An Empirical Null for Contrastive Principal Component Analysis

Muhammad Salman Kabir
Plekhanov Russian University of Economics
Moscow, Russia

Artem A. Badarin
Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology
Lytkino, Russia

Elena N. Pitsik
Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology
Lytkino, Russia

Vladimir S. Khorev
Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology
Lytkino, Russia

Semen A. Kurkin
Federal Research and Clinical Center of Intensive Care Medicine and Rehabilitology
Lytkino, Russia

2026 10th Scientific School Dynamics of Complex Networks and their Applications (DCNA) | 979-8-3195-2599-4/26/$31.00 ©2026 IEEE | DOI: 10.1109/DCNA72004.2026.11733414

***Abstract*—Contrastive principal component analysis (cPCA) uncovers structure enriched in a foreground dataset relative to a background dataset. The method, however, returns components whether or not such structure exists, and nothing in its output distinguishes the two cases. Here we use the background dataset as an empirical null distribution. A subset of the background, treated as a pseudo-foreground, contains no foreground-specific structure, so refitting cPCA on many such subsets gives a null distribution for its output. We used this empirical null to obtain a criterion for selecting the contrast parameter α, to test whether foreground-specific structure is present, to obtain a calibrated component spectrum and to perform a per-sample salience test. On the corrupted MNIST dataset, the global test returned the smallest attainable p-value and the per-sample salience test achieved an AUC above 0.999 with a false-positive rate of 4.7% against a nominal 5%.**

***Index Terms*—Contrastive principal component analysis, dimensionality reduction, empirical null distribution, parallel analysis, resampling, case–control studies**

I. Introduction

Principal component analysis (PCA) finds directions of maximum variance in a dataset [1]. This is a limitation when the variance of interest is not the largest. In case–control studies, for instance, patients and controls share many sources of variance, e.g., anatomy, acquisition protocols and batch effects. PCA applied to the pooled data recovers this shared structure rather than the difference between groups. This situation is typical of studies of functional brain networks [2], [3]. In major depressive disorder (MDD), for example, reported connectivity alterations depend on the study, the preprocessing pipeline and the network thresholding [4]–[6], and identifying the features that drive multivariate classification remains an open problem [7], [8].

Contrastive PCA (cPCA) addresses this problem using a second dataset [9]. Given a foreground X_f and a background X_b in the same feature space, cPCA takes the leading eigenvectors of $C_f - \alpha C_b$, where C_f and C_b are the sample covariances and $\alpha \geq 0$ controls how strongly background directions are suppressed. cPCA has been applied to gene expression, single-cell and neuroimaging data [10], and contrastive learning has revealed salient features of resting-state EEG networks in autism spectrum disorder [11].

Subsequent work on cPCA has focused on the choice of α [10]. Because C_b is positive semidefinite, every eigenvalue of $C_f - \alpha C_b$ is non-increasing in α; the eigenvalues therefore change scale with α, are not comparable across its values and cannot be used to select it. Abid *et al.* [9] suggested computing subspaces over a logarithmic grid of α, clustering them by principal angle and presenting a few representatives. Probabilistic cPCA formulates the problem through a contrastive likelihood, which guides the choice of the parameter [12]. Generalized contrastive PCA (gcPCA) removes the parameter by replacing the difference $C_f - \alpha C_b$ with a ratio objective [13].

cPCA, however, returns components whether or not the foreground contains structure that the background lacks. Given two datasets that differ in no relevant way, it still returns a subspace with large leading eigenvalues. Three questions therefore arise in practice: (i) whether any foreground-specific structure is present, (ii) how many components carry it and (iii) which foreground samples express it. We refer to question (i) as the global test. Hawke *et al.* [14] addressed (i) and (ii) with a contrastive bootstrap test and an estimator of the contrastive dimension; both rely on intrinsic-dimension estimates for each dataset, and the test is conservative by design.

