments with empirical evaluations on two real-world datasets, namely the standard embedding datasets MovieLens (Harper and Konstan 2015) and US Congress (Gentzkow et al. 2018).

5.1 Simulated data

In the simulation study, we use a set of randomly generated word and context vectors. For each $w \in W$, the word and context vectors are simulated from

$$\rho_w \sim \mathcal{N}(\mathbf{0}, \frac{\varepsilon^2}{K} \mathbf{I}), \quad \alpha_w \sim \mathcal{N}(\mathbf{0}, \frac{\varepsilon^2}{K} \mathbf{I}), \quad (19)$$

where K is the dimensionality of the embedding, and $\varepsilon = 1$ the signal-to-noise ratio. The embedding is scaled inversely to the dimensionality so that the dot products, and thus the signal-to-noise ratio, remain similar in magnitude regardless of the dimensionality K .

For SGNS, we generate random word pairs $(w_i, v_i) \in W \times W$ by sampling uniformly. A Bernoulli random variable, representing whether the pair is a positive or negative sample, X_i is then sampled for $i \in \{1, \dots, N\}$

$$\begin{aligned} (w_i, v_i) &\sim \text{Uniform}(W \times W) \\ X_i &\sim \text{Bernoulli}(\sigma(\alpha_{v_i}^\top \rho_{w_i})) \end{aligned} \quad (20)$$

Each resulting dataset is then estimated using MAP estimation, VI, HMC, and the Gibbs sampler, with the true K and λ . For the CBoW experiments, a context window of size 2 will be generated uniformly, and the rest of the simulation will follow the same procedure as above.

In natural language, words often have frequencies proportional to their inverse rank in the vocabulary, i.e. words tend to follow Zipf's law (Zipf 1935; Piantadosi 2014). As another experiment, we simulate this by generating w_i and v_i with

$$\begin{aligned} p(w_i) &\propto \frac{1}{\text{rank}(w_i)^a + b}, \\ p(v_i) &\propto \frac{1}{\text{rank}(v_i)^a + b}, \end{aligned} \quad (21)$$

² <https://mc-stan.org/>

³ <https://anonymous.4open.science/r/embedding-uncertainty-estimation-46ED/>

Table 1 The mean effective sample size of the Gibbs sampler by N for the 1000 samples after a warmup of 1000

N	ESS (SGNS)	ESS (CBoW)
1, 000	963.8	964.7
2, 000	963.6	964.9
5, 000	964.0	962.3
10, 000	963.3	957.5
20, 000	958.9	943.0
50, 000	934.2	857.0
100, 000	880.2	706.8
200, 000	763.6	669.9
300, 000	679.1	683.6
400, 000	632.4	682.8
500, 000	603.1	659.0

Simulated data, $K = 20$, $V = 200$

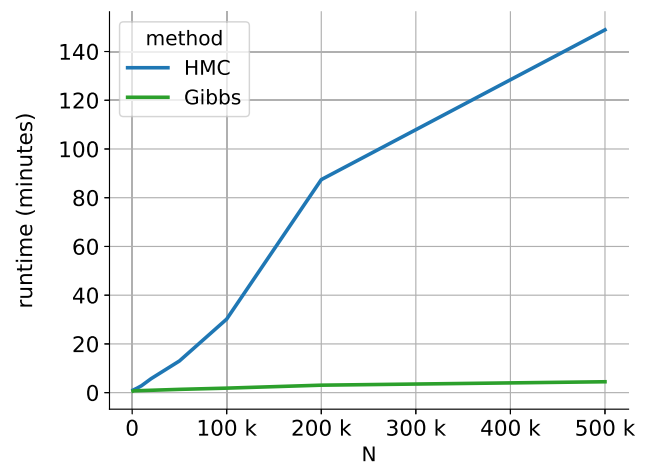
where $\text{rank}(\cdot)$ is an enumeration of W , and a and b are positive constants. We set $a = 1$ and $b = 2.7$, as suggested by Mandelbrot et al. (1953).

5.1.1 Sampling rate and effective sample size

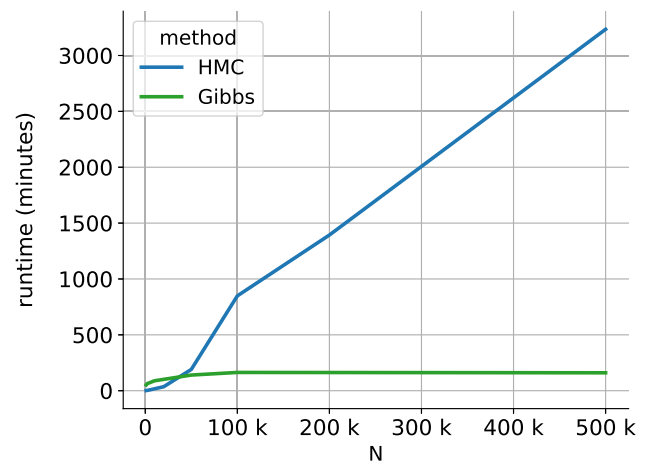
The goal of the simulation is to investigate two aspects of the Gibbs sampler: its efficiency and accuracy. First, the wall-clock sampling rates of NUTS HMC and the Gibbs sampler are illustrated in Fig. 2.

The Gibbs sampler outperforms HMC on all data sizes for SGNS, and almost all data sizes for CBoW, and the difference increases as N grows. Stan HMC scales roughly linearly wrt. N , while the Gibbs sampler scales sublinearly. This is because the cost of factorising $S_{(\cdot)}^{-1}$ is constant wrt. N , and while the calculation of $R^T \text{diag}(\omega) R$ and $Q^T \text{diag}(\omega) Q$ is asymptotically linear in N , it scales better than that on real data sizes. In addition to the already large performance gap, the Gibbs sampler could be easily run on the GPU. As we will see, running HMC on large datasets is practically not possible due to poor scaling with respect to model and data size.

In addition to the absolute number of samples, we calculated the autocorrelation-based effective sample size (Gelman et al. 2013, p. 286). Since the cosine similarities are the most common downstream usage of word embeddings, we calculated the cosine similarity $\text{cossim}(\rho_w, \rho_v)$ for all word pairs $(w, v) \in W \times W$ for each sample. The results are presented in Table 1. For all data sizes, the Gibbs sampler achieves an effective sample rate (ESS per no. of samples) of 60% or more, with smaller data sizes yielding more effective samples per raw sample.



(a) SGNS



(b) CBoW

Fig. 2 Time required to generate 2000 samples with Stan-HMC and the Gibbs sampler by the number of observations N for $K = 20$, $V = 200$. Both algorithms run using the CPU

5.1.2 Convergence, coverage and cosine similarities

As the first measure of accuracy, we study the convergence of the estimated embeddings $\hat{\theta}$ to the true value θ_{true} . To do this, we use the divergence RMSE_{co} defined in Equation (14). Figure 3 shows the convergence of the different estimation methods in terms of RMSE_{co} . HMC and the Gibbs sampler perform similarly, converging toward the true values, while MFVI appears to converge slightly slower as N increases. With small sample sizes, MAP underperforms the other methods, while on large sample sizes, it approaches the performance of HMC and the Gibbs sampler. As expected, HMC, MAP and the Gibbs sampler are asymptotically very

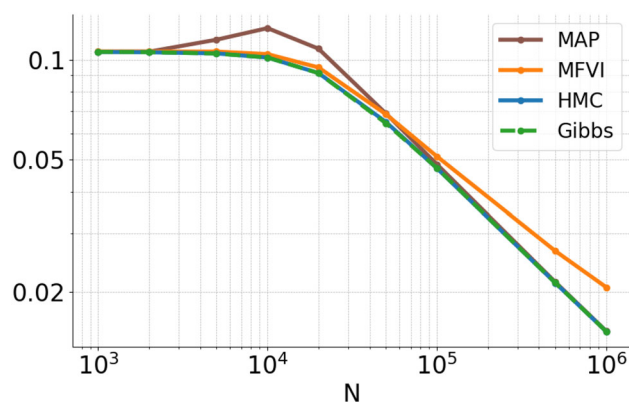


Fig. 3 Log-log plotted RMSE_{co} for $K = 5$ and $V = 100$. Averaged over 10 simulated datasets. Uniform simulation

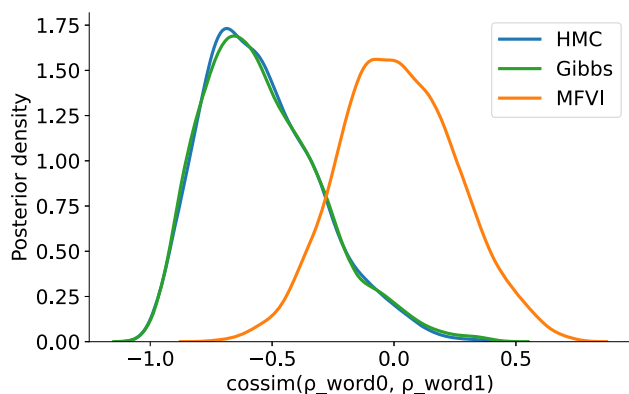


Fig. 4 Cosine similarity between ρ_1 and ρ_2 . SGNS, $K = 5$, $V = 100$, $N = 10^5$. Gibbs sampler in green, HMC in blue and MFVI in orange

close. Different model sizes and the Zipf's law experiment yielded similar results and are plotted in Appendix B.

