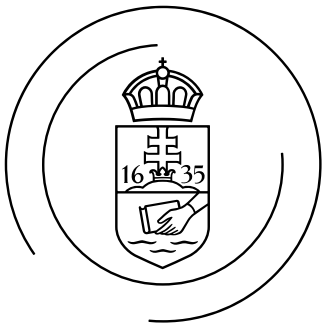
High Entropy Projections in Machine Learning Models

Applied Mathematics MSc Project

Emma Lukács

Supervisor: Adrián Csiszárik



EÖTVÖS LORÁND
UNIVERSITY | BUDAPEST

Eötvös Loránd University
2025/1

1 Introduction

In the field of machine learning, extracting meaningful and informative representations from data is essential for effective model performance. Traditionally, many models and dimensionality reduction techniques, such as Principal Component Analysis (PCA), have focused on variance maximization to identify the most significant directions of data variability. The method proves to be sensitive to outiers and is based on the unimodal Gaussian setting, so while this approach effectively captures directions with the highest variance, it may overlook subtler, yet informative, structures within complex and non-Gaussian datasets. This work revolves around exploring a method that provides the maximum entropy subspace or direction within a given scenario and implementing it in an existing method like Autoencoder based Outlier Detection.

Entropy, a fundamental concept from information theory introduced by Shannon [6], quantifies the uncertainty or randomness in a system. In order to capture the most diverse and informative aspects of the data I experimented with locating subspaces that maximize entropy, thereby enriching the representations beyond what variance maximization alone can achieve. This Maximum Entropy approach, specifically through Maximum Entropy PCA [2], presents a promising direction that is capable of uncovering richer data structures and enhancing model robustness while being free from any distribution assumptions.

By integrating MaxEnt-PCA into ML models, my goal was to explore its effectiveness and improve model performance in tasks where traditional PCA may fall short. Specifically, I apply MaxEnt-PCA to Autoencoders for Outlier Detection. Incorporating MaxEnt-PCA into the latent space of Autoencoders enhances their ability to detect anomalies. The maximized entropy promotes diverse and informative representations, enabling the model to more accurately distinguish between normal and outlier data points.

Through the application, this study demonstrates the practical benefits of the MaxEnt-PCA approach, showcasing its ability to enhance model performance in tasks requiring high-quality representations and effective anomaly detection extending the effectiveness of already existing methods. The subsequent sections delve into the theoretical foundations of entropy maximization, the structure of the MaxEnt-PCA algorithm, and its integration into Autoencoder architectures. Experimental results highlight the effectiveness of this method, offering insights into its potential as a tool in the Machine Learning arsenal.

2 Entropy: Definition and Significance

Definition 1 (Entropy)**.** Let X be a discrete random variable with probability mass function $P_X(x)$, $x \in \mathcal{X}$. The *entropy* (or *Shannon entropy*) of X is

$$H(X) = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_{x\in\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}$$

$$= \int_{\mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}\, dx.$$

Definition 2 (Rényi's Quadratic Entropy)**.** The *Rényi entropy of order 2* [5] (also known as the *quadratic entropy*) of X is defined as

$$H_2(X) = -\log \sum_{x\in\mathcal{X}} P_X(x)^2.$$

Let X be a continuous random variable with probability density function $f_X(x)$, $x \in \mathcal{X}$. The *Rényi entropy of order 2* (quadratic entropy) of X is defined as

$$H_2(X) = -\log \int_{\mathcal{X}} f_X(x)^2 \, dx.$$

Entropy, in the context of information theory, quantifies the uncertainty or randomness in a system. Introduced by Shannon [6], entropy serves as a measure of information content and is pivotal in various machine learning algorithms, including autoencoders, where it aids in optimizing the latent space representations.