To answer these questions, we compare the output of cPCA with its output on data that contain no foreground-specific structure, and we obtain this reference (null) distribution from the background dataset itself. If foreground and background observations are exchangeable, then a subset of the background treated as a pseudo-foreground contains no foreground-specific structure, and refitting cPCA on many such subsets gives an empirical null distribution for its output. Our approach follows Horn's parallel analysis [15], [16], which retains PCA components whose eigenvalues exceed those obtained from data with no structure, and resembles the permutation nulls used for group comparisons of brain networks [17]. It requires

Authorized licensed use limited to: Kant Baltic Federal Univ. Downloaded on October 23,2026 at 15:23:24 UTC from IEEE Xplore. Restrictions apply.

neither intrinsic-dimension estimates nor a parametric model.

Our work aimed to construct an empirical null for cPCA by resampling the background, to use it to select α and to test whether foreground-specific structure is present (global test), and to use it to describe the component spectrum and to test individual foreground samples for salience.

II. Methodology

A. Contrastive PCA

Let $X_f \in \mathbb{R}^{n_f \times d}$ and $X_b \in \mathbb{R}^{n_b \times d}$ be the foreground and background datasets with separately centered sample covariances C_f and C_b. cPCA computes the eigendecomposition of

$$C(\alpha) = C_f - \alpha C_b \tag{1}$$

and retains the leading eigenvectors.

B. Construction of the null

We fixed a grid $\mathcal{A}$ of values of α and a number k of components. In each of R replicates we drew two disjoint index sets $\mathcal{I}_f$ and $\mathcal{I}_b$ from the background, of sizes n_f and n_b', and computed the k leading eigenvalues of $\hat{C}[X_b(\mathcal{I}_f)] - \alpha\hat{C}[X_b(\mathcal{I}_b)]$ for every $\alpha \in \mathcal{A}$. Because each pseudo-background contained only n_b' samples, the observed spectrum was also computed with C_b estimated from a random subset of n_b' background samples, so that observed and null statistics share the same sample sizes.

The two index sets must not exhaust the background. An exhaustive split fixes the pooled covariance of the two sides, so variance assigned to the pseudo-foreground is absent from the pseudo-background. Conditionally on the observed background, the two estimates become anti-correlated and the null eigenvalues come out too small. The real foreground originates outside the background and has no such dependence. We therefore left at least n_f background samples out of both sides in each replicate.

Standardizing the observed eigenvalues against the null gives

$$Z_j(\alpha) = \frac{\lambda_j(\alpha) - \mu_j(\alpha)}{\sigma_j(\alpha)}, \tag{2}$$

where μ_j and σ_j are the mean and standard deviation of the null eigenvalues at that α.

C. Contrast parameter α and global test

The observed and null values of $\lambda_1(\alpha)$ both change with α. Their standardized difference $Z_1(\alpha)$, however, has zero mean and unit variance under the null at every α and is therefore comparable across values of α. We therefore selected

$$\alpha^\star = \arg\max_{\alpha \in \mathcal{A}} Z_1(\alpha). \tag{3}$$

Because $\alpha^\star$ was obtained by searching the grid, the global test at $\alpha^\star$ must account for that search. We therefore compared the observed maximum with the null distribution of the same maximized statistic,

$$p_{\text{global}} = \frac{1 + \left|\{r : \max_\alpha Z_1^{(r)}(\alpha) \geq Z_1(\alpha^\star)\}\right|}{R+1}, \tag{4}$$

where $Z_1^{(r)}$ is (2) evaluated on replicate r and the unit terms keep the p-value valid for a finite number of replicates [18]. This tests the null hypothesis that the foreground contains no contrastive structure beyond that expected under exchangeability with the background.

