Statistical Foundations of DIME: Risk Estimation for Practical Index Selection

Giulio D'Erasmo[1], Cesare Campagnano[3], Antonio Mallia[3], Pierpaolo Brutti[1], Nicola Tonellotto[2], Fabrizio Silvestri[1]

[1]Sapienza University of Rome, [2]University of Pisa, [3]Seltz, San Francisco, US

Correspondence: g.derasmo@diag.uniroma1.it

Abstract

High-dimensional dense embeddings have become central to modern Information Retrieval, but many dimensions are noisy or redundant. Recently proposed DIME (Dimension IMportance Estimation), provides query-dependent scores to identify informative components of embeddings. DIME relies on a costly grid search to select a priori a dimensionality for all the query corpus's embeddings. Our work provides a statistically grounded criterion that directly identifies the optimal set of dimensions for each query at inference time. Experiments confirm achieving parity of effectiveness and reduces embedding size by an average of $\sim 50\%$ across different models and datasets at inference time[1].

1 Introduction

In Information Retrieval (IR) Dense Retrieval (DIR) represents a significant advancement that is driven by the use of high-dimensional embedding spaces to represent queries and documents. DIR models, for example Devlin et al. (2019) and Khattab and Zaharia (2020), have demonstrated a competitive trade-off between effectiveness and efficiency in numerous benchmarks. However, observations have indicated that not all dimensions are equally useful, and some can even be harmful (Kovaleva et al., 2021; Puccetti et al., 2022; Takeshita et al., 2025).

Research in this area has explored compressing encoder embeddings by projecting them onto a lower-dimensional manifold. For instance, one can use Principal Component Analysis (PCA) or autoencoders designed to learn and represent such a manifold (Liu et al., 2022; Siciliano et al., 2024; Acquavia et al., 2023; Chang et al., 2024). Although these approaches reduce retrieval performance, they do contribute to improved efficiency.

In contrast, Faggioli et al. (2024) theorise the existence of a query-dependent lower-dimensional manifold within the original embedding space, where retrieval performance also increases. Their work introduces Dimension Importance Estimation (DIME), a method that scores each dimension by fusing the embeddings of the query and a relevant document through element-wise multiplication. These scores can then be used to rank dimensions, retaining only those most likely to contribute to effective retrieval. One major drawback of this approach is that the exact number of dimensions to retain cannot be fixed in advance, and previous work often resorts to evaluating a grid of candidate dimensionalities or simply keeping a fixed percentage of dimensions.

The current study addresses this limitation. In Section 3, we prove that DIME estimates the squared latent signal, that is, the true information need underlying the query, which the embedding aims to capture. Unlike prior approaches, which do not specify the number of dimensions in advance, our Risk Dimension Importance Estimation (RDIME) provides a theoretically grounded criterion for identifying the dimensions to be retained. We further generalize DIME using a kernel-based formulation, which allows for rigorous weighted contributions from relevant documents and thereby improves the estimator's robustness. We validate this criterion in Section 5.

2 Background

DIME In recent years, Faggioli et al. (2024) conjectured the Manifold Clustering Hypothesis, which posits that high-dimensional embeddings lie in a query-dependent low-dimensional manifold where retrieval is more effective. To discover this manifold, Dimension Importance Estimation (DIME) assigns importance scores to each embedding dimension using a query-dependent func-

[1]Code available at https://github.com/giulio-derasmo/RDIME

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 722–730

tion $\boldsymbol{u}_q$. The importance scores should estimate the contribution of each dimension to retrieving relevant documents for the given query $\boldsymbol{q}$.

They propose different methods for estimating the importance of dimensions in dense retrieval systems. These methods are based on reinforcing the query signal with a specific document $\boldsymbol{p} \in \mathbb{R}^p$ as:

$$\boldsymbol{u}_q = \boldsymbol{q} \odot \boldsymbol{p}.$$

Two examples are PRF and LLM DIMEs. The first one, inspired by Rocchio (1971) Pseudo-Relevance Feedback (PRF), assumes that the top M documents $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M$ retrieved by a similarity measure such as BM25 (Robertson and Zaragoza, 2009) are likely relevant to the query. We shall refer to them as pseudo-relevant documents. Their centroid is used to reinforce the original signal. The second one, instead, uses a synthetic document generated by a Large Language Model (LLM).

Although PRF can enhance the retrieval effectiveness of IR systems, it suffers from several limitations. D'Erasmo et al. (2025) address this by designing a scoring function that incorporates irrelevant feedback. In parallel work, Campagnano et al. (2025) propose a different weighting scheme to avoid the non-tunable parameter k.

Top-k Thresholding Following DIME, a Top-k thresholding strategy is employed to select the most informative dimensions. Specifically, given the importance scores $\boldsymbol{u}_q$,

$$S = \text{Top-}\,k(\boldsymbol{u}_q) \subseteq \{1, \ldots, p\} \tag{1}$$

where S contains the indices of the k dimensions with highest importance score in $\boldsymbol{u}_q$.

However, k must be fixed globally for all queries and cannot be tuned using a validation set, limiting the method's adaptability.

