When Can We Work in Embedding Space? What Text Embeddings Preserve

Simon Freyaldenhoven
*Federal Reserve Bank of Philadelphia**

August 26, 2026

Abstract

When do text embeddings work as inputs to empirical analysis? Their use rests on an assumption: that we can trade text for its low-dimensional embedding, and lose little in doing so. I make that assumption precise under a generative model in which documents are mixtures of latent topics. I study two uses—*clustering* units in embedding space and *controlling* for high-dimensional text. A cluster of embeddings is a set of documents with similar topic mixtures; controlling for the embedding is equivalent to controlling for the topic mixture, so validity reduces to whether that mixture captures the confounding. In an application to 363 U.S. metropolitan areas, embedding-based clusters of LLM-generated economic descriptions recover interpretable economic archetypes and separate local employment dynamics more sharply than clustering on model residuals, or on a curated set of industry and demographic covariates.

JEL-Classification: C21, C38, C45, C55

KEYWORDS: Text embeddings, topic models, text as data, clustering, high-dimensional controls, Word2Vec, Large Language Models

*The views expressed herein are those of the author and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia, or the Federal Reserve System. Email: simon.freyaldenhoven@phil.frb.org

1 Introduction

Text embeddings are now a standard tool in empirical economics. Researchers turn product descriptions (Bajari et al., 2025; Bach et al., 2025), central bank communication (Casella et al. (2026)), labor-market histories (Vafa et al., 2022, 2025), and word meanings (Kozlowski et al., 2019) into vectors (embeddings) and use those in subsequent analysis. The appeal is dimension: While text itself is extremely high-dimensional, its embedding is (relatively) low-dimensional. The justification, either explicitly or implicitly, is that little should be lost by working in embedding space if the information contained in a text can be represented in a low-dimensional space.[1] I make this argument precise and study when text embeddings work as inputs to empirical analysis.

Any reduction of text to a vector may discard something the analysis needed. Existing theoretical work rules this out via a high level sufficiency assumption: that a low-dimensional attribute vector exists and the embedding proxies it (Christensen and Compiani, 2026), or that whatever the representation discards induces a bias vanishing faster than $n^{-1/2}$ (Vafa et al., 2025). But how plausible is such an assumption? In order to answer this, my starting point is a simple generative model of text, which allows me to derive exactly what embeddings preserve.

I first show that, under a standard model in which documents are mixtures of K latent topics (Blei et al., 2003; Hofmann, 1999), the corpus's matrix of word co-occurrence ratios, centered at independence, is positive semi-definite of rank exactly $K-1$ (Proposition 1). Any embedding that matches this matrix inherits the topic-loading geometry — plainly speaking: words with similar loadings receive similar embeddings (Theorem 1). Standard embedding methods do not target this matrix directly, and instead factorize a *logarithmic* transform of the co-occurrence ratios (Table 1), which is generically full rank. Focusing on one leading embedding — Word2Vec (Mikolov et al., 2013a,b), in its skip-gram-with-negative-sampling (SGNS) form — I then prove that it recovers the same geometry up to a distortion (Proposition 2).

At the document level, averaging word embeddings yields an invertible linear image of the topic mixture (Proposition 3). I then consider two concrete use cases. The first *groups* units by latent type: clustering in embedding space discretizes unobserved heterogeneity into peer groups. Because distinct topic mixtures cannot share an embedding, a cluster is a set of documents with similar mixtures, and its

[1] For example, Bajari et al. (2025) states that the success of their approach "depends on the existence of parsimonious structures behind images and text" and that they "believe that information in images and sentences can be effectively represented in a much lower-dimensional space".

centroid corresponds to a well-defined mixture (Corollary 1). Recovering the latent partition exactly is a stronger claim: when documents concentrate tightly enough around a small number of common mixture-types, that partition is a fixed point of k-means (Proposition 4), but separation alone does not make it the unique optimum. The second *conditions* on textual embeddings: the embedding enters a regression as a control. Adjusting for the embedding is equivalent to adjusting for the topic mixture, so the high-level assumption that "the embedding is a sufficient control" reduces to a transparent condition that the topic mixture captures the confounding. (Corollary 2).

To make the grouping use concrete, consider a researcher who wants to model local employment dynamics. A natural starting point is an autoregressive model for log employment in location i, e.g.:

$$y_{it} = \alpha_i + \rho_1^i y_{i,t-1} + \cdots + \rho_p^i y_{i,t-p} + \varepsilon_{it}, \tag{1}$$

where y_{it} is log employment in year t for a given Core-Based Statistical Area (CBSA) i. Pooling across all locations imposes implausible homogeneity, while estimating each city separately throws away cross-sectional information and can result in very noisy estimates. I propose the following compromise: cluster units by their textual description. Concretely, for each unit CBSA (i) generate a 500-word economic narrative using a large language model, (ii) embed the narrative in $\mathbb{R}^r$, and (iii) apply k-means in embedding space. Our theoretical results from Section 2 provide a lens into this algorithm: If the corpus can be reasonably approximated by a topic model, k-means clustering recovers peer groups with similar topic mixtures. I find that the peer groups that emerge indeed encode similarity (e.g., "deindustrialized Eds-and-Meds", "Sunbelt growth", "energy and resource extraction"), and that the resulting clusters capture meaningful heterogeneity in employment dynamics which clustering on outcome residuals, or on observed covariates, misses.

Related Literature. This paper connects several literatures. The generative model builds on the topic model literature, including probabilistic latent semantic indexing (Hofmann, 1999) and Latent Dirichlet Allocation (Blei et al., 2003). Our use of a topic model as the data-generating process is not only analytical convenience: a recent literature argues that large language models themselves behave as implicit latent-variable models, inferring a latent concept from context and generating text conditional on it (Xie et al., 2022; Wang et al., 2023).

On the embedding side, the Word2Vec model was introduced by Mikolov et al. (2013a,b), and Levy and Goldberg (2014) established that skip-gram with negative sampling implicitly factorizes a shifted pointwise mutual information matrix when the embedding dimension is unrestricted. Arora et al. (2016) model the writing of a

document as a random walk over a latent topic vector in a way that justifies what Word2Vec and GloVe (Pennington et al., 2014) estimate. I instead take the topic model as the data-generating process—a model that is easy to interpret and that economists already use to summarize text—and ask what it implies for embeddings. Dieng et al. (2020) put words and topics in the same embedding space by assumption, giving each topic its own vector. I derive what they assume: under our model each topic has a centroid in embedding space, and document embeddings lie in the simplex those centroids span (Proposition 3)—for an embedding estimated without any knowledge of the topics. Finally, Li et al. (2023) show that the attention layers of a transformer can also recover topic information, but under the restrictive assumption that each word belongs to a single topic, so that topics have no vocabulary in common.

Veitch et al. (2020) use text embeddings as a control for confounding. They construct an embedding that captures the confounding by explicitly supervising it on both treatment and outcome, while I consider "standard" unsupervised embeddings in this paper. Other papers studying causal inference with text include Roberts et al. (2020) and Egami et al. (2022). Finally, Battaglia et al. (2024), Vafa et al. (2025) and Christensen and Compiani (2026) study the gap between the representation a researcher has and the object the model requires, and each closes it downstream: a bias correction for a noisy estimate of the right target, a characterization of the omitted-variable bias induced by coarsening, and a correction for an imperfect proxy, respectively. Our paper is complementary to this literature: there, the object the representation recovers is taken as given; here, I ask what an unsupervised embedding recovers in the first place.

2 Theoretical Results

2.1 Generative Model

Suppose we generate a corpus of text, meaning a collection of D documents from a vocabulary of size V. Each document d consists of N_d terms: $\{w_{d,1}, \ldots, w_{d,N_d}\}$. Within document d, each term is drawn i.i.d. according to the column-stochastic distribution $\Pi_{\bullet d}$, where Π is a column-stochastic $V \times D$ matrix; that is, Π_{vd} denotes the probability that a randomly drawn term in document d equals word v. Throughout, I use $P(\cdot)$ exclusively as the probability operator and reserve Π for the matrix of word–document probabilities. For a matrix A, $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ denote its largest and smallest singular values.

I further assume that $\Pi = B\Theta$, where B is $V \times K$ (word-topic matrix) and Θ

is $K \times D$ (topic-document matrix); I take $K \geq 2$ throughout to avoid degeneracy. Both B and Θ have non-negative entries, and each column of Π and B sums to 1. The probability for a given term v is thus given by $P(w = v|d) = (B\Theta_{\bullet d})_v$. and does not depend on the position in the document.

Following Hofmann (1999), I assume that for each document d

$$N_{\bullet d}|(B, \Theta) \sim \text{Multinomial}\left(N_d, B\Theta_{\bullet d}\right), \tag{2}$$

where $N_{\bullet d} = (N_{1d}, \ldots, N_{Vd})^\top$ is the vector of word counts in document d, and the vectors of counts $N_{\bullet d}$ are independent across documents, conditional on (B, Θ).

Once we add a Dirichlet prior on the per-document topic distribution we obtain a proper generative model for new documents in the form of standard LDA (Blei et al., 2003). On the other hand, treating the topic mixture θ_d as a fixed (unknown) parameter rather than a random variable, this has been studied in the Non-negative Matrix Factorization (NMF) literature (Lee and Seung, 1999; Donoho and Stodden, 2003; Arora et al., 2013).

Notation. Consider two words v and u appearing in the same document d. Under the i.i.d.-given-document model, conditional on d a target word and a separately drawn context word are independent, so:

$$P(w = v, c = u|d) = \Pi_{vd} \cdot \Pi_{ud} = (B\Theta_{\bullet d})_v (B\Theta_{\bullet d})_u. \tag{3}$$

Define the *word co-occurrence matrix* $M \in \mathbb{R}^{V \times V}$ by aggregating over documents:

$$M_{vu} \;=\; P(w = v, c = u) \;=\; \sum_{d=1}^{D} p_d (B\Theta_{\bullet d})_v (B\Theta_{\bullet d})_u, \tag{4}$$

where $p_d = N_d / \sum_{d'} N_{d'}$ weights documents by length and $\sum_d p_d = 1$. Equivalently, with $p = [p_1, \ldots, p_D]^\top$ and $G := \sum_d p_d \Theta_{\bullet d} \Theta_{\bullet d}^\top = \Theta \,\mathrm{diag}(p)\, \Theta^\top \in \mathbb{R}^{K \times K}$ denoting the corpus second-moment matrix of topic mixtures,

$$M \;=\; B\, G\, B^\top. \tag{5}$$

Let $\bar{\theta} := \sum_d p_d \Theta_{\bullet d}$ denote the corpus-mean topic mixture, let $q := B\bar{\theta}$ collect the marginal word probabilities, $q_v = P(w{=}v) = P(c{=}v)$, and let $D_q := \mathrm{diag}(q)$ denote the corresponding diagonal matrix. Define the *probability-ratio matrix*

$$R \;:=\; D_q^{-1}\, M\, D_q^{-1} \;\in\; \mathbb{R}_{>0}^{V \times V}, \qquad R_{vu} \;=\; \frac{P(w = v,\, c = u)}{P(w = v)\, P(c = u)}. \tag{6}$$

Finally, define the *marginal-normalized loading matrix* $\tilde{B} := D_q^{-1} B \in \mathbb{R}^{V \times K}$ and the *topic-mixture covariance*

$$\Sigma_\Theta \;:=\; \sum_{d=1}^{D} p_d \,(\Theta_{\bullet d} - \bar{\theta})(\Theta_{\bullet d} - \bar{\theta})^\top = G - \bar{\theta}\bar{\theta}^\top \;\in\; \mathbb{R}^{K \times K}. \tag{7}$$

2.2 Word Embeddings

This section derives our theoretical results on word embeddings. I first introduce a matrix I call the centered probability ratio and show that it is positive semi-definite (PSD) of rank $K-1$ (Proposition 1). I then construct an explicit embedding β whose entries are written directly in the topic-model primitives B and Θ (Lemma 1) that factorize the centered probability ratio. Our main word-embedding result (Theorem 1) then shows that any embedding $\hat{\beta}$ matching this factorization inherits the marginal-normalized topic-loading geometry. Simply stated: words with similar topic loadings receive similar embeddings.

I maintain the following regularity conditions.

Assumption 1 (Topic regularity)**.** I maintain the following regularity conditions on the topic model throughout:

(a) $\operatorname{rank} B = K$;

(b) $\operatorname{rank} \Sigma_\Theta = K - 1$;

(c) $q_v > 0$ for all $v \in [V]$.

