Sparse Principal Component Analysis

D3S Session 1 - Project B

Théo Druilhe, Nathan Pizzetta, Sigurd Saue

Master 2 Data Science for Social Sciences

December 2024

Table of Contents

Introduction

Dimensionality Reduction:

- Essential for analyzing high-dimensional datasets.
- Techniques: Principal Component Analysis (PCA) and Sparse PCA (SPCA).

Dataset: Labeled Faces in the Wild (LFW)

- Benchmark dataset for facial image analysis.
- **Size:** 1,288 grayscale images of 7 individuals.
- **Resolution:** 50×37 pixels flattened to 1,850-dimensional vectors.
- Challenge: High dimensionality and multi-class structure.

Goals and Evaluation

Goals:

- Compare PCA and SPCA for meaningful feature extraction.
- Analyze interpretability, accuracy, and efficiency.
- Evaluate classification performance using Support Vector Machine (SVM).

Methods:

- **PCA:** Captures maximum variance but lacks interpretability.
- **SPCA:** Introduces sparsity, highlighting localized features (e.g., eyes, mouth).
- **SSPCA:** Incorporates domain-specific structures for improved interpretability.

Acknowledgements

Acknowledgement of Sources:

- Based on seminal works:
  - *Zou et al. (2006)*: Sparse Principal Component Analysis [**?**].
  - *Jenatton et al. (2010)*: Structured Sparse PCA [**?**].
- Provides mathematical foundation and optimization techniques.

Focus:

- Theoretical insights.
- Practical application and evaluation on the LFW dataset.

Exploratory Data Analysis: Dataset Overview

Dataset: Labeled Faces in the Wild (LFW)

- Grayscale images of 7 well-known individuals.
- **Total Samples:** 1,288 images.
- **Image Dimensions:** 50×37 pixels.
- **Classes:**
  - Ariel Sharon, Colin Powell, Donald Rumsfeld, George W. Bush, Gerhard Schroeder, Hugo Chavez, Tony Blair.

Sample Images

Understanding the Dataset:

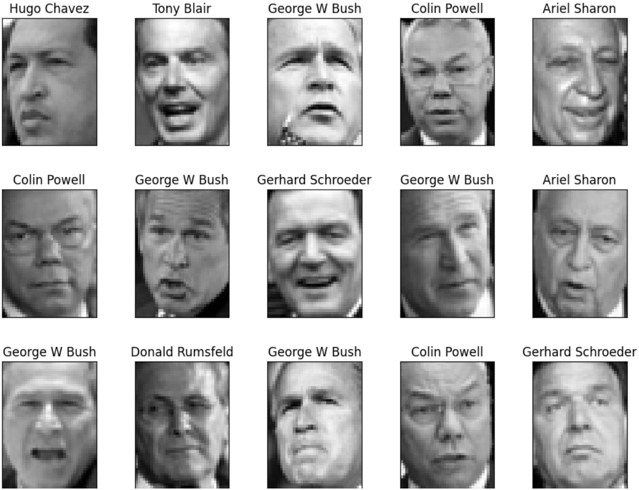
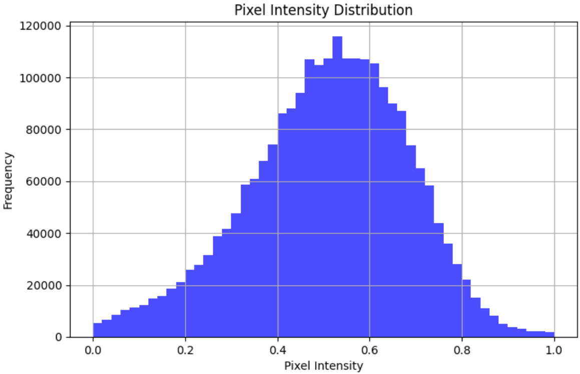


Figure: Sample images from the LFW dataset.

Pixel Intensity Distribution

Analysis:

- Pixel intensity values range from 0 (black) to 1 (white).
- Histogram indicates:
  - Concentration around mid-range values.
  - Few very dark or bright pixels.
  - Well-lit facial images.



Class Distribution

Observations:

- Slight imbalance in class distribution.
- *George W. Bush* has the most samples.
- Class imbalance could impact model performance.

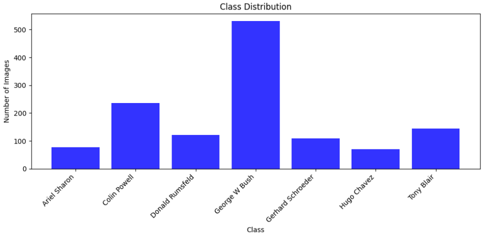


Figure: Number of images per individual in the dataset.

Image Statistics: Mean and Standard Deviation

Key Metrics:

- **Mean Image:** Highlights shared features (e.g., eyes, mouth).
- **Standard Deviation Image:** Shows areas of high variability (e.g., hairline, facial contours).

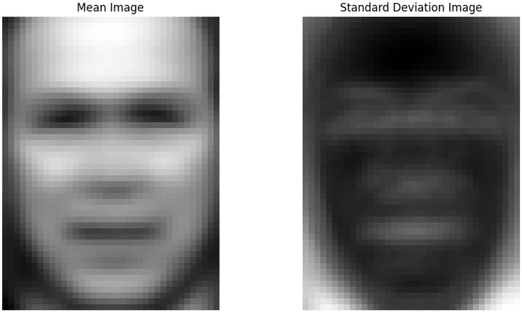


Figure: (Left) Mean image. (Right) Standard deviation image.