As seen in Fig. 3, a loglinear pattern in RMSE_{co} emerges once N is sufficiently large. This holds for all values of V and K , as well as the Zipf's law generated data (see details in Appendix B). Log-log linearity corresponds to $O(N^{-D})$ convergence of RMSE_{co} for some constant D . For all combinations of K and V , numerical estimates of D were close to -0.5 , i.e. $\text{RMSE}_{co}(N) \approx O(N^{-0.5})$. Since Bayesian point estimates $\hat{\beta}$ approach a $1/\sqrt{N}$ scaled Gaussian as $N \rightarrow \infty$ (Kleijn 2012), the squared error $\mathbb{E}[(\hat{\beta} - \beta)^2]$ is $O(N^{-1})$. Thus, its square root is $O(N^{-0.5})$, which matches the convergence rate that we observe for HMC, MAP, and the Gibbs sampler, but interestingly, not for MFVI. Moreover, the increase in RMSE_{co} given an increase in V was very close to 1, implying a convergence rate of $O(N^{0.5}/V)$.

According to Bernstein-von-Mises theorem, under relatively mild assumptions, credible intervals should have valid coverage proportions as the number of observations $N \rightarrow \infty$ (Kleijn 2012). Hence, we study the coverage intervals (90%) of $\alpha_v^T \rho_w$ as a function of data size, keeping V and K static.

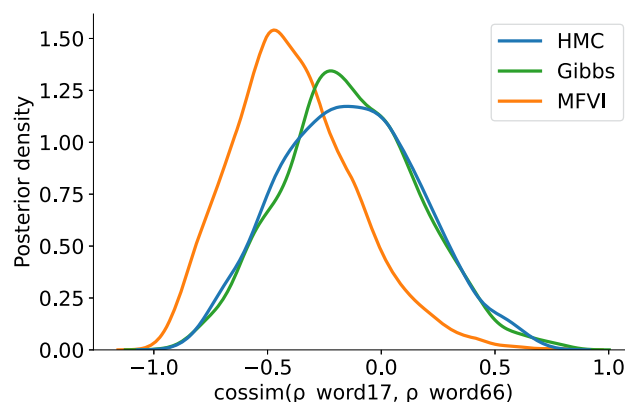


Fig. 5 Cosine similarity between ρ_{18} and ρ_{67} . CBoW, $K = 5$, $V = 100$, $N = 10^5$. Gibbs sampler in green, HMC in blue and MFVI in orange

The results over all word pairs and 10 simulated datasets are shown in Table 2. HMC and the Gibbs sampler cover the true value roughly 90% of the time regardless of K , V or N . For MFVI, the confidence intervals are too wide for small values of N , and for large values of N , MFVI only covers the true value under 80.0%-85.0% of the time. This issue worsens as N grows and follows a similar pattern across all combinations of K and V . The tendency of MFVI to underestimate uncertainty is in line with theoretical results on mean-field variational inference (Wang and Blei 2018).

Finally, we compare the posterior distributions of cosine similarities, a common use case for word embedding in applied settings (Garg et al. 2018; Eisenstein 2018; Kozłowski et al. 2019). As seen in Figs. 4 and 5, the posterior distributions of the cosine similarities obtained from HMC and the Gibbs sampler are very similar. On the other hand, the posteriors obtained from MFVI differ drastically from these. This is not surprising, given the use of an MFVI posterior approximation, and it clearly shows the problem of using MFVI for uncertainty estimation in SGNS embeddings.

5.2 MovieLens dataset

Rudolph et al. (2016) embed the MovieLens data (Harper and Konstan 2015) using a probabilistic embedding to find thematically similar movies. Subsequently, we use the MovieLens dataset as the basis for experiments with non-text embedding models Rudolph et al. (2016) Some results of the posterior of cosine similarities is illustrated in Fig. 6.

MovieLens is a relatively small dataset, allowing us to compare the performance of HMC and the Gibbs sampler. For smaller embedding models ($K = 20$), HMC is still competitive, as shown in Tables 3 and 4. However, HMC's efficiency drastically drops as dimensionalities and dataset size increase. For SGNS, HMC only produced a handful of samples before the program was terminated due to hitting

Table 2 Coverage (%) of the true co-occurrence probability $P(w \wedge v \mid \theta_{true})$ of the 90%-credible interval of $P(w \wedge v \mid \theta)$ in the simulated experiment for different data sizes

Dataset	Method	1k	2k	5k	10k	20k	50k	100k	500k	1M
$K = 5, V = 100$	HMC	89.9	89.8	89.7	89.7	89.7	89.9	89.9	90.2	89.9
	MFVI	90.2	88.3†	84.0‡	84.5‡	84.2‡	85.8‡	86.6‡	83.3‡	79.5‡
	Gibbs	89.7	89.6	89.7	89.5	89.6	89.9	89.6	90.1	89.7
$K = 10, V = 100$	HMC	90.0	90.0	90.0	90.0	90.0	90.1	90.2	90.0	90.1
	MFVI	92.5‡	91.9‡	90.4	88.0‡	83.7‡	85.4‡	85.6‡	85.1‡	83.0‡
	Gibbs	89.9	89.9	89.9	89.9	89.9	89.9	89.6	89.9	90.1
$K = 20, V = 100$	HMC	90.0	89.9	90.0	89.9	90.0	90.0	89.9	89.7	89.9
	MFVI	93.3‡	93.2‡	92.8‡	92.3‡	90.9	87.3‡	83.5‡	85.3‡	84.5‡
	Gibbs	89.4	89.4	89.4	89.4	89.4	89.4	89.4	89.7	90.0
$K = 5, V = 200$	HMC	90.2	90.2	90.2	90.3	90.2	90.2	90.1	90.1	90.0
	MFVI	91.5‡	90.6	87.7‡	84.0‡	84.5‡	83.9‡	86.2‡	85.4‡	83.1‡
	Gibbs	90.1	90.1	90.1	90.1	90.0	90.0	89.9	89.7	89.9
$K = 10, V = 200$	HMC	88.7	88.8	88.8	88.8	88.9	89.0	89.3	90.4	90.2
	MFVI	92.3‡	92.1‡	91.4‡	90.0	87.4‡	82.6‡	84.4‡	85.6‡	84.9‡
	Gibbs	89.5	89.5	89.5	89.6	89.7	89.7	89.7	89.7	89.7
$K = 20, V = 200$	HMC	90.1	90.1	90.2	90.1	90.2	90.2	90.1	90.1	90.1
	MFVI	93.5‡	93.5‡	93.5‡	93.2‡	92.6‡	90.5	87.7‡	85.1‡	84.4‡
	Gibbs	90.1	90.1	90.1	90.1	90.1	90.1	90.0	89.8	89.8

Averaged over 10 simulated datasets and all word pairs $(w, v) \in W \times W$. A deviation of over 1% is denoted by † and over 5% by ‡

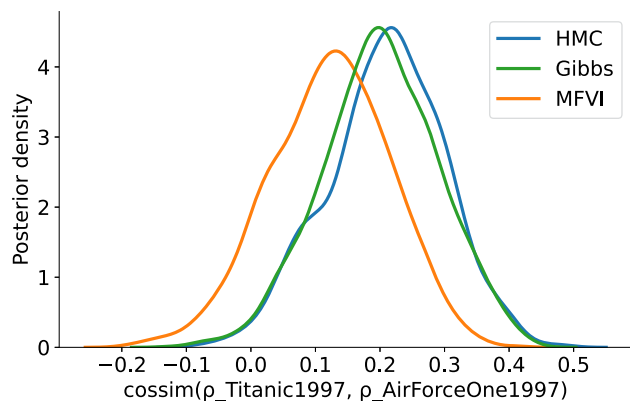


Fig. 6 Cosine similarity between *Titanic1998* and *AirForceOne1998* in the MovieLens dataset. CBoW, $K = 100$, $V = 603$, $N = 1.2 \cdot 10^5$

the time limit. While the Gibbs sampler also yields significantly fewer samples per second under these conditions, its performance remains acceptable.

Another advantage of posterior sampling is that it permits the use of the posterior mean as a point estimator for the word embeddings, rather than relying on the MAP estimator. The posterior mean is the Bayes estimator under squared-error loss. Thus, employing the posterior mean yields point estimates that are optimal under a more appropriate loss for continuous Euclidean parameters, providing estimators that

Table 3 Samples per second for the Movielens dataset for the SGNS model

$K = 20$		
Method	Samples/s	Words/s
HMC	0.833	502.5
Gibbs	12.55	7566.8
$K = 100$		
Method	Samples/s	Words/s
HMC	< 0.0069*	< 4.19*
Gibbs	1.010	662.64

Above: $K = 20$, $V = 603$, $N = 5 \cdot 10^4$; below: $K = 100$, $V = 603$, $N = 10^6$. Words per second is samples per second normalised by V . Gibbs results shown for $S = 10$. Both algorithms were run on the CPU. Runs with an asterisk (*) did not finish before a time limit

are both more principled and more natural for embeddings in $\mathbb{R}^K$ (Bernardo et al. 1994).