The full column rank of B rules out redundant topics. Because the columns of Θ lie on the simplex Δ^{K-1}, $\Sigma_\Theta \mathbf{1}_K = 0$, so $\operatorname{rank} \Sigma_\Theta \leq K-1$. Assumption 1(b) is thus a maximal rank condition that is implied by the existence of K documents whose topic mixtures are affinely independent in Δ^{K-1}. The marginal-positivity condition simply rules out terms with zero marginal probability and is used to invert D_q.[2]

Proposition 1 (Exact rank-$(K-1)$ factorization)**.** *Under Assumption 1, the centered probability ratio $(R - \mathbf{1}_V \mathbf{1}_V^\top)$ is PSD of rank $K-1$, such that the top-$(K-1)$ eigendecomposition $R - \mathbf{1}_V \mathbf{1}_V^\top = \Phi \Lambda \Phi^\top$ yields a factor $\beta_{\mathrm{SVD}} := \Phi\, \Lambda^{1/2} \in \mathbb{R}^{V \times (K-1)}$, with*

$$\beta_{\mathrm{SVD}}\; \beta_{\mathrm{SVD}}^\top \;=\; R - \mathbf{1}_V \mathbf{1}_V^\top.$$

[2]Note that, if there exists a term v with $q_v = 0$, this term is not used in any document. Removing any unused terms from the dictionary and using the smaller vocabulary V' immediately implies that $q_v > 0$ for all $v \in [V']$.

Proof. By (5), $M = BGB^\top$, hence

$$R = D_q^{-1}(BGB^\top)D_q^{-1} = \tilde{B}G\tilde{B}^\top. \tag{8}$$

Substituting $G = \Sigma_\Theta + \bar{\theta}\bar{\theta}^\top$ and using $\tilde{B}\bar{\theta} = D_q^{-1}q = \mathbf{1}_V$,

$$R - \mathbf{1}_V\mathbf{1}_V^\top \;=\; \tilde{B}\,\Sigma_\Theta\,\tilde{B}^\top. \tag{9}$$

Since Σ_Θ is a covariance matrix it is PSD, hence $\tilde{B}\Sigma_\Theta\tilde{B}^\top$ is PSD. By Assumptions 1(a) and (c), $\tilde{B}$ has full column rank K.It follows that

$$\operatorname{rank}(R - \mathbf{1}_V\mathbf{1}_V^\top) \;=\; \operatorname{rank}(\Sigma_\Theta) \;=\; K - 1$$

by Assumption 1(b). The factorization follows. □

Proposition 1 delivers an embedding β_{SVD} from the centered probability ratio $R - \mathbf{1}_V\mathbf{1}_V^\top$ alone, with no reference to the topic-model primitives. The next result gives a second, equivalent representative whose entries are written directly in those primitives (B, Θ).

Lemma 1 (Topic-primitive representation)**.** *Under Assumption 1, take the compact eigendecomposition of the topic-mixture covariance defined in* (7)*:*

$$\Sigma_\Theta \;=\; \Phi_\Theta\,\Lambda_\Theta\,\Phi_\Theta^\top.^{3}$$

For each $v \in [V]$, define

$$\beta_v \;:=\; \frac{\Lambda_\Theta^{1/2}\,\Phi_\Theta^\top\,B_{v\bullet}^\top}{q_v} \;\in\; \mathbb{R}^{K-1}. \tag{10}$$

Then, the matrix $\beta \in \mathbb{R}^{V\times(K-1)}$ with rows $\beta_v^\top$ satisfies

$$R - \mathbf{1}_V\mathbf{1}_V^\top \;=\; \beta\,\beta^\top, \tag{11}$$

and is related to β_{SVD} from Proposition 1 by an orthogonal transformation: there exists an orthogonal Q ($Q^\top Q = I_{K-1}$) with $\beta = \beta_{\mathrm{SVD}}\,Q$.

[3]Formally, $\Phi_\Theta \in \mathbb{R}^{K\times(K-1)}$ holds orthonormal eigenvectors of the nonzero eigenvalues ($\Phi_\Theta^\top\Phi_\Theta = I_{K-1}$) and $\Lambda_\Theta \in \mathbb{R}^{(K-1)\times(K-1)}$ is the diagonal matrix of corresponding (positive) eigenvalues.

Proof. For any $v, u \in [V]$,

$$\beta_v^\top \beta_u \;=\; \frac{B_{v\bullet}\,\Phi_\Theta \Lambda_\Theta \Phi_\Theta^\top\, B_{u\bullet}^\top}{q_v q_u} \;=\; \frac{B_{v\bullet}\,\Sigma_\Theta\, B_{u\bullet}^\top}{q_v q_u} \;=\; (\tilde{B}\,\Sigma_\Theta\,\tilde{B}^\top)_{vu} \;=\; (R - \mathbf{1}_V \mathbf{1}_V^\top)_{vu},$$

where the last step follows from (9). Both β and β_{SVD} are full-column-rank $V \times (K-1)$ factors of the same rank-$(K-1)$ PSD matrix, so they coincide up to an orthogonal transformation. □

The inner product $\beta_v^\top \beta_u = \frac{P(w=v,\, c=u)}{P(w=v)\,P(c=u)} - 1$ can be interpreted as the lift of the joint above independence: zero if w and c are independent, positive if they co-occur more than under independence, and negative if less. I next show that *any* $(K-1)$-dimensional embedding matching this factorization inherits the marginal-normalized topic-loading geometry.

Theorem 1 (Topic-loading geometry)**.** *Suppose Assumption 1 holds and let* $\hat{\beta} \in \mathbb{R}^{V \times (K-1)}$ *satisfy* $\hat{\beta}\,\hat{\beta}^\top = R - \mathbf{1}_V \mathbf{1}_V^\top$*. Then, for all* $v, u \in [V]$*,*

$$\|\hat{\beta}_v - \hat{\beta}_u\|^2 \;=\; (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet})^\top \Sigma_\Theta\, (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet}). \tag{12}$$

In particular, $B_{v\bullet} = \lambda B_{u\bullet}$ *for some* $\lambda > 0$ *implies* $\hat{\beta}_v = \hat{\beta}_u$*: words with proportional topic loadings receive identical embeddings.*

Proof. By Lemma 1, the topic-primitive representative β satisfies $\beta\beta^\top = R - \mathbf{1}_V \mathbf{1}_V^\top = \hat{\beta}\hat{\beta}^\top$, so both are full-column-rank $V \times (K-1)$ factors of the same rank-$(K-1)$ PSD matrix and coincide up to an orthogonal transformation: $\hat{\beta} = \beta Q$ with $Q^\top Q = I_{K-1}$. From the closed form, $\beta_v - \beta_u = \Lambda_\Theta^{1/2} \Phi_\Theta^\top (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet})$, so

$$\|\beta_v - \beta_u\|^2 \;=\; (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet})^\top \Phi_\Theta \Lambda_\Theta \Phi_\Theta^\top (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet}) \;=\; (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet})^\top \Sigma_\Theta\, (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet}),$$

using $\Sigma_\Theta = \Phi_\Theta \Lambda_\Theta \Phi_\Theta^\top$. Finally, $\|\hat{\beta}_v - \hat{\beta}_u\| = \|\hat{\beta}_v Q - \hat{\beta}_u Q\| = \|(\hat{\beta}_v - \hat{\beta}_u) Q\| = \|\beta_v - \beta_u\|$ since orthogonality of Q preserves the Euclidean norm. This completes the proof of (12). The proportional case follows from $B_{v\bullet} = \lambda B_{u\bullet} \Rightarrow \tilde{B}_{v\bullet} = \tilde{B}_{u\bullet} \Rightarrow \beta_v = \beta_u \Rightarrow \hat{\beta}_v = \hat{\beta}_u$. □

I call the metric on the right of (12) — the Σ_Θ-weighted distance between marginal-normalized topic loadings — the *topic-loading geometry*. Theorem 1 provides an algorithm-agnostic reading: any procedure that matches the centered probability ratio $R - \mathbf{1}_V \mathbf{1}_V^\top$ in dimension $K-1$ recovers the topic-loading geometry, regardless of the construction. The direct example is the rank-$(K-1)$ singular value decomposition of $R - \mathbf{1}_V \mathbf{1}_V^\top$ (Proposition 1). I next analyze some of the word embeddings commonly used in the literature.

Remark 1 (Dimensions above $K-1$)**.** Theorem 1 is stated at $r = K-1$, but nothing in it requires equality. Let $\hat{\beta} \in \mathbb{R}^{V\times r}$ with $r \geq K-1$ satisfy $\hat{\beta}\hat{\beta}^\top = R - \mathbf{1}_V\mathbf{1}_V^\top$. The right-hand side has rank $K-1$ by Proposition 1, so $\hat{\beta} = \beta Q$ for some $Q \in \mathbb{R}^{(K-1)\times r}$ with orthonormal rows, $QQ^\top = I_{K-1}$, and

$$\|\hat{\beta}_v - \hat{\beta}_u\|^2 \;=\; (\beta_v - \beta_u)^\top QQ^\top(\beta_v - \beta_u) \;=\; \|\beta_v - \beta_u\|^2,$$

so (12) holds unchanged. The proof above is the special case $r = K-1$, where Q is square and orthogonal. The extra $r-(K-1)$ coordinates are identically zero and drop out of every distance.

Thus, any $r \geq K-1$ delivers the same geometry. This matters because K is rarely known in practice. Section 4 sets $r = 50$ on this basis.

Remark 2 (Identification-invariance)**.** The factorization $\Pi = B\Theta$ is not unique absent further structure: for any invertible $K \times K$ matrix A (that preserves the column-stochasticity of both factors), the reparameterization $B' = BA$, $\Theta' = A^{-1}\Theta$ generates the same observable joint distribution (see, e.g., Donoho and Stodden (2003), Fu et al. (2019)). However, under any such reparameterization the quadratic form in Theorem 1 transforms as

$$\begin{aligned}(\tilde{B}'_{v\bullet} - \tilde{B}'_{u\bullet})^\top \Sigma'_\Theta\,(\tilde{B}'_{v\bullet} - \tilde{B}'_{u\bullet}) &= (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet})^\top A\,(A^{-1}\Sigma_\Theta A^{-\top})\,A^\top(\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet}) \\ &= (\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet})^\top \Sigma_\Theta\,(\tilde{B}_{v\bullet} - \tilde{B}_{u\bullet}),\end{aligned}$$

so it is invariant. Theorem 1 therefore holds for any valid factorization, and the geometry it characterizes is identification-free.

2.3 Relationship to Standard Algorithms

Word2Vec (Mikolov et al., 2013a,b) and GloVe (Pennington et al., 2014) are some of the most widely used word embeddings in practice, and this section asks what they recover under our model. I treat three objectives: skip-gram with a full softmax (the idealized version that is costly to estimate), its negative-sampling approximation SGNS (Levy and Goldberg, 2014) (which is what practitioners usually run), and GloVe.[4]

None of these fit $R - \mathbf{1}_V\mathbf{1}_V^\top$. Each scores a word–context pair (v, u) by an inner product $X_{vu} = w_v^\top \tilde{w}_u$ between a target vector w_v and a context vector $\tilde{w}_u$, the rows of $W, \tilde{W} \in \mathbb{R}^{V\times r}$, and fits that score to the corpus. They differ in only three ingredients:

[4]Word2Vec's other architecture, CBOW, generally admits no closed-form target, and I return to it at the end of the section.

	Response t_{vu}	Generator ϕ	Weight ω_{vu}	Target $\phi'(t_{vu})$	Free offsets
β_{SVD}	$R_{vu} - 1$	squared	1	$R - \mathbf{1}_V \mathbf{1}_V^\top$	—
Full softmax	$P(w{=}v \mid c{=}u)$	entropy	q_u	$\log R + \log q\, \mathbf{1}_V^\top$	$\mathbf{1}_V g^\top$
SGNS	$R_{vu}/(R_{vu} + \nu)$	binary entropy	$q_v q_u (R_{vu} + \nu)$	$\log R - \log \nu\, \mathbf{1}_V \mathbf{1}_V^\top$	—
GloVe	$\log M_{vu}$	squared	$f(M_{vu})$	$\log R$	$b\mathbf{1}_V^\top + \mathbf{1}_V \tilde{b}^\top$

Table 1: The four objectives as weighted low-rank fits. The response t_{vu} and the generator ϕ determine the target $\phi'(t_{vu})$; the weights ω_{vu} do not. The generators squared, entropy and binary entropy give squared-error, Kullback–Leibler and Bernoulli losses. "Free offsets" are the row- or column-constant terms the objective leaves undetermined: the softmax column gauge g and GloVe's per-word biases $b, \tilde{b}$. $\nu \geq 1$ is the negative-sampling rate and f is GloVe's weighting function. Row one attains its target exactly at rank $K - 1$ (Proposition 1); the other three attain theirs only where the rank constraint does not bind. For more detail, see Appendix A.

a *response* t_{vu} read off the corpus, a strictly convex *generator* ϕ that fixes the loss, and *weights* ω_{vu} on pairs. Where the rank constraint of the inner product does not bind, the fitted scores are $\phi'(t_{vu})$ entrywise. I therefore call that matrix the algorithm's *target*. Table 1 lists the three ingredients and the target for each algorithm, alongside β_{SVD} from the previous section. I derive its entries in Appendix A.