3 Principal Component Analysis (PCA)

Definition 3 (Principal Component Analysis)**.** According to Jolliffe [3], PCA is defined as an orthogonal linear transformation on a real inner product space that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.
Consider an $n \times p$ data matrix, $\mathbf{X}$, with column-wise zero empirical mean. The transformation is defined by a set of size l of p-dimensional vectors of weights or coefficients $\mathbf{w}(k) = (w_{1(k)}, \ldots, w_{p(k)})^\top$ that map each row vector $\mathbf{X}(i) = (x_{1(i)}, \ldots, x_{p(i)})$ of $\mathbf{X}$ to a new vector of principal component scores $\mathbf{t}(i) = (t_{1(i)}, \ldots, t_{l(i)})$, given by

$$t_k(i) = \mathbf{X}(i) \cdot \mathbf{w}(k) \quad \text{for } i = 1, \ldots, n \text{ and } k = 1, \ldots, l,$$

in such a way that the individual variables $t_1, \ldots, t_l$ of $\mathbf{t}$ considered over the data set successively inherit the maximum possible variance from $\mathbf{X}$, with each coefficient vector $\mathbf{w}$ constrained to be a unit vector (where l is usually selected to be strictly less than p to reduce dimensionality).
The above may equivalently be written in matrix form as

$$\mathbf{T} = \mathbf{X}\mathbf{W},$$

where $T_{ik} = t_k(i)$, $X_{ij} = x_j(i)$, and $W_{jk} = w_{j(k)}$.
In order to maximize variance, the first weight vector $\mathbf{w}(1)$ thus has to satisfy

$$\mathbf{w}(1) = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (\mathbf{X}(i) \cdot \mathbf{w})^2 \right\}$$

Equivalently, writing this in matrix form gives

$$\mathbf{w}(1) = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\mathbf{X}\mathbf{w}\|^2 \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} \right\}$$

4 Maximum Entropy PCA (MaxEnt-PCA)

MaxEnt-PCA extends traditional PCA by incorporating entropy maximization to ensure that the principal components capture the maximum uncertainty possible given the data constraints. The method was introduced by He et al. [2].

Consider Rényi's quadratic entropy of a random variable X with probability density function (P.D.F.) $f_X(x)$ defined by

$$H(X) = -\log\left(\int f_X^2(x)\,dx\right).$$

If $f_X(x)$ is a Gaussian distribution, the estimate of Rényi's quadratic entropy is obtained by:

$$H(X) = \frac{1}{2}\log(|\Sigma|) + \frac{d}{2}\log(2\pi) + \frac{d}{2},$$

where $|\cdot|$ denotes the absolute value of the determinant.

If the Parzen window method is used to estimate the P.D.F. $f_X(x)$, $f_X(x)$ can be obtained by

$$\hat{f}_X(x) = \frac{1}{n}\sum_{j=1}^{n} G(x - x_j, \sigma^2),$$

where $G(x - x_j, \sigma^2)$ is the Gaussian kernel with bandwidth $\Sigma = \sigma^2 I$:

$$G(x - x_j, \sigma^2) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - x_j)^\top \Sigma^{-1}(x - x_j)\right).$$

By substituting $f_X(x)$ in the entropy definition with the Parzen window estimate, the entropy estimate by the Parzen window method can be formulated as:

$$H(X) = -\log\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(x_i - x_j, \sigma^2)\right).$$

Thus, the MaxEnt-PCA optimization problem can be expressed as:

$$\max_U \left(-\log\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(U^\top x_i - U^\top x_j, \sigma^2)\right)\right)$$

subject to $U^\top U = I$, where U is the projection matrix [4].

Proposition 1 (Optimal Solution of MaxEnt-PCA)**.** The optimal solution of MaxEnt-PCA is given by the eigenvectors of the following generalized eigen-decomposition problem:

$$XL(U)X^\top U = 2U\Lambda,$$

where

$$L(U) = D(U) - W(U),$$

$$W_{ij}(U) = \frac{G(U^\top x_i - U^\top x_j, \sigma^2)}{\sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{n} G(U^\top x_i - U^\top x_j, \sigma^2)},$$

and

$$D_{ii}(U) = \sum_{j=1}^{n} W_{ij}(U).$$

Proof: To meet the orthonormal constraint on U, we define the Lagrangian function as follows:

$$J_H = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} G(U^\top x_i - U^\top x_j, \sigma^2) - \mathrm{Tr}(\Lambda(U^\top U - I)),$$

where $\mathrm{Tr}(\cdot)$ denotes the matrix trace operation. The Karush-Kuhn-Tucker (KKT) condition for the optimal solution specifies that the derivative of J_H with respect to U must be zero:

$$\frac{\partial J_H}{\partial U} = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(U)(x_i - x_j)(x_i^\top - x_j^\top)U - 2U\Lambda = 0.$$