In Table 5, we compare the posterior mean (Bayes) estimator $\hat{\theta}_{PM}$ to the MAP estimator $\hat{\theta}_{MAP}$ and the MFVI mean estimator $\hat{\theta}_{MFVI}$ on the MovieLens data. The metric used is the SGNS likelihood on held-out data. On all sample sizes (N), the posterior mean estimator outperforms the MAP estimator, giving an empirical argument for the use of a Bayes word embedding estimator. However, as expected, the main benefit can be seen in smaller data.

Table 4 Samples per second for the Movielens dataset for the CBoW model

$K = 20$		
Method	Samples/s	Words/s
HMC	0.0790	47.63
Gibbs	0.1538	92.77
$K = 100$		
Method	Samples/s	Words/s
HMC	0.0218	13.16
Gibbs	0.0881	53.10

Above: $K = 20$; below: $K = 100$. $V = 603$, $N = 1.2 \cdot 10^6$. Words per second is samples per second normalized by V

Table 5 Averaged hold-out log likelihood on the MovieLens data by estimator and sample size

	$\hat{\theta}_{Bayes}$	$\hat{\theta}_{MAP}$	$\hat{\theta}_{MFVI}$
$N = 85,000$	-0.7320	-0.7801	-1.1142
$N = 170,000$	-0.6814	-0.7247	-1.1201
$N = 425,000$	-0.6458	-0.6721	-0.8715
$N = 850,000$	-0.6382	-0.6571	-0.6609

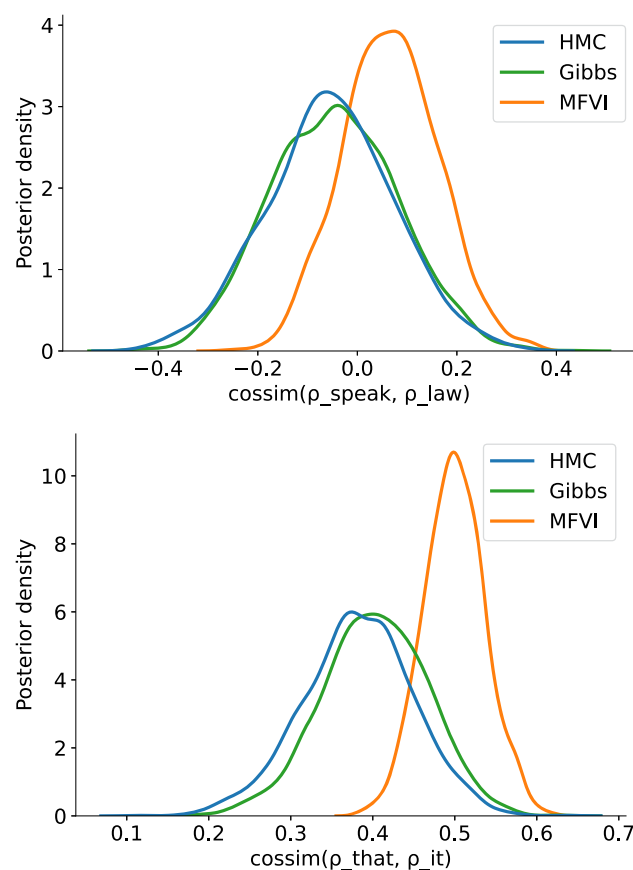
SGNS, $K = 20$, $\lambda = 1$

5.3 US Congress dataset

For the text data experiment, we use the US Congress speech corpus of Gentzkow et al. (2018), a well-known text dataset popular in the social sciences (Myrick 2021; Weiss and Zoorob 2021; Card et al. 2022). We restrict the vocabulary to the 5000 most common words in the dataset, and use a subset of 1 – 10 million words. Further hyperparameters were the context window size $M = 2$, and the number of negative samples $n_s = 1$. After this preprocessing, we obtained a sample of 10 million observations. We set $K = 100$ and $\lambda = 1$, common value in applications.

Our experiments are threefold. First, we compare the Gibbs sampler to HMC and MFVI in terms of posteriors of cosine similarities. Second, we illustrate how these posteriors of cosine similarities affect a common downstream measurement of association tests. Finally, we compare the VI mean, MAP and HMC and Gibbs sampler posterior mean point estimates in terms of test set hold-out performance. For the larger data sizes, we do not use HMC as it was infeasibly slow.

The posteriors of the cosine similarities between some word pairs are illustrated in Fig. 7, on a low dimensionality $K = 50$ where all three algorithms could be run. The Gibbs sampler and HMC yield similar posteriors for this metric. On the other hand, MFVI yields substantially different results from HMC and the Gibbs sampler, both in terms of the posterior mean, and a smaller posterior variance. The latter

**Fig. 7** Posterior distribution of the cosine similarity between words in the US Congress data. Above: 'speak' and 'law', below: 'that' and 'it'. SGNS, $K = 50$, $N = 500,000$. Gibbs sampler in green, HMC in blue and MFVI in orange

result is in line with theoretical results on MFVI (Wang and Blei 2018). The results are similar when $K = 100$, which is illustrated in Appendix B. Since cosine similarity is one of the primary ways of analysis of word embeddings, these differences argue against the use of mean-field approximation for word embeddings in scientific settings where proper uncertainty quantification is needed.

A common downstream use of word embeddings is to quantify the association between a target word and two contrasting groups of words (Caliskan et al. 2017; Garg et al. 2018; Ma et al. 2025). Given a target word w and two sets of words G^+ and G^- , an association score is defined as the difference in mean cosine similarity,

$$\Delta_{(G^+, G^-)}(w) = \frac{1}{|G^+|} \sum_{v \in G^+} \text{cossim}(\rho_w, \rho_v) - \frac{1}{|G^-|} \sum_{v \in G^-} \text{cossim}(\rho_w, \rho_v).$$

A positive value of $\Delta_{(G^+, G^-)}(w)$ indicates that w is more closely associated with G^+ than G^- in the embedding space.

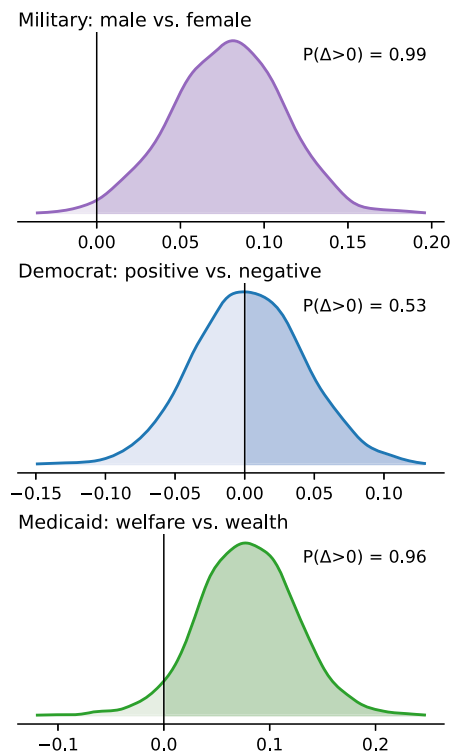


Fig. 8 Posterior distribution of the association score $\Delta_{(G^+, G^-)}(w)$ in three cases for the US Congress data, computed from posterior samples of the Gibbs sampler ($N = 10,000,000$, $K = 100$)

Table 6 Hold-out log likelihood on the US Congress data

$K = 50$	$\hat{\theta}_{PM-Gibbs}$	$\hat{\theta}_{PM-HMC}$	$\hat{\theta}_{MAP}$
$N = 10^5$	-0.7648	-0.7584	-0.8927
$N = 5 \cdot 10^5$	-0.7556	-0.7574	-0.8648
$N = 10^6$	-0.72549	-0.7374	-0.7896
$N = 5 \cdot 10^6$	-0.6551	NA	-0.6625
$K = 100$	$\hat{\theta}_{PM-Gibbs}$	$\hat{\theta}_{PM-HMC}$	$\hat{\theta}_{MAP}$
$N = 5 \cdot 10^5$	-0.7374	NA	-0.8209
$N = 10^6$	-0.7136	NA	-0.7836
$N = 5 \cdot 10^6$	-0.6594	NA	-0.6948
$N = 10^7$	-0.6360	NA	-0.6357

SGNS, $\lambda = 1$, $M = 2$, $n_s = 1$. PM-Gibbs is the posterior mean estimator for the Gibbs sampler, PM-HMC is the posterior mean estimator for the HMC

In isolation, a point estimate of $\Delta_{(G^+, G^-)}(w)$ with no uncertainty is difficult to interpret. With posterior samples of all the embeddings, however, we can make natural probabilistic statements about $\Delta_{(G^+, G^-)}(w)$. For example, one can ask for $P(\Delta_{(G^+, G^-)}(w) > 0 \mid \mathcal{D})$, the posterior probability that the target word is more associated with G^+ compared to G^- . For illustrative purposes, we conduct three such analyses, with results presented in Fig. 8. The details of the analyses can be found in Appendix F.