Since none of these targets is $R - \mathbf{1}_V \mathbf{1}_V^\top$, Theorem 1 does not apply to the embeddings that fit them. However, I still obtain the following result on how word embeddings vary with the loadings, analogous to Theorem 1 for the SGNS target.

Assumption 2 (Word co-occurrence)**.** $R_{vu} > 0$ for all $v, u \in [V]$.

Assumption 2 requires every pair of words to co-occur and guarantees that every entry of $\log R$ is finite. Write $R_{\min} := \min_{v,u} R_{vu}$, $R_{\max} := \max_{v,u} R_{vu}$ and $\kappa_R := R_{\max}/R_{\min}$, so that $\log \kappa_R$ is the range of entries in $\log R$.

Proposition 2 (Recovery from the SGNS target)**.** *Let Assumptions 1 and 2 hold. Suppose* $\hat{W}\hat{\tilde{W}}^\top = \log R - \log \nu\, \mathbf{1}_V \mathbf{1}_V^\top$*, the SGNS target of Table 1, with* $\hat{\tilde{W}} \in \mathbb{R}^{V \times r}$ *of full column rank. Then, for all* $v, v' \in [V]$*,*

$$\frac{\sigma_{\min}(\beta)}{R_{\max}\, \sigma_{\max}(\hat{\tilde{W}})} \, \|\beta_v - \beta_{v'}\| \;\leq\; \|\hat{w}_v - \hat{w}_{v'}\| \;\leq\; \frac{\sigma_{\max}(\beta)}{R_{\min}\, \sigma_{\min}(\hat{\tilde{W}})} \, \|\beta_v - \beta_{v'}\|, \tag{13}$$

where $\|\beta_v - \beta_{v'}\|^2 = (\tilde{B}_{v\bullet} - \tilde{B}_{v'\bullet})^\top \Sigma_\Theta \, (\tilde{B}_{v\bullet} - \tilde{B}_{v'\bullet})$ *is the topic-loading distance of Theorem 1.*

In particular, $B_{v\bullet} = \lambda B_{v'\bullet}$ *for some* $\lambda > 0$ *implies* $\hat{w}_v = \hat{w}_{v'}$*.*

Proof. Since the shift $\log \nu \, \mathbf{1}_V \mathbf{1}_V^\top$ is constant and cancels in row differences, $(\log R)_{v\bullet} - (\log R)_{v'\bullet} = (\hat{w}_v - \hat{w}_{v'})\hat{\tilde{W}}^\top$. All entries of R lie in $[R_{\min}, R_{\max}]$, so the mean value theorem applied to log gives $|R_{vu} - R_{v'u}|/R_{\max} \le |\log R_{vu} - \log R_{v'u}| \le |R_{vu} - R_{v'u}|/R_{\min}$ for each u. Thus, for the entire vector:

$$\frac{1}{R_{\max}} \|R_{v\bullet} - R_{v'\bullet}\| \;\le\; \|(\hat{w}_v - \hat{w}_{v'})\hat{\tilde{W}}^\top\| \;\le\; \frac{1}{R_{\min}} \|R_{v\bullet} - R_{v'\bullet}\|.$$

By Lemma 1, $\beta\beta^\top = R - \mathbf{1}_V \mathbf{1}_V^\top$. All rows of $\mathbf{1}_V \mathbf{1}_V^\top$ are equal, so they cancel in the row difference: $R_{v\bullet} - R_{v'\bullet} = (\beta_v - \beta_{v'})^\top \beta^\top$. Since β has full column rank $K-1$ (Proposition 1), $\sigma_{\min}(\beta) > 0$ and

$$\sigma_{\min}(\beta) \, \|\beta_v - \beta_{v'}\| \;\le\; \|R_{v\bullet} - R_{v'\bullet}\| \;\le\; \sigma_{\max}(\beta) \, \|\beta_v - \beta_{v'}\|.$$

Combining with the display above gives

$$\frac{\sigma_{\min}(\beta)}{R_{\max}} \|\beta_v - \beta_{v'}\| \;\le\; \|(\hat{w}_v - \hat{w}_{v'})\hat{\tilde{W}}^\top\| \;\le\; \frac{\sigma_{\max}(\beta)}{R_{\min}} \|\beta_v - \beta_{v'}\|.$$

Finally, full column rank of $\hat{\tilde{W}}$ gives $\sigma_{\min}(\hat{\tilde{W}})\|x\| \le \|x\hat{\tilde{W}}^\top\| \le \sigma_{\max}(\hat{\tilde{W}})\|x\|$ for every x. Substituting into the display above gives (13).

The proportional case follows from $B_{v\bullet} = \lambda B_{v'\bullet} \Rightarrow \tilde{B}_{v\bullet} = \tilde{B}_{v'\bullet} \Rightarrow \beta_v = \beta_{v'}$, which makes the right-hand side of (13) zero. □

Proposition 2 is the analogue of Theorem 1 for the SGNS target: any $\hat{W}, \hat{\tilde{W}}$ that attain that target induce a metric equivalent to the one β induces, reproducing the topic-loading geometry up to a distortion of at most $\kappa_R \operatorname{cond}(\beta) \operatorname{cond}(\hat{\tilde{W}})$. In particular, words with proportional topic loadings receive identical embeddings under any such factorization, exactly as in Theorem 1.

Remark 3 (Target versus embedding)**.** Proposition 2 is a statement about the SGNS target, not about the trained embedding. Since generically the SGNS target is full rank, the trained embedding can only approximate the target, and is instead the weighted low-rank fit of Table 1, with the weights determining which approximation (Levy and Goldberg, 2014). The proposition's practical content therefore depends on the quality of this approximation, and hence on the spectrum of the target.[5]

[5]To bound the spectrum theoretically requires strong assumptions I deem unlikely to hold in practice. Two examples: i) Call v an *anchor word* for topic k if $B_{vk} > 0$ and $B_{vk'} = 0$ for all $k' \neq k$ (Donoho and Stodden, 2003; Arora et al., 2013). If every word is an anchor, then $\log R$, and hence the target, has at most rank K; this is the population analogue of the block structure Li et al. (2023) obtain for transformers under disjoint topics. ii) Write $\varrho := \|R - \mathbf{1}_V \mathbf{1}_V^\top\|_\infty$ for the maximum distance from independence. Then, $\varrho < 1$ gives $\log R = \beta\beta^\top + O(\varrho^2)$ entrywise, so the target is of rank at most K up to an $O(V\varrho^2)$ perturbation.

I explore this further in Section 3, where I compute the spectrum exactly under our simulation designs and find the target approximately low rank. In our application in Section 4 I simply set $r = 50$ to ensure $r > K$ and that the rank constraint is not too binding.

Remark 4 (The other log targets)**.** Proposition 2 uses only that the SGNS target is $\log R$ up to a constant, so it covers any target that is $\log R$ up to an offset constant down each column. However, both full-softmax skip-gram and GloVe carry an offset constant along each *row*, which does not cancel in a row difference. For example, for full softmax that offset is $\log q\, \mathbf{1}_V^\top$, so $R_{v\bullet} = R_{v'\bullet}$ gives

$$(\hat{w}_v - \hat{w}_{v'})\hat{\tilde{W}}^\top \;=\; (\log q_v - \log q_{v'})\, \mathbf{1}_V^\top,$$

and the two embeddings coincide if and only if $q_v = q_{v'}$: same-loading words agree up to a single marginal direction. Lemma A.5 gives the general statement: for any entrywise transform of R and any offsets constant along a row or a column, words with proportional loadings have target rows differing by the row-offset difference alone.

CBOW (Mikolov et al., 2013a) predicts the target word from the *average* of its context embeddings. With a one-word context the average is trivial: under negative sampling its target is again $\log R - \log \nu\, \mathbf{1}_V \mathbf{1}_V^\top$, so Proposition 2 applies verbatim. The averaging is what removes the theory: with more than one context word the score depends on the context only through the averaged embedding, so no closed-form $V \times V$ target exists. Given its prevalence in the literature, I nevertheless treat CBOW empirically in Sections 3–4.

2.4 Document Embeddings

Word embeddings can be aggregated to create document representations. A natural approach is to represent each document by the average of its word embeddings:

$$\bar{h}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} h_{w_{dj}}, \tag{14}$$

where h_w is the embedding assigned to word w—for instance, the $\hat{\beta}_w \in \mathbb{R}^{K-1}$ of Theorem 1, with explicit representative β_w given by Lemma 1—and w_{dj} is the j-th word in document d.

First, note that averaging is the natural aggregator under the model of Section 2.1, since the tokens $w_{d,1}, \ldots, w_{d,N_d}$ are i.i.d. draws from the document's word distribution

$\Pi_{\bullet d}$. Averaging is also the operation behind a broad class of document- and sentence-embedding methods that are used in practice. Unweighted averages of static word vectors are a widely used baseline (Wieting et al., 2016; Iyyer et al., 2015); Arora et al. (2017) reweight the average by *smooth inverse frequency* (down-weighting frequent words), and Joulin et al. (2017) average word- and n-gram vectors for text classification. The same pooling extends to contextual models, where transformer sentence encoders mean-pool token embeddings into a fixed-length vector (Reimers and Gurevych, 2019).[6]

Intuitively, under our topic model, documents with similar topic mixture θ_d should have similar document embeddings, since they draw words from similar distributions over the vocabulary. I now formalize this intuition. Define the *expected document embedding* for a document with topic mixture θ_d as:

$$\mu_d = \mathbb{E}[\bar{h}_d|\theta_d] = \sum_{v=1}^{V} P(w=v|d)h_v = \sum_{v=1}^{V}(B\Theta_{\bullet d})_v h_v. \tag{15}$$

Further, because the words within document d, $w_{d,1}, \ldots, w_{d,N_d}$ are i.i.d. conditional on θ_d, clearly $\bar{h}_d = \frac{1}{N_d}\sum_j h_{w_{dj}} \xrightarrow{p} \mu_d$ as $N_d \to \infty$ by the law of large numbers. Finally, define the *topic centroid embeddings*

$$c_k \;:=\; \sum_{v=1}^{V} B_{vk}\, h_v, \qquad k = 1, \ldots, K, \tag{16}$$

and let $C \equiv [c_1 \;\cdots\; c_K]$ denote the matrix with the centroids as columns. I obtain the following result.

Proposition 3 (Linearity in the topic mixture)**.** *Suppose* $\Pi = B\Theta$ *and the topic model is identified.*[7] *Then, for every document* d*:*

(i) the expected document embedding is a linear function of the topic mixture,

$$\mu_d \;=\; C\,\Theta_{\bullet d}; \tag{17}$$

[6]The main alternative learns document vectors directly rather than by averaging (Paragraph Vector, or doc2vec; Le and Mikolov, 2014), which I do not treat here.

[7]The anchor-word condition — every topic has at least one word exclusive to it — is one sufficient condition for identification (Donoho and Stodden (2003); Arora et al. (2013); Fu et al. (2019)), also see the discussion in Freyaldenhoven et al. (2025). However, it is not necessary; see the recent work of Chen et al. (2022) that uses the *sufficiently-scattered* condition in Huang et al. (2013) and Huang et al. (2016).

(ii) μ_d lies in the centroid simplex, $\mu_d \in \mathrm{conv}\{c_1, \ldots, c_K\}$.