Modulation Estimators The recovery of true signals from noisy observations is a central problem in sparse and redundant representations theory (Beran and Dümbgen, 1998; Elad, 2010). One approach to address this challenge is the modulation estimators framework. In this framework, a random vector $\boldsymbol{X} \in \mathbb{R}^p$ represents the noisy observation, where the dimensions X_i are independent with $\mathbb{E}[X_i] = \xi_i$ and $\text{Var}(X_i) = \sigma^2$ for every $i \in \{1, \ldots, p\}$. Given a modulator function $(f_i)_{i=1}^p \in [0,1]^p$, the goal is to produce a linear estimator with components $f_i X_i$ for the corresponding true signal components ξ_i, such that the ℓ^2 risk, $\mathbb{E}\big[\sum_{i=1}^p (\xi_i - f_i X_i)^2\big]$, is minimised.

Under the assumption that only a few dimensions carry signal information, Donoho and Johnstone (1994) proposed the hard-threshold estimator, defined component-wise as:

$$f_i X_i = \mathbf{1}_S(X_i) X_i, \quad S = \{x \in \mathbb{R} \,:\, |x| > \lambda\}$$

for some threshold $\lambda > 0$.

We shall show that DIME can be interpreted as a modulation estimator of the query's latent signal, which represents a user's information need.

3 Method

In this section, we formalize the problem of estimating latent information needs from noisy query embeddings. We introduce a hard thresholding approach that retains only the most informative components, providing a practical strategy for denoising queries.

3.1 IR for Modulation Estimators

Let $\boldsymbol{\theta} \in \mathbb{R}^p$ denote the query's latent signal, or in other words, the user's information need. The query embedding $\boldsymbol{q} \in \mathbb{R}^p$, which is obtained through an encoder, can be modeled as

$$\boldsymbol{q} = \boldsymbol{\theta} + \varepsilon \boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(0, I_p),$$

where ε controls the noise level and $\boldsymbol{z}$ is standard Gaussian noise. The query embedding is then interpreted as a noisy sample of the underlying information need.

Theorem 1 (Hard Thresholding Estimator)**.** *Given the noisy embedding $\mathbf{q}$, the optimal hard-threshold estimator $\hat{\boldsymbol{\theta}}(S^\star)$ of the latent signal $\boldsymbol{\theta}$ under ℓ_2 risk is given by:*

$$\hat{\theta}_i(S^\star) = q_i \mathbf{1}_{S^\star}(i) \quad \forall i \in \{1, \ldots, p\},$$

where

$$S^\star = \{i \in \{1, \ldots, p\} \mid \theta_i^2 > \varepsilon^2\}. \tag{2}$$

Proof. We want to find the best subset $S^\star$ that minimises the ℓ_2 risk function:

$$\begin{aligned} S^\star &= \operatorname*{argmin}_S \sum_{i=1}^p \mathbb{E}\big[\hat{\theta}_i(S) - \theta_i\big]^2 \\ &= \operatorname*{argmin}_S \Big\{ \sum_{i \in S} \mathbb{E}\big[q_i - \theta_i\big]^2 + \sum_{i \notin S} \mathbb{E}\big[\theta_i^2\big] \Big\} \\ &= \operatorname*{argmin}_S \Big\{ |S|\varepsilon^2 + \sum_{i \notin S} \theta_i^2 \Big\} \end{aligned}$$

Each coordinate i contributes ε^2 to the risk if retained, and θ_i^2 if discarded. Therefore, to minimise the risk, we should include i in $S^\star$ if and only if $\theta_i^2 > \varepsilon^2$, which yields (2). □

In order to be applied, Theorem 1 requires knowledge of the set $S^\star$ defined in Eq. 2, which depends on the unknown true signal $\boldsymbol{\theta}$. We show in the next section that DIME serves as an estimator of $\boldsymbol{\theta}^2$ and it can then be directly applied in Eq. 2.

3.2 DIME as an Estimator of Squared Latent Signal

Let the query $\boldsymbol{q} \in \mathbb{R}^p$ and $\mathcal{D}^M = \{\boldsymbol{d}^{(i)}\}_{i=1}^M \subset \mathbb{R}^p$ its top-M (pseudo) relevant documents. According to the Probability Ranking Principle (Robertson, 1977), an IR system should rank documents in decreasing order of relevance to the query. Here, we interpret relevance as inversely related to variance: highly relevant documents are modeled as having lower noise. In comparison, less relevant documents have higher noise. Formally, we model each document $\boldsymbol{d}^{(i)}$ as:

$$\boldsymbol{d}^{(i)} = \boldsymbol{\theta} + \sigma_i \boldsymbol{z}^{(i)}, \quad \boldsymbol{z}^{(i)} \overset{i.i.d}{\sim} \mathcal{N}(0, I_p),$$

where $\sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_k$ reflects increasing uncertainty (or decreasing relevance) as the rank i increases. By construction, higher-ranked documents are more likely to be truly relevant, and thus are distributed near the latent information need $\boldsymbol{\theta}$.

To formalize DIME as a statistical estimator with a controlled bias-variance trade-off, we introduce a kernel function $K(\boldsymbol{q}, \boldsymbol{d}^{(i)})$ that measures query documents similarity. This kernel determines how much each document contributes to the final estimate, with more similar documents receiving higher weight. This formulation allows us to generalize standard DIME variants within a unified framework called Kernel DIME.

