Whitening transformation

Shaghayegh Sadeghi
Supervisor: Dr. Jianguo Lu

School of Computer Science
University of Windsor

March 3, 2024

What is Whitening

- Whitening is a linear transformation.
- Transformation is a fancy word for function: it gets some inputs (vectors) and returns some output (vectors).
  - Transformation: $Z = W(X - \mu)$
    - The mean is centred at the origin,
    - Covariances are eliminated,
    - The variance is normalized to an identity matrix.
    - This transformation is achieved using a matrix W with unit diagonal “white” covariance $var(Z) = I$ [4, 7, 3]

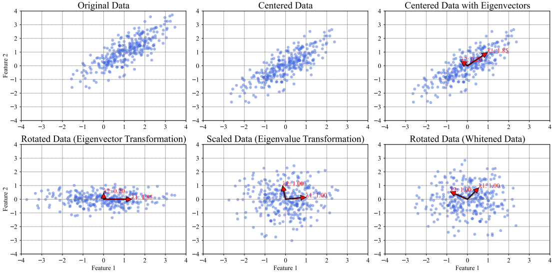


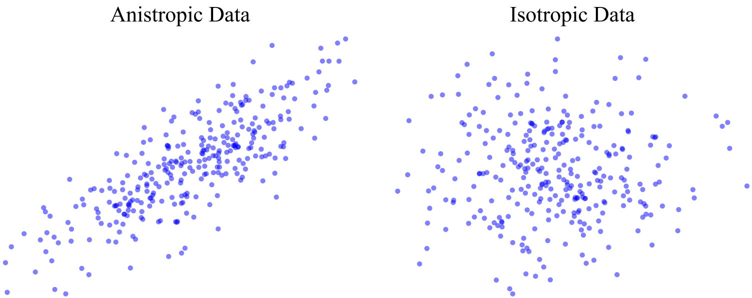
Figure: The Step-by-step ZCA Whitening on a Correlated Anisotropic Data

Why Whitening: Anisotropy and Feature Correlation

- A data is not isotropic if: their variance is not uniformly distributed across dimensions
- Anisotropy is the property of being directionally dependent, as opposed to isotropy, which means homogeneity in all directions.
- In NLP, anisotropy confines the embeddings within a restricted area of the vector space, known as a "narrow cone" [5, 2, 1].

$\Downarrow$ Solution

Isotropization methods (Whitening)



Mahalonobis Distance

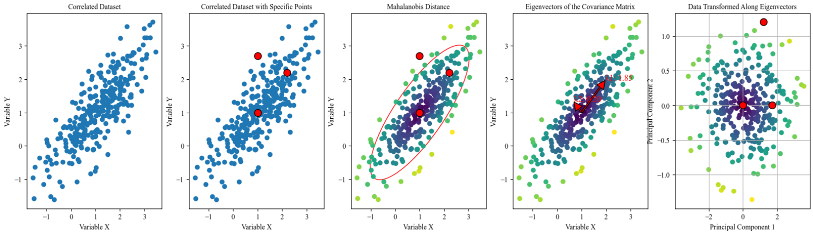


Figure: Mahalonobis Distance

$$D_M(\mathbf{p}) = \sqrt{(\mathbf{p} - \mu)^T \mathbf{S}^{-1} (\mathbf{p} - \mu)}$$

- D_M is the Mahalanobis distance from p to the mean of the distribution D,
- p is the point (a vector) for which the distance from the distribution is being calculated,
- mu is the mean of the distribution D
- $S^{-1}$is the inverse of the covariance matrix S of the distribution D.