Finally, Table 6 presents the results of the hold-out performance experiment on the US congress data. The Bayes estimators using HMC and Gibbs both outperform the MAP estimator and behave similarly, as expected. However, on the full training data $N = 10^7$, MAP is on par with the Bayes estimators. This is to be expected both based on theory and the simulation results: the Bayes estimator and MAP converge to the same result as $N \rightarrow \infty$.

6 Conclusion

In this article, we introduced the first scalable and accurate Bayesian posterior sampler for probabilistic word embeddings leveraging Polya–Gamma–augmented Gibbs sampling. By addressing the long-standing identifiability problem in SGNS and CBoW word embeddings, we established conditions under which the likelihood becomes identifiable and demonstrated that this correction is essential for obtaining meaningful posterior summaries. Our experiments showed that the proposed sampler attains accuracy comparable to NUTS–HMC while scaling substantially better to realistic corpus sizes. In addition, it yields calibrated uncertainty estimates that MFVI consistently fails to provide. These results establish the proposed sampler as a practical and principled alternative to existing inference techniques for high-dimensional word embedding models.

Beyond uncertainty quantification, posterior sampling enables the use of the posterior mean as a Bayes estimator under squared-error loss, providing more stable and statistically appropriate point estimates for embeddings in Euclidean space. Across both simulated and real-world datasets, these posterior-mean estimators consistently outperformed MAP and MFVI in predictive evaluations, particularly in moderate-data regimes. Taken together, our theoretical and empirical findings demonstrate that fully Bayesian inference for word embeddings is both feasible and beneficial, opening the door to more reliable statistical analysis of text and other co-occurrence data.

A Proofs

A. 1 Proofs of Propositions 1 and 2

Proof (Proposition 1) For any data $\mathcal{D}$, the likelihood only depends on the parameters via the product $\alpha \rho^\top$, i.e.

$$p_{SGNS}(\mathcal{D} \mid \rho, \alpha) \equiv g_{SGNS}(\mathcal{D}, \alpha \rho^\top) \quad (22)$$

$$p_{CBoW}(\mathcal{D} \mid \rho, \alpha) \equiv g_{CBoW}(\mathcal{D}, \alpha \rho^\top) \quad (23)$$

which in turn means that for both models

$$\alpha \rho^\top = \alpha'(\rho')^\top \implies p(\mathcal{D} \mid \rho, \alpha) = p(\mathcal{D} \mid \rho', \alpha') \quad (24)$$

for any data $\mathcal{D}$. Thus, we need to show that the matrices $\alpha \rho^\top$ and $\hat{\alpha} \hat{\rho}^\top$ are always equal

$$\alpha \rho^\top = \begin{bmatrix} \mathbf{M} \\ \hat{\alpha}_J \hat{\alpha}_I^{-1} \mathbf{M} \end{bmatrix} (\hat{\rho}(\hat{\alpha}_I^{-1} \mathbf{M})^{-\top})^\top \quad (25)$$

$$= \begin{bmatrix} \mathbf{M} \\ \hat{\alpha}_J \hat{\alpha}_I^{-1} \mathbf{M} \end{bmatrix} (\hat{\alpha}_I^{-1} \mathbf{M})^{-1} \hat{\rho} \quad (26)$$

$$= \begin{bmatrix} \mathbf{M} \\ \hat{\alpha}_J \hat{\alpha}_I^{-1} \mathbf{M} \end{bmatrix} \mathbf{M}^{-1} \hat{\alpha}_I \hat{\rho}^\top \quad (27)$$

$$= \begin{bmatrix} \mathbf{M} \mathbf{M}^{-1} \hat{\alpha}_I \\ \hat{\alpha}_J \hat{\alpha}_I^{-1} \mathbf{M} \mathbf{M}^{-1} \hat{\alpha}_I \end{bmatrix} \hat{\rho}^\top \quad (28)$$

$$= \begin{bmatrix} \hat{\alpha}_I \\ \hat{\alpha}_J \end{bmatrix} \hat{\rho}^\top = \hat{\alpha} \hat{\rho}^\top \quad (29)$$

and thus $p(\mathcal{D} \mid \hat{\rho}, \hat{\alpha}) \equiv p(\mathcal{D} \mid \rho, \alpha)$.

Proof (Proposition 2)

$$\theta \neq \theta' \implies p(\mathcal{D} \mid \theta) \neq p(\mathcal{D} \mid \theta') \quad (30)$$

By contraposition, this is equivalent to

$$p(\mathcal{D} \mid \theta') = p(\mathcal{D} \mid \theta) \implies \theta = \theta' \quad (31)$$

The likelihood is a function of the product

$$p(\mathcal{D} \mid \theta) \equiv g(\mathcal{D}, \alpha \rho^\top) \quad (32)$$

which means that two likelihoods are identical for any data $\mathcal{D}$ iff $\alpha \rho^\top = \hat{\alpha} \hat{\rho}^\top$. Given that $\alpha_I = \hat{\alpha}_I = \mathbf{M}$, the equality becomes

$$\alpha \rho^\top = \begin{pmatrix} \alpha_I \rho_I^\top & \alpha_I \rho_J^\top \\ \alpha_1 \rho_I^\top & \alpha_1 \rho_J^\top \end{pmatrix} \quad (33)$$

$$= \begin{pmatrix} \hat{\alpha}_I \rho_I^\top & \hat{\alpha}_I \rho_J^\top \\ \alpha_1 \rho_I^\top & \alpha_1 \rho_J^\top \end{pmatrix} \quad (34)$$

$$= \begin{pmatrix} \hat{\alpha}_I \hat{\rho}_I^\top & \hat{\alpha}_I \hat{\rho}_J^\top \\ \hat{\alpha}_1 \hat{\rho}_I^\top & \hat{\alpha}_1 \hat{\rho}_J^\top \end{pmatrix} \quad (35)$$

Since α_I^{-1} exists

$$\hat{\alpha}_I \rho_I^\top = \hat{\alpha}_I \hat{\rho}_I^\top \implies \rho_I = \hat{\rho}_I \quad (36)$$

$$\hat{\alpha}_I \rho_J^\top = \hat{\alpha}_I \hat{\rho}_J^\top \implies \rho_J = \hat{\rho}_J \quad (37)$$

and thus

$$\alpha_J \hat{\rho}_I^\top = \hat{\alpha}_J \hat{\rho}_I^\top \implies \alpha_J = \hat{\alpha}_J \quad (38)$$

i.e., $\rho = \hat{\rho}$ and $\alpha = \hat{\alpha}$. $\square$

A. 2 Inflated variance under skew transformations

Given a likelihood, defined as all parameter values that fulfil $\alpha \rho^\top = A$, the posterior becomes

$$p(\rho, \alpha \mid \alpha \rho^\top = A) \quad (39)$$

$$\propto \exp(-\lambda (\|W \rho_{\text{ref}}\|^2 + \|W^{-\top} \alpha_{\text{ref}}\|^2)) \quad (40)$$

where $\rho = \mathbf{P} \rho_{\text{ref}}$, $\alpha = \mathbf{P}^{-\top} \alpha_{\text{ref}}$ and $\rho_{\text{ref}}, \alpha_{\text{ref}}$ are some reference embeddings that fulfill $\alpha \rho^\top = A$. The cosine similarity given $\mathbf{P}$ is then

$$\text{cossim}(x, y; \mathbf{P}) = \frac{\rho_{\text{ref},x}^\top \mathbf{P}^\top \mathbf{P} \rho_{\text{ref},y}}{\|\mathbf{P} \rho_{\text{ref},x}\| \|\mathbf{P} \rho_{\text{ref},y}\|} \quad (41)$$

Let $\mathcal{W}_{x,y,s} = \{\mathbf{P} \in \mathbb{R}^{K \times K} \mid \text{cossim}(x, y; \mathbf{P}) = s\}$. Conditional on the likelihood being $\alpha \rho^\top = A$, the cosine similarity then has the conditional posterior

$$p_{\text{cossim}}(s \mid \alpha \rho^\top = A) \propto \quad (42)$$

$$\int_{\mathcal{W}_{x,y,s}} \exp(-\lambda (\|\mathbf{P} \rho_{\text{ref}}\|^2 + \|\mathbf{P}^{-\top} \alpha_{\text{ref}}\|^2)) d\mathbf{P}$$

The marginal posterior of the cosine similarity then becomes

$$p_{\text{cossim}}(s) = \quad (43)$$

$$\int p_{\text{cossim}}(s \mid \alpha \rho^\top = A) p(\alpha \rho^\top = A) d(\alpha \rho^\top)$$

i.e. the marginalizing the conditional distribution $s \mid \mathbf{P}, \alpha \rho^\top$ over the distributions of $\mathbf{P}$ and $\alpha \rho^\top$. This is in contrast to a fixed model, where the marginalization only happens over $\alpha \rho^\top$ given one specific value of $\mathbf{P}$. This leads to an inflated variance of the cosine similarity of the unfixed model in comparison to the fixed.