Definition 1 (Kernel DIME)**.** *Kernel DIME is the solution of the local kernel-weighted least squares problem:*

$$\min_{\boldsymbol{u} \in \mathbb{R}^p} \sum_{i=1}^M w_i \left\| \boldsymbol{d}^{(i)} \odot \boldsymbol{q} - \boldsymbol{u} \right\|_2^2.$$

where the weights $w_1, \ldots, w_k$ are given by

$$w_i = \frac{K(\boldsymbol{q}, \boldsymbol{d}^{(i)})}{\sum_{j=1}^M K(\boldsymbol{q}, \boldsymbol{d}^{(j)})},$$

and $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ is a kernel function which measures the similarity between the query $\boldsymbol{q}$ and the documents $\boldsymbol{d}^{(i)}$.

In this case, Kernel DIME has a closed-form solution, which generalises DIMEs as:

$$\boldsymbol{u}_q = \boldsymbol{q} \odot \sum_{i=1}^M w_i \boldsymbol{d}^{(i)}. \tag{3}$$

Several meaningful choices of weights recover existing DIME variants: (i) PRF DIME, where $w_i = 1/k$; and (ii) SWC DIME, where for every dimension i, $w_i = \mathrm{softmax}(\boldsymbol{s})_i$; where $\boldsymbol{s} = \left[\boldsymbol{d}^{(i)} \boldsymbol{q}\right]_{i=1}^M$.

Theorem 2. *Kernel DIME with uniform weights $w_i = 1/k$ is an unbiased estimator of $\boldsymbol{\theta}^2$ for each component.*

Proof. Using independence between $\boldsymbol{q}$ and $\boldsymbol{d}^{(i)}$s, we have:

$$\begin{aligned}\mathbb{E}[(\boldsymbol{u}_q)_j] &= \sum_{i=1}^M \frac{1}{M} \mathbb{E}\left[d_j^{(i)} q_j\right] \\ &= \sum_{i=1}^M \frac{1}{M} \mathbb{E}[d_j^{(i)}] \mathbb{E}[q_j] = \theta_j^2.\end{aligned}$$

□

While uniform weighting yields unbiasedness, the choice of kernel influences the bias–variance trade-off of Kernel DIME as an estimator of $\boldsymbol{\theta}^2$. Appendix A.1 discusses strategies for selecting kernel weights to control document noise aggregation. It should be noted that we have not established formal theoretical guarantees for all kernel functions. Nevertheless, our experimental results in Section 5 suggest that several kernel choices perform well empirically.

Corollary 1 (RDIME: Risk Dimension Importance Estimation)**.** *Given a query $\boldsymbol{q}$, we can use the Kernel DIME representation $\boldsymbol{u}_q$, as defined in Eq. 3, for the Hard Thresholding Estimator described in Eq. 2:*

$$\begin{cases} \hat{S} = \{i \in \{1, \ldots, p\} \mid (\boldsymbol{u}_q)_i > \hat{\varepsilon}^2\} \\ \hat{\varepsilon}^2 = \frac{1}{p} \sum_{i=1}^p (q_i^2 - (\boldsymbol{u}_q)_i). \end{cases}$$

This corollary demonstrates that the hard thresholding estimator can be implemented directly from DIME scores. In contrast Eq. 1, which must explore a grid of candidate dimensionalities to determine how many dimensions to retain, RDIME provides a statistically grounded criterion for directly identifying the optimal dimension set. Crucially, this selection is performed in a query-dependent manner, allowing us to adapt the dimensionality per query rather than relying on a single fixed value for the entire collection (see Appendix A.2).

Model	Filter	DL '19					DL '20					RB '04				
		0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$
ANCE	$\boldsymbol{u}^{LLM}$	0.570	0.660	**0.663**	0.656* (0.94)	-1.05	0.533	0.622	**0.658**	0.651* (0.93)	-1.06	0.257	0.381	**0.392**	0.387* (0.94)	-1.27
	$\boldsymbol{u}^{PRF}$	0.552	0.640	0.657	0.650* (0.93)	-1.06	0.541	0.616	0.650	0.645* (0.93)	-0.92	0.269	0.357	0.383	0.386* (0.92)	0.78
	$\boldsymbol{u}^{SWC}$	0.558	0.645	0.653	0.652* (0.93)	-0.15	0.546	0.616	0.651	0.645* (0.93)	-0.77	0.265	0.361	0.381	0.384* (0.92)	0.79
	Baseline	-	-	-	0.645 (1.0)	1.04	-	-	-	0.646 (1.0)	-0.19	-	-	-	0.384 (1.0)	-0.08
Contriever	$\boldsymbol{u}^{LLM}$	0.736	0.733	**0.745**	0.741* (0.59)	-0.54	0.695	0.697	0.689	0.700* (0.63)	0.43	**0.527**	0.519	0.518	0.524* (0.57)	-0.57
	$\boldsymbol{u}^{PRF}$	0.684	0.692	0.689	0.695* (0.53)	0.43	0.701	**0.704**	0.693	0.687 (0.55)	-2.41	0.489	0.491	0.492	0.481 (0.46)	-2.23
	$\boldsymbol{u}^{SWC}$	0.678	0.683	0.683	0.682* (0.52)	-0.146	0.690	0.691	0.687	0.688* (0.54)	-0.43	0.496	0.495	0.495	0.494* (0.44)	-0.40
	Baseline	-	-	-	0.677 (1.0)	0.87	-	-	-	0.666 (1.0)	3.24	-	-	-	0.48 (1.0)	2.79
TAS-B	$\boldsymbol{u}^{LLM}$	0.763	0.758	0.757	**0.766*** (0.54)	0.39	0.697	0.696	0.705	0.702* (0.60)	-0.42	0.466	0.469	0.467	**0.472*** (0.54)	0.64
	$\boldsymbol{u}^{PRF}$	0.737	0.738	0.735	0.738* (0.51)	0.00	0.705	0.712	0.706	0.707* (0.52)	-0.70	0.442	0.448	0.448	0.444* (0.43)	-0.89
	$\boldsymbol{u}^{SWC}$	0.729	0.732	0.728	0.726* (0.51)	-0.68	0.712	0.717	0.712	**0.720*** (0.52)	0.42	0.432	0.442	0.447	0.436 (0.44)	-2.46
	Baseline	-	-	-	0.719 (1.0)	1.02	-	-	-	0.685 (1.0)	5.09	-	-	-	0.428 (1.0)	1.87