Proof. For (i), simply expanding the definition of μ_d in (15) and rearranging the sums yields

$$\begin{aligned}\mu_d \;&=\; \sum_{v=1}^{V} (B\Theta_{\bullet d})_v \, h_v \;=\; \sum_{v=1}^{V}\sum_{k=1}^{K} B_{vk}\,\theta_{dk}\,h_v \\ &=\; \sum_{k=1}^{K} \theta_{dk} \sum_{v=1}^{V} B_{vk}\,h_v \;=\; \sum_{k=1}^{K} \theta_{dk}\, c_k = C\,\Theta_{\bullet d}.\end{aligned}$$

For (ii), the topic mixture satisfies $\theta_{dk} \geq 0$ and $\sum_k \theta_{dk} = 1$, which immediately implies $\mu_d \in \mathrm{conv}\{c_1, \ldots, c_K\}$. □

Note that, beyond the topic-model decomposition $\Pi = B\Theta$, Proposition 3 imposes no further conditions on the word embeddings h_v: the linear decomposition $\mu_d = C\,\Theta_{\bullet d}$ and the simplex containment follow from linearity of averaging and hold for arbitrary—even random—embeddings. Proposition 3 characterizes the *space* in which document embeddings live: a low-dimensional simplex spanned by the topic centroids, with each document's position encoding its topic mixture. Corollary 1 below characterizes the *metric* on that space—the pairwise distances between document embeddings when the word embeddings satisfy the factorization of Theorem 1.[8]

Corollary 1 (Document-embedding metric)**.** *Suppose Assumption 1 holds, $\Pi = B\Theta$, and the topic model is identified. If the word embeddings $H = [h_1, \ldots, h_V]^\top$ satisfy $HH^\top = R - \mathbf{1}_V\mathbf{1}_V^\top$ (such as the explicit construction in Lemma 1), the pairwise distance between expected document embeddings is given by*

$$\|\mu_d - \mu_{d'}\|_2^2 \;=\; (\Pi_{\bullet d} - \Pi_{\bullet d'})^\top\,(R - \mathbf{1}_V\mathbf{1}_V^\top)\,(\Pi_{\bullet d} - \Pi_{\bullet d'}). \tag{19}$$

Equivalently,

$$\|\mu_d - \mu_{d'}\|_2^2 \;=\; (\Theta_{\bullet d} - \Theta_{\bullet d'})^\top\,G_B\,\Sigma_\Theta\,G_B\,(\Theta_{\bullet d} - \Theta_{\bullet d'}), \tag{20}$$

where $G_B := B^\top D_q^{-1} B$. Moreover, this distance is strictly positive whenever $\Theta_{\bullet d} \neq \Theta_{\bullet d'}$.

[8] Note that linearity of $\mu_d = C\,\Theta_{\bullet d}$ already implies a Lipschitz bound between expected document embeddings and topic mixtures: for any two documents,

$$\|\mu_d - \mu_{d'}\|_2 \;\leq\; \|C\|_{\mathrm{op}}\,\|\Theta_{\bullet d} - \Theta_{\bullet d'}\|_2, \tag{18}$$

with Lipschitz constant equal to the operator norm of the centroid matrix. Corollary 1 sharpens this bound once the word embeddings satisfy $HH^\top = R - \mathbf{1}_V\mathbf{1}_V^\top$.

Proof. Using $\mu_d = H^\top \Pi_{\bullet d}$, we have

$$\begin{aligned}\|\mu_d - \mu_{d'}\|_2^2 &= (H^\top \Pi_{\bullet d} - H^\top \Pi_{\bullet d'})^\top (H^\top \Pi_{\bullet d} - H^\top \Pi_{\bullet d'}) \\ &= (\Pi_{\bullet d} - \Pi_{\bullet d'})^\top H H^\top (\Pi_{\bullet d} - \Pi_{\bullet d'}) \\ &= (\Pi_{\bullet d} - \Pi_{\bullet d'})^\top (R - \mathbf{1}_V \mathbf{1}_V^\top) (\Pi_{\bullet d} - \Pi_{\bullet d'}).\end{aligned}$$

Using (9) to substitute $R - \mathbf{1}_V \mathbf{1}_V^\top = D_q^{-1} B \, \Sigma_\Theta \, B^\top D_q^{-1}$ and using $\Pi_{\bullet d} = B \, \Theta_{\bullet d}$ yields

$$\begin{aligned}\|\mu_d - \mu_{d'}\|_2^2 &= (\Pi_{\bullet d} - \Pi_{\bullet d'})^\top (R - \mathbf{1}_V \mathbf{1}_V^\top) (\Pi_{\bullet d} - \Pi_{\bullet d'}) \\ &= (\Theta_{\bullet d} - \Theta_{\bullet d'})^\top B^\top D_q^{-1} B \, \Sigma_\Theta \, B^\top D_q^{-1} B \, (\Theta_{\bullet d} - \Theta_{\bullet d'}).\end{aligned}$$

For the final claim, write $x := \Theta_{\bullet d} - \Theta_{\bullet d'} \neq 0$, and note $\mathbf{1}_K^\top x = 0$ because both mixtures lie on the simplex. Since $\operatorname{rank} \Sigma_\Theta = K - 1$ and $\Sigma_\Theta \mathbf{1}_K = 0$, we have $\ker \Sigma_\Theta = \operatorname{span}(\mathbf{1}_K)$, so (20) vanishes if and only if $G_B x = \lambda \mathbf{1}_K$ for some $\lambda \in \mathbb{R}$. In that case $x = \lambda G_B^{-1} \mathbf{1}_K$, and $\mathbf{1}_K^\top x = 0$ gives $\lambda \, \mathbf{1}_K^\top G_B^{-1} \mathbf{1}_K = 0$. Under Assumption 1, G_B is positive definite, hence so is G_B^{-1} and $\mathbf{1}_K^\top G_B^{-1} \mathbf{1}_K > 0$; therefore $\lambda = 0$ and $G_B x = 0$, so $x = 0$, a contradiction. □

Equation (19) expresses the document-embedding distance through the centered probability ratio $R - \mathbf{1}_V \mathbf{1}_V^\top$ and the word-distribution profiles $\Pi_{\bullet d}$ with no reference to the topic-model primitives.[9] The equivalent form (20) rewrites the same distance in latent topic coordinates as weighted topic-mixture differences. Those weights depend on the topic-mixture covariance Σ_Θ and the overlap among the topic loadings, G_B. Taken together, Proposition 3 and Corollary 1 show that the document embedding is a linear—and, under Assumption 1, invertible—encoding of the latent topic mixture $\Theta_{\bullet d}$. In other words, μ_d is a sufficient statistic for $\Theta_{\bullet d}$, so a cluster of nearby embeddings is a set of documents with similar mixtures, and the map from an embedding-space centroid to a mixture is well defined. It requires no assumption about how documents are distributed over the simplex.

I next apply the results in this section to two uses of embeddings: one *groups* units by latent type, where I show that k-means in embedding space[10] recovers structure on the topic mixtures, Section 2.5. The other *conditions* on them, where I show that adjusting for the embedding is equivalent to adjusting for the topic mixture, Section 2.6.

[9] Note that, because $R - \mathbf{1}_V \mathbf{1}_V^\top$ has rank $K - 1$ (Proposition 1), the distance depends only on the projection of $\Pi_{\bullet d} - \Pi_{\bullet d'}$ onto the $(K-1)$-dimensional topic subspace; word-frequency variation orthogonal to it leaves the embedding distance unchanged.

[10] Throughout, I call the Euclidean space in which the word embeddings h_v, document embeddings μ_d, and centroids c_k live the embedding space — $\mathbb{R}^{K-1}$ under the exact factorization of Proposition 1.

2.5 Grouping: Clustering Documents

I next ask when k-means clustering recovers meaningful structure. I begin with the simplest case—cluster centers at the topic centroids—and then generalize.

Lemma 2 (Voronoi-cell membership for concentrated documents)**.** *Suppose Assumption 1 holds, the topic model is identified, and the word embeddings satisfy $HH^\top = R - \mathbf{1}_V \mathbf{1}_V^\top$ (such as the explicit construction in Lemma 1). Let $\mathcal{V}_k$ denote the Voronoi cell of c_k under the Euclidean metric in embedding space,*[11] *and $k^*(d) = \arg\max_k \theta_{dk}$. Define*

$$\eta \;:=\; \frac{1}{2}\frac{\min_{k\neq k'} \|c_k - c_{k'}\|_2}{\max_{k\neq k'} \|c_k - c_{k'}\|_2} \;\in\; (0, 1/2]. \tag{21}$$

Then, for any document d with dominant-topic weight $\theta_{d,k^*(d)} > 1 - \eta$, $\mu_d \in \mathcal{V}_{k^*(d)} \cap \operatorname{conv}\{c_1, \ldots, c_K\}$.

Proof. The centroids are distinct: $c_k = Ce_k$ is the embedding of a pure-topic document, and $e_k \neq e_{k'}$ for $k \neq k'$, so Corollary 1 gives $\|c_k - c_{k'}\|_2 > 0$. Hence $\eta \in (0, 1/2]$ is well defined.

Next, fix document d and let $k^* = k^*(d)$. Write $\Theta_{\bullet d} = (1-\delta)e_{k^*} + \delta z$ with $\delta = 1 - \theta_{d,k^*} \in [0,1]$ and $z \in \Delta_{K-1}$. With $\mu_d = C\,\Theta_{\bullet d}$ (Proposition 3),

$$\begin{aligned}\mu_d - c_{k^*} &= C\Theta_{\bullet d} - c_{k^*} = C(1-\delta)e_{k^*} + C\delta z - c_{k^*} = (1-\delta)c_{k^*} + C\delta z - c_{k^*} \\ &= \delta(Cz - c_{k^*}).\end{aligned}$$

Since $z \mapsto \|Cz - c_{k^*}\|_2$ is convex on the simplex Δ_{K-1}, its maximum is attained at a vertex; hence $\|Cz - c_{k^*}\|_2 \le \max_{k'} \|c_{k'} - c_{k^*}\|_2 \le C^{\max}$, where $C^{\max} := \max_{k\neq k'} \|c_k - c_{k'}\|_2$. Therefore, since $\delta = 1 - \theta_{d,k^*}$,

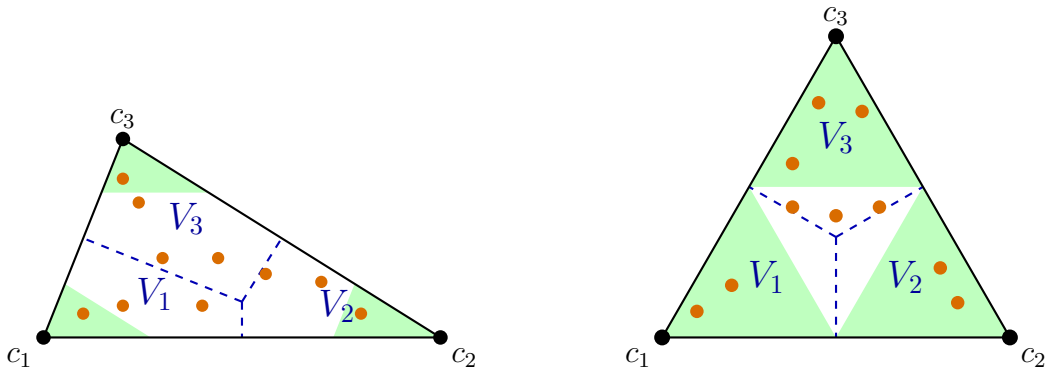
$$\|\mu_d - c_{k^*}\|_2 \;\le\; \delta\, C^{\max} \;=\; (1 - \theta_{d,k^*})\, C^{\max}. \tag{22}$$

For any $k' \neq k^*$, the triangle inequality gives

$$\|\mu_d - c_{k'}\|_2 \;\ge\; \|c_{k^*} - c_{k'}\|_2 - \|\mu_d - c_{k^*}\|_2 \;\ge\; \Delta^* - \delta C^{\max},$$

where $\Delta^* := \min_{k\neq k'} \|c_k - c_{k'}\|_2 > 0$.

[11] Given a simplex with vertices $c_1, \ldots, c_K$, the *Voronoi cell* of c_k is $\mathcal{V}_k = \{x : \|x - c_k\|_2 \le \|x - c_{k'}\|_2 \text{ for all } k' \neq k\}$, the points at least as close to c_k as to any other vertex. The cells $\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ partition the simplex into convex polytopes meeting at the perpendicular bisectors of the vertex pairs; assigning each point to its nearest vertex is the nearest-centroid rule used by k-means with fixed centers.



(a) Unequal centroid triangle: $\eta \approx 0.27$.

(b) Equilateral centroid triangle: $\eta = 1/2$ (maximal).