A notable special case is when asymptotically the likelihood concentrates around the MAP estimate. Then $p(\alpha \rho^\top = A)$ approaches a δ distribution. In that case, the posterior of the cosine similarities of an unrestricted model still retains the conditional variance of $p_{\text{cossim}}(s \mid \alpha \rho^\top = A)$, while the rotation fixed model makes the cosine similarity converge. In other words, the unrestricted model will have posterior variance of the cosine similarities even in the limit $N \rightarrow \infty$.

B Additional results and illustrations

In higher dimensions than $K = 2$, the hypersphere resulting from the orthogonal symmetries $\mathcal{O}$ is not a circle, but a sphere or a hypersphere instead. Hence, the 2D projection of the

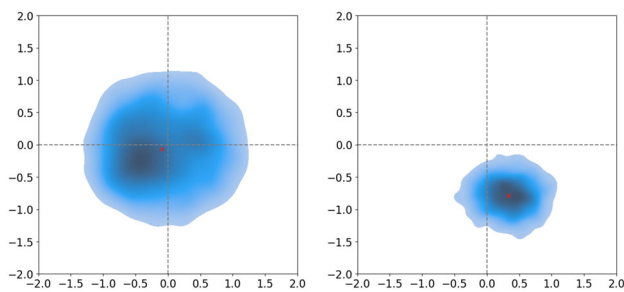


Fig. 9 $\rho_{4,1}$ and $\rho_{4,2}$ for $K = 5$, $V = 100$, $N = 50,000$. The left figure demonstrates the rotational symmetry which is eliminated by fixing α_I to a MAP estimate, shown in the right figure

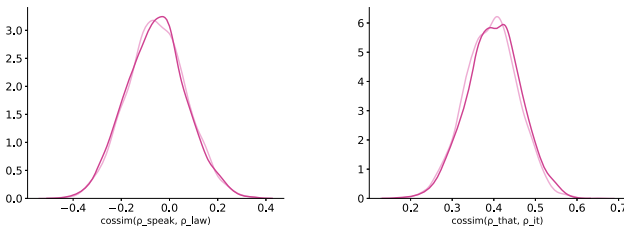


Fig. 10 Posterior distribution of the cosine similarity between words. In dark, a Gibbs sampler with the last K context vectors fixed, and in light, randomly selected K words fixed to a MAP estimate. Left: 'speak' and 'law', right: 'that' and 'it'. $K = 50$, $N = 500,000$, $K = 50$

circular shell is closer to standard Gaussian than a torus, as in the $K = 2$ case. The symmetries for $K = 5$ are illustrated in Fig. 9.

To check the sensitivity of the results to the rotational fix, we calculated posteriors for the cosine similarity of some word vector pairs. In one run, we fixed the last K context vectors as we do in the article. In another run, K words were selected randomly from the vocabulary and their corresponding context vectors were fixed to the values of the MAP estimate. The results are shown in Fig. 10. Visually, the resulting posteriors are very similar. Moreover, the $\hat{R}$ measures were very close to 1: $\hat{R} = 1.0012$ for both inspected word pairs, which is an indication of no difference between the two approaches.

Figure 11 shows RMSE_{co} for different simulated datasets. The RMSE_{co} is plotted by N/V for the different data sizes, using the Gibbs sampler. Averaged over 10, uniform simulation.

Figure 12 shows the convergence of the different estimation methods in terms of RMSE_{co} for the Zipf's law simulation.

The posterior of the cosine similarities, using $K = 100$, can be seen in Fig. 13.

B. 1 Metropolis-adjusted Langevin dynamics (MALA)

As another baseline, we ran the Metropolis-adjusted Langevin Dynamics (MALA) sampler. For each iteration, MALA sam-

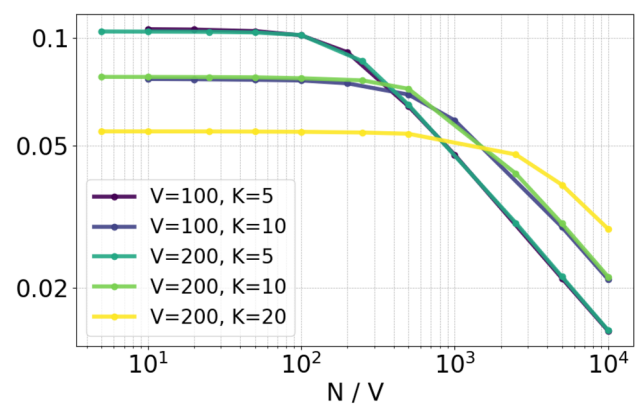


Fig. 11 Log-log plotted RMSE_{co} by N/V for the different data sizes, using the Gibbs sampler. Averaged over 10 simulated datasets, uniform simulation

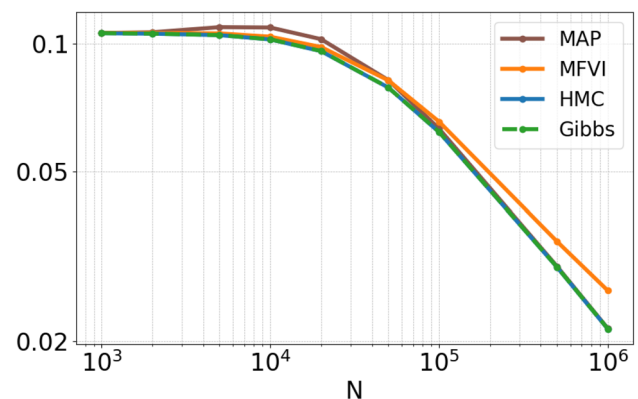


Fig. 12 Log-log plotted RMSE_{co} for $K = 5$ and $V = 100$. Averaged over 10 simulated datasets. Zipf's law simulation

ples a momentum in a similar fashion as HMC, and then takes one leapfrog step combining the momentum and the gradient. It can be viewed as a special case of HMC where the number of leapfrog steps is set to 1 instead of adapting it at each iteration like NUTS does.

Due to the theoretical connection, we implemented the MALA algorithm using Stan by restricting `max_treedepth` to 1. The same Stan code and preprocessing was used as for NUTS-HMC.

The algorithm was tested on the simulated data sets. Due to a lower number of gradient evaluations, it ran substantially faster than NUTS-HMC. However, the samples it generated were very highly autocorrelated. This meant that the effective sample size was extremely low. For example, on the simulated data with $V = 200$, $K = 20$, the median ESS for 1000 samples was 4.67, being under 1.5 for some parameters. The low effective sample sizes make the MALA sampler unusable in these settings. Our observations are in line with previous theoretical work, which finds that MALA requires at least $O(d^{1/3})$ gradient evaluations for mixing, in contrast with $O(d^{1/4})$ for HMC (Chen et al. 2023).

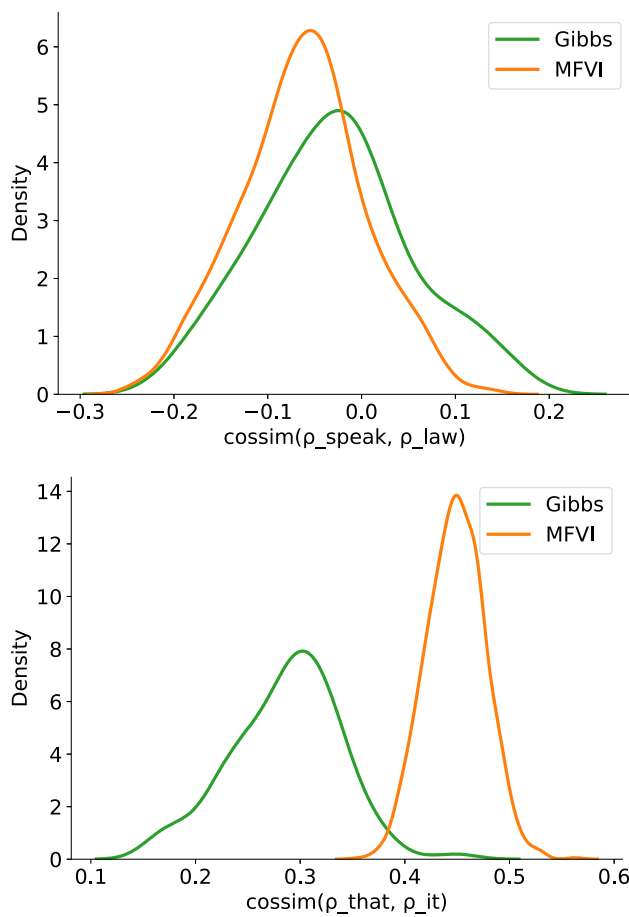


Fig. 13 Posterior distribution of the cosine similarity between words in the US Congress data. Above: 'speak' and 'law', below: 'that' and 'it'. SGNS, $K = 100$, $N = 500,000$. Gibbs sampler in green, HMC in blue and MFVI in orange

C Derivation of the Gibbs sampler

We consider binary observations $x_i \in \{0, 1\}$, representing positive negative context label, with associated linear predictors η_i . The likelihood is

$$x_i \mid \eta_i \sim \text{Bernoulli}(\sigma(\eta_i)). \quad (44)$$

Following Polson et al. (2013), the Bernoulli likelihood admits the representation

$$\frac{\exp(x_i \eta_i)}{1 + \exp(\eta_i)} \propto \exp(\kappa_i \eta_i) \int \exp\left(-\frac{1}{2} \omega_i \eta_i^2\right) p(\omega_i \mid 1, 0) d\omega_i, \quad (45)$$

where $\kappa_i := x_i - \frac{1}{2}$ and $\omega_i \sim PG(1, 0)$.

