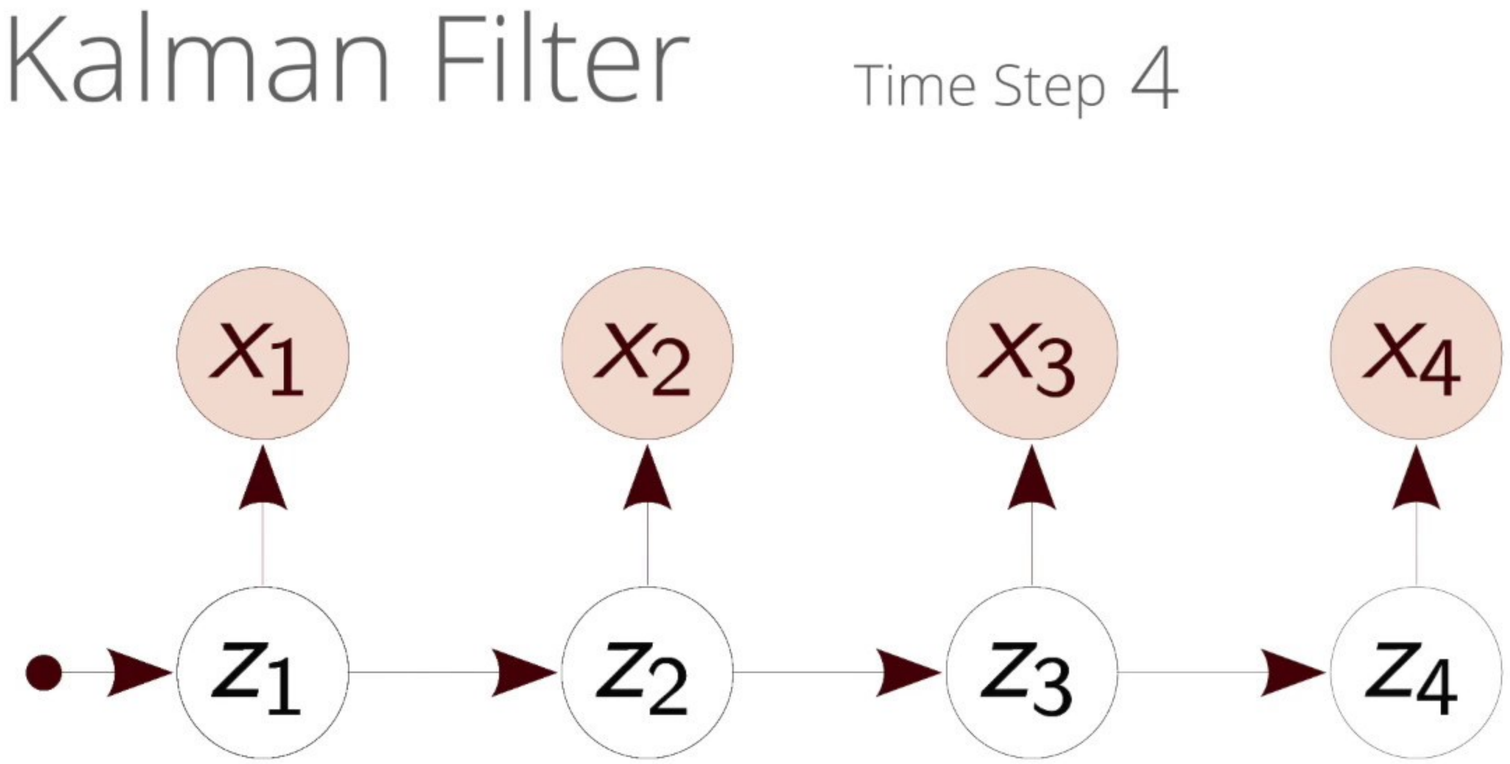
Dynamic Word Embeddings

Robert Bamler Stephan Fandt

Ming Feng Wan CSC2611

Introduction

A cool clip the authors made (00:20 - 01:50)



Introduction

Problem with current model

- Current approaches to learning word embeddings in a dynamic context rely on grouping the data into time bins and training the embeddings separately
- This approach, however, raises three fundamental problems.

Introduction

Problem with current model

1. word embedding models are non-convex, training them twice on the same data will lead to different results.

   - Thus, embedding vectors at successive times can only be approximately related to each other, and only if the embedding dimension is large (Hamilton et al., 2016).

2. dividing a corpus into separate time bins may lead to training sets that are too small to train a word embedding model.

   - Runs the risk of overfitting

3. due to the finite corpus size the learned word embedding vectors are subject to random noise.

   - difficult to disambiguate this noise from systematic semantic drifts between subsequent times

Introduction

Circumvent these problems by introducing a dynamic word embedding model

- Derive a probabilistic state space model where word and context embeddings evolve in time according to a diffusion process.
    - This leads to continuous embedding trajectories, smoothes out noise in the word-context statistics, and allows us to share information across all times.
- Propose two scalable black-box variational inference algorithms for filtering and smoothing. These algorithms find word embeddings that generalize better to held-out data.
- Analyze three massive text corpora that span over long periods of time.
    - Automatically find the words whose meaning changes the most, with smooth word embedding trajectories.