Rearranging terms, we obtain:

$$XL(U)X^\top U = 2U\Lambda,$$

where

$$L(U) = D(U) - W(U),$$

$$W_{ij}(U) = \frac{G(U^\top x_i - U^\top x_j, \sigma^2)}{\sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{n} G(U^\top x_i - U^\top x_j, \sigma^2)},$$

$$D_{ii}(U) = \sum_{j=1}^{n} W_{ij}(U).$$

Intuitively, an optimal U is the eigenvectors of the symmetric matrix $XL(U)X^\top$, and the Lagrangian multipliers Λ become a diagonal matrix: $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ [2]. □

4.1 Algorithm 1. MaxEnt-PCA

Algorithm 1 MaxEnt-PCA

Input: Data matrix X, initial orthonormal matrix U, small positive value ϵ
Output: Orthonormal matrix U

1: `converged` ← `FALSE`
2: **while** not `converged` **do**
3: Calculate $L(U)$ according to the definitions above
4: Solve the generalized eigen-decomposition $XL(U)X^\top U = 2U\Lambda$
5: Update U to be the eigenvectors corresponding to the top m eigenvalues
6: Check for convergence: if the change in entropy is smaller than ϵ, set `converged` ← `TRUE`
7: **end while**

4.2 Empirical Characteristics of MaxEnt-PCA

To illustrate the distinctive behavior of MaxEnt-PCA, I conducted an empirical analysis using a minimal dataset consisting of three points in a two-dimensional plane. By iterating over 180 degrees I calculated the entropy estimate in each angle and picked the highest value in a brute-force way and compared it to the result of the MaxEnt-PCA algorithm. The results shown on Figure 1 demonstrate that the maximum entropy projection coincides with the angle that maximizes the separation between the two furthest points,

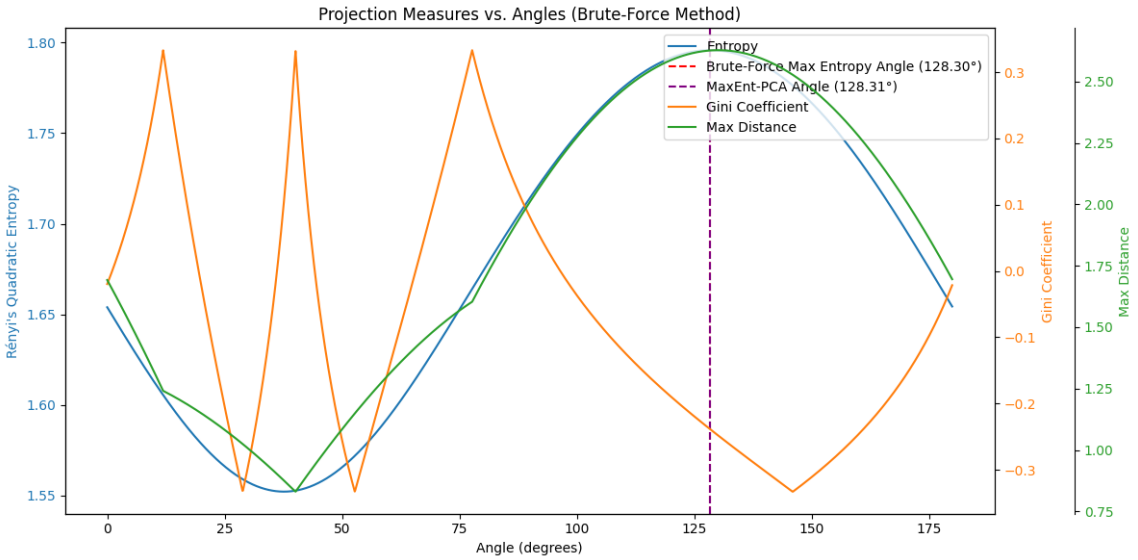


Figure 1: Main characteristics of the maximum entropy projection

effectively aligning the projection vector with the direction that renders these points as distant as possible in the projected space. This alignment ensures that the projected distribution achieves maximum entropy by spreading the data points uniformly along the one-dimensional axis. Additionally, the Gini coefficient of the nearest neighbors in the projected space was found to be low, indicating a uniform distribution with small inequality among the projected distances. These characteristics—maximizing inter-point distances and promoting uniformity—underscore the effectiveness of MaxEnt-PCA in capturing the essential geometric structure of the data, even in simple configurations.