Introducing Polya-Gamma latent variables $\omega = (\omega_1, \dots, \omega_N)$ yields the augmented posterior

$$p(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\omega} \mid x) \propto p(\boldsymbol{\rho}) p(\boldsymbol{\alpha}) \prod_{i=1}^N p(\omega_i \mid 1, 0) \exp\left(\kappa_i \eta_i - \frac{1}{2} \omega_i \eta_i^2\right). \quad (46)$$

Conditioning on $(\boldsymbol{\rho}, \boldsymbol{\alpha})$ gives

$$\omega_i \mid \boldsymbol{\rho}, \boldsymbol{\alpha}, x \sim PG(1, \eta_i). \quad (47)$$

The form of the augmented posterior (46) is shared by both SGNS and CBoW, but they differ in the linear predictor η_i .

C.1 SGNS Gibbs sampler

Each observation $i = 1, \dots, N$ consists of the triplet (w_i, v_i, x_i) . The linear predictor is

$$\eta_i^{\text{SGNS}} := \boldsymbol{\alpha}_{v_i}^\top \boldsymbol{\rho}_{w_i}. \quad (48)$$

Substituting η_i^{SGNS} into (46) yields a standard Polya-Gamma augmented logistic regression model. Conditional on $(\boldsymbol{\alpha}, \boldsymbol{\omega})$, the updates for $\boldsymbol{\rho}$ coincide with those of Bayesian logistic regression with regressors $\boldsymbol{\alpha}_{v_i}$, and conditional on $(\boldsymbol{\rho}, \boldsymbol{\omega})$, the updates for $\boldsymbol{\alpha}$ coincide with those obtained using regressors $\boldsymbol{\rho}_{w_i}$ (Polson et al. 2013).

C.1.1 Conditional distribution of $\boldsymbol{\rho}_w$.

Fix a word type $w \in W$ and let $I_w = \{i : w_i = w\}$. Collect and define,

$$\mathbf{A}_w := \begin{pmatrix} \boldsymbol{\alpha}_{v_{i_1}}^\top \\ \vdots \\ \boldsymbol{\alpha}_{v_{i_{|I_w|}}}^\top \end{pmatrix}, \quad (49)$$

$$\boldsymbol{\Omega}_w := \text{diag}(\omega_{i_1}, \dots, \omega_{i_{|I_w|}}), \quad (50)$$

$$\boldsymbol{\kappa}_w := (\kappa_{i_1}, \dots, \kappa_{i_{|I_w|}})^\top. \quad (51)$$

Assume the prior $\boldsymbol{\rho}_{w,0} \sim \mathcal{N}(\boldsymbol{\mu}_{\rho_w}, \boldsymbol{\Sigma}_{\rho_w})$. The update for $\boldsymbol{\rho}_w$ is then

$$\boldsymbol{\rho}_w \mid \boldsymbol{\alpha}, \boldsymbol{\omega}, x \sim \mathcal{N}(\mathbf{m}_{\rho_w}, \mathbf{S}_{\rho_w}), \quad (52)$$

with

$$\mathbf{S}_{\rho_w} = (\boldsymbol{\Sigma}_{\rho_w}^{-1} + \mathbf{A}_w^\top \boldsymbol{\Omega}_w \mathbf{A}_w)^{-1}, \quad (53)$$

$$\mathbf{m}_{\rho_w} = \mathbf{S}_{\rho_w} (\boldsymbol{\Sigma}_{\rho_w}^{-1} \boldsymbol{\mu}_{\rho_w} + \mathbf{A}_w^\top \boldsymbol{\kappa}_w). \quad (54)$$

C. 1.2 Conditional distribution of α_v .

Similarly to ρ_w , fix a context word type $v \in W$ and let $I_v = \{i : v_i = v\}$ and define,

$$\mathbf{R}_v := \begin{pmatrix} \rho_{w_{i_1}}^\top \\ \vdots \\ \rho_{w_{i_{|I_v|}}}^\top \end{pmatrix}, \quad (55)$$

$$\mathbf{\Omega}_v := \text{diag}(\omega_{i_1}, \dots, \omega_{i_{|I_v|}}), \quad (56)$$

$$\kappa_v := (\kappa_{i_1}, \dots, \kappa_{i_{|I_v|}})^\top. \quad (57)$$

Assume the prior $\alpha_v \sim \mathcal{N}(\mu_{\alpha_v}, \Sigma_{\alpha_v})$, the update for α_v is

$$\alpha_v \mid \rho, \omega, x \sim \mathcal{N}(\mathbf{m}_{\alpha_v}, \mathbf{S}_{\alpha_v}), \quad (58)$$

with

$$\mathbf{S}_{\alpha_v} = (\Sigma_{\alpha_v}^{-1} + \mathbf{R}_v^\top \mathbf{\Omega}_v \mathbf{R}_v)^{-1}, \quad (59)$$

$$\mathbf{m}_{\alpha_v} = \mathbf{S}_{\alpha_v} (\Sigma_{\alpha_v}^{-1} \mu_{\alpha_v} + \mathbf{R}_v^\top \kappa_v). \quad (60)$$

C. 1.3 Aggregated updates

Further the SGNS likelihood can be efficiently expressed in terms of aggregated counts n_{wv}^+ and n_{wv}^- for each word-context pair (w, v) , yielding

$$\log p(\mathcal{D} \mid \rho_w, \alpha) = \sum_{v \in W} \left(n_{wv}^+ \log \sigma(\alpha_v^\top \rho_w) + n_{wv}^- \log \sigma(-\alpha_v^\top \rho_w) \right). \quad (61)$$

This corresponds to $n_{wv} = n_{wv}^+ + n_{wv}^-$ independent Bernoulli observations with linear predictor $\eta_{wv} = \alpha_v^\top \rho_w$ and $\kappa_{wv} = \frac{1}{2}(n_{wv}^+ - n_{wv}^-)$. By the additive property of the Polya-Gamma distribution the n_{wv} latent variables $\omega_i \sim PG(1, 0)$ associated with these observations can be aggregated into a single variable

$$\omega_{wv} \sim PG(n_{wv}, 0), \quad (62)$$

yielding the representation

$$\begin{aligned} p(\mathcal{D} \mid \rho_w, \alpha) \\ \propto \prod_{v \in W} \exp(\kappa_{wv} \alpha_v^\top \rho_w) \\ \mathbb{E}_{\omega_{wv}} \left[\exp \left(-\frac{1}{2} \omega_{wv} (\alpha_v^\top \rho_w)^2 \right) \right]. \end{aligned} \quad (63)$$

Further, for the α and ρ updates we aggregate $\mathbf{A}_w$ and $\mathbf{R}_v$. For ρ , fix a context word $v \in W$ and define the set of target

words

$$V_w := \{v \in W : n_{wv} > 0\}.$$

we get,

$$\mathbf{A}_w = \begin{pmatrix} \alpha_{v_1}^\top \\ \vdots \\ \alpha_{v_{|V_w|}}^\top \end{pmatrix}, \quad (64)$$

$$\mathbf{\Omega}_w = \text{diag}(\omega_{wv})_{v \in V_w}, \quad (65)$$

$$\kappa_w = (\kappa_{wv})_{v \in V_w}. \quad (66)$$

Similarly for α ,

$$W_v := \{w \in W : n_{wv} > 0\}.$$

The corresponding aggregated quantities are

$$\mathbf{R}_v = \begin{pmatrix} \rho_{w_1}^\top \\ \vdots \\ \rho_{w_{|W_v|}}^\top \end{pmatrix}, \quad (67)$$

$$\mathbf{\Omega}_v = \text{diag}(\omega_{wv})_{w \in W_v}, \quad (68)$$

$$\kappa_v = (\kappa_{wv})_{w \in W_v}. \quad (69)$$

Inserting these expressions into the associated posterior conditionals gives the aggregated Gibbs updates.