University of Windsor

ZCA Whitening

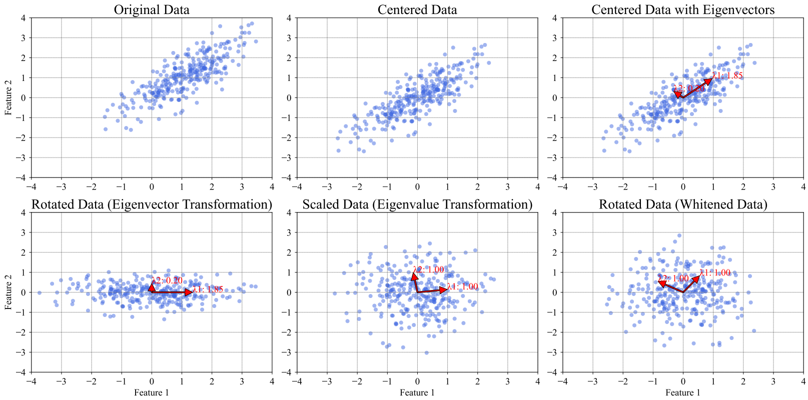


Figure: The Step-by-step ZCA Whitening on a Correlated Anisotropic Data

Whitening Algorithm

Algorithm Whitening Operations

Input: Embeddings $\{x_i\}_{i=1}^{N}$
Output: Transformed embeddings $\{\tilde{x}_i\}_{i=1}^{N}$
Compute the mean μ of $\{x_i\}_{i=1}^{N}$
Compute the covariance matrix Σ of $\{x_i\}_{i=1}^{N}$
Compute the correlation matrix P of $\{x_i\}_{i=1}^{N}$
Perform $U, \Lambda, U^T = \text{SVD}(\Sigma)$
Perform $V, \Theta, V^T = \text{SVD}(P)$
Perform $LL^T = \text{Chol}(\Sigma^{-1})$
Transform $\tilde{x}_i = (x_i - \mu)W$ using eq. 1

$$W = \begin{cases} U\Lambda^{-\frac{1}{2}}U^T & \text{ZCA} \\ U\Lambda^{-\frac{1}{2}} & \text{PCA} \\ L^T & \text{Cholskey} \\ V\Theta^{-\frac{1}{2}}V^T & \text{ZCA-Cor} \\ V\Theta^{-\frac{1}{2}} & \text{PCA-Cor} \end{cases} \tag{1}$$

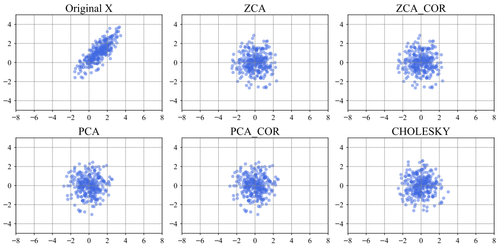


Figure: Different Whitenings

Eigendecomposition

- Eigendecomposition is a form of matrix decomposition where a given matrix A can be factorized in terms of its eigenvalues and eigenvectors. For a general matrix A, the eigendecomposition is given by:

$$A = Q\Lambda Q^{-1} \tag{2}$$

- when A is symmetric:

$$A = Q\Lambda Q^{T} \tag{3}$$

- When dealing with covariance matrices:

$$\Sigma = U\Lambda U^{T} \tag{4}$$

SVD

- SVD is applicable to any $m \times n$ matrix, not just square matrices.
- Although a covariance matrix is always a square matrix. It decomposes a matrix Σ into three matrices:

$$A = USV^T \tag{5}$$

- For a symmetric matrix like the covariance matrix Σ, SVD and eigendecomposition yield similar structures:

$$\Sigma = USU^T \tag{6}$$

- S corresponds to the diagonal matrix of singular values.

University of Windsor

Eigendecomposition VS. SVD

$$
\begin{aligned}
\Sigma &= X^T X && \text{(Definition of } \Sigma\text{)} && (7)\\
&= V S^T U^T (U S V^T) && \text{(Substitute } X = USV^T\text{)} && (8)\\
&= V S^T (U^T U) S V^T && \text{(Associate } U^T U \text{ (identity))} && (9)\\
&= V S^T I S V^T && \text{(Since } U^T U = I\text{)} && (10)\\
&= V S^T S V^T && \text{(Simplify } S^T I S = S^T S\text{)} && (11)\\
&= V S^2 V^T && \text{(Since } S^T = S \text{ (diagonal matrix))} && (12)
\end{aligned}
$$

Now, if we simply apply SVD on the original matrix instead of the covariance matrix $X = USV^T$, we can use the V^T and S matrix as follows:

$$\Sigma = (V^T)^T S^2 V^T \tag{13}$$

Eigendecomposition VS. SVD

- It is important to note that eigenvectors are not unique.
- Eigenvectors are determined only up to a scalar multiple. If u is an eigenvector of a matrix Σ, then any scalar multiple of u(i.e., cu, where c is a non-zero scalar) is also an eigenvector of Σ. This is because eigenvectors are directions in a vector space, and scaling doesn't change the direction(Figure 5).