Introduction

“Smooth”

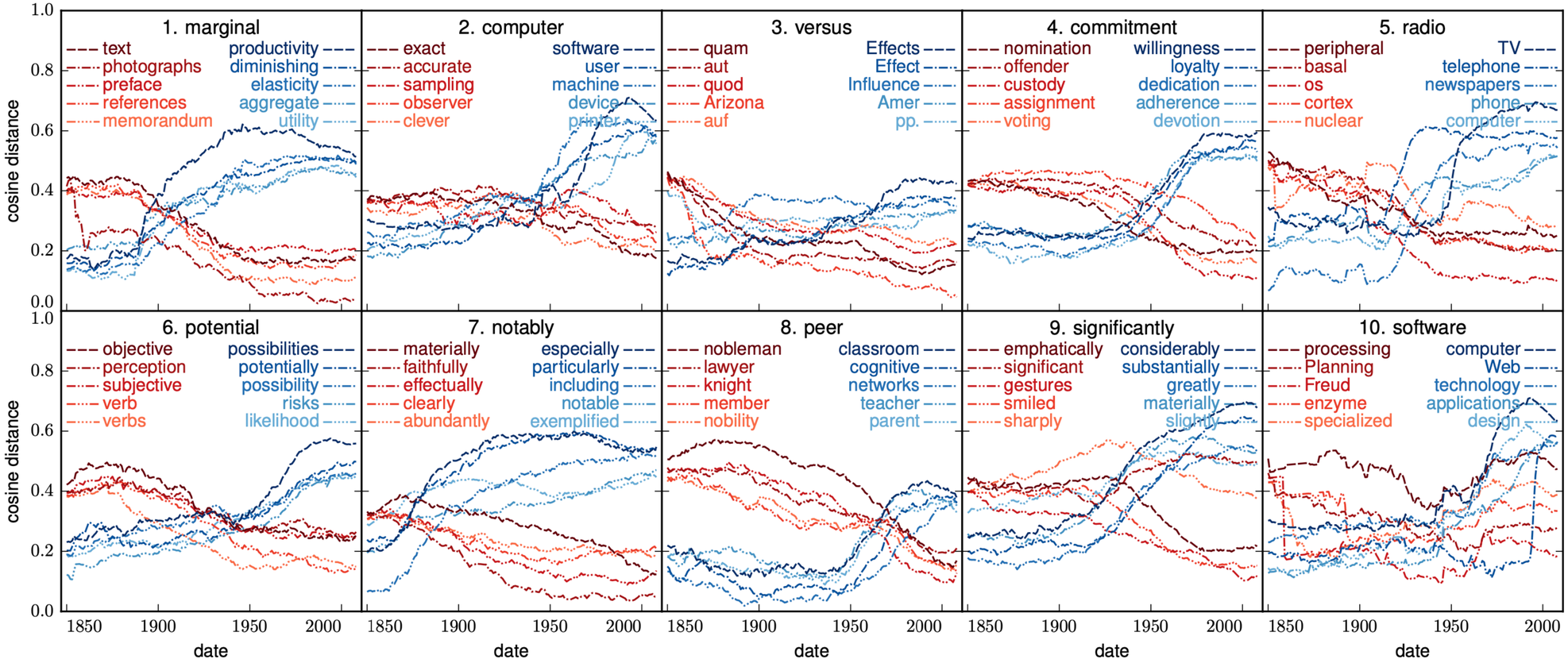


Figure 1. Evolution of the 10 words that changed the most in cosine distance from 1850 to 2008 on Google books, using skip-gram filtering (proposed). Red (blue) curves correspond to the five closest words at the beginning (end) of the time span, respectively.

Model
(Skip related work)

- Dynamic skip-gram is a probabilistic model which combines a Bayesian version of the skip-gram model with a latent time series.
-

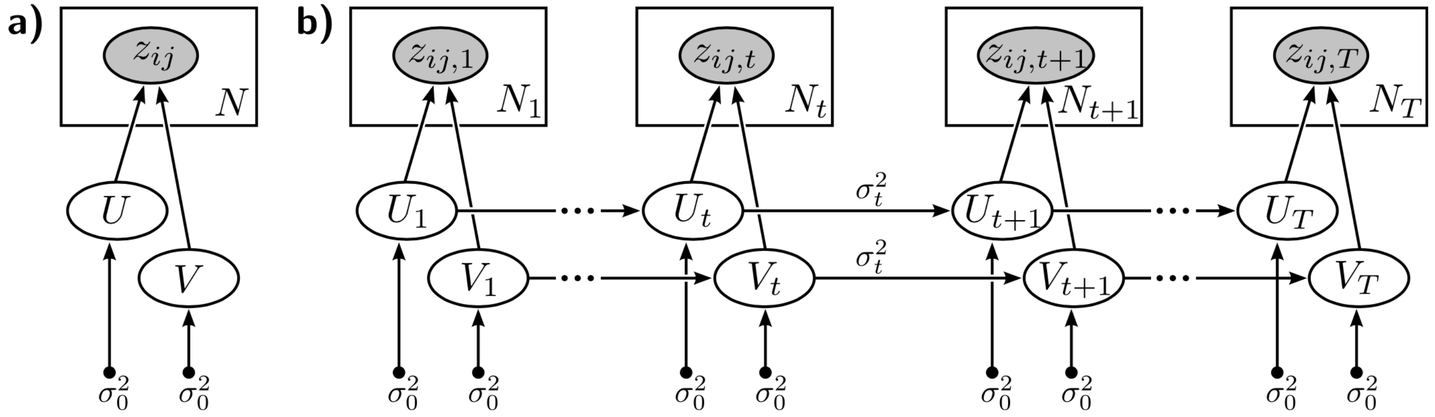
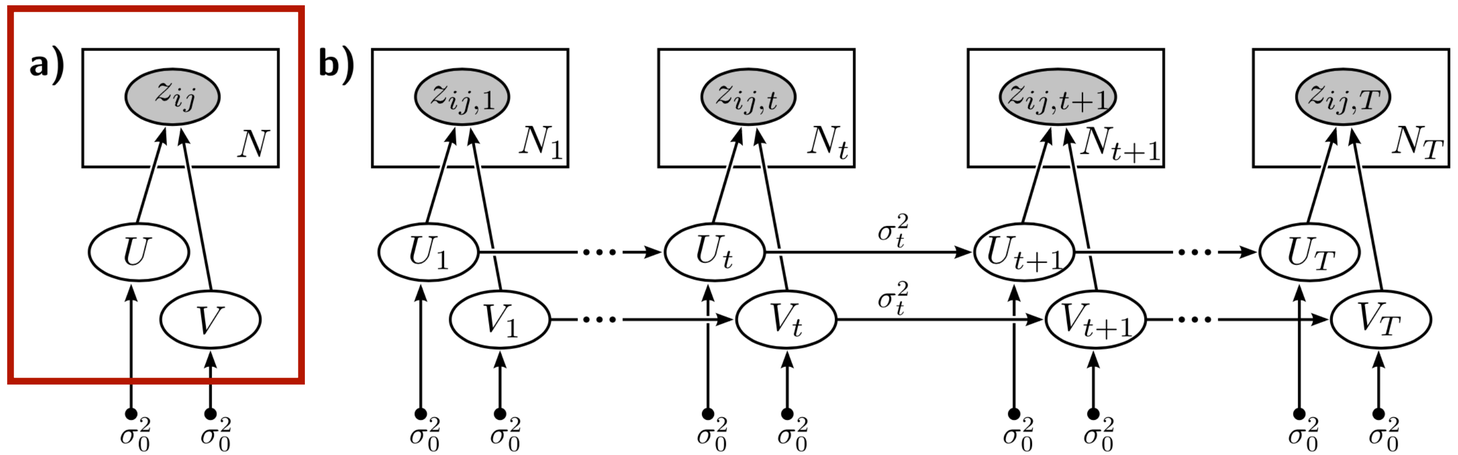