Figure 1: Illustration of Lemma 2 for $K = 3$. In each panel, the topic centroids c_1, c_2, c_3 (black) span the convex hull $\text{conv}\{c_1, c_2, c_3\}$ (solid triangle). The perpendicular bisectors of the centroid pairs (dashed) partition the triangle into three Voronoi cells $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3$, with $\mathcal{V}_k$ containing the points closer to c_k than to any other centroid. Green shading marks the regions where the lemma guarantees correct assignment, $\theta_{d,k^*(d)} > 1 - \eta$. Orange dots illustrate document embeddings $\mu_d = C\,\Theta_{\bullet d}$ for ten documents with varying topic mixtures.

It follows that, whenever $\delta C^{\max} < \Delta^* - \delta C^{\max}$, or equivalently, $\delta < \Delta^*/(2C^{\max}) = \frac{\min_{k \neq k'} \|c_k - c_{k'}\|_2}{2 \max_{k \neq k'} \|c_k - c_{k'}\|_2} = \eta$, we have:

$$\|\mu_d - c_{k^*}\|_2 \leq \delta C^{\max} < \Delta^* - \delta C^{\max} \leq \|\mu_d - c_{k'}\|_2 \quad \text{for all } k' \neq k^*.$$

Since $\delta < \eta$, and $\theta_{d,k^*} > 1 - \eta$ are equivalent, $\theta_{d,k^*} > 1 - \eta$ thus implies $\mu_d \in \mathcal{V}_{k^*}$. Finally, since, by Proposition 3, $\mu_d \in \text{conv}\{c_1, \ldots, c_K\}$ for all d, this completes the proof. □

In words, Lemma 2 states that documents whose mass is concentrated on a single topic vertex of Δ_{K-1} end up in the Voronoi cell of the corresponding centroid. The constant η in Lemma 2 is defined as half the ratio of minimum to maximum pairwise centroid distance. $\eta = 1/2$ is achieved when the centroids are equidistant, i.e., when they form an equilateral K-simplex; For the other extreme, $\eta \to 0$ when some pair of centroids becomes much closer than the others. I illustrate in Figure 1.

Lemma 2 is the vertex case: documents close to a vertex will be assigned to the corresponding Voronoi cell. But the logic behind this is not special to the vertices: Below, I establish the more general result that, for any set of "archetypes" $\mu_1^*, \ldots, \mu_L^*$,

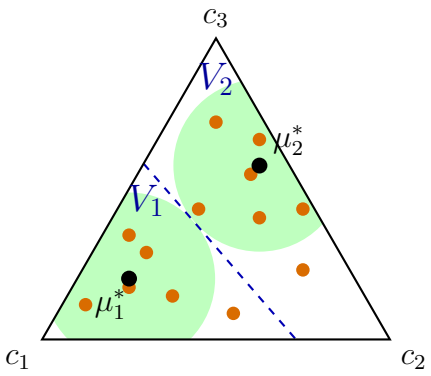


Figure 2: Illustration of the nearest-archetype membership for $K = 3$ and two archetypes ($L = 2$). The convex hull $\mathrm{conv}\{c_1, c_2, c_3\}$ (solid black) is partitioned by the perpendicular bisectors of archetype pairs (dashed) into Voronoi cells V'_ℓ of the archetype embeddings μ^*_ℓ (black dots in the interior or on the hull boundary). Green shading marks the guaranteed regions $\{\mu_d : \|\mu_d - \mu^*_\ell\|_2 < \Delta'^*/2\}$ — disks of radius $\Delta'^*/2$ around each archetype, clipped to the hull. Orange dots illustrate document embeddings $\mu_d = C\,\Theta_{\bullet d}$. Documents inside a green disk are guaranteed by the triangle-inequality argument to lie in the corresponding Voronoi cell.

documents whose embeddings are sufficiently close to an archetype μ^*_ℓ will be assigned to the Voronoi cell of μ^*_ℓ. Thus, the recovered clusters then reflect *mixture-type* similarity rather than dominant-topic similarity, and the number of clusters L need not equal the number of topics K. Figure 2 illustrates the archetype geometry. Formally, extending the argument from a single document to the full corpus, I obtain the following proposition.

Proposition 4 (k-means identification under archetype concentration)**.** *Suppose Assumption 1 holds. Let $\pi^*_1, \ldots, \pi^*_L \in \Delta_{K-1}$ be $L \geq 2$ distinct* archetype mixtures*—points in Δ_{K-1} around which documents concentrate—with minimum separation*

$$\Delta'^* \;:=\; \min_{\ell \neq \ell'} \|C(\pi^*_\ell - \pi^*_{\ell'})\|_2 \;>\; 0,$$

and let $\ell(d) := \arg\min_\ell \|C(\Theta_{\bullet d} - \pi^*_\ell)\|_2$ denote the nearest-archetype assignment in topic-mixture space. If every document satisfies

$$\|C(\Theta_{\bullet d} - \pi^*_{\ell(d)})\|_2 \;<\; \Delta'^*/4,$$

then the partition $\mathcal{P}^* := \{d : \ell(d) = \ell\}_{\ell=1}^L$ is a fixed point of k-means: assigning each document to the nearest cell mean of $\mathcal{P}^*$ returns $\mathcal{P}^*$, and recomputing the means

returns the same means. Equivalently, $\mathcal{P}^*$ is simultaneously the Voronoi partition of the archetype embeddings $\{\mu_\ell^*\}$ and of its own cell means.

Proof. Let $\mathcal{G}_\ell := \{d : \ell(d) = \ell\}$, $\mu_\ell^* := C\,\pi_\ell^*$ and $\mu_d = C\,\Theta_{\bullet d}$. Within-cluster pairwise distances then satisfy, for $d, d' \in \mathcal{G}_\ell$,

$$\|\mu_d - \mu_{d'}\|_2 \;\le\; \|\mu_d - \mu_\ell^*\|_2 + \|\mu_{d'} - \mu_\ell^*\|_2 \;<\; \tfrac{\Delta'^*}{4} + \tfrac{\Delta'^*}{4} \;=\; \tfrac{\Delta'^*}{2},$$

while for $d \in \mathcal{G}_\ell, d' \in \mathcal{G}_{\ell'}$ with $\ell \neq \ell'$,

$$\|\mu_d - \mu_{d'}\|_2 \;\ge\; \|\mu_\ell^* - \mu_{\ell'}^*\|_2 - \|\mu_d - \mu_\ell^*\|_2 - \|\mu_{d'} - \mu_{\ell'}^*\|_2 \;\ge\; \Delta'^* - 2\,\tfrac{\Delta'^*}{4} \;=\; \tfrac{\Delta'^*}{2}.$$

so within-cluster pairwise distances are strictly smaller than cross-cluster ones.

For the fixed-point claim, let $m_\ell := \bar{\mu}_{\mathcal{G}_\ell}$ denote the cell means. Each m_ℓ is a convex combination of points within $\Delta'^*/4$ of μ_ℓ^*, so $\|m_\ell - \mu_\ell^*\|_2 < \Delta'^*/4$. For $d \in \mathcal{G}_\ell$,

$$\|\mu_d - m_\ell\|_2 \;\le\; \|\mu_d - \mu_\ell^*\|_2 + \|\mu_\ell^* - m_\ell\|_2 \;<\; \tfrac{\Delta'^*}{2},$$

while for $\ell' \neq \ell$,

$$\|\mu_d - m_{\ell'}\|_2 \;\ge\; \|\mu_\ell^* - \mu_{\ell'}^*\|_2 - \|\mu_d - \mu_\ell^*\|_2 - \|\mu_{\ell'}^* - m_{\ell'}\|_2 \;>\; \Delta'^* - \tfrac{\Delta'^*}{2} \;=\; \tfrac{\Delta'^*}{2}.$$

Every document is therefore strictly closer to its own cell mean than to any other, so the assignment step reproduces $\mathcal{P}^*$; the update step then returns the same means. □

At the simplex vertices, Proposition 4 specializes to the dominant-topic case, translating the embedding-distance condition back into a sufficient threshold on θ. It is worth noting that Proposition 4 is a stability statement, not a recovery statement.[12] Global optimality requires even stronger conditions. However, even the weaker separation condition in Proposition 4 is not satisfied in our application. I therefore claim no recovery of a latent partition in our application. Instead, I rely on Corollary 1, which connects embedding distance to the difference of topic mixtures. A cluster is thus a set of documents with similar mixtures. This is what the interpretation of the clusters in Section 4 rests on, and it requires no concentration assumption.

[12]Note that the archetypes in Proposition 4 need not be fixed in advance. A natural choice is the k-means centers at convergence: by construction they lie in $\mathrm{conv}\{c_1, \ldots, c_K\}$ and equal $C\pi_\ell^*$ for some $\pi_\ell^* \in \Delta_{K-1}$. Proposition 4 then applies to these centers: if they are separated and every document lies within $\Delta'^*/4$ of its assigned center, the converged partition is a k-means fixed point.

2.6 Conditioning: Embeddings as Controls

The second use of document embeddings I consider is as a *control*. A researcher who wants to control for a high-dimensional text X_d—a firm disclosure, a court filing, a job posting—cannot condition on the raw text and instead conditions on its embedding, implicitly assuming that adjustment for the embedding suffices. Proposition 3, and in particular the identity $\mu_d = C\,\Theta_{\bullet d}$, immediately makes the content of that assumption precise.

Formally, let $Z_d \in \{0, 1\}$ be a treatment, $Y_d(0), Y_d(1)$ the potential outcomes for unit d, and $Y_d = Y_d(Z_d)$ the observed outcome.

Assumption 3 (Topic unconfoundedness)**.** $\{Y_d(0), Y_d(1)\} \perp Z_d \mid \Theta_{\bullet d}$, and $0 < P(Z_d = 1 \mid \Theta_{\bullet d}) < 1$.

Assumption 3 states that the document's topic mixture captures all confounding between treatment and potential outcomes: given $\Theta_{\bullet d}$, the realized words carry no further information about $(Z_d, Y_d(0), Y_d(1))$. This is the explicit content of the informal claim that the text is a sufficient control.

Corollary 2 (Embedding adjustment)**.** *Suppose Assumption 1 holds, the topic model is identified, and C is injective on Δ_{K-1}. Then, controlling for document embedding is valid if and only if controlling for the topic mixture is. In particular, under Assumption 3, the average treatment effect is identified by*

$$\tau \;=\; \mathbb{E}\big[\,\mathbb{E}[Y_d \mid Z_d = 1, \mu_d] - \mathbb{E}[Y_d \mid Z_d = 0, \mu_d]\,\big].$$

Proof. Since $\Pi = B\Theta$ and the topic model is identified, $\mu_d = C\,\Theta_{\bullet d}$ (Proposition 3). Since C is injective on Δ_{K-1}, μ_d and $\Theta_{\bullet d}$ are one-to-one functions of each other: conditioning on one is the same as conditioning on the other. Adjustment on the embedding is therefore valid exactly when adjustment on the topic mixture is. In particular, under Assumption 3, $\{Y_d(0), Y_d(1)\} \perp Z_d \mid \mu_d$ and $0 < P(Z_d = 1 \mid \mu_d) < 1$, so τ is identified by adjustment on μ_d (Rosenbaum and Rubin, 1983). □

Corollary 2 is an identification statement at the population level; it is exact only under three conditions. First, it uses the expected embedding μ_d. With finite document length, both μ_d and $\Theta_{\bullet d}$ will be measured with error (e.g., $\bar{h}_d = \mu_d + O_p(N_d^{-1/2})$; see the discussion in Section 2.4). Also see Battaglia et al. (2024) for a more general discussion of measurement error in AI/ML-generated variables. Second, it requires embedding dimension $r \geq K - 1$: with $r < K - 1$ the map $\mu_d = C, \Theta_{\bullet d}$ cannot be injective and conditions on only a projection of $\Theta_{\bullet d}$, leaving residual confounding regardless of N_d. However, for the high-dimensional embeddings used

in practice I believe this bound is slack. Third, C must be injective on Δ_{K-1}. For an embedding satisfying $HH^\top = R - \mathbf{1}_V \mathbf{1}_V^\top$ this holds by construction (Corollary 1). For Word2Vec, GloVe, or pretrained embeddings, it is a maintained assumption; see the discussion in Section 2.3.

The point is not that embedding-based adjustment is valid, but that its validity reduces to a transparent condition on the topic mixture. This connects to a growing literature on causal inference with text (Roberts et al., 2020; Egami et al., 2022); relative to that work, I derive the sufficiency condition from the generative model rather than assert it.