5 Equivalence in the Gaussian setting

Based on the paper by Tsur, Goldfeld, and Greenewald [7], an important connection for Gaussian data was formally revealed. This connection extends to Rényi's quadratic entropy.

Proposition 2 (Equivalence of MaxEnt-PCA and PCA for Gaussian Data)**.** *In the case where the data X follows a Gaussian distribution $X \sim N(0, \Sigma)$, the optimal solution of MaxEnt-PCA coincides with the solution of traditional PCA. Specifically, the projection matrix U obtained from MaxEnt-PCA is identical to the matrix containing the top k eigenvectors of Σ.*

Proof. Assume that the data $X \in \mathbb{R}^d$ follows a multivariate Gaussian distribution:

$$X \sim N(0, \Sigma),$$

where $\Sigma \in \mathbb{R}^{d \times d}$ is the covariance matrix of full rank. Under the Gaussian assumption, the Parzen window estimate of the probability density function (P.D.F.) $f_X(x)$ is:

$$\hat{f}_X(x) = \frac{1}{n} \sum_{j=1}^{n} G(x - x_j, \sigma^2),$$

where $G(x - x_j, \sigma^2)$ is the Gaussian kernel with bandwidth $\Sigma = \sigma^2 I$:

$$G(x - x_j, \sigma^2) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - x_j)^\top \Sigma^{-1}(x - x_j)\right).$$

Given that X is Gaussian, the density $f_X(x)$ is exactly:

$$f_X(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}x^\top \Sigma^{-1} x\right).$$

Thus, the Parzen window estimate becomes an empirical estimate of the Gaussian density. For Gaussian data, Rényi's quadratic entropy is given by:

$$H(X) = \frac{1}{2}\log|\Sigma| + \frac{d}{2}\log(2\pi) + \frac{d}{2}.$$

When projecting onto a subspace defined by $U \in \mathbb{R}^{d\times k}$ with $U^\top U = I$, the projected data $Y = U^\top X$ follows:

$$Y \sim N(0, U^\top \Sigma U).$$

Thus, the Rényi's quadratic entropy of the projection is:

$$H(Y) = \frac{1}{2}\log|U^\top \Sigma U| + \frac{k}{2}\log(2\pi) + \frac{k}{2}.$$

MaxEnt-PCA seeks to maximize the entropy of the projected data Y:

$$\max_U H(Y) = \max_U \left(\frac{1}{2}\log|U^\top \Sigma U| + \frac{k}{2}\log(2\pi) + \frac{k}{2}\right).$$

Since $\frac{k}{2}\log(2\pi) + \frac{k}{2}$ is constant with respect to U, the optimization simplifies to:

$$\max_U \log|U^\top \Sigma U|.$$

Traditional PCA aims to find the projection matrix U that maximizes the variance of the projected data, which can be formulated as:

$$\max_U \operatorname{tr}(U^\top \Sigma U),$$

subject to $U^\top U = I$.

To relate the two objectives, observe that:

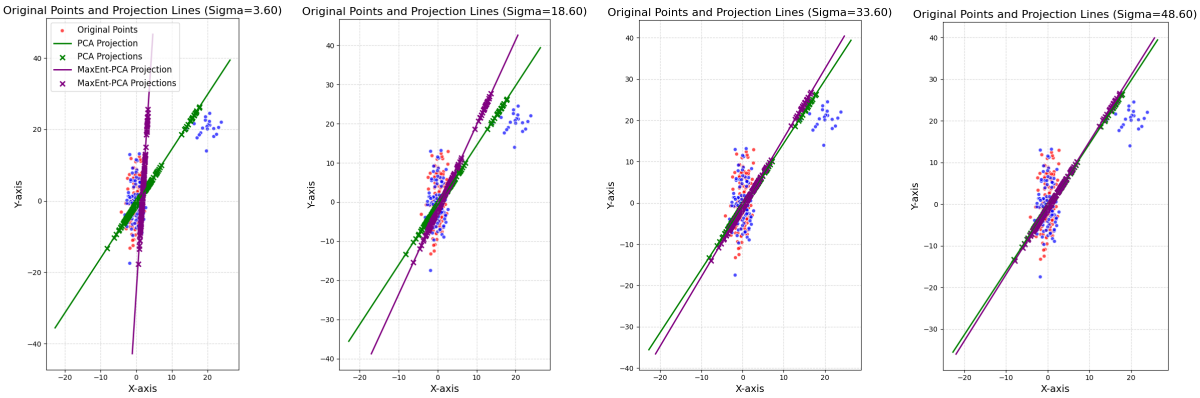
$$\log|U^\top \Sigma U| = \log\left(\prod_{i=1}^k \lambda_i(U^\top \Sigma U)\right) = \sum_{i=1}^k \log\lambda_i(U^\top \Sigma U),$$

where $\lambda_i(U^\top \Sigma U)$ are the eigenvalues of $U^\top \Sigma U$.
Maximizing $\sum_{i=1}^k \log\lambda_i(U^\top \Sigma U)$ is equivalent to maximizing the product $\prod_{i=1}^k \lambda_i(U^\top \Sigma U)$, which is the determinant $|U^\top \Sigma U|$. On the other hand, PCA maximizes $\sum_{i=1}^k \lambda_i(U^\top \Sigma U)$, the trace of $U^\top \Sigma U$.
Under the Gaussian assumption, maximizing the determinant $|U^\top \Sigma U|$ and the trace $\operatorname{tr}(U^\top \Sigma U)$ are aligned in their optimal solutions. Specifically, both objectives are maximized when U spans the subspace corresponding to the top k eigenvectors of Σ. This

is because the arithmetic mean is always greater than or equal to the geometric mean, with equality if and only if all λ_i are equal. However, in the context of maximizing both the trace and the determinant, the optimal configuration occurs when U captures the directions of maximum variance, i.e., the top eigenvectors. Therefore, in the Gaussian setting where $X \sim N(0, \Sigma)$, the MaxEnt-PCA objective:

$$\max_U \log |U^\top \Sigma U|$$

has its optimal solution U aligned with the top k eigenvectors of Σ, which is exactly the solution obtained by traditional PCA. □



(a) σ=0.60 (b) σ=5.60 (c) σ=10.60 (d) σ=15.60

Figure 2: Comparison of PCA and MaxEnt-PCA projections across different sigma values.

I experimented further with comparing the methods in a simple setting. The goal was to achieve a better understanding of the relationship between the projections with differing hyperparameter values. The empirical analysis in a two-dimensional multimodal Gaussian setting demonstrates that MaxEnt-PCA converges significantly to the traditional PCA as the bandwidth parameter σ increases. Specifically, the projection vectors obtained through MaxEnt-PCA gradually align with those derived from PCA when σ is incrementally raised. This convergence is visually illustrated in Figure 2, where higher values of σ result in MaxEnt-PCA projections closely matching the PCA projections.