Figure 2. a) Bayesian skip-gram model (Barkan, 2017). b) The dynamic skip-gram model (proposed) connects T copies of the Bayesian skip-gram model via a latent time series prior on the embeddings.

Model

(Non-Bayesian) Skip-Gram Model



- For each pair of words i, j in the vocabulary, the model assigns probabilities that word i appears in the context of word j.

- The generative model assumes that many word-word pairs (i, j) are uniformly drawn from the vocabulary and tested for being a word-context pair

Model

Some notations for (Non-Bayesian) Skip-Gram Model

- $u_i, v_i \in R^d$ for each word i in the vocabulary, where d is the embedding dimension.
  - u_i is the word embedding vector
  - v_i is the context embedding vector.
- sigmoid function $\sigma(x) = 1/(1 + e^{-x})$.
- Let $z_{ij} \in \{0,1\}$ be an indicator variable that denotes a draw from that probability distribution, hence $p(z_{ij} = 1) = \sigma(u_i^\top v_j)$.
  - collect evidence of word-word pairs for which $z_{ij} = 1$
  - n_{ij}^+, the number of times that a word-context pair (i, j) is observed in the corpus.

Model

(Non-Bayesian) Skip-Gram Model

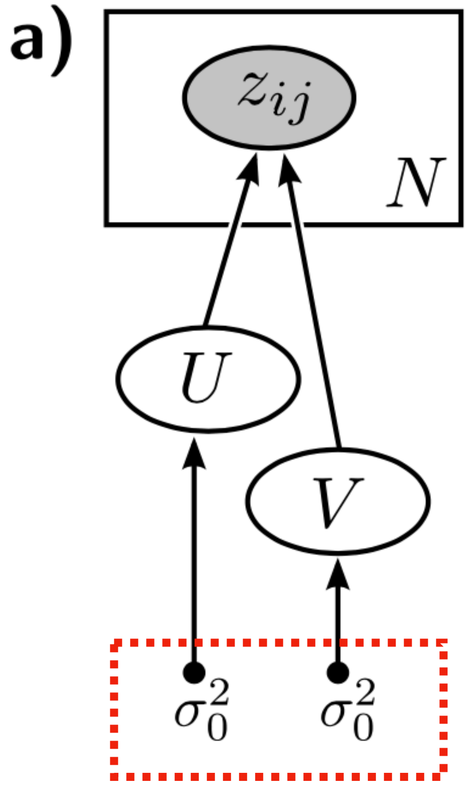
- But, we also need negative sampling, the possibility to reject word-context pairs if $z_{ij} = 0$.
- $n_{ij}^{-} \propto P(i)P(j)^{3/4}$, $P(i)$ is the frequency of word i in the training corpus.
- $U = (u_1, \ldots, u_L) \in R^{d \times L}$ is the matrix of all word embedding vectors, and V is defined analogously for the context vectors

$$p(n^+, n^- | U, V) = \prod_{i,j=1}^{L} \sigma(u_i^\top v_j)^{n_{ij}^+} \sigma(-u_i^\top v_j)^{n_{ij}^-}.$$

Model

(Non-Bayesian & Bayesian) Skip-Gram Model

- $n^{\pm} = (n^{+}, n^{-})$, the combination of both positive and negative examples.
- $$\log p(n^{\pm} \mid U, V) = \sum_{i,j=1}^{L} (n_{ij}^{+} \log \sigma(u_i^{\top} v_j) + n_{ij}^{-} \log \sigma(-u_i^{\top} v_j))$$
- Barkan (2017) gives an approximate Bayesian treatment of the model with Gaussian priors on the embeddings.