Although Corollary 2 is stated for the topic model, its two structural requirements are general. Suppose the confounding is driven by a latent summary of the text of dimension s (here $s = K - 1$, the topic mixture). Embedding-based control then requires, first, an embedding of dimension $r \geq s$—otherwise it encodes only a projection of that summary and leaves residual confounding; and second, that the embedding actually *recover* the summary, so that its low-dimensional geometry coincides with the one the confounding lives in. Dimension alone thus does not suffice—an embedding can be low-dimensional in a geometry that has little to do with the text's, and adjusting for it then controls for the wrong object.

3 Numerical Illustration

I next simulate data from the model described in Section 2.1 and compare two embedding constructions: (i) the closed-form SVD-β embedding of Proposition 1, and (ii) skip-gram with negative sampling (SGNS) (Mikolov et al., 2013b) as implemented in gensim (Řehůřek and Sojka, 2010) (cf. Table 1). Theorem 1 predicts that SVD-β inherits the topic-loading geometry, and Proposition 2 carries the prediction over to the SGNS target.[13]

Both embeddings use dimension $r = K$. By Proposition 1, $r = K - 1$ is the minimum sufficient embedding dimension for SVD-β. The extra dimension should have population eigenvalue zero and be genuine slack, and I confirm this in the figures below. The SGNS target is the approximately rank-$(K-1)$ $\log R$ shifted by the constant $-\log \nu$. We quantify that approximation in the last paragraph of this section. Because the shift is constant, it raises the rank of the target without spreading words apart, so it does not appear in the centered projections plotted below.

[13] Appendix C.4 repeats the exercise with continuous bag-of-words (CBOW). The results are qualitatively similar to SGNS.

To illustrate our theoretical results, consider two simulation designs on a common $V = 16$ vocabulary of fruits and vegetables: a two-topic case (Design 1) and a three-topic extension (Design 2). Documents are generated with topic mixtures drawn from a symmetric Dirichlet ($\alpha = 1$) prior on the simplex, with $D = 363$ documents and $N_d = 487$ words per document, matching the dimensions of the application in Section 4, and the full document as context.[14]

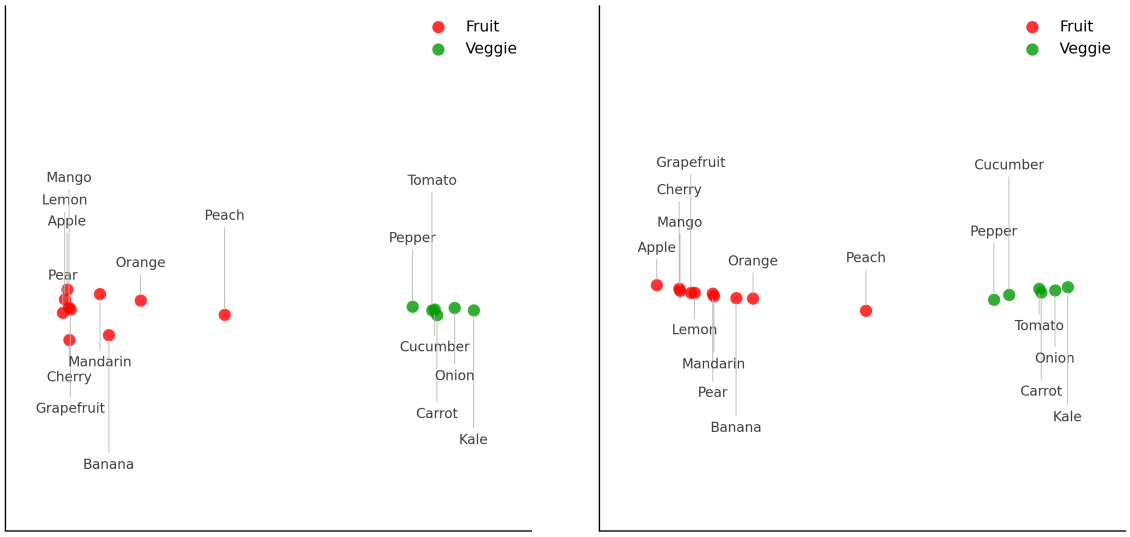
Design 1: Two Topics ($K = 2$). The topic-word matrix is

$$B = \begin{pmatrix} \mathbf{0.08} & 0.00 \\ \mathbf{0.08} & 0.01 \\ \mathbf{0.08} & 0.00 \\ \mathbf{0.12} & 0.00 \\ \mathbf{0.06} & 0.04 \\ \mathbf{0.07} & 0.00 \\ \mathbf{0.10} & 0.02 \\ \mathbf{0.12} & 0.00 \\ \mathbf{0.10} & 0.01 \\ \mathbf{0.13} & 0.00 \\ 0.01 & \mathbf{0.10} \\ 0.02 & \mathbf{0.18} \\ 0.00 & \mathbf{0.30} \\ 0.01 & \mathbf{0.05} \\ 0.01 & \mathbf{0.12} \\ 0.01 & \mathbf{0.17} \end{pmatrix} \quad \text{with rows corresponding to:} \quad \begin{matrix} \text{Apple} \\ \text{Banana} \\ \text{Pear} \\ \text{Mango} \\ \text{Peach} \\ \text{Cherry} \\ \text{Orange} \\ \text{Lemon} \\ \text{Mandarin} \\ \text{Grapefruit} \\ \text{Cucumber} \\ \text{Tomato} \\ \text{Kale} \\ \text{Pepper} \\ \text{Carrot} \\ \text{Onion} \end{matrix} . \tag{23}$$

Apple, Pear, Mango, Cherry, Lemon, and Grapefruit are anchor words for topic 1 (fruits) with varying prevalence, and Kale is an anchor word for topic 2 (vegetables). The remaining words load on both topics with varying intensity.

Figure 3 shows the resulting word embeddings. Both methods recover the same qualitative structure: words separate cleanly along their dominant topic, with fruits (red) and vegetables (green) forming distinct clusters. The SGNS and SVD-β geometries are nearly identical up to rotation and scaling. Both clouds are one-dimensional, which is what the theory implies: SVD-β has population rank $K-1 = 1$, and SGNS's second direction is the common offset, which the centered projection removes. Proposition 2 predicts only that the two induce equivalent metrics on the words; here the distortion it permits is evidently small. Within each cluster, the relative positions

[14] Appendix C repeats this exercise with two further specifications: i) a larger corpus, $D = 1{,}000$ and $N_d = 10{,}000$ (Appendix C.1); and ii) a more concentrated topic-mixture distribution ($\alpha = 0.01$) that puts more mass near the simplex vertices (Appendix C.2).



(a) SVD-β **(b)** SGNS

Figure 3: Word embeddings for Design 1 ($K = 2$): SGNS vs. the closed-form SVD-β construction of Proposition 1, both computed on the same corpus ($D = 363$, $N_d = 487$). Colors mark the dominant topic from the true B matrix.

of words reflect their loadings: anchor words (e.g., Mango, Grapefruit, Kale) lie at the extremes, while mixed words (e.g., Peach, Orange) sit closer to the boundary between clusters, with their positions reflecting their relative loadings. This pattern holds exactly for SVD-β, and approximately for SGNS. On the other hand, the SVD-β embedding appears somewhat noisier (further from rank 1), perhaps reflecting the difficulty in directly matrix-factorizing the noisy sample co-occurrence structure.

Design 2: Three Topics ($K = 3$). I next split the "fruit" category into a "citrus" topic (orange) and a "non-citrus" topic (red), while keeping the same

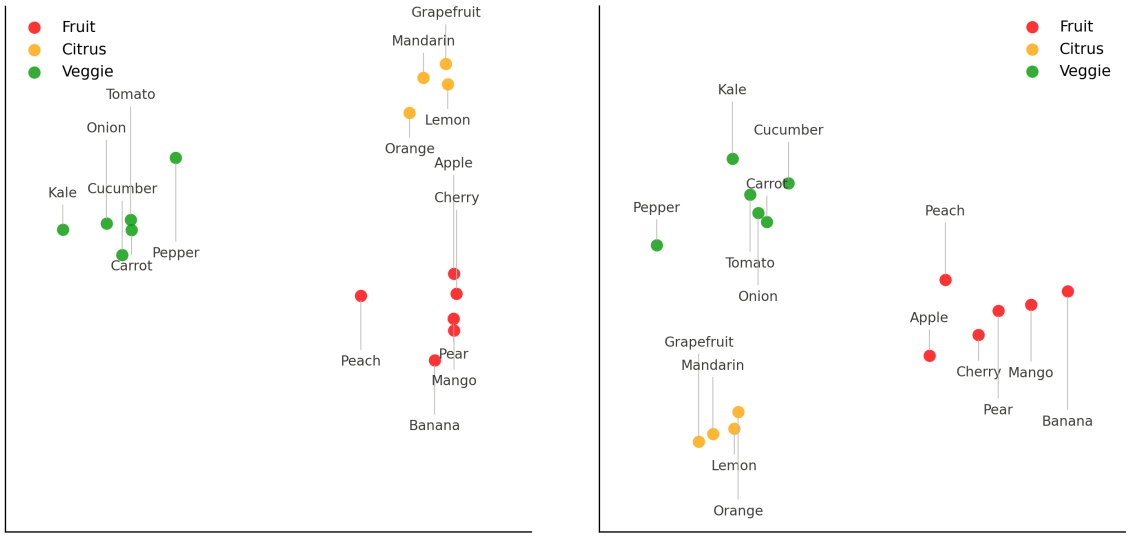
vegetable topic (green). The topic-word matrix is

$$B = \begin{pmatrix} \mathbf{0.10} & 0.05 & 0.00 \\ \mathbf{0.15} & 0.02 & 0.01 \\ \mathbf{0.13} & 0.04 & 0.00 \\ \mathbf{0.19} & 0.05 & 0.00 \\ \mathbf{0.10} & 0.03 & 0.04 \\ \mathbf{0.10} & 0.04 & 0.00 \\ 0.05 & \mathbf{0.15} & 0.02 \\ 0.05 & \mathbf{0.18} & 0.00 \\ 0.03 & \mathbf{0.16} & 0.01 \\ 0.04 & \mathbf{0.22} & 0.00 \\ 0.02 & 0.00 & \mathbf{0.10} \\ 0.02 & 0.02 & \mathbf{0.18} \\ 0.00 & 0.00 & \mathbf{0.30} \\ 0.00 & 0.02 & \mathbf{0.05} \\ 0.01 & 0.01 & \mathbf{0.12} \\ 0.01 & 0.01 & \mathbf{0.17} \end{pmatrix} \quad \text{with rows corresponding to:} \quad \begin{matrix} \text{Apple} \\ \text{Banana} \\ \text{Pear} \\ \text{Mango} \\ \text{Peach} \\ \text{Cherry} \\ \text{Orange} \\ \text{Lemon} \\ \text{Mandarin} \\ \text{Grapefruit} \\ \text{Cucumber} \\ \text{Tomato} \\ \text{Kale} \\ \text{Pepper} \\ \text{Carrot} \\ \text{Onion} \end{matrix} \tag{24}$$

Kale is an anchor word for topic 3 (vegetables), but topics 1 (non-citrus fruits) and 2 (citrus) have no anchor words, though they differ in their relative loadings on the same set of words.

Figure 4 shows the resulting word embeddings. With $r = K = 3$, the embeddings live in $\mathbb{R}^3$, and I project to two dimensions by simply retaining the first two coordinates. For SVD-β, $\beta_{\bullet 1}$ and $\beta_{\bullet 2}$ span the two largest-variance directions and the third column corresponds to a zero eigenvalue in population. For SGNS, the coordinate basis is chosen by the gradient-based optimizer and is arbitrary. Dropping the third coordinate projects out one mixture of the three coordinates.

Both methods produce embeddings in which the three topics are clearly separated, with words organized into three distinguishable groups corresponding to non-citrus fruits (red), citrus fruits (orange), and vegetables (green). The anchor word for vegetables (Kale) sits on the edge of the convex hull, consistent with the theory: anchor words receive embeddings proportional to a single topic centroid, while non-anchor words are pulled toward mixtures of centroids in proportion to their topic loadings. Notably, the absence of anchor words for topics 1 and 2 does not prevent the embeddings from separating these two groups—the differing relative loadings of shared words across the two topics are sufficient to recover the distinction. The SGNS and SVD-β embeddings again yield qualitatively similar geometries, providing empirical support for the theoretical connections established in Section 2.3.