Table 1: Comparison of nDCG@10 between RDIME and fixed Top-k thresholding strategies ($k \in \{0.4, 0.6, 0.8\}$), across different query sets and DIR models. $\Delta(\%)$ represents the relative improvement of RDIME over the best performing Top-k strategy. The value in parentheses indicates the average proportion of dimensions retained by RDIME. **Bold** indicates the best result per configuration. The superscript indicates that our criterion is not statistically significantly different from Top-k thresholding methods.

4 Experimental Section

We follow prior work in DIME (D'Erasmo et al., 2025; Campagnano et al., 2025; Faggioli et al., 2024) to set up our experimental evaluation.

Datasets. We evaluate on three passage retrieval benchmarks: TREC Deep Learning 2019 (Craswell et al., 2020, DL '19), TREC Deep Learning 2020 (Craswell et al., 2021, DL '20), and the Deep Learning Hard set (Mackie et al., 2021, DL HD). We assess out-of-domain robustness on TREC Robust 2004 collection (Voorhees, 2004, RB '04).

Models. Experiments use three 768-dimensional DIR models: ANCE (Xiong et al., 2021), Contriever (Izacard et al., 2021), and TAS-B (Hofstätter et al., 2021). We shall refer to these models at full dimensionality as Baseline.

Evaluation. We report mean Average Precision (AP) and nDCG@10. Statistical significance is tested using a paired Student's t-test (Student, 1908) (or one-sided Wilcoxon signed-rank (Wilcoxon, 1945) if non-normal) at level $\alpha = 0.05$ using Holm–Bonferroni corrections.

DIMEs. To validate our criterion, we use three DIME variants: PRF-DIME, LLM-DIME, and SWC-DIME. The hyperparameters are kept fixed, since the goal is not to tune performance but to demonstrate that our method correctly identifies the optimal number of dimensions. We set $M = 2$ for PRF-DIME and $M = 10$ for SWC-DIME.

5 Results

Table 1 reports nDCG@10 scores for DIME variants across collections and bi-encoders, comparing Top-k thresholding ($k \in \{0.4, 0.6, 0.8\}$, as per existing literature), with our proposed RDIME method. The last column shows the improvement of our method over the best Top-k baseline. The value in parentheses indicates the average proportion of dimensions retained by RDIME.

RDIME adapts dimensionality per query without hyperparameter tuning, whereas Top-k methods (see 1) require a grid search to set the threshold. Since no validation set is available, this tuning is often done on the test set, and the optimal k varies widely across collections and encoders. For example, $k = 0.8$ in one setup and $k = 0.4$ in another, leading to unpredictable performance.

Our criterion achieves performance that are not statistically different from the baselines in most settings. Where differences occur, they are small (0.15% on DL '19 to 2.46% on RB '04), demonstrating robust performance without hyperparameter tuning. Further results are provided in Appendix A.2 (effects of query-specific dimension selection) and A.3 (AP and full dimensionality comparisons).

Taken together, these findings demonstrate that our theoretical criterion enables adaptive, query-specific dimensionality reduction without validation-set tuning, while maintaining performance competitive with prior method.

6 Conclusion

In this work, we introduced RDIME, a statistical criterion for query-dependent dimension selection in dense retrieval. Unlike previous approaches, which rely on grid search to set a single global dimensionality and cannot adapt at inference time, our method estimate the optimal dimensions for each query. We tested this framework using several DIMEs and observed consistent improvements in

retrieval effectiveness, accompanied by substantial reductions in dimensionality.

This work opens the door to designing new DIMEs variants that leverage kernel functions to better structure the embedding space.

7 Limitations

We acknowledge two main limitations. First, our experimental evaluation focuses primarily on Standard DIMEs (PRF and LLM DIMEs), which can be seen as a special case of Kernel DIME. While Section 5 demonstrates the effectiveness of SWC DIME, the performance of other kernel functions (e.g., Radial Basis Function, Sigmoid) remains to be evaluated empirically.

Second, our theoretical analysis establishes unbiasedness only for uniform weights (Theorem 2). Extending these guarantees to general kernel functions presents additional challenges. In particular, when weights depend on both the query and retrieved documents, they become random variables rather than fixed constants. Analyzing the statistical properties of Kernel DIME under such random weighting schemes, requires a more sophisticated theoretical treatment and is left for future work. Appendix A.1 provides preliminary guidance for kernel selection under the Probability Ranking Principle.

References

Antonio Acquavia, Craig Macdonald, and Nicola Tonellotto. 2023. Static pruning for multi-representation dense retrieval. In *Proceedings of the ACM Symposium on Document Engineering 2023*, DocEng '23, New York, NY, USA. Association for Computing Machinery.