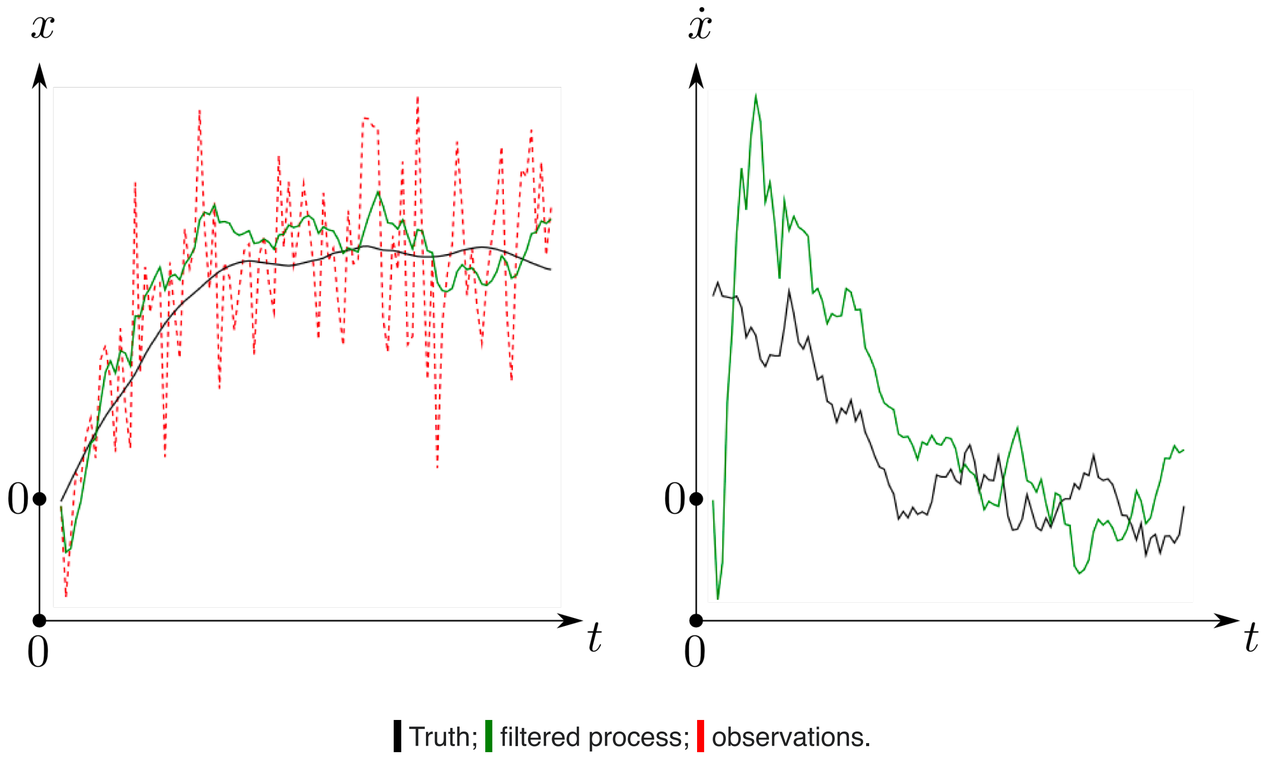
Dynamic Skip-Gram Model

Kalman filter to “smooth”

Consider a truck on frictionless, straight rails.

Initially, the truck is stationary at position 0, but it is buffeted this way and that by random uncontrolled forces.

We measure the position of the truck every Δt seconds, but these measurements are imprecise; we want to maintain a model of the truck's position and velocity.

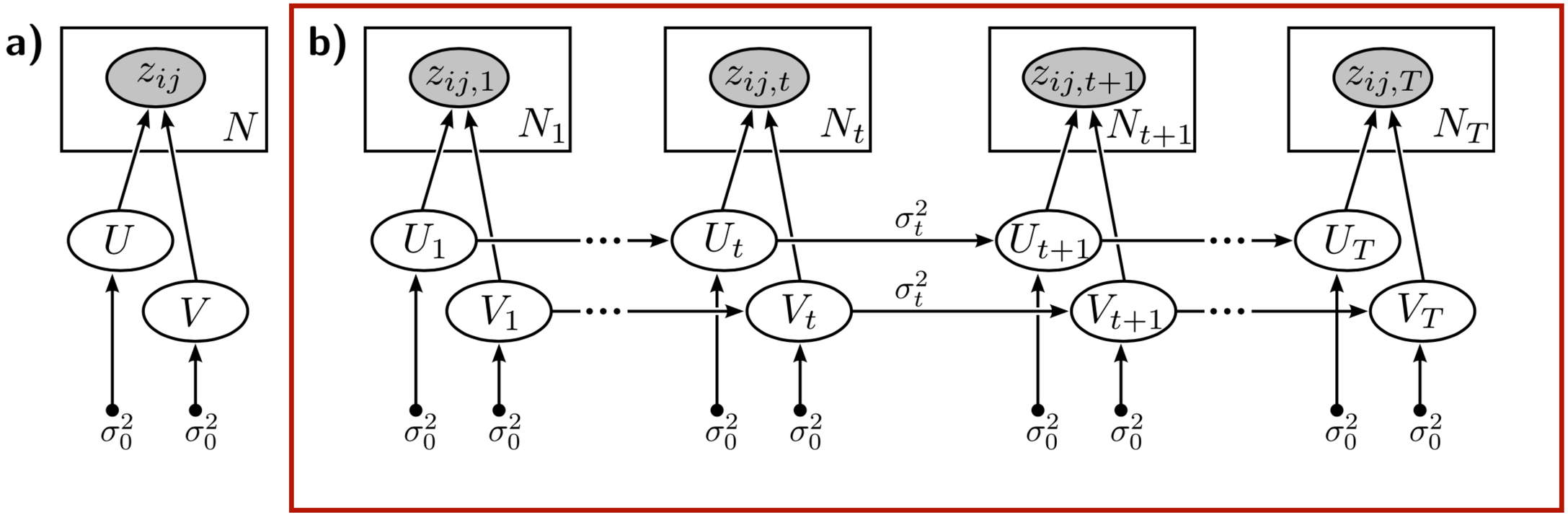


Truth; filtered process; observations.