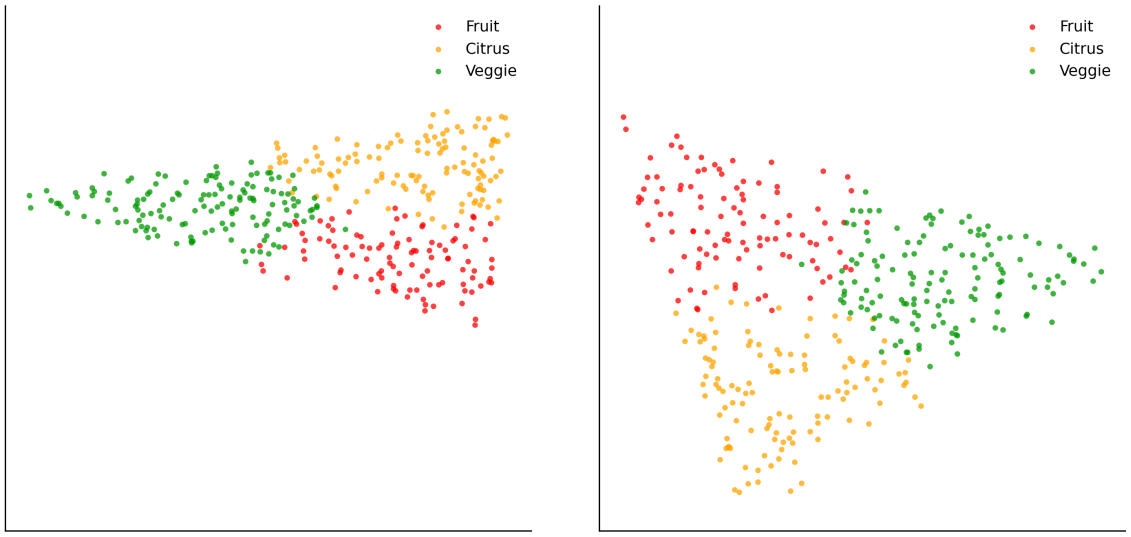
(a) SVD-β **(b)** SGNS

Figure 4: Word embeddings for Design 2 ($K = 3$), projected into two dimensions. Both embeddings recover topic structure even though topic 3 lacks an anchor word.

Document Embeddings. Next, I compute document embeddings as the average of word embeddings within each document (cf. Equation (14)). By Proposition 3, documents with similar topic mixtures $\Theta_{\bullet d}$ should have similar embeddings, and document embeddings should lie in the $(K-1)$-dimensional simplex spanned by the topic centroids. Figure 5 shows the document embeddings for Design 2 ($K = 3$), colored by dominant topic (projected into 2-d following the same step as in Figure 4).[15] With Dirichlet concentration parameter $\alpha = 1$, the topic mixtures are well spread across the simplex, and the document embeddings fill out the simplex spanned by the three topic centroids. The two embedding methods yield similar document-embedding geometries. Appendix C.3 shows additional document-embedding figures for a version that replaces the Dirichlet prior on $\Theta_{\bullet d}$ with an archetype mixture in the interior of the topic simplex, with documents concentrated around the two archetypes.

I note that all of the figures above are based on a single corpus. However, I found the results to be representative across multiple realizations.

[15] Alternatively, Appendix C.5 presents a UMAP-projected version of Figure 4. UMAP's nonlinear transformation distorts the geomerty and "warps" the simplex, but the results are qualitatively similar. I use UMAP projections in our empirical application in Section 4, where the dimensionality is much larger.



(a) SVD-β **(b)** SGNS

Figure 5: Document embeddings for Design 2 ($K = 3$), projected in two dimensions. Color indicates the dominant topic.

The rank of the SGNS target. Remark 3 reduces the gap between the SGNS target and the trained embedding to a property of the spectrum of $T = \log R - \log \nu \, \mathbf{1}_V \mathbf{1}_V^\top$. The shift is a known constant and cancels in the row differences, so the question is how much of $\log R$ the remaining $K - 1$ directions capture. I measure that by

$$\pi_{K-1} = \frac{\sum_{i \le K-1} |\lambda_i(\log R)|}{\sum_i |\lambda_i(\log R)|}, \tag{25}$$

the share of spectral mass in the leading $K - 1$ eigenvalues. Under the two designs above the spectrum is available exactly — $R = \tilde{B} G \tilde{B}^\top$ from the design B and the closed-form Dirichlet moments — and gives $\pi_{K-1} = 0.88$ for Design 1 and 0.91 for Design 2. In these designs we can therefore think of Proposition 2 approximately describing the trained embedding and not only its target, in line with the empirical results above. Table 2 reports π_{K-1} on a grid of vocabulary sizes and topic counts under a generic data-generating process: each column of B and each topic mixture $\Theta_{\bullet d}$ is an independent draw from a symmetric Dirichlet($\alpha = 1$), uniform on its simplex. Across the grid, a rank-K approximation remains close to the SGNS target.[16]

[16]It is worth noting that the quality of the approximation degrades for smaller values of α (cf. Appendix C.6.

V	$K=2$	$K=3$	$K=5$	$K=10$	$K=25$	$K=50$
100	0.909	0.926	0.946	0.968	0.987	0.995
250	0.905	0.924	0.943	0.966	0.985	0.993
500	0.905	0.923	0.943	0.964	0.984	0.992
1,000	0.905	0.924	0.943	0.964	0.983	0.991
2,500	0.905	0.923	0.943	0.963	0.983	0.991

Table 2: Share of the spectral mass of $\log R$ carried by its leading $K-1$ eigenvalues, $\pi_{K-1} = \sum_{i \le K-1} |\lambda_i| / \sum_i |\lambda_i|$, at $\alpha = 1$. Each column of B and each topic mixture $\Theta_{\bullet d}$ is drawn from a symmetric Dirichlet(α); R is then formed in population from B and G, so no corpus is sampled. Means over 8 draws; the largest standard deviation in any cell is 0.004. The negative-sampling shift contributes one further eigenvalue, of size $V \log \nu$, and is excluded.

4 Application: Clustering Metropolitan Economies

I now apply the embedding-then-cluster pipeline to 363 U.S. Core-Based Statistical Areas (CBSAs). The empirical setup is straightforward: generate a textual economic description of each CBSA, embed the descriptions, and apply k-means.

The theory in Section 2 provides a principled reading of what this pipeline does. Corollary 1 gives the pairwise document-embedding distance in closed form:

$$\|\mu_d - \mu_{d'}\|_2^2 \;=\; (\Pi_{\bullet d} - \Pi_{\bullet d'})^\top \, (R - \mathbf{1}_V \mathbf{1}_V^\top) \, (\Pi_{\bullet d} - \Pi_{\bullet d'}).$$

Running k-means on these embeddings groups CBSAs whose descriptions imply similar mixtures. The clusters therefore reflect *narrative similarity*, and each centroid corresponds to a well-defined mixture.

Our main empirical analysis considers five embedding models: (i) the closed-form SVD-β embedding of Proposition 1; (ii) corpus-trained skip-gram with negative sampling (SGNS; Proposition 2); (iii) corpus-trained continuous bag-of-words (CBOW)[17]; (iv) pretrained Word2Vec averaging, using the Google News vectors; and (v) a transformer-based variant using OpenAI's `text-embedding-3-large`.[18] Detailed method descriptions are collected in Appendix B.2.

[17]Both corpus-trained arms and SVD-β use the *full document* as context. Note that the model from Section 2.1 suggests this choice: since the words of document d are drawn from the same mixture $\Pi_{\bullet d}$, positions within a document are exchangeable, so R does not depend on the window at all in population and every within-document pair is another draw on the same object. Widening the window therefore reduces estimation variance without moving the target, and taking all pairs is the efficient choice. I show in Appendix B.8 that results are robust to shorter context windows, although SVD-β tends to deteriorate for very short windows J.

[18]Alternatively, one could ask an LLM to form the clusters and assign labels directly, skipping

Remark 5 (Theoretical coverage of the five methods)**.** The five differ in how much of Section 2 applies to them. Method (i) is the theory's own construction: it computes the rank-$(K-1)$ factorization of Proposition 1 directly, and averaging within a document returns exactly the μ_d of Proposition 3. Method (ii) is the corpus-trained arm the theory covers: the SGNS target is $\log R$ up to a constant, so by Proposition 2 it recovers the topic-loading geometry up to a distortion bounded by κ_R, the exponentiated spread of $\log R$ (Section 2.3). Method (iii), CBOW, is not covered: it has no closed-form $V \times V$ target once the context holds more than one word (Section 2.3), and the context here is the full document. Method (iv) runs the same architecture on a different corpus, so the geometry it carries is that of the corpus it was trained on—the object of interest only if that corpus and ours share their topic structure. Method (v) falls outside the framework altogether: it is a contextual transformer rather than an average of word vectors. I include it as a benchmark for a modern state-of-the-art approach to embeddings.

For each Core-Based Statistical Area (CBSA), a 500-word economic description is generated by Claude Opus 4.8. I provide the exact prompt in Appendix B.1. I then apply k-means with $L = 5$ clusters to group the 363 CBSAs. Table 3 summarizes the five approaches.

Method	L	Cluster sizes	Max share	Within-cluster share
Corpus CBOW	5	57/62/68/85/91	25%	0.90
Corpus SGNS	5	60/71/109/34/89	30%	0.78
SVD-β ($r = 50$)	5	97/154/75/21/16	42%	0.77
Pretrained (Google News)	5	89/33/109/64/68	30%	0.83
LLM Embeddings	5	25/114/59/101/64	31%	0.90

Table 3: Comparison of clustering approaches (363 CBSAs). Cluster sizes are member counts by k-means label index. "Within-cluster share" is the fraction of embedding variance within clusters.

All five identify recognizable archetypes: energy dependence, university and government anchoring, deindustrialization, and diversified growth. I give per-cluster interpretations for all methods in Appendix B.10. The number of clusters is chosen to balance interpretability and flexibility. I consider alternative values for L in the Appendix (e.g., Online Appendix Figures OA.1 and OA.3). The qualitative finding

the embedding step. This did not work at this scale: prompted with all 363 descriptions, the model returned partitions covering only part of the sample and including locations absent from the input list.

that these groupings separate economic narratives (and employment dynamics, cf Section 4.1) holds across different choices of L.

4.1 Local Employment Dynamics

I return to our motivating example in the introduction: modelling local employment dynamics. In order to do so, I estimate autoregressive models for log employment across 363 CBSAs and examine whether the text-based clusters we derived above capture meaningful heterogeneity in employment dynamics. Specifically, for each of the five embedding-based clustering methods (SVD-β, corpus-trained SGNS, corpus-trained CBOW, pretrained Word2Vec averaging, and LLM embeddings), I estimate:

$$y_{it} = \alpha_i + \rho_1^{(g_i)} y_{i,t-1} + \rho_2^{(g_i)} y_{i,t-2} + \varepsilon_{it}, \tag{26}$$

where y_{it} is log total employment in year t at the CBSA level, and $g_i \in \{1, \ldots, L\}$ denotes the cluster membership of CBSA i.[19] I take $p = 2$ as the primary specification based on our discussion in Appendix B.7. I thus allow the AR(2) slope coefficients in (1) to vary across groups, but remain common *within* each group—using our text-based cluster assignments. The hope is that similarity in economic conditions between CBSAs translates into similar local employment dynamics. Under the topic model of Section 2, I can make this condition more precise: that similarity in topic shares between CBSAs translates into similar local employment dynamics. As a concrete example, this might entail that descriptions with a relatively large share of words about manufacturing decline (treating this as a "topic"), are associated with similar employment dynamics.

Figure 6 depicts the group-specific implied impulse response functions (IRF) over a 10-year horizon under each clustering method. The top four panels use embedding-based groupings. The bottom row gives the two groupings that do not use the text. The bottom left uses residual k-means, which is formed by applying k-means clustering to the residuals from a pooled AR(2) specification—analogous to the first step of Bonhomme et al. (2022). The bottom right is observables k-means, applied to 1970 covariates (standardized industry mix, government and military employment shares, education, foreign-born and minority shares and population density;

[19] I construct a panel of geographically-harmonized metropolitan areas for which I can measure employment at an annual frequency starting in 1969. I use 2019 as the last year of analysis to avoid complications from the COVID-19 pandemic and recession. I focus on wage and salary employment, which is constructed by the Bureau of Economic Analysis (BEA) using administrative data and is less sensitive than self employment to changes over time in the tax code. The covariates used for the observables benchmark in Section 4.1 are 1970 cross-sections from the same source.