C. 2 CBoW Gibbs sampler

Each observation now consists of a context $C_i \subset W$, instead of a context word. The aggregated context embedding and linear predictor is

$$\alpha_{C_i} = \sum_{v \in C_i} \alpha_v, \quad \eta_i^{\text{CBoW}} := \alpha_{C_i}^\top \rho_{w_i}. \quad (70)$$

We substitute η_i^{CBoW} into (46) to get the augmented posterior for CBoW.

C. 2.1 Conditional distribution of ρ_w .

Fix $w \in W$ and let I_w , $\mathbf{\Omega}_w$ and κ_w be the same as earlier for SGNS. Define,

$$\mathbf{Q}_w := \begin{pmatrix} \alpha_{C_{i_1}}^\top \\ \vdots \\ \alpha_{C_{i_{|I_w|}}}^\top \end{pmatrix}, \quad (71)$$

Assuming $\rho_{w,0} \sim \mathcal{N}(\mu_{\rho_w}, \Sigma_{\rho_w})$, the update is similar to SGNS ρ with $\mathbf{Q}_w$ defining the regressors, we get

$$\rho_w \mid \alpha, \omega, x \sim \mathcal{N}(\mathbf{m}_{\rho_w}, S_{\rho_w}), \quad (72)$$

with

$$S_{\rho_w} = (\Sigma_{\rho_w}^{-1} + \mathbf{Q}_w^\top \Omega_w \mathbf{Q}_w)^{-1}, \quad (73)$$

$$\mathbf{m}_{\rho_w} = S_{\rho_w} (\Sigma_{\rho_w}^{-1} (\mu_{\rho_w} + \mathbf{Q}_w^\top \kappa_w)). \quad (74)$$

C. 2.2 Conditional distribution of α_v .

Fix $v \in W$. For each $i \in I_v$ define

$$\alpha_{C_i}^{(-v)} := \sum_{u \in C_i \setminus \{v\}} \alpha_u, \quad (75)$$

$$\delta_i^{(v)} := (\alpha_{C_i}^{(-v)})^\top \rho_{w_i}. \quad (76)$$

Let $\mathbf{R}_v, \Omega_v, \kappa_v$ be as for SGNS, and further let

$$\delta_v = (\delta_{i_1}^{(v)}, \dots, \delta_{i_{|I_v|}}^{(v)})^\top. \quad (77)$$

Assuming the prior $\alpha_{w,0} \sim \mathcal{N}(\mu_{\alpha_w}, \Sigma_{\alpha_w})$, we form the conditional

$$p(\alpha_v \mid \rho, \omega, \alpha_{-v}, x) \propto p(\alpha_v) \prod_{i \in I_v} \exp\left(\kappa_i \eta_i^{\text{CBoW}} - \frac{1}{2} \omega_i (\eta_i^{\text{CBoW}})^2\right), \quad (78)$$

where $\eta_i^{\text{CBoW}} = \rho_{w_i}^\top \alpha_v + \delta_i^{(v)}$ for $i \in I_v$ and the remaining factors in (46) are constant with respect to α_v . Writing the terms for $i \in I_v$ in stacked form using $(\mathbf{R}_v, \Omega_v, \kappa_v, \delta_v)$ gives the α_v -dependent part of the log-density, up to constants,

$$-\frac{1}{2} (\alpha_v - \mu_{\alpha_v})^\top \Sigma_{\alpha_v}^{-1} (\alpha_v - \mu_{\alpha_v}) + \kappa_v^\top (\mathbf{R} \alpha_v + \delta_v) - \frac{1}{2} (\mathbf{R} \alpha_v + \delta_v)^\top \Omega_v (\mathbf{R} \alpha_v + \delta_v). \quad (79)$$

Expanding the quadratic term and collecting terms in α_v gives

$$-\frac{1}{2} \alpha_v^\top (\Sigma_{\alpha_v}^{-1} + \mathbf{R}_v^\top \Omega_v \mathbf{R}_v) \alpha_v + \alpha_v^\top \left(\Sigma_{\alpha_v}^{-1} \mu_{\alpha_v} + \mathbf{R}_v^\top (\kappa_v - \Omega_v \delta_v) \right). \quad (80)$$

Completing the square yields

$$\alpha_v \mid \rho, \omega, x \sim \mathcal{N}(\mathbf{m}_{\alpha_v}, S_{\alpha_v}), \quad (81)$$

with

$$S_{\alpha_v} = (\Sigma_{\alpha_v}^{-1} + \mathbf{R}_v^\top \Omega_v \mathbf{R}_v)^{-1}, \quad (82)$$

$$\mathbf{m}_{\alpha_v} = S_{\alpha_v} \left(\Sigma_{\alpha_v}^{-1} \mu_{\alpha_v} + \mathbf{R}_v^\top (\kappa_v - \Omega_v \delta_v) \right). \quad (83)$$

Note S_{α_v} only depends on the prior covariance Σ_{α_v} , ρ and ω , which means it can be calculated, inverted and decomposed independently and potentially in parallel for each α_v .

D Computational complexity

Let S be the posterior variance of either ρ or α .

D. 1 SGNS

The Gibbs sampler consists of four steps which are not negligible in terms of computational cost. The sampling cost per word can be expressed in terms of the dimensionality K , number of unique co-occurring words n_u .

1. Polya-Gamma sampling: $O(n_u)$ (CPU)
2. Calculating $S_{(\cdot)}^{-1}$ via matmul: $O(n_u \cdot K^2)$
3. Inverting $S_{(\cdot)}^{-1}$: $O(K^3)$
4. Cholesky decomposition $S_{(\cdot)} = LL^\top$: $O(S \cdot K^3)$

While step 3. can be skipped if triangular solves are used instead of , this has no effect on the asymptotic complexity. To get the full complexity, most of these are just multiplied by the vocabulary size V . However, n_u depends on the specific word. However, it occurs as a linear term, so $O(n_u \cdot V)$ evaluates to $O(N_u)$, where N_u is the total number of unique word pairs in the data.

Thus, the total asymptotic complexity can be dominated by one of the following three factors

1. Polya-Gamma random variable sampling: $O(N_u)$ (when the CPU dominates)
2. Calculating the precision matrix S^{-1} : $O(N_u \cdot K^2)$
3. Inverting and decomposing the precision matrix: $O(V \cdot K^3)$

In practice, the `polygamma` package seems to be fast enough that the total complexity is

$$O(N_u K^2 + V K^3) \quad (84)$$

Note that by definition, $N_u \leq V^2$.

D. 2 CBoW

The Gibbs sampler consists of four steps which are not negligible in terms of computational cost. The sampling cost per word can be expressed in terms of the dimensionality K ,

number of unique co-occurring words N , and the number of co-occurring words in each batch n_b .

1. Polya-Gamma sampling: $O(N)$ (CPU)
2. Calculating $S_{(\cdot)}^{-1}$ via matmul: $O(n \cdot K^2 + n_b \cdot K^2 B^2)$
3. Cholesky decomposition $LL^T = S_{(\cdot)}^{-1}$: $O(K^3 B^3)$

To get the full complexity, most of these are just multiplied by the vocabulary size V . However, n_b depends on the batch. However, it occurs as a linear term, so $O(n_b \cdot K^2 B^2)$ evaluates to $O(N_b K^2 B^2)$, where N_b is the total number of co-occurrences of words in the same batch.

Thus, the total asymptotic complexity can be dominated by one of the following three factors

1. Polya-Gamma random variable sampling: $O(N)$ (when the CPU dominates)
2. Calculating the precision matrix block-diagonal $S_{(\cdot)}^{-1}$: $O(N \cdot K^2)$
3. Calculating the precision matrix off-diagonal $S_{(\cdot)}^{-1}$: $O(N_b \cdot K^2 B^2)$
4. Cholesky decomposing the precision matrix: $O(V/B \cdot K^3 B^3) = O(V K^3 B^2)$

In practice, the `polyagamma` package seems to be fast enough that the total complexity is

$$O(NK^2 + N_b B^2 K^2 + V K^3 B^3) \quad (85)$$

In the specific case where $B = 1$, this becomes

$$O(NK^2 + V K^3) \quad (86)$$

i.e. roughly the same as SGNS for small N , and scaling similarly as SGNS but in terms of N and not the number of unique word occurrences N_u .

E MovieLens preprocessing

We use the MovieLens 100K dataset. User ratings on a five-point scale are binarized by setting a threshold at three, such that ratings of three, four, or five are treated as positive and ratings below three as negative. The vocabulary is defined by the movie titles. To control dataset size, we keep only the 600 movies with the largest number of ratings and limit each user to at most 25 rated movies and at least two ratings.

E. 1 SGNS MovieLens

We construct pairwise word-context observations from user co-ratings. For each user, movies rated with the same binary

rating label are treated as co-occurring. All pairs of such movies are included as data points in both directions, yielding word-context pairs (w_i, v_i) with a positive co-occurrence label $x_i = 1$.

Negative samples are generated following the standard negative sampling procedure by sampling the target word from the empirical distribution of movies.