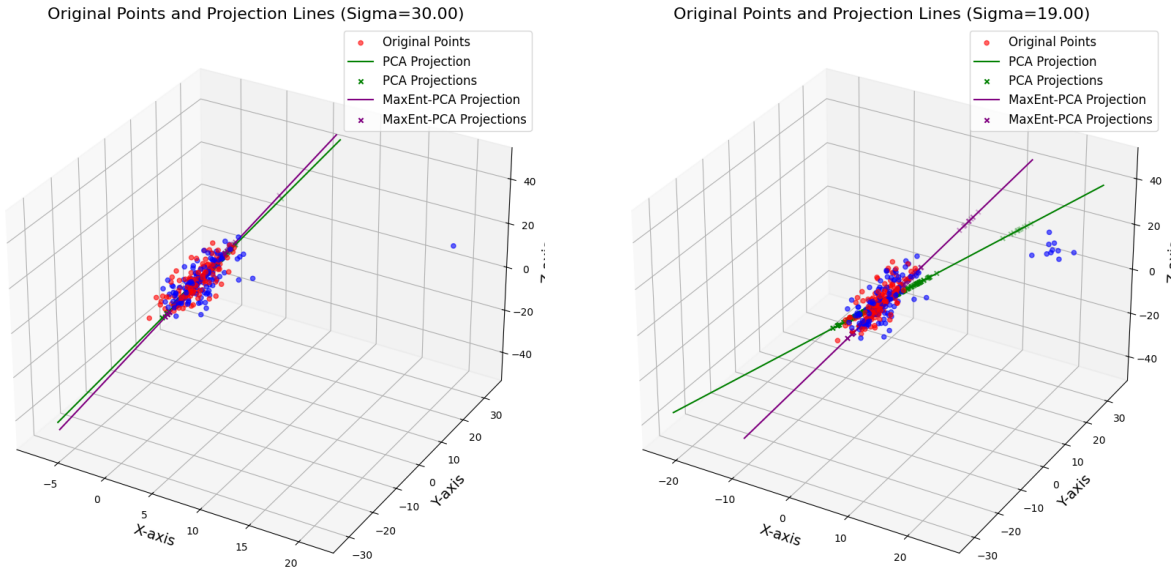
6 Entropy Maximization in Autoencoders

Autoencoders are neural network architectures composed of an encoder and a decoder that collaboratively learn to represent input data through a compressed latent space. Mathematically, the encoder function $f_\theta : \mathbb{R}^n \to \mathbb{R}^m$ maps an input vector $\mathbf{x}$ to a lower-dimensional latent representation $\mathbf{z} = f_\theta(\mathbf{x})$, where $m < n$. The decoder function $g_\phi : \mathbb{R}^m \to \mathbb{R}^n$ reconstructs the input as $\hat{\mathbf{x}} = g_\phi(\mathbf{z})$. The latent space $\mathcal{Z}$ thus serves as a lower-dimensional manifold that captures the essential features of the data, preserving its intrinsic geometric and topological properties. Training is achieved by minimizing a reconstruction loss, typically defined as $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$, which ensures that the latent representation effectively encodes the information necessary for accurate data reconstruction. [1]

6.1 Outlier Detection

To evaluate the effectiveness of MaxEnt-PCA in identifying outliers, I conducted a comparative analysis against standard PCA using synthetic three-dimensional datasets. These datasets comprised both normally distributed data points and outliers, generated in varying proportions to simulate diverse anomaly scenarios.

The outlier detection framework commenced with the synthesis of datasets where normal data points were sampled from a multivariate Gaussian distribution with randomly selected means and covariances. Specifically, for each dataset instance, the normal data points $\mathbf{x}_i \in \mathbb{R}^3$ were drawn from $\mathcal{N}(\boldsymbol{\mu}_{\text{normal}}, \boldsymbol{\Sigma}_{\text{normal}})$, where $\boldsymbol{\mu}_{\text{normal}}$ and $\boldsymbol{\Sigma}_{\text{normal}}$ are the mean vector and covariance matrix, respectively. Outliers were introduced by sampling from a distinct Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_{\text{outlier}}, \boldsymbol{\Sigma}_{\text{outlier}})$, with $\|\boldsymbol{\mu}_{\text{outlier}} - \boldsymbol{\mu}_{\text{normal}}\| \geq \delta$ to ensure significant separation between normal data and outliers, where δ is a predefined threshold. For each dataset example, both standard PCA and MaxEnt-PCA were employed to project the high-dimensional data onto a one-dimensional subspace.



(a) Synthetic dataset with 1 anomaly point (b) Synthetic dataset with 10 anomaly points

Figure 3: Comparison of PCA and MaxEnt-PCA projections represented as lines on the original data distrbution.

The results suggest that MaxEnt-PCA is indeed insensitive to anomaly points compared to PCA, providing a more robust tool. An important example is shown on Figure 3 where the PCA projection is highly influenced by the additional variance introduced by the increasing instance of the anomalious pattern.