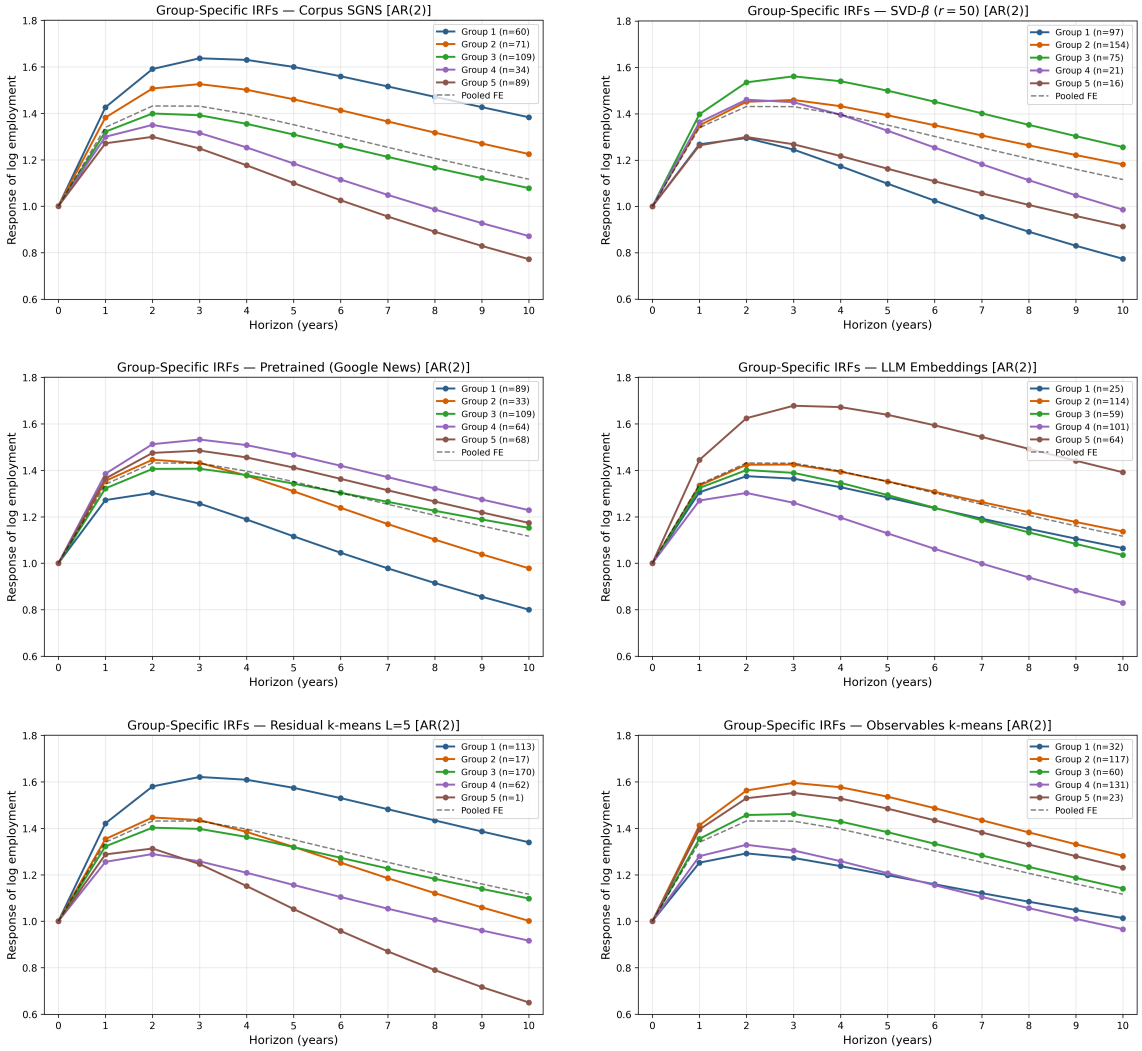


Figure 6: Group-specific AR(2) impulse responses under each clustering ($L = 5$), with the pooled FE estimate (dashed) for reference. The top four panels are embedding-based groupings; corpus CBOW is omitted here for space and appears in Table 4. The bottom row gives the two groupings that do not use the text—residual k-means and observables k-means, on the 1970 covariates.

see Appendix B.9 for more detail). I use $L = 5$ throughout to match the number of clusters used by the text-based methods and report alternative specifications in Appendix B.5.

To quantify how well each clustering method captures heterogeneity in employ-

ment dynamics, I compare the between-group variance of the 363 unit-specific IRFs across clustering methods in Table 4. I do this in two ways: first, I compute the ratio of between-group variance to total variance, which is a standard measure of how well a grouping explains variation in a variable of interest (column 1). Since the total variance includes estimation error in the unit-specific responses, which is substantial in our application, I also include a version that normalizes the between-group variance relative to a permutation floor, which is the mean between-group variance over 2,000 random relabelings that hold the realized group sizes fixed (column 2). This is the level of between-group variance a random partition of the same sizes would deliver and provides a reference point for evaluating whether a given partition captures real heterogeneity in employment dynamics. The last two columns count how frequently the grouping exceeds the two non-text based benchmarks when I vary the number of clusters $L \in \{2, 3, 4, 5, 6, 8, 10\}$.

Grouping	% Between	Ratio to floor	Exceeds benchmark, of 7 L values	
			Residual k-means	Observables k-means
Corpus SGNS	28.3%	25.7×	7	6
LLM Embeddings	27.1%	24.7×	7	6
SVD-β ($r = 50$)	24.3%	21.8×	7	6
Corpus CBOW	22.5%	20.4×	7	5
Pretrained (Google News)	18.8%	16.8×	6	5
Observables k-means	16.7%	15.5×	7	—
Residual k-means	13.9%	12.6×	—	0

Table 4: Grouping-level summary at $L = 5$, AR(2). "% Between" is the share of total cross-city IRF variance explained by between-group differences in group-mean IRFs. "Ratio to floor" is the between-group variance of the 363 unit-specific impulse responses divided by that grouping's own size-matched permutation floor. The last two columns count how frequently the grouping exceeds each benchmark out of 7 cluster counts we consider. The upper block is the four embedding-based groupings, the lower the two that do not use the text.

At $L = 5$, every grouping separates employment dynamics far more sharply than chance, and the five text-based groupings tend to separate it more sharply than the two benchmarks.[20] Varying L does not change the main finding: The textual

[20]I date the covariates to 1970, so that they are effectively predetermined for a panel beginning in 1969. Repeating the exercise at later vintages raises its ratio monotonically, from 15.5× at 1970 to 18.6× at 2000. Nevertheless four of the five text methods exceed the benchmark at every vintage. I acknowledge that the same objection (not being predetermined) applies to the textual descriptions, which span 1979–2019 and are thus concurrent with the outcome window. This is because it is

descriptions seem to carry real information about employment dynamics, more so than the two benchmarks I consider. Among the text-based methods, pretrained Word2Vec averaging tends to perform worst, while SVD-β, corpus-trained SGNS, and the LLM embeddings tend to perform best, trading top spots at different values of L (cf. Appendix B.5). It is perhaps worth noting that a simple closed-form factorization computed from the corpus in seconds matches a commercial transformer embedding on this task.

In-sample vs Out-of-sample. Our exercise in this section is entirely in-sample.I read the evidence as establishing that narrative text encodes economically meaningful structure about local labor markets—more of it than either benchmark recovers, including the curated covariate set—but not as establishing that text-based peer groups are the preferred way to model this panel for forecasting purposes.

5 Conclusion

Embeddings are ubiquitous in NLP and increasingly used in economics. This paper asks what replacing a text with its embedding preserves, under a generative model in which documents are mixtures of K latent topics.

At the word level, an embedding built from how often words appear together reflects the topic geometry: words that play the same role across topics end up close together. Extra dimensions do no harm, so a researcher need not know how many topics a corpus contains. At the document level, proximity in embedding space is proximity in topic mixture. This means that a cluster of document embeddings corresponds to a set of documents with similar mixtures, and that adjusting for the embedding is equivalent to adjusting for the topic mixture.

Another lesson is that dimension alone is not the relevant property. Two alignments matter: the embedding must recover the text's low-dimensional structure, and that structure must be the one the analysis requires. My theory delivers the first. At the same time, connecting embeddings to the parameters of an underlying topic model makes the second an assumption about the economics rather than the algorithm. Take the control use: the informal premise that "the embedding is a sufficient control" reduces to the transparent condition that the topic mixture captures

challenging to obtain a good textual description of the economy in the 1960s for some of the smaller CBSAs in the sample. Perhaps the modest increase in performance from using later vintages of observables, well short of the margin by which text exceeds the two benchmarks, is reassuring in that regard.

the confounding—an assumption with economic content that a practitioner can defend. Directly assuming that d-dimensional embedding vectors in Euclidean space are a sufficient control is much harder to justify on economic grounds.

In my application, I find that clustering LLM-generated descriptions of 363 U.S. metropolitan areas yields interpretable economic types, and those types separate local employment dynamics more sharply than a curated set of industry and demographic covariates. The comparisons line up with the theory: SVD-β and corpus-trained SGNS, which target closely related population objects, perform similarly, while pretrained averaging, which carries the geometry of a different corpus, trails at almost every number of clusters. Remarkably, our closed-form embedding matches a commercial transformer embedding on this task, at a fraction of the cost and with a guarantee attached.

Several extensions merit investigation. Transformer embeddings currently fall outside our theoretical framework. Characterizing what transformer architectures learn about topic structure—building on Li et al. (2023)—would bring the embeddings practitioners actually use inside the theory. A second extension is dynamic topic models (Blei and Lafferty, 2006): characterizing how embeddings track changes in word meaning over time.

Finally, the pipeline studied here—generate text with a language model, embed it, then use the embedding downstream—retains the representational power of a large model while leaving every downstream step inspectable and reproducible. That division of labour seems a reasonable way to incorporate such models into empirical work.

References

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288. PMLR, 2013.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

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

Philipp Bach, Victor Chernozhukov, Sven Klaassen, Martin Spindler, Jan Teichert-Kluge, and Suhas Vijaykumar. Adventures in demand analysis using AI. *arXiv preprint arXiv:2501.00382*, 2025.

Patrick Bajari, Zhihao Cen, Victor Chernozhukov, Manoj Manukonda, Suhas Vijaykumar, Jin Wang, Ramon Huerta, and Junbo Li. Hedonic prices and quality adjusted price indices powered by AI. *Journal of Econometrics*, 2025.

Laura Battaglia, Timothy Christensen, Stephen Hansen, and Szymon Sacher. Inference for regression with variables generated by AI or machine learning. *arXiv preprint arXiv:2402.15585*, 2024.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. Discretizing unobserved heterogeneity. *Econometrica*, 90(2):625–643, 2022.

Sara Casella, Jesús Fernández-Villaverde, Stephen Hansen, Ryohei Oishi, and Minchul Shin. Structural estimation with unstructured data. Working paper, 2026.

Yinyin Chen, Shishuang He, Yun Yang, and Feng Liang. Learning topic models: Identifiability and finite-sample analysis. *Journal of the American Statistical Association*, pages 1–16, 2022.

Timothy Christensen and Giovanni Compiani. From unstructured data to demand counterfactuals: Theory and practice. *arXiv preprint arXiv:2601.05374*, 2026.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, volume 16, 2003.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. How to make causal inferences using texts. *Science Advances*, 8 (42):eabg2652, 2022.

Simon Freyaldenhoven, Shikun Ke, Dingyi Li, and José Luis Montiel Olea. On the testability of the anchor-words assumption in topic models. 2025.

Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, and Wing-Kin Ma. Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications. *IEEE Signal Process. Mag.*, 36(2):59–80, 2019.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2013.

Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1681–1691, 2015.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, pages 427–431, 2017.

Austin C Kozlowski, Matt Taddy, and James A Evans. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949, 2019.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR, 2014.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27, pages 2177–2185, 2014.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pages 19689–19729. PMLR, 2023.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, 2013b.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010. ELRA.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.

Margaret E. Roberts, Brandon M. Stewart, and Richard A. Nielsen. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4): 887–903, 2020.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Keyon Vafa, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei. CAREER: A foundation model for labor sequence data. *arXiv preprint arXiv:2202.08370*, 2022.

Keyon Vafa, Susan Athey, and David M. Blei. Estimating wage disparities using foundation models. *Proceedings of the National Academy of Sciences*, 2025. doi: 10.1073/pnas.2427298122.

Victor Veitch, Dhanya Sridhar, and David M. Blei. Adapting text embeddings for causal inference. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 919–928, 2020.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36:15614–15638, 2023.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations*, 2016.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *International Conference on Learning Representations*, 2022.