Source: Wikipedia

Dynamic Skip-Gram Model

Kalman filter as a prior for the time-evolution of the latent embeddings



Dynamic Skip-Gram Model

Kalman filter as a prior for the time-evolution of the latent embeddings

- The variance of the transition kernel $\sigma_t^2 = D(\tau_{t+1} - \tau_t)$, where D is a global diffusion constant and $(\tau_{t+1} - \tau_t)$ is the time between subsequent observations
- $p(U_{t+1} \mid U_t) \propto \mathcal{N}(U_t, \sigma_t^2)\mathcal{N}(0,\sigma_0^2)$. $\mathcal{N}(0,\sigma_0^2)$ is an additional Gaussian that prevents the word embeddings from growing very large. Same for V_t
- $p(U_1 \mid U_0) \equiv \mathcal{N}(0,\sigma_0^2 I)$
- $$p(n^{\pm}, U, V) = \prod_{t=0}^{T-1} p(U_{t+1}|U_t)\, p(V_{t+1}|V_t) \times \prod_{t=1}^{T} \prod_{i,j=1}^{L} p(n_{ij,t}^{\pm}|u_{i,t}, v_{j,t})$$

Inference

Bayesian inference

$$p(U, V|n^{\pm}) = \frac{p(n^{\pm}, U, V)}{\int p(n^{\pm}, U, V)\, dU\, dV}$$

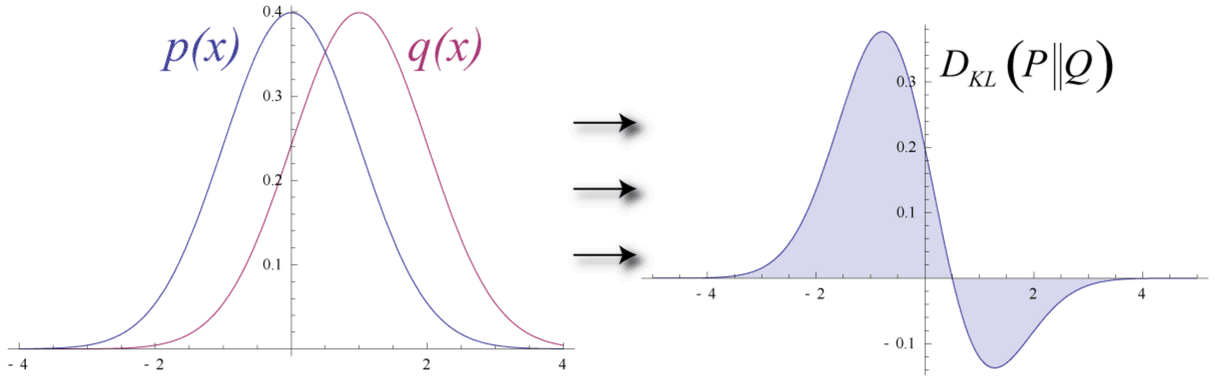
- Problem: normalization is intractable :(
- KL divergence / ELBO to approximate $p(U, V \,|\, n^{\pm})$ with $q_{\lambda}(U, V \,|\, n^{\pm})$
- Note: the paper use $q_{\lambda}(U, V)$, but I prefer the notation $q_{\lambda}(U, V \,|\, n^{\pm})$

Inference

KL Divergence

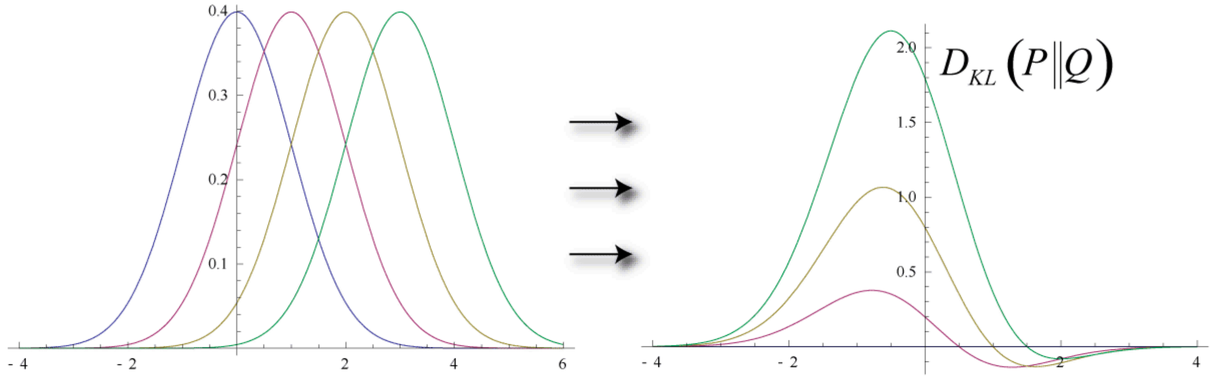
$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

A measure of how one probability distribution is different from a second, reference probability distribution



Original Gaussian PDF's

KL Area to be Integrated



Source: Wikipedia

Inference
ELBO

A distribution Q over unobserved variables Z is optimized as an approximation to the true posterior $P(Z|X)$, given observed data X.

$$D_{\mathrm{KL}}(Q \parallel P(Z|X)) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \left[\log \frac{Q(\mathbf{Z})P(\mathbf{X})}{P(\mathbf{Z},\mathbf{X})} \right]$$

$$D_{\mathrm{KL}}(Q \parallel P) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \left[\log \frac{Q(\mathbf{Z})}{P(\mathbf{Z},\mathbf{X})} + \log P(\mathbf{X}) \right]$$