D. Component spectrum

Z_j grows with sample size: σ_j is the sampling standard deviation of the null eigenvalues and shrinks approximately as $n^{-1/2}$, whereas a fixed excess $\lambda_j - \mu_j$ does not. For large n, every component with a positive excess therefore becomes significant. The quantile criterion, which retains component j if $\lambda_j(\alpha^\star)$ exceeds a fixed upper quantile of its null distribution and is the direct analogue of Horn's test, behaves in the same way. Neither criterion therefore gives a usable cutoff in this setting. Instead, we reported the relative excess

$$E_j = \frac{\lambda_j(\alpha^\star) - \mu_j(\alpha^\star)}{\lambda_1(\alpha^\star) - \mu_1(\alpha^\star)} \tag{5}$$

and the same excess normalized by its total over the k components examined,

$$Q_j = \frac{\lambda_j(\alpha^\star) - \mu_j(\alpha^\star)}{\sum_{\ell=1}^{k}\left(\lambda_\ell(\alpha^\star) - \mu_\ell(\alpha^\star)\right)}. \tag{6}$$

E. Per-sample salience

Beyond the global test, we also tested individual foreground samples for salience. Let V contain the first k components at $\alpha^\star$, and define the salience score $s(x) = \|V^\top(x - \bar{x}_b)\|^2$, where $\bar{x}_b$ is the background mean. A background subset X_b^{ref}, held out from all fitting, provides the reference distribution

$$p_i = \frac{1 + |\{x \in X_b^{\text{ref}} : s(x) \geq s(x_i)\}|}{|X_b^{\text{ref}}| + 1}. \tag{7}$$

The test thus treats salient samples as outliers relative to the background, similarly to outlier-based seizure detection from EEG [19], [20].

Because V is fitted to the foreground, the samples used to fit it have inflated projections onto it by construction. To avoid this bias, we split the foreground into halves, fitted V on one half and scored the other, and then reversed the roles of the two halves.

III. Evaluation

We evaluated the proposed method on a corrupted MNIST dataset [21], constructed similarly to that of Abid *et al.* [9]. Foreground images were MNIST digits 0–2 superimposed on images of grass, and images of grass alone formed the background. The foreground contained $n_f = 3000$ images: 900 per digit and 300 grass-only images included to evaluate the per-sample salience test. The background contained $n_b = 13500$ images. All images were flattened to $d = 784$. Digit and grass labels were used for evaluation only and were not available to the procedure.

We projected both datasets onto the leading 200 principal components of the background alone, which retained 77.6% of the background variance. We used $R = 500$ replicates, $k = 50$

Authorized licensed use limited to: Kant Baltic Federal Univ. Downloaded on October 23,2026 at 15:23:24 UTC from IEEE Xplore. Restrictions apply.

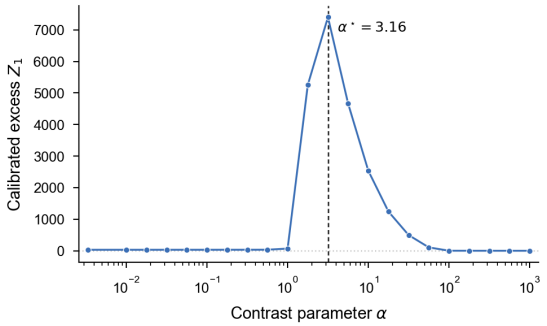


Fig. 1. **Contrast parameter selection.** Standardized excess $Z_1(\alpha)$ over the grid ($\alpha = 0$ is plotted at the left edge). Z_1 is standardized against the null, so the null lies at 0 ± 1 by construction (dotted line). The peak is at $\alpha^\star = 3.162$.

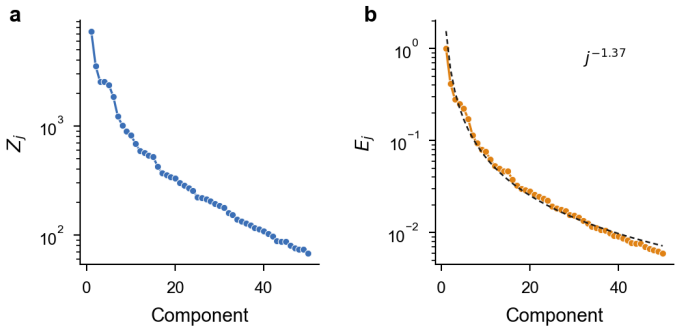


Fig. 2. **Component spectrum at $\alpha^\star$.** (a) Z_j on a logarithmic scale; it remains above 67 for all 50 components. (b) E_j on a logarithmic scale with the fitted power law $j^{-1.37}$ ($R^2 = 0.987$, dashed).

components and the grid $\mathcal{A} = \{0\} \cup \{10^{-2}, 10^{-1.75}, \ldots, 10^{3}\}$, i.e., zero and 21 logarithmically spaced values.

