

Transformers as Provable Approximators of Sparse Principal Component Analysis

Qinyan Liu¹ Yan Chen² Canhong Wen³

¹School of the Gifted Young, University of Science and Technology of China (USTC)

²Dept. of Statistics and Actuarial Science, HKU

³Dept. of Statistics and Finance, USTC

Paper on
OpenReview



Background and Motivation

Transformers can provably implement statistical learning algorithms in context [1], and recent work shows that multi-layer Transformers can emulate spectral algorithms like PCA [2].

However, existing work largely considers non-sparse settings and often relies on ground-truth or algorithm-generated labels for pre-training, even for unsupervised tasks. This motivates us to investigate constructive approximation under **sparsity constraints** and learning directly from **unlabeled data**.

We study this through **Sparse Principal Component Analysis (SPCA)** [3], which enables interpretable dimensionality reduction through sparse feature selection. We formulate it sequentially as:

$$\hat{\mathbf{v}}_j \in \arg \max_{\|\mathbf{v}\|_2=1, \|\mathbf{v}\|_0 \leq r} \mathbf{v}^\top \mathbf{A}_j \mathbf{v}, \quad j \in [k].$$

Transformer for SPCA

TPower as constructive template. For component i , one of τ TPower iterations consists of: (Trunc _{r} : top- r hard-thresholding)

$$\mathbf{u}_i^{(t)} = \mathbf{A}_i \mathbf{v}_i^{(t-1)}, \quad \hat{\mathbf{u}}_i^{(t)} = \text{Trunc}_r(\mathbf{u}_i^{(t)}), \quad \mathbf{v}_i^{(t)} = \hat{\mathbf{u}}_i^{(t)} / \|\hat{\mathbf{u}}_i^{(t)}\|_2.$$

Iterative deflation. After extracting $\mathbf{v}_i = \mathbf{v}_i^{(\tau)}$, we update

$$\mathbf{A}_1 = \mathbf{A}, \quad \mathbf{A}_{i+1} = (\mathbf{I} - \mathbf{v}_i \mathbf{v}_i^\top) \mathbf{A}_i (\mathbf{I} - \mathbf{v}_i \mathbf{v}_i^\top),$$

and repeat the same procedure to recover the leading k PCs.

Transformer realization. Auxiliary tokens act as memory for the covariance matrix, initializations, and intermediate iterates. Stacking the following modules emulates the full TPower procedure.

Multi-head ReLU attention $\Rightarrow \mathbf{X} \mathbf{X}^\top, \mathbf{A}_i \mathbf{v}_i^{(t-1)}$, normalization, deflation

Multi-head ReLU attention and ReLU MLP $\Rightarrow \text{Trunc}_r(\cdot)$.

End-to-end unsupervised training. The model is trained directly with an SPCA objective based on explained variance, sparsity, and orthogonality, without pre-training labels.

Main Theoretical Results

Setting and assumptions. We consider the noisy covariance model $\mathbf{A} = \mathbf{A}^* + \mathbf{E}$, where the noise is controlled by the eigengap.

The theoretical analysis requires assumptions of aligned initialization, a restricted sparse eigengap, and a truncation margin under controlled noise.

Constructive approximation theorem. There exists a multi-layer ReLU Transformer under architecture bounds emulating τ TPower iterations for k PCs, such that

$$\|\hat{\mathbf{v}}_j - \mathbf{v}_j^*\|_2 \lesssim \underbrace{\varepsilon_{\text{trunc}} + \varepsilon_{\text{norm}}}_{\text{Transformer approximation}} + \underbrace{\mu^\tau}_{\text{finite iterations}} + \Delta_{\text{noise}}, \quad \mu \in (0, 1).$$

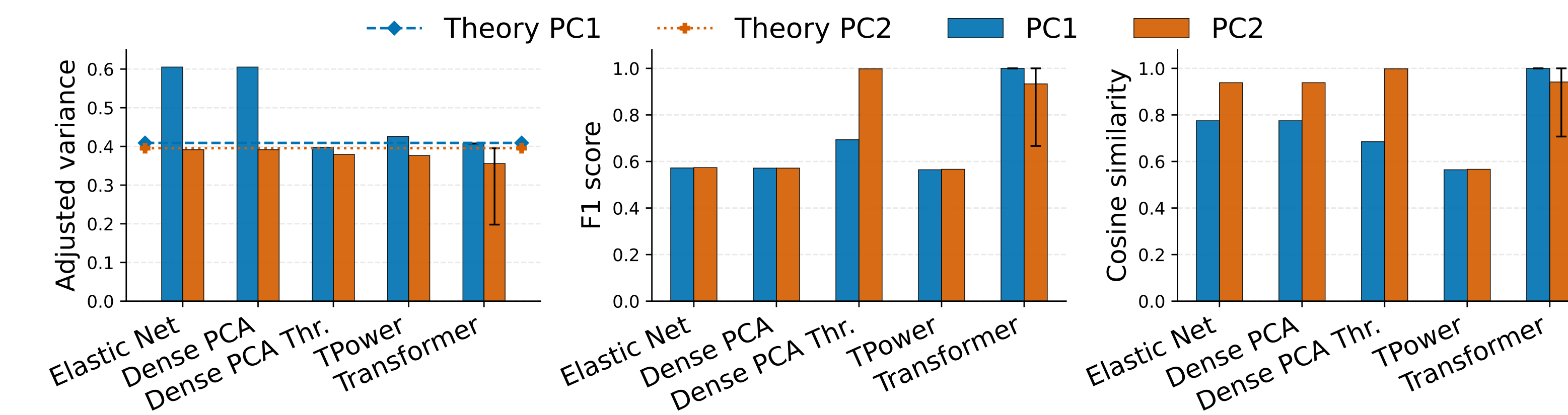
Thus, the iterative error decreases geometrically to a noise-dependent floor.