Rudolf Beran and Lutz Dümbgen. 1998. Modulation of estimators and confidence sets. *Annals of Statistics*, pages 1826–1856.

Cesare Campagnano, Antonio Mallia, and Fabrizio Silvestri. 2025. Unveiling dime: Reproducibility, generalizability, and formal analysis of dimension importance estimation for dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3367–3376, New York, NY, USA. Association for Computing Machinery.

Xuejun Chang, Debabrata Mishra, Craig Macdonald, and Sean MacAvaney. 2024. Neural passage quality estimation for static pruning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 174–185, New York, NY, USA. Association for Computing Machinery.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *Preprint*, arXiv:2003.07820.

Giulio D'Erasmo, Giovanni Trappolini, Fabrizio Silvestri, and Nicola Tonellotto. 2025. Eclipse: Contrastive dimension importance estimation with pseudo-irrelevance feedback for dense retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 147–154, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

David L Donoho and Iain M Johnstone. 1994. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455.

Michael Elad. 2010. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media.

Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. 2024. Dimension importance estimation for dense information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1318–1328, New York, NY, USA. Association for Computing Machinery.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Preprint*, arXiv:2104.06967.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Zhenghao Liu, Han Zhang, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Xiaohua Li. 2022. Dimension reduction for efficient dense retrieval via conditional autoencoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5692–5698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iain Mackie, Jeffrey Dalton, and Andrew Yates. 2021. How deep is your learning: the dl-hard annotated deep learning dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2335–2341, New York, NY, USA. Association for Computing Machinery.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1286–1304, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Stephen E Robertson. 1977. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304.

J.J. Rocchio. 1971. *Relevance Feedback in Information Retrieval*. Prentice Hall, Englewood Cliffs, New Jersey.

Federico Siciliano, Francesca Pezzuti, Nicola Tonellotto, and Fabrizio Silvestri. 2024. Static pruning in dense retrieval using matrix decomposition. *Preprint*, arXiv:2412.09983.

Student. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25.

Sotaro Takeshita, Yurina Takeshita, Daniel Ruffinelli, and Simone Paolo Ponzetto. 2025. Randomly removing 50% of dimensions in text embeddings has minimal impact on retrieval and classification tasks. *Preprint*, arXiv:2508.17744.

Ellen M. Voorhees. 2004. Overview of the trec 2004 robust track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD. NIST Special Publication 500-261, National Institute of Standards and Technology (NIST).

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

A Appendix

A.1 Optimal weights under heteroskedastic noise in IR

We aim to show that, as $M \to \infty$, the Kernel DIME estimator is not a consistent estimator, meaning that it does not converge in probability to the true unknown value $\boldsymbol{\theta}^2$ as the number of documents grows. Nevertheless, its Mean Squared Error (MSE) can be reduced through an appropriate choice of the weighting scheme, i.e, by selecting an optimal kernel function. For the sake of the proof, we assume that each weight w_i is deterministic.

First, we can express the Kernel DIME explicitly as follows:

$$\begin{aligned}(\boldsymbol{u}_q)_j &= q_j \sum_{i=1}^{M} w_i d_j^{(i)} \\ &= (\theta_j + \varepsilon z_j)\Big[\sum_{i=1}^{M} w_i(\theta_j + \sigma_i z_j^{(i)})\Big] \\ &= \theta_j^2 + \theta_j \sum_{i=1}^{M} \sigma_i w_i z_j^{(i)} + \varepsilon\theta_j z_j \\ &\quad + \varepsilon z_j \sum_{i=1}^{M} \sigma_i w_i z_j^{(i)}.\end{aligned}$$

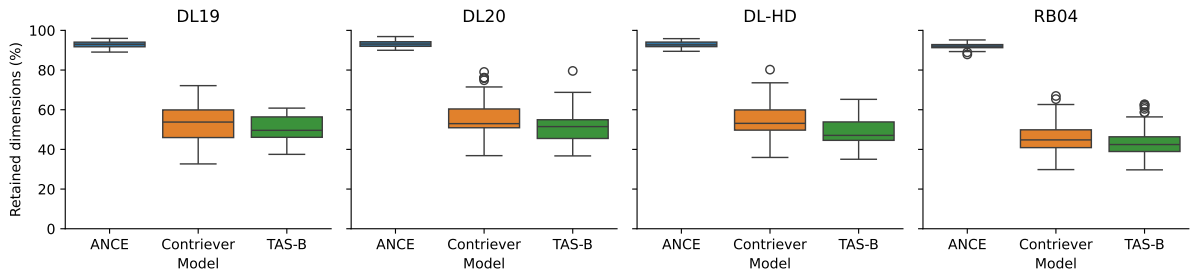
This allows us to plug the estimator into the MSE and, following the same reasoning as in Theorem 2, simplify the terms in the MSE, since the weights w_i are constant. Hence,

$$\begin{aligned}\mathrm{MSE}\Big((\boldsymbol{u}_q)_j\Big) &= \mathbb{V}\Big((\boldsymbol{u}_q)_j\Big) + \Big(\mathbb{E}[(\boldsymbol{u}_q)_j] - \theta_j^2\Big)^2 \\ &= \mathbb{E}\Big[\Big((\boldsymbol{u}_q)_j - \theta_j^2\Big)^2\Big] + 0 \\ &= \mathbb{E}\Big[\Big(\theta_j\eta_{w,j} + \varepsilon\theta_j z_j + \varepsilon z_j\eta_{w,j}\Big)^2\Big]\end{aligned}$$