IV. Results and Discussion

A. Contrast parameter and global test

Fig. 1 shows $Z_1(\alpha)$ over the grid. The curve was flat for $\alpha \leq 1$, rose sharply, peaked at $\alpha^\star = 3.162$ and returned to the null level for $\alpha \geq 10^2$. At $\alpha^\star$ the leading eigenvalue exceeded its null mean by 2.48, and none of the 500 null replicates reached the observed maximized statistic. The resulting $p_{\text{global}} = 1/501 \approx 0.002$ is the smallest value attainable with 500 replicates. These results indicate the presence of foreground-specific structure.

B. Component spectrum

Fig. 2 shows the component spectrum at $\alpha^\star$. Z_j remained above 67 for all 50 components, so both the standardized and the quantile criteria retained every component tested: with $n_f = 3000$, the null spread is small and any positive excess is significant.

E_j decayed from 1.000 to 0.006 over the 50 components and followed a power law, $E_j \propto j^{-1.37}$ ($R^2 = 0.987$). All values of E_j were positive. The contrastive excess is therefore distributed over many dimensions without a sharp cutoff. The cumulative shares Q_j were 28% for the leading component, 48% for the first three, 67% for the first six and 77% for the first ten.

Repeating the analysis with 100 instead of 200 retained principal components gave shares of 29%, 49%, 68% and 79%, an exponent of -1.50 instead of -1.37 and $\alpha^\star = 1.778$, the adjacent grid value.

C. Per-sample salience test

Table I compares the per-sample results with the withheld labels. Of the 2700 digit-bearing images, 2699 were flagged as salient at $p < 0.05$. Of the 300 grass-only images, 14 were flagged, which gives a false-positive rate of 4.7%. The single unflagged digit-bearing image had $p \approx 0.2$ (Fig. 4), so the separation was not perfect, but the AUC exceeded 0.999. Fig. 3 shows the two groups in the salient subspace, and Fig. 4 shows the per-sample p-values.

TABLE I
Per-sample salience against withheld labels

Foreground images	n	Flagged	Rate
Digit-bearing	2700	2699	99.96%
Grass-only	300	14	4.7%

A test that merely separated the foreground from the background would flag all 3000 images, whereas a test detecting nothing would flag approximately 150. The observed count of 2713 matched the 2715 expected from complete detection with nominal error control on the grass-only images. Grass-only images were present in the half of the foreground used to estimate the subspace and were nevertheless not flagged in the other half. These results indicate that the split-half scoring separates samples within the foreground rather than separating the foreground from the background as a whole.

D. Implications for neurophysiological data

The target application is case–control analysis of functional brain networks, e.g., in MDD [6], [7], [22] and autism spectrum disorder [11], where a per-sample salience test can indicate which patients express group-specific structure. The same procedure applies to EEG, MEG and fNIRS recordings during cognitive and motor tasks [23]–[28], in which task recordings or one group of participants can serve as the foreground and baseline recordings or a reference group as the background.

V. Conclusion

We showed that the background dataset used by contrastive PCA can also serve as an empirical null for its output. Resampling the background as a pseudo-foreground and refitting the method gives this null directly, without a parametric model for the eigenvalues. On the corrupted MNIST dataset, the null selected the contrast parameter α and confirmed the presence of foreground-specific structure. The per-sample salience test recovered all but one digit-bearing image while flagging grass-only images at 4.7% against a nominal 5%. The calibrated spectrum remained informative where the standardized and quantile criteria retained every component tested. The null requires only an estimator that is cheap to refit, so it should

Authorized licensed use limited to: Kant Baltic Federal Univ. Downloaded on October 23,2026 at 15:23:24 UTC from IEEE Xplore. Restrictions apply.