Generalization guarantee. Let $\mathcal{E}_{\text{pop}}$ denote the expected squared subspace error. Then

$$\mathcal{E}_{\text{pop}} \lesssim \varepsilon_{\text{app}} + \varepsilon_{\text{stat}} + \tilde{O}\left(\sqrt{\frac{P_\Theta}{n}}\right) + \varepsilon_{\text{opt}}.$$

Here P_Θ measures model complexity. Balancing approximation and complexity yields $\tilde{O}(n^{-2/(q+4)})$ when $P_\Theta(\varepsilon) \lesssim \varepsilon^{-q}$.

Synthetic Experiments

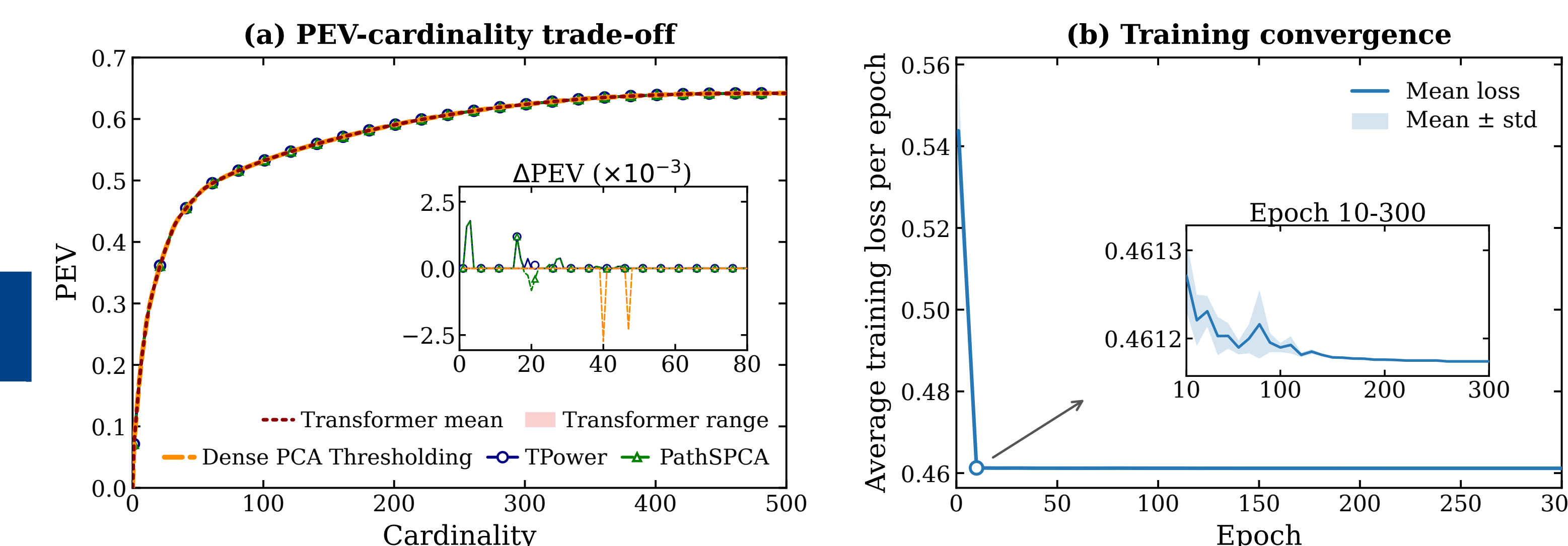


Sparse spiked synthetic data. For both PCs, Transformer SPCA

achieves near-population adjusted variance and **leading performance** of support F1 and cosine similarity.

Means and min-max ranges over five model seeds.

Real-data Experiments



Colon Cancer gene-expression data. $N = 62$ samples and $d = 2000$ genes (top 500 selected by variance). The Transformer **closely matches** classical SPCA baselines along the PEV-cardinality curve and exhibits **stable** unsupervised training.

Takeaways

- **Unsupervised:** Transformers trained directly with an SPCA objective without labeled pre-training.
- **Provable:** Constructed Transformers emulate TPower, with recovery and generalization bounds under noise.
- **Efficient:** competitive recovery with batched inference.

References:

- [1] Bai et al., *Transformers as Statisticians*, NeurIPS, 2023.
- [2] He et al., *Learning Spectral Methods by Transformers*, arXiv:2501.01312, 2025.
- [3] Zou et al., *Sparse Principal Component Analysis*, JCGS, 2006.
- [4] Yuan and Zhang, *Truncated Power Method for Sparse Eigenvalue Problems*, JMLR, 2013.

Main Contributions

- **Unsupervised learning.** We construct and train Transformers that emulate the Truncated Power Method (TPower) [4] for SPCA, without relying on pre-training labels.
- **Statistical guarantees.** We establish approximation and high-probability generalization bounds, separating noise, sampling, model capacity, and optimization effects.
- **Empirical validation.** Transformers extract PCs competitively on synthetic and high-dimensional gene-expression datasets, while enabling efficient batched inference.

Method Overview