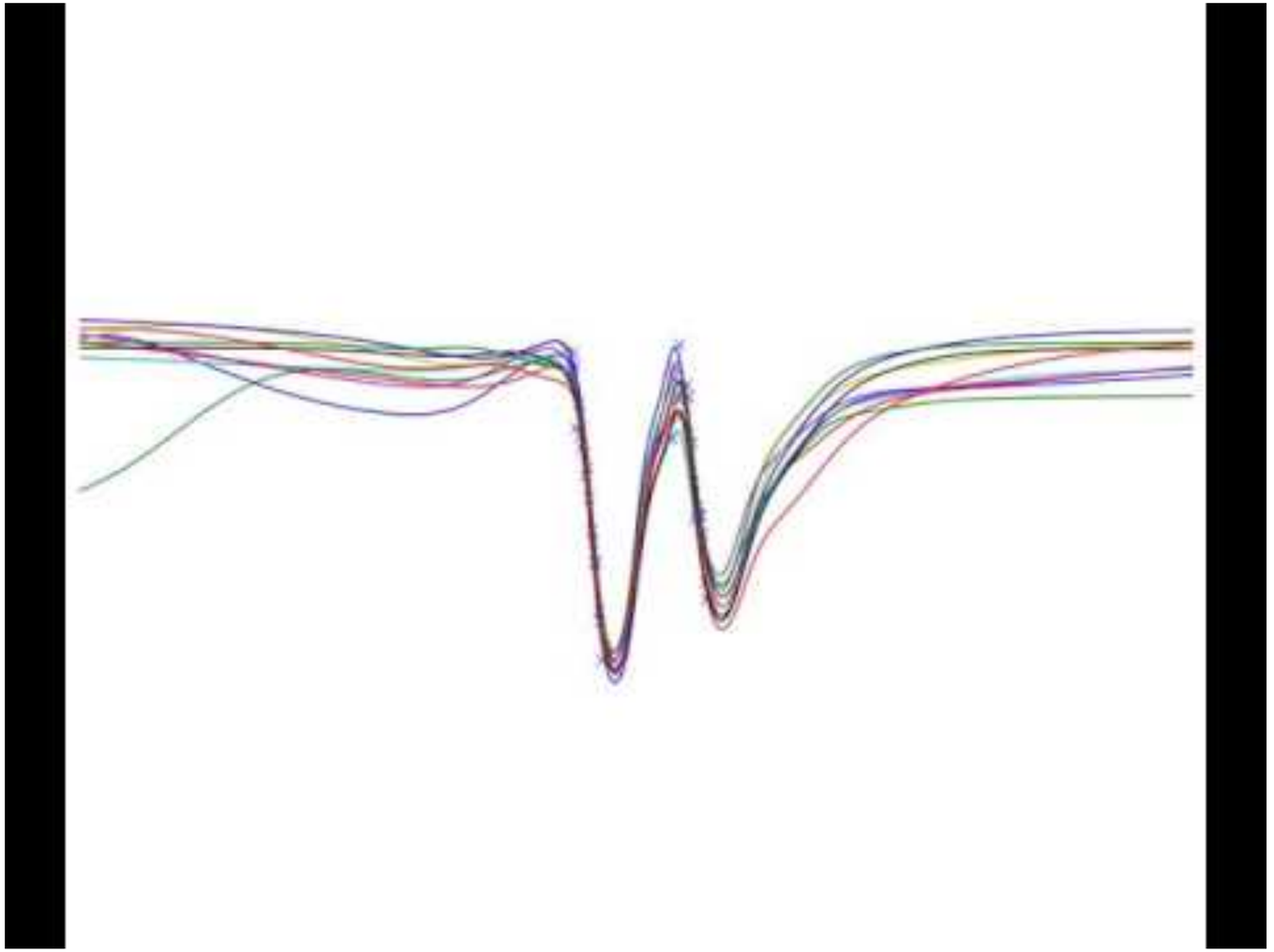
$$D_{\mathrm{KL}}(Q \parallel P) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log Q(\mathbf{Z}) - \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log P(\mathbf{Z},\mathbf{X}) + \log P(\mathbf{X})$$

$$\log P(\mathbf{X}) - D_{\mathrm{KL}}(Q \parallel P) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log P(\mathbf{Z},\mathbf{X}) - \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log Q(\mathbf{Z}) = L(X)$$

Source: Wikipedia

Inference

A cool video demo



Inference

Skip-Gram Filtering

- $p(U_t, V_t \mid n^{\pm}_{1:t}) \propto p(n^{\pm}_t \mid U_t, V_t) p(U_t, V_t \mid n^{\pm}_{1:t-1})$
- Approximate $p(U_{t-1}, V_{t-1} \mid n^{\pm}_{1:t-1})$ from ELBO
- $p(U_t, V_t \mid n^{\pm}_{1:t-1}) \equiv E_{p(U_{t-1}, V_{t-1} \mid n^{\pm}_{1:t-1})}[p(U_t, V_t \mid U_{t-1}, V_{t-1})] \approx E_{q(U_{t-1}, V_{t-1} \mid n^{\pm}_{1:t-1})}[p(U_t, V_t \mid U_{t-1}, V_{t-1})]$
- $p(U_t, V_t \mid U_{t-1}, V_{t-1})$ from Kalman filter

Inference

Skip-Gram Smoothing

- $q(U_{1:T} \mid n^{\pm}_{1:T}) = \Pi_{i=1}^{L} \Pi_{k=1}^{d} q(u_{ik,1:T} \mid n^{\pm}_{ik,1:T})$
- Same for V
- Fitted jointly to all time steps; no longer restricted to a variational distribution that factorizes in time
- This approach results in smoother trajectories and typically higher likelihoods than with filtering, because evidence is used from both future and past observations.