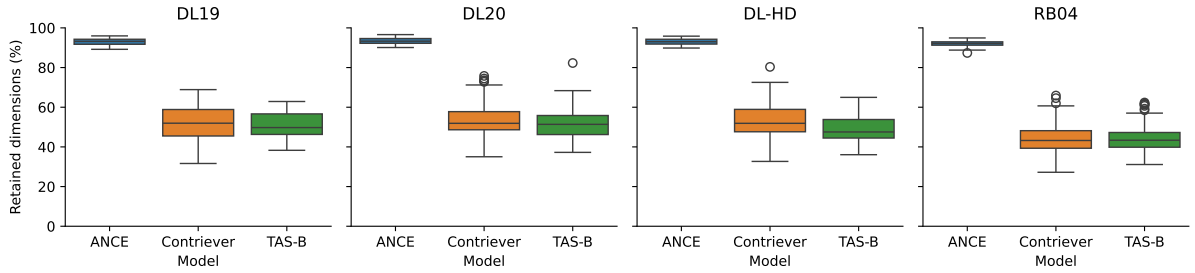
where $\eta_{w,j} = \sum_{i=1}^{M} \sigma_i w_i \mathbf{z}_j^{(i)}$. Since $\mathbf{z}_j$ is independent of each $\mathbf{z}_j^{(i)}$, we have

$$\mathrm{cov}(z_j, \eta_{w,j}) = \sum_{i=1}^{M} \sigma_i w_i \,\mathrm{cov}(z_j, z_j^{(i)}) = 0.$$

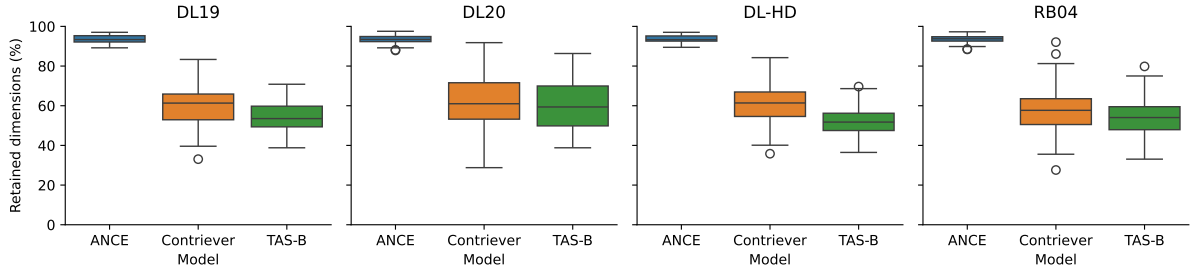
Furthermore, because $(\mathbf{z}_j, \eta_{w,j})$ is jointly gaussian, zero covariance implies independence. Consequently, $\mathbf{z}_j$ is independent from $\eta_{w,j}$. By the



(a) PRF@2 DIME



(b) SWC@10 DIME



(c) LLM DIME

Figure 1: Percentage of dimensions retained per query across three bi-encoders (ANCE, Contriever, TAS-B) and four collections (DL '19, DL '20, DL HD, RB '04) using our risk-based criterion with three DIME variants. Boxplots show median, interquartile range, and outliers, revealing substantial query-dependent variation.

standard property that measurable functions of independent random variables remain independent, it follows that $\mathbf{z}_j$ is also independent of $\eta_{w,j}^2$. Therefore,

$$\mathbb{E}[z_j \eta_{w,j}^2] = \mathbb{E}[z_j]\mathbb{E}[\eta_{w,j}^2] = 0.$$

An identical argument shows that z_j^2 is independent of $\eta_{w,j}$. The same for z_j^2 and $\eta_{w,j}^2$.

Expanding the squared term in the MSE and removing the cross product using the argument we prove before,

$$\begin{aligned}
\mathrm{MSE}\Big((\boldsymbol{u}_q)_j\Big) &= \theta_j^2 \mathbb{E}[\eta_{w,j}^2] + \varepsilon^2 \theta_j^2 \mathbb{E}[z_j^2] \\
&\quad + \varepsilon^2 \mathbb{E}[z_j^2 \eta_{w,j}^2] \\
&= \theta_j^2 \sum_{i=1}^{M} \sigma_i^2 w_i^2 + \varepsilon^2 \theta_j^2 \\
&\quad + \varepsilon^2 \mathbb{E}[z_j^2]\mathbb{E}[\eta_{w,j}^2] \\
&= (\theta_j^2 + \varepsilon^2) \sum_{i=1}^{M} \sigma_i^2 w_i^2 + \varepsilon^2 \theta_j^2.
\end{aligned}$$

We can see that if we let the kernel weights to be uniform, the series $\lim_{M\to\infty} \sum_{i=1}^{M} \frac{\sigma_i^2}{M^2}$ does not converge.