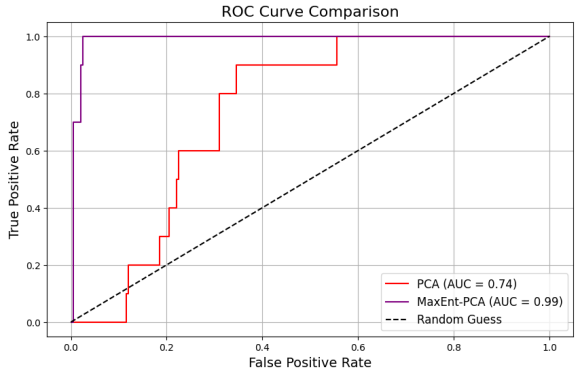
Anomaly scores were derived from reconstruction errors, quantified as the Euclidean distance between each original data point $\mathbf{x}_i$ and its projection $\hat{\mathbf{x}}_i$ onto the subspace:

$$\text{Anomaly Score}_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2 = \|\mathbf{x}_i - (\mathbf{w}\mathbf{w}^\top)\mathbf{x}_i\|_2.$$

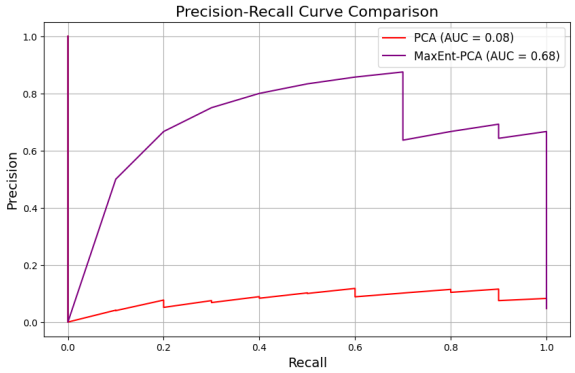
Higher reconstruction errors corresponded to higher anomaly likelihoods, flagging potential outliers.

The performance of both PCA and MaxEnt-PCA was assessed using Receiver Operating Characteristic Area Under the Curve (ROC AUC), Precision-Recall Area Under the Curve

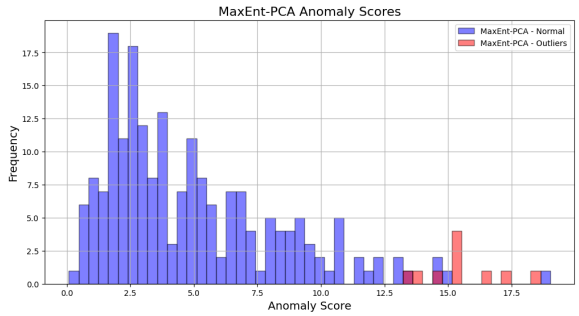
(PR AUC), and the Gini coefficient. The ROC AUC evaluates the trade-off between true positive rate and false positive rate, while PR AUC focuses on the trade-off between precision and recall, particularly relevant in imbalanced datasets. The Gini coefficient provides a measure of inequality in the distribution of anomaly scores, with lower values indicating a more uniform distribution conducive to reliable threshold-based detection.



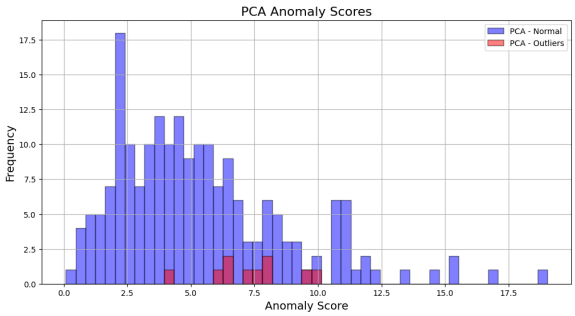
(a) ROC Curve Comparison



(b) Precision-Recall Curve Comparison



(c) MaxEnt-PCA Anomaly Scores



(d) PCA Anomaly Scores

Figure 4: Comparative analysis of PCA and MaxEnt-PCA projections and performance. (a) and (b) show ROC and Precision-Recall curves, respectively, while (c) and (d) illustrate anomaly score distributions for MaxEnt-PCA and PCA, respectively.