Each observation therefore consists of a target movie, a context movie, and a binary co-occurrence label, (w_i, v_i, x_i) .

E. 2 CBoW MovieLens

For the CBoW model, user ratings are first sorted by time stamp. For each rated movie we construct a context consisting of the nearest positively rated movies in time. If the target movie is positively rated, it is excluded from the set of context candidates; if it is negatively rated, the context is still formed only from positively rated movies but the resulting observation is assigned co-occurrence label $x = 0$. The latter case is introduced to ensure that we also include movies negatively rated, which are much fewer.

Negative samples are generated following the standard negative sampling procedure by sampling the target word from the empirical distribution of movies.

Each observation now consists of a target movie, a set of context movies, and the co-occurrence label, (w_i, C_i, x_i) .

F Word association analyses

Using the US Congress data, we use the posterior distribution of the association score to draw conclusions about a word w 's association between two contrasting attribute groups G^+ and G^- . The posterior is computed from the Gibbs sampler ($N = 10,000,000$, $K = 100$). We examine three contrasts.

Military: male vs. female.

$w = \text{military}$, $G^+ = \{\text{man, men, male, he, him, his, father, son, brother, husband, boy, gentleman}\}$, $G^- = \{\text{woman, women, she, her, herself, mother, daughter, sister, wife, girl, gentlewoman}\}$.

Democrat: positive vs. negative.

$w = \text{democrat}$, $G^+ = \{\text{good, great, best, success, strong, honor, fair, safe, freedom, respect, hope}\}$, $G^- = \{\text{bad, worst, violent, violence, threat, danger, fail, corrupt, fraud, terror, war}\}$.

Medicaid: welfare vs. wealth.

$w = \text{medicaid}$, $G^+ = \{\text{welfare, poverty, poor, low-income, afdc, food, benefit, unemployment}\}$, $G^- = \{\text{rich, wealth, wealthy, profit, bank, investor, corporate, market}\}$.

We found, for example, that the posterior probability that *military* is more associated with *male*-words than *female*-words is 99%, as shown in Fig. 8.

Acknowledgements The research was partially funded by the Swedish Research Council through grant 2022-03381, as well as by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Author contributions Deriving the sampler and formulating the proofs, conducting the experiments and simulations, and writing the article and appendices were carried out by V.Y. and I.B. The literature review was performed by V.Y., I.B., and M.M. The manuscript was reviewed and suggestions for improvement were provided by M.M., while J.J. checked the proofs and derivations for correctness. Overall decision-making and supervision of the project were led by M.M. and J.J. All authors read and approved the final manuscript.

Funding Open access funding provided by Uppsala University.

Data Availability The data that support the findings of this study are openly available in the MovieLens data repository at <https://grouplens.org/datasets/movielens/>, as well as the Congressional Record for the 43rd–114th Congresses: Parsed Speeches and Phrase Counts data at https://data.stanford.edu/congress_text.

Code availability All code necessary to replicate the results is provided on GitHub at <https://anonymous.4open.science/r/embedding-uncertainty-estimation-46ED/>.

Declarations

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abramson, C.M., Li, Z., Prendergast, T., Sánchez-Jankowski, M.: Inequality in the origins and experiences of pain: What "big (qualitative) data" reveal about social suffering in the united states. *RSF: The Russell Sage Foundation. J. Soc. Sci.* **10**(5), 34–65 (2024)
- Aceves, P., Evans, J.A.: Human languages with greater information density have higher communication speed but lower conversation breadth. *Nat. Hum. Behav.* **8**(4), 644–656 (2024)
- Abmann, C., Boysen-Hogrefe, J., Pape, M.: Bayesian analysis of static and dynamic factor models: an ex-post approach towards the rotation problem. *J. Econom.* **192**(1), 190–206 (2016)
- Bamler, R., Mandt, S.: Dynamic word embeddings. In: International conference on Machine learning, pp. 380–389. (2017). (PMLR)
- Bernardo, J.M., Smith, A.F., Berliner, M.: Bayesian theory, Vol. 586. Wiley Online Library (1994)
- Bickel, P.J., Doksum, K.A.: Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1. Chapman and Hall/CRC (2015)
- Bodell, M.H., Arvidsson, M., Magnusson, M.: Interpretable word embeddings via informative priors, (2019). [arXiv:1909.01459](https://arxiv.org/abs/1909.01459) arXiv preprint
- Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. *Science* **356**(6334), 183–186 (2017)
- Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., Jurafsky, D.: Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proc. Natl. Acad. Sci.* **119**(31), e2120510119 (2022)
- Chen, Y., Gatmiry, K., Jiang, M.: When does metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm? (2023). [arXiv:2304.04724](https://arxiv.org/abs/2304.04724) arXiv preprint
- Durrheim, K., Schuld, M., Mafunda, M., Mazibuko, S.: Using word embeddings to investigate cultural biases. *Br. J. Soc. Psychol.* **62**(1), 617–629 (2023)
- Eisenstein, J.: Natural Language Processing. Jacob Eisenstein, 507 (2018)
- Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci.* **115**(16), E3635–E3644 (2018)
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis. CRC Press (2013)
- Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992)
- Gentzkow, M., Shapiro, J.M., Taddy, M.: Congressional Record for the 43rd–114th Congresses: Parsed Speeches and Phrase counts. In <https://data.stanford.edu/> (2018)
- Geweke, J., Zhou, G.: Measuring the pricing error of the arbitrage pricing theory. *Rev. Financ. Stud.* **9**(2), 557–587 (1996)
- Harper, F.M., Konstan, J.A.: The movielens datasets: history and context. *Acm Trans. Interact. Intell. Syst. (tiis)* **5**(4), 1–19 (2015)
- Klami, A.: Polya-gamma augmentations for factor models. In: Asian Conference on Machine Learning, pp. 112–128. (2015). (PMLR)
- Kleijn, B.J., A.W.: Van der Vaart. 2012. The bernstein-von-mises theorem under misspecification
- Kozłowski, A.C., Taddy, M., Evans, J.A.: The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**(5), 905–949 (2019)
- Liu, Y., Zhao, L., Liu, G., Lu, X., Gao, P., Li, X.L., Jin, Z.: Dynamic bayesian logistic matrix factorization for recommendation with implicit feedback. In *IJCAI* **18**, 3463–3469 (2018)
- Ma, Y., Peng, Y., Zhu, Y.: Word embeddings track social group changes across 70 years in china, (2025). [arXiv:2504.12327](https://arxiv.org/abs/2504.12327) arXiv preprint
- Mandelbrot, B., et al.: An informational theory of the statistical structure of language. *Commun. Theory* **84**, 486–502 (1953)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems p. 26. (2013)
- Mu, C., Yang, G., Zheng, Y.: Revisiting skip-gram negative sampling model with rectification. In *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 1*, pp. 485–497. Springer (2019)
- Myrick, R.: Do external threats unite or divide? security crises, rivalries, and polarization in american foreign policy. *Int. Organ.* **75**(4), 921–958 (2021)
- Neal, R.M., et al.: MCMC Using Hamiltonian Dynamics. In: Handbook of Markov Chain Monte Carlo, Vol. 2(11), p. 2. (2011)
- Papastamoulis, P., Ntzoufras, I.: On the identifiability of bayesian factor analytic models. *Stat. Comput.* **32**(2), 23 (2022)

- Peeters, C.F.: Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika* **77**(2), 288–292 (2012)
- Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. (2014)
- Piantadosi, S.T.: Zipf’s word frequency law in natural language: a critical review and future directions. *Psychon. Bull. Rev.* **21**, 1112–1130 (2014)
- Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using pólya-gamma latent variables. *J. Am. Stat. Assoc.* **108**(504), 1339–1349 (2013)
- Roberts, G.O., Sahu, S.K.: Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **59**(2), 291–317 (1997)
- Rodman, E.: A timely intervention: tracking the changing meanings of political concepts with word vectors. *Polit. Anal.* **28**(1), 87–111 (2020). <https://doi.org/10.1017/pan.2019.23>
- Rudolph, M., Blei, D.: Dynamic Bernoulli embeddings for language evolution, (2017). [arXiv:1703.08052](https://arxiv.org/abs/1703.08052) arXiv preprint
- Rudolph, M., Ruiz, F., Athey, S., Blei, D.: Structured embedding models for grouped data. *Advances in neural information processing systems* 30, (2017)
- Rudolph, M., Ruiz, F., Mandt, S., Blei, D.: Exponential family embeddings. *Adv. Neural. Inf. Process. Syst.* 29, (2016)
- Wang, Y., Blei, D.M.: Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc.*, (2018)
- Weiss, M., Zoorob, M.: Political frames of public health crises: discussing the opioid epidemic in the us congress. *Soc. Sci. Med.* **281**, 114087 (2021)
- Yrjänäinen, V., Magnusson, M.: Probabilistic embeddings with laplacian graph priors, (2022). [arXiv:2204.01846](https://arxiv.org/abs/2204.01846) arXiv preprint
- Zipf, G.K.: *The Psychobiology of Language*, (1935)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.