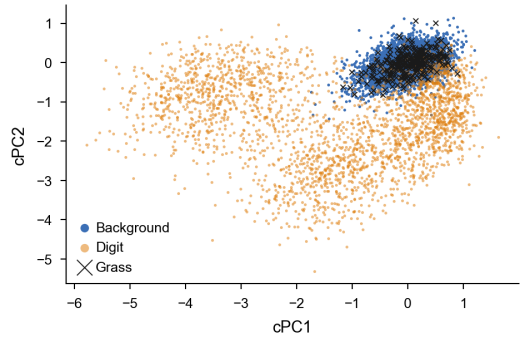


Fig. 3. **Salient subspace.** Background images (blue) and foreground images projected onto the first two components at $\alpha^\star$. Grass-only foreground images (crosses) remain within the background cloud, whereas digit-bearing images (orange) spread along both components.

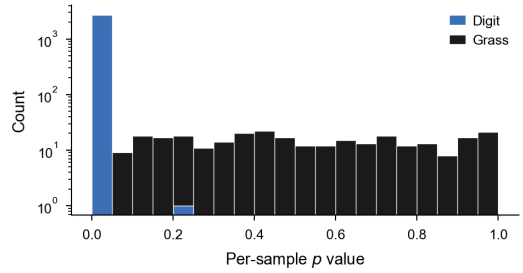


Fig. 4. **Per-sample p-values for the 3000 foreground images.** Digit-bearing images (blue) and grass-only images (black), stacked; the labels were withheld from the method. Counts are on a logarithmic scale. Of the 2700 digit-bearing images, 2699 fall below $p = 0.05$. The 300 grass-only images are approximately uniformly distributed, with 14 below $p = 0.05$.

also apply to hyperparameter-free contrastive decompositions such as gcPCA. We believe that our work could contribute to the ongoing efforts to develop interpretable contrastive methods for high-dimensional data.

References

[1] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Comput. Geosci.*, vol. 19, no. 3, pp. 303–342, Mar. 1993.

[2] E. Bullmore and O. Sporns, "Complex brain networks: Graph theoretical analysis of structural and functional systems," *Nat. Rev. Neurosci.*, vol. 10, no. 3, pp. 186–198, 2009.

[3] A. E. Hramov, N. S. Frolov, V. A. Maksimenko, S. A. Kurkin, V. B. Kazantsev, and A. N. Pisarchik, "Functional networks of the brain: From connectivity restoration to dynamic integration," *Phys.-Usp.*, vol. 64, no. 6, pp. 584–616, 2021.

[4] R. H. Kaiser, J. R. Andrews-Hanna, T. D. Wager, and D. A. Pizzagalli, "Large-scale network dysfunction in major depressive disorder: A meta-analysis of resting-state functional connectivity," *JAMA Psychiatry*, vol. 72, no. 6, pp. 603–611, 2015.

[5] V. Borchardt *et al.*, "Preprocessing strategy influences graph-based exploration of altered functional networks in major depression," *Hum. Brain Mapp.*, vol. 37, no. 4, pp. 1422–1442, 2016.

[6] A. N. Pisarchik *et al.*, "Topology switching during window thresholding fMRI-based functional networks of patients with major depressive disorder: Consensus network approach," *Chaos*, vol. 33, no. 9, p. 093122, 2023.

[7] A. V. Andreev, S. A. Kurkin, D. Stoyanov, A. A. Badarin, R. Paunova, and A. E. Hramov, "Toward interpretability of machine learning methods for the classification of patients with major depressive disorder based on functional network measures," *Chaos*, vol. 33, no. 6, p. 063140, 2023.

[8] E. N. Pitsik *et al.*, "The topology of fMRI-based networks defines the performance of a graph neural network for the classification of patients with major depressive disorder," *Chaos Solitons Fractals*, vol. 167, p. 113041, 2023.

[9] A. Abid, M. J. Zhang, V. K. Bagaria, and J. Zou, "Exploring patterns enriched in a dataset with contrastive principal component analysis," *Nat. Commun.*, vol. 9, no. 1, p. 2134, May 2018.