The results consistently demonstrated that MaxEnt-PCA outperformed standard PCA in outlier detection across all synthetic datasets. Specifically, MaxEnt-PCA achieved higher ROC AUC and PR AUC scores, indicating its superior ability to distinguish between normal data points and outliers. Additionally, the Gini coefficient was generally lower for MaxEnt-PCA, suggesting a more uniform distribution of anomaly scores and enhancing the reliability of threshold-based detection by reducing false positives and negatives.
MaxEnt-PCA's optimization of Rényi's quadratic entropy facilitated a more informative latent representation, effectively capturing the uncertainty within the data distribution and enabling clearer separation between normal and anomalous points. This is visually supported by the distinct clustering patterns observed in the anomaly score distributions, where MaxEnt-PCA exhibited more pronounced class separation. Furthermore, the ROC and Precision-Recall curves validated MaxEnt-PCA's robustness, with steeper curves and higher AUC values compared to standard PCA.
These comprehensive findings underscore the efficacy of MaxEnt-PCA as a dimensionality reduction technique for outlier detection, particularly in complex data distributions where traditional variance-based methods may be insufficient. The mathematical optimization framework of MaxEnt-PCA ensures that the latent space preserves the intrinsic geometric

structure of the data, thereby enhancing its discriminative power in identifying anomalies.

7 Conclusion

In this study, I experimented with Maximum Entropy Principal Component Analysis (MaxEnt-PCA) as an extension to traditional Principal Component Analysis (PCA), aiming to enhance dimensionality reduction by incorporating entropy maximization. Theoretical analysis, supported by Proposition 2, established that MaxEnt-PCA aligns with PCA under Gaussian assumptions, ensuring that both methods yield identical projection vectors in such settings. Empirical evaluations on synthetic datasets demonstrated that MaxEnt-PCA indeed converges to PCA as the bandwidth parameter σ increases.

Further analysis proved that MaxEnt-PCA outperforms PCA in outlier detection tasks. Specifically, MaxEnt-PCA achieved higher ROC AUC and PR AUC scores while maintaining lower Gini coefficients, indicating more uniform anomaly score distributions and robust data representations. These findings highlight MaxEnt-PCA's ability to capture essential geometric structures and promote uniformity, making it a superior choice for complex, multimodal data distributions where traditional PCA may fall short.

However, this study has limitations, including reliance on synthetic datasets and the computational complexity associated with higher-dimensional data. Additionally, the necessity for careful tuning of the bandwidth parameter σ presents a challenge for practical applications. Future research will focus on applying MaxEnt-PCA to real-world datasets and exploring its integration with more complex models such as Generative Adversarial Networks (GANs). By comparing latent space structures between models with and without MaxEnt-PCA, we aim to further elucidate its benefits and enhance its applicability across diverse machine learning frameworks.

Overall, MaxEnt-PCA emerges as a powerful dimensionality reduction technique that not only preserves the intrinsic geometric structure of the data but also enhances representation informativeness and robustness. Its flexibility in adapting through entropy maximization makes it particularly suited for intricate and non-Gaussian datasets, thereby expanding the toolkit available for advanced machine learning applications.

References

[1] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021.

[2] Ran He, Baogang Hu, XiaoTong Yuan, and Wei-Shi Zheng. Principal component analysis based on non-parametric maximum entropy. *Neurocomputing*, 73(10):1840–1852, 2010. Subspace Learning / Selected papers from the European Symposium on Time Series Prediction.

[3] Ian T Jolliffe. *Principal component analysis*. Springer, 2002.

[4] Jose Principe and John Iii. Information-theoretic learning. *Advances in unsupervised adaptive filtering*, 09 2000.

[5] Alfréd Rényi. On measures of information and entropy. In *Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561. University of California Press, 1961.

[6] Claude E Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[7] Ori Tsur, Benjamin Goldfeld, and Paul Greenewald. Max sliced entropy: A principal component analysis for non-gaussian data. *Journal of Machine Learning Research*, 24(1):1–25, 2023.