Instead we can minimize with respect to the

weights resolving the optimization problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w}) = \min_{\boldsymbol{w}\in\mathbb{R}^p} \sum_{i=1}^{M} \sigma_i^2 w_i^2$$

restricted on $E = \{\boldsymbol{w} \in \mathbb{R}^M \mid w_i \geq 0 \ \ \forall i \ , \sum_{i=1}^{M} w_i = 1\}$. This is a well-posed problem, so we can derive the solution using the Method of Lagrange Multipliers having

$$w_i^\star = \frac{\sigma_i^{-2}}{\sum_{j=1}^{M} \sigma_j^{-2}}.$$

With these weights, we conclude that for large M the $\text{MSE}\Big((\boldsymbol{u}_q)_j\Big) \to \varepsilon^2\theta_j^2$.

The optimal weights $w_i \propto \frac{1}{\sigma_i^2}$ demonstrate that effective DIME estimation requires inverse-variance weighting: documents with lower noise variance should contribute more heavily to the final estimate. However, practical implementations must operate without direct access to the true noise variances σ_i^2.

This theoretical result provides a justification for the weighting scheme employed in SWC DIME. In this variant, documents with higher similarity scores to the query receive greater weight. Under the Probability Ranking Principle, higher-ranked (more similar) documents are assumed to have lower noise variance. Therefore, the SWC DIME weighting scheme can be interpreted as an approximation to inverse-variance weighting, where similarity scores serve as a proxy for the unobserved inverse noise variances.

A.2 Importance of Query-Wise Dimension Selection

To observe how important it is to select a unique set of dimensions for each query, Figure 1 presents the distribution of retained dimensions across different bi-encoders and query sets when applying our criterion to each DIME variant.

The boxplots reveal substantial variation in the optimal proportion of retained dimensions across queries, as reflected by the wide interquartile ranges and presence of outliers across all datasets and models. This demonstrates that query-dependent selection is necessary: while some queries require high dimensionality to preserve effectiveness, others can be adequately represented using approximately half the original dimensions.

The results reveal distinct behaviors across encoder models. ANCE consistently retains the highest proportion of dimensions (typically >90%) regardless of dataset or DIME variant, suggesting that its representations distribute information more uniformly across the embedding space. In contrast, Contriever and TAS-B are more prone to dimensional reduction, with the majority of queries retaining between 30% and 70% of dimensions. This indicates that these models encode information more sparsely, concentrating signal in a subset of dimensions.

A.3 Full Comparison

In Table 2 we show the complete results for all the collections, namely DL '19, DL '20, DL HD and RB '04, along with performance using AP e nDCG@10 metrics. We compare Top-k thresholding ($k \in \{0.4, 0.6, 0.8\}$, as per existing literature), with our proposed RDIME method. The last column shows the improvement of our method over the best Top-k baseline. The value in parentheses indicates the average proportion of dimensions retained by RDIME.

Our criterion achieves performances that are not statistically different from baselines in most settings. We notice that we can see a more consistent improvement in AP, proving the robustness of our method. In nDCG@10, where the differences occur, they are small.

In general, RDIME can be successfully used as a replacement for DIME-based methods, with the advantage of being hyperparameter-agnostic.

Model	Filter	DL '19										DL '20									
		AP					nDCG@10					AP					nDCG@10				
		0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$
ANCE	$\boldsymbol{u}^{LLM}$	0.267	0.351	**0.370**	0.367 (0.94)	-1.0	0.570	0.660	**0.663***	0.656 (0.94)	-1.19	0.281	0.372	**0.397**	0.395 (0.93)	-0.35	0.533	0.622	**0.658**	0.651* (0.93)	-1.0
	$\boldsymbol{u}^{PRF}$	0.253	0.339	**0.370**	0.368 (0.93)	-0.54	0.552	0.640	0.657	0.650* (0.93)	-1.04	0.287	0.364	0.390	0.393 (0.93)	0.92	0.541	0.616	0.651	0.645* (0.93)	-0.94
	$\boldsymbol{u}^{SWC}$	0.255	0.340	**0.370**	0.369 (0.93)	-0.14	0.558	0.645	0.653	0.652* (0.93)	-0.09	0.289	0.366	0.390	0.392 (0.93)	0.67	0.546	0.616	0.65	0.645* (0.93)	-0.72
	Baseline	-	-	-	0.361 (1.0)	2.22	-	-	-	0.645 (1.0)	1.04	-	-	-	0.392 (1.0)	0.15	-	-	-	0.646 (1.0)	-0.19
Contriever	$\boldsymbol{u}^{LLM}$	0.523	0.521	0.519	**0.526** (0.59)	0.65	0.736	0.733	**0.745**	0.741* (0.59)	-0.51	0.500	**0.505**	0.501	0.503 (0.63)	-0.28	0.695	0.696	0.689	0.700* (0.63)	0.52
	$\boldsymbol{u}^{PRF}$	0.507	0.511	0.509	0.513 (0.53)	0.35	0.684	0.692	0.689	0.695* (0.53)	0.38	0.489	0.497	0.495	0.495 (0.55)	-0.42	0.701	**0.704**	0.693	0.687 (0.55)	-2.4
	$\boldsymbol{u}^{SWC}$	0.512	0.513	0.509	0.514* (0.52)	0.21	0.678	0.683	0.683	0.682* (0.52)	-0.06	0.497	0.495	0.495	0.497 (0.54)	0.04	0.69	0.691	0.687	0.688 (0.54)*	-0.48
	Baseline	-	-	-	0.494 (1.0)	4.03	-	-	-	0.677 (1.0)	0.87	-	-	-	0.478 (1.0)	3.85	-	-	-	0.666 (1.0)	3.24
TAS-B	$\boldsymbol{u}^{LLM}$	**0.527**	0.52	0.514	0.526 (0.54)	-0.25	0.763	0.758	0.757	**0.766*** (0.54)	0.42	0.495	0.494	0.495	**0.497** (0.6)	0.34	0.697	0.696	0.705	0.702* (0.6)	-0.38
	$\boldsymbol{u}^{PRF}$	0.509	0.512	0.507	0.511 (0.51)	-0.06	0.737	0.738	0.735	0.738* (0.51)	-0.01	0.486	0.489	0.489	0.490 (0.52)	0.12	0.705	0.712	0.706	0.707* (0.52)	-0.67
	$\boldsymbol{u}^{SWC}$	0.508	0.506	0.503	0.508 (0.51)	-0.08	0.729	0.732	0.728	0.726* (0.51)	-0.78	0.490	0.492	0.492	0.494 (0.52)	0.24	0.712	0.717	0.712	**0.720*** (0.52)	0.46
	Baseline	-	-	-	0.476 (1.0)	6.68	-	-	-	0.719 (1.0)	1.02	-	-	-	0.476 (1.0)	3.81	-	-	-	0.685 (1.0)	5.09