[10] S. Hawke, E. Zhang, J. Chen, and D. Li, "Contrastive dimension reduction: A systematic review," *WIREs Comput. Stat.*, vol. 18, no. 2, p. e70065, Jun. 2026.

[11] M. S. Kabir, S. Kurkin, G. Portnova, O. Martynova, Z. Wang, and A. Hramov, "Contrastive machine learning reveals in EEG resting-state network salient features specific to autism spectrum disorder," *Chaos Solitons Fractals*, vol. 185, p. 115123, 2024.

[12] D. Li, A. Jones, and B. Engelhardt, "Probabilistic contrastive dimension reduction for case-control study data," *Ann. Appl. Stat.*, vol. 18, no. 3, pp. 2207–2229, Sep. 2024.

[13] E. F. De Oliveira, P. Garg, J. Hjerling-Leffler, R. Batista-Brito, and L. Sjulson, "Identifying patterns differing between high-dimensional datasets with generalized contrastive PCA," *PLoS Comput. Biol.*, vol. 21, no. 2, p. e1012747, Feb. 2025.

[14] S. Hawke, Y. Ma, and D. Li, "Contrastive dimension reduction: When and how?" in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 37, 2024, pp. 74 034–74 057.

[15] J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185, Jun. 1965.

[16] A. Buja and N. Eyuboglu, "Remarks on parallel analysis," *Multivariate Behav. Res.*, vol. 27, no. 4, pp. 509–540, 1992.

[17] A. Zalesky, A. Fornito, and E. T. Bullmore, "Network-based statistic: Identifying differences in brain networks," *NeuroImage*, vol. 53, no. 4, pp. 1197–1207, 2010.

[18] B. Phipson and G. K. Smyth, "Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn," *Stat. Appl. Genet. Mol. Biol.*, vol. 9, no. 1, p. Article 39, 2010.

[19] O. E. Karpov *et al.*, "Extreme value theory inspires explainable machine learning approach for seizure detection," *Sci. Rep.*, vol. 12, no. 1, p. 11474, 2022.

[20] V. V. Grubov *et al.*, "Two-stage approach with combination of outlier detection method and deep learning enhances automatic epileptic seizure detection," *IEEE Access*, vol. 12, pp. 122 168–122 182, 2024.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] D. Stoyanov *et al.*, "Resting-state functional connectivity impairment in patients with major depressive episode," *Int. J. Environ. Res. Public Health*, vol. 19, no. 21, p. 14045, 2022.

[23] M. V. Khramova *et al.*, "Monitoring the cortical activity of children and adults during cognitive task completion," *Sensors*, vol. 21, no. 18, p. 6021, 2021.

[24] S. A. Kurkin, V. V. Grubov, V. A. Maksimenko, E. N. Pitsik, M. V. Khramova, and A. E. Hramov, "System for monitoring and adjusting the learning process of primary schoolchildren based on the EEG data analysis," *Inf. Upravl. Sist. [Inf. Control Syst.]*, no. 5, pp. 50–61, 2020.

[25] S. Kurkin, A. Hramov, P. Chholak, and A. Pisarchik, "Localizing oscillatory sources in a brain by MEG data during cognitive activity," in *Proc. 4th Int. Conf. CINE*, 2020, pp. 1–4.

[26] A. A. Badarin *et al.*, "Brain compensatory mechanisms during the prolonged cognitive task: fNIRS and eye-tracking study," *IEEE Trans. Cogn. Dev. Syst.*, vol. 17, no. 2, pp. 303–314, 2024.

[27] A. Badarin *et al.*, "Psychophysiological parameters predict the performance of naive subjects in sport shooting training," *Sensors*, vol. 23, no. 6, p. 3160, 2023.

[28] S. Kurkin *et al.*, "Characteristics of brain functional networks specific for different types of tactile perception," *Eur. Phys. J. Spec. Top.*, pp. 1–6, 2023.

Authorized licensed use limited to: Kant Baltic Federal Univ. Downloaded on October 23,2026 at 15:23:24 UTC from IEEE Xplore. Restrictions apply.