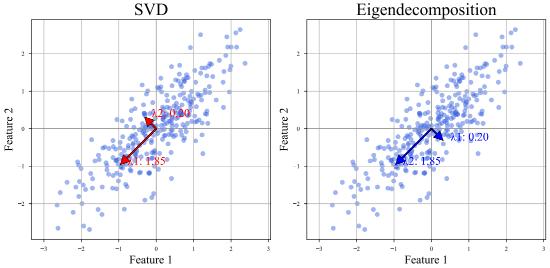


Figure: The Difference Between SVD and Eigendecomposition in Finding Eigen Vectors and Eigen Values

Isotropy

- A distribution is isotropic if its variance is uniformly distributed across all dimensions.
- Namely, the covariance matrix of an isotropic distribution is proportional to the identity matrix.

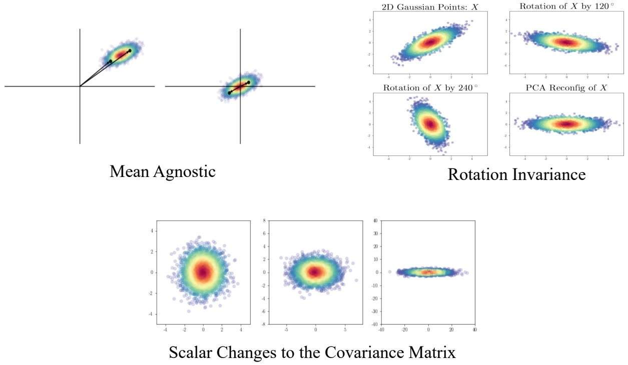


Figure: Essential Properties of Isotropy

Steps to compute IsoScore [6]

1. PCA-reorientation of data set: *Performing PCA reorients the axes of X so that the i'th coordinate accounts for the i'th greatest variance. Further, it eliminates all correlation between dimensions making the covariance matrix diagonal.*
2. Compute variance vector of reoriented data
3. Length normalization of variance vector.
4. Compute the distance between the covariance matrix and identity matrix
5. Use the isotropy defect to compute percentage of dimensions isotropically utilized.
6. Linearly scale percentage of dimensions utilized to obtain IsoScore.

References I

[1] Xingyu Cai et al. "Isotropy in the contextual embedding space: Clusters and manifolds". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/pdf?id=xYGNO86OWDH.

[2] Kawin Ethayarajh. "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings". In: *arXiv preprint arXiv:1909.00512* (2019). URL: https://arxiv.org/pdf/1909.00512.pdf.

[3] Takoua Jendoubi and Korbinian Strimmer. "A whitening approach to probabilistic canonical correlation analysis for omics data integration". In: *BMC bioinformatics* 20 (2019), pp. 1–13. URL: https://doi.org/10.1186/s12859-018-2572-9.

[4] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. "Optimal whitening and decorrelation". In: *The American Statistician* 72.4 (2018), pp. 309–314. URL: https://doi.org/10.1080/00031305.2016.1277159.

[5] Bohan Li et al. "On the sentence embeddings from pre-trained language models". In: *arXiv preprint arXiv:2011.05864* (2020). URL: https://arxiv.org/pdf/2011.05864.pdf.

[6] William Rudman et al. "IsoScore: Measuring the uniformity of embedding space utilization". In: *arXiv preprint arXiv:2108.07344* (2021). URL: https://arxiv.org/pdf/2108.07344.pdf.

References II

[7] Ziyuan Wang and You Wu. "Investigating the Effectiveness of Whitening Post-processing Methods on Modifying LLMs Representations". In: *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE. 2023, pp. 813–820. URL: https://doi.org/10.1109/ICTAI59109.2023.00124.