Model	Filter	DL HD										RB '04									
		AP					nDCG@10					AP					nDCG@10				
		0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$	0.4	0.6	0.8	RDIME	$\Delta(\%)$
ANCE	$\boldsymbol{u}^{LLM}$	0.125	0.172	0.186	0.183* (0.94)	-1.35	0.253	0.329	0.346	0.336* (0.94)	-2.66	0.082	0.137	0.148	**0.149*** (0.94)	0.61	0.257	0.381	**0.392**	0.387* (0.94)	-1.28
	$\boldsymbol{u}^{PRF}$	0.128	0.176	0.18	0.183* (0.93)	1.39	0.278	0.340	0.335	0.331* (0.93)	-2.53	0.084	0.134	0.148	**0.149*** (0.92)	0.47	0.269	0.357	0.383	0.386* (0.92)	0.57
	$\boldsymbol{u}^{SWC}$	0.125	0.174	0.184	**0.185*** (0.93)	0.22	0.27	**0.342**	0.339	0.336* (0.93)	-1.87	0.083	0.135	0.147	**0.149*** (0.92)	1.57	0.265	0.361	0.381	0.384* (0.92)	0.95
	Baseline	-	-	-	0.18 (1.0)	2.38	-	-	-	0.334 (1.0)	0.51	-	-	-	0.146 (1.0)	1.98	-	-	-	0.384 (1.0)	-0.08
Contriever	$\boldsymbol{u}^{LLM}$	0.259	**0.262**	0.261	0.259* (0.61)	-1.14	0.376	0.39	0.392	0.387* (0.61)	-1.07	**0.263**	0.261	0.259	**0.263*** (0.57)	0.08	**0.527**	0.519	0.518	0.524* (0.57)	-0.4
	$\boldsymbol{u}^{PRF}$	0.254	0.252	0.251	0.257* (0.54)	0.9	**0.393**	0.387	0.384	**0.393*** (0.54)	-0.05	0.254	0.256	0.257	0.253* (0.46)	-1.4	0.489	0.491	0.492	0.481 (0.46)	-2.4
	$\boldsymbol{u}^{SWC}$	0.253	0.248	0.249	0.25* (0.53)	-0.99	0.387	0.383	0.382	0.383* (0.53)	-0.98	0.256	0.257	0.258	0.257* (0.44)	-0.43	0.496	0.495	0.495	0.494* (0.44)	-0.4
	Baseline	-	-	-	0.241 (1.0)	3.73	-	-	-	0.375 (1.0)	2.27	-	-	-	0.239 (1.0)	7.67	-	-	-	0.48 (1.0)	2.79
TAS-B	$\boldsymbol{u}^{LLM}$	0.261	**0.265**	0.260	0.264* (0.52)	-0.19	0.402	**0.408**	0.395	0.405* (0.52)	-0.71	0.214	0.218	0.218	0.218* (0.54)	-0.09	0.466	0.469	0.467	**0.472*** (0.54)	0.62
	$\boldsymbol{u}^{PRF}$	0.243	0.250	0.256	0.238* (0.49)	-6.77	0.383	0.388	0.394	0.382* (0.49)	-3.05	0.219	**0.224**	0.222	0.220 (0.43)	-1.74	0.442	0.448	0.448	0.444* (0.43)	-0.87
	$\boldsymbol{u}^{SWC}$	0.243	0.243	0.241	0.245* (0.49)	0.78	0.389	0.388	0.381	0.393* (0.49)	1.0	0.212	0.217	0.216	0.213 (0.44)	-1.8	0.432	0.442	0.447	0.436* (0.44)	-2.46
	Baseline	-	-	-	0.238 (1.0)	2.85	-	-	-	0.374 (1.0)	5.05	-	-	-	0.197 (1.0)	7.85	-	-	-	0.428 (1.0)	1.87

Table 2: Comparison of retrieval effectiveness between RDIME and fixed Top-k thresholding strategies ($k \in \{0.4, 0.6, 0.8\}$), across different query sets and DIR models. $\Delta(\%)$ represents the relative improvement of RDIME over the best performing Top-k strategy. The value in parentheses indicates the average proportion of dimensions retained by RDIME. **Bold** indicates the best result per configuration. The superscript indicates that our criterion is not statistically significantly different from Top-k thresholding methods.