Experiments

Let’s see some cool graphs

- SGI: non-Bayesian skip-gram model with independent random initializations of word vectors
- SGP denotes the same approach as above, but with word and context vectors being pre-initialized with the values from the previous year
- DSG-F: dynamic skip-gram filtering (proposed).
- DSG-S: dynamic skip-gram smoothing (proposed).

Experiments

Data they use

- Google books corpus (Michel et al., 2011) from the last two centuries (T = 209). This amounts to 5 million digitized books and approximately 1010 observed words.
  - The corpus consists of n-gram tables with n $\in$ {1, . . . , 5}, annotated by year of publication.
  - Considered the years from 1800 to 2008 (the latest available).
  - In 1800, the size of the data is approximately ~70m words. Used the 5-gram counts, resulting in a context window size of 4.
- “State of the Union” (SoU) addresses of U.S. presidents, which spans more than two centuries, resulting in T = 230 different time steps and approximately 106 observed words.
  - Some presidents gave both a written and an oral address; if these were less than a week apart, concatenate them and use the average date.
  - Constructed the positive sample counts n_{ij}^{+}using a context window size of 4.
- Twitter corpus of news tweets for 21 randomly drawn dates from 2010 to 2016.
  - Used a context window size of 4

Experiments

Hyperparameters

- The vocabulary for each corpus was constructed from the 10,000 most frequent words through-out the given time period.
    - In the Google books corpus, the number of words per year grows by a factor of 200 from the year 1800 to 2008.
    - To avoid that the vocabulary is dominated by modern words, normalize the word frequencies separately for each year before adding them up.

Experiments

Qualitative results

- Results in smooth word embedding trajectories on all three corpora
- Can automatically detect words that undergo significant semantic changes over time

Experiments

Let’s see some cool graphs

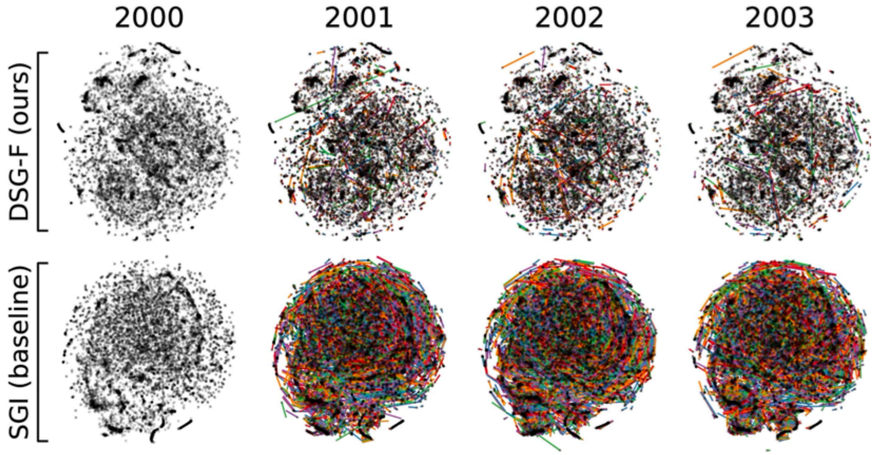


Figure 3. Word embeddings over a sequence of years trained on Google books, using DSG-F (proposed, top row) and compared to the static method by Hamilton et al. (2016) (bottom). We used dynamic t-SNE (Rauber et al., 2016) for dimensionality reduction. Colored lines in the second to fourth column indicate the trajectories from the previous year. Our method infers smoother trajectories with only few words that move quickly. Figure 4 shows that these effects persist in the original embedding space.

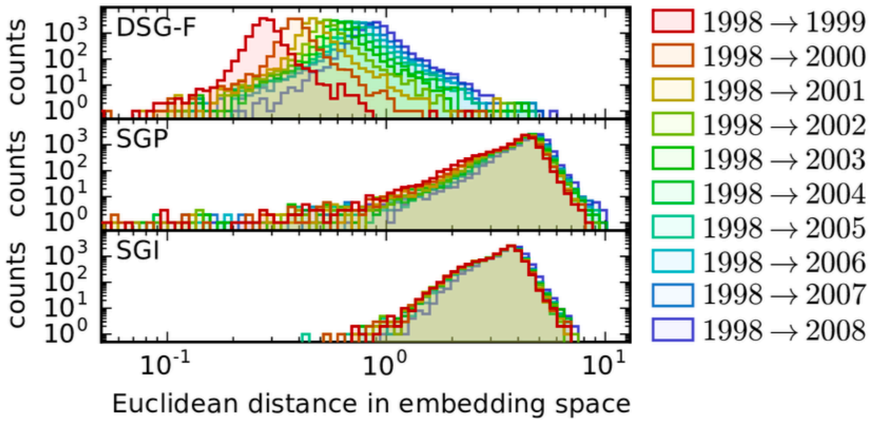


Figure 4. Histogram of distances between word vectors in the year 1998 and their positions in subsequent years (colors). DSG-F (top panel) displays a continuous growth of these distances over time, indicating a directed motion. In contrast, in SGP (middle) (Kim et al., 2014) and SGI (bottom) (Hamilton et al., 2016), the distribution of distances jumps from the first to the second year but then remains largely stationary, indicating absence of a directed drift; i.e. almost all motion is random.

Experiments

Let’s see some cool graphs

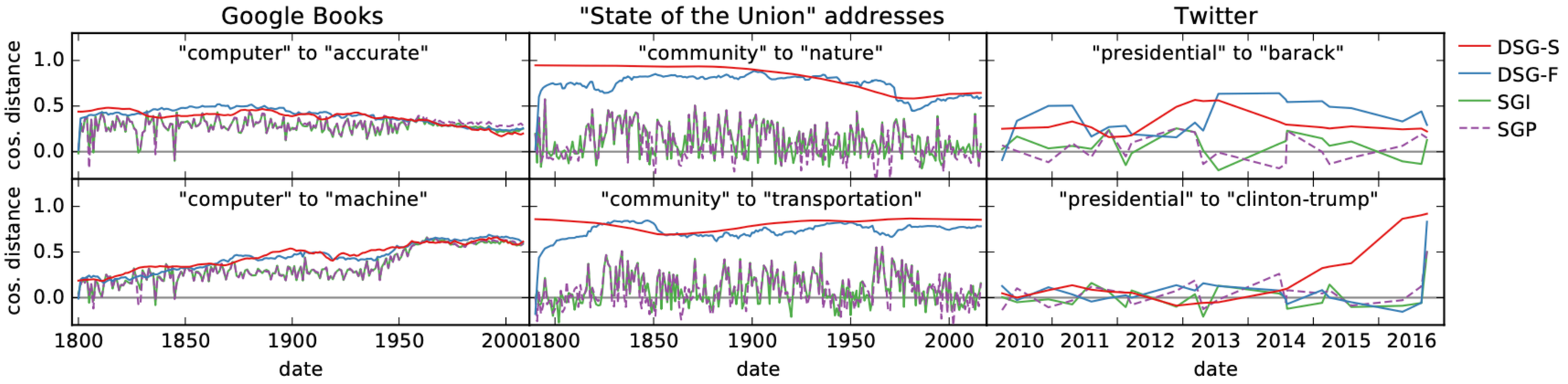


Figure 5. Smoothness of word embedding trajectories, compared across different methods. We plot the cosine distance between two words (see captions) over time. High values indicate similarity. Our methods (DSG-S and DSG-F) find more interpretable trajectories than the baselines (SGI and SGP). The different performance is most pronounced when the corpus is small (SoU and Twitter).

Experiments

Quantitative results

- Generalizes better to unseen data
- Analyze $\frac{1}{|n^{\pm}|} \log p(n_t^{\pm} \mid \tilde{U}_t, \tilde{V}_t)$, the predictive likelihoods on word-context pairs at a given time t, where t is excluded from the training set
- For SGI and SGP, used the embeddings $\tilde{U}_t = U_{t-1}$ and $\tilde{V}_t = V_{t-1}$ from the previous step
- For DSG-F, set $\tilde{U}_t$ and $\tilde{V}_t$ to be the modes U_{t-1}, V_{t-1} of the approximate posterior at the previous time step
- For DSG-S, hold out 10%, 10% and 20% of the documents from the Google books, SoU, and Twitter corpora for testing, respectively.
    - After training, estimate the word (context) embeddings $\tilde{U}_t, \tilde{V}_t$ by linear interpolation between the values of U_{t+1}, V_{t+1} and U_{t-1}, V_{t-1} in the mode of the variational distribution, taking into account that the time stamps τ_t in general are not equally spaced.

Experiments

Let’s see some cool graphs

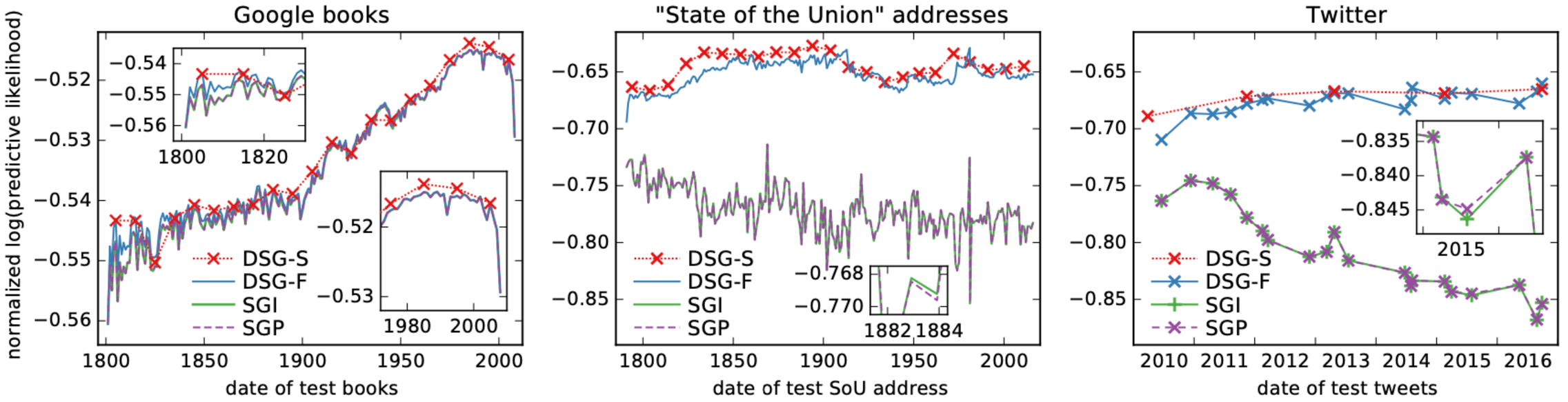


Figure 6. Predictive log-likelihoods (Eq. 16) for two proposed versions of the dynamic skip-gram model (DSG-F & DSG-S) and two competing methods SGI (Hamilton et al., 2016) and SGP (Kim et al., 2014) on three different corpora (high values are better).

Conclusion

TL;DR

- Presented the dynamic skip-gram model: a Bayesian probabilistic model that combines word2vec with a latent continuous time series.
- Showed experimentally that both dynamic skip-gram filtering (which conditions only on past observations) and dynamic skip-gram smoothing (which uses all data) lead to smoothly changing embedding vectors that are better at predicting word-context statistics at held-out time steps.
- The benefits are most drastic when the data at individual time steps is small, such that fitting a static word embedding model is hard. This approach may be used as a data mining and anomaly detection tool when streaming text on social media, as well as a tool for historians and social scientists interested in the evolution of language.