Image Flattening and Representation

Flattening Process:

- Images of 50×37 pixels reshaped into 1-dimensional vectors of size 1,850.
- Facilitates compatibility with machine learning models.

Matrix Representation:

$$\mathbf{X} \in \mathbb{R}^{1{,}288 \times 1{,}850}, \quad \mathbf{Y} = \left[y_1, y_2, \ldots, y_{1288}\right]^\top, \quad y_i \in \{1, \ldots, 7\}.$$

Key Points:

- Enables dimensionality reduction.
- Removes spatial structure; can be restored through advanced techniques.

Principal Component Analysis (PCA)

Overview:

- Widely used dimensionality reduction technique.
- Extracts and encodes significant features in high-dimensional data.
- Retains essential patterns while discarding noise and redundancy.

Application to Face Recognition:

- Encodes facial images into low-dimensional feature spaces.
- Reduces dimensions while preserving critical features for classification.

PCA: How It Works

Principal Components (PCs):

- Linear combinations of original variables (pixel intensities).
- Capture maximum variance in the data.
- PCs are uncorrelated, ensuring efficient feature representation.

Computation Using Singular Value Decomposition (SVD):

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

- **U**: Left singular vectors (principal components scaled by variance).
- **D**: Diagonal matrix of singular values (variance explained by each PC).
- **V**: Right singular vectors (loadings of the PCs).

PCA in Face Recognition

Eigenfaces:

- PCs derived from facial data are often called "eigenfaces."
- Represent key patterns and variations in facial features.

Dimensionality Reduction:

- Retain only the leading PCs, reducing dimensions.
- Enables efficient storage and classification.

Limitations of PCA

- **Lack of Sparsity:**
  - Each PC is a combination of all features.
  - Challenging to interpret in the context of specific facial features.
- **Interpretability Issues:**
  - Dense components make it difficult to associate PCs with distinct image regions.
- **Potential Information Loss:**
  - Retains maximal variance but does not prioritize task-specific features.

Loadings of the First Six Principal Components

Table 1. Pitprops Data: Loadings of the First Six Principal Components

Variable	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
topdiam	−0.404	0.218	−0.207	0.091	−0.083	0.120
length	−0.406	0.186	−0.235	0.103	−0.113	0.163
moist	−0.124	0.541	0.141	−0.078	0.350	−0.276
testsg	−0.173	0.456	0.352	−0.055	0.356	−0.054
ovensg	−0.057	−0.170	0.481	−0.049	0.176	0.626
ringtop	−0.284	−0.014	0.475	0.063	−0.316	0.052
ringbut	−0.400	−0.190	0.253	0.065	−0.215	0.003
bowmax	−0.294	−0.189	−0.243	−0.286	0.185	−0.055
bowdist	−0.357	0.017	−0.208	−0.097	−0.106	0.034
whorls	−0.379	−0.248	−0.119	0.205	0.156	−0.173
clear	0.011	0.205	−0.070	−0.804	−0.343	0.175
knots	0.115	0.343	0.092	0.301	−0.600	−0.170
diaknot	0.113	0.309	−0.326	0.303	0.080	0.626
Variance (%)	32.4	18.3	14.4	8.5	7.0	6.3
Cumulative variance (%)	32.4	50.7	65.1	73.6	80.6	86.9

Source: Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265-286.

Loadings of the First Six Sparse PCs by SPCA

Table 3. Pitprops Data: Loadings of the First Six Sparse PCs by SPCA. Empty cells have zero loadings.

Variable	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*
topdiam	−0.477					
length	−0.476					
moist		0.785				
testsg		0.620				
ovensg	0.177		0.640			
ringtop			0.589			
ringbut	−0.250		0.492			
bowmax	−0.344	−0.021				
bowdist	−0.416					
whorls	−0.400					
clear				−1		
knots		0.013			−1	
diaknot			−0.015			1
Number of nonzero loadings	7	4	4	1	1	1
Variance (%)	28.0	14.4	15.0	7.7	7.7	7.7
Adjusted variance (%)	28.0	14.0	13.3	7.4	6.8	6.2
Cumulative adjusted variance (%)	28.0	42.0	55.3	62.7	69.5	75.8

Source: Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265-286.

The Lasso

Overview:

- Adds an L_1 penalty to the regression coefficients.
- Encourages sparsity by shrinking some coefficients to zero.

Optimization Problem:

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

- λ_1: Regularization parameter controlling sparsity.
- Shrinks less relevant features, selecting the most important ones.

Limitation:

- Struggles with correlated predictors.

Elastic Net

Overview:

- Combines L_1 (Lasso) and L_2 (Ridge) penalties.
- Addresses limitations of the Lasso for correlated predictors.

Optimization Problem:

$$\hat{\boldsymbol{\beta}}_{\text{en}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

- λ_1: Sparsity parameter (controls Lasso effect)
- λ_2: Stabilization parameter for correlated features (controls Ridge effect).

Sparse Principal Component Analysis (SPCA)

Overview:

- Introduced by *Zou et al. (2006)*.
- Improves PCA by introducing **sparsity** in the loadings matrix $\mathbf{V}$.
- Ensures that only a subset of features contributes to each principal component (PC).

Benefits:

- Improved **interpretability**.
- Reduces the number of used variables, enhancing computational efficiency.

PCA as a Regression Problem

Regression Framework:

- Each PC in PCA can be recovered by solving a regression problem.
- Ridge regression formulation:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Z}_i - \mathbf{X}\beta\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

- λ: Ridge penalty for stability and uniqueness.
- The normalized ridge solution is:

$$\hat{\mathbf{v}} = \frac{\hat{\boldsymbol{\beta}}_{\text{ridge}}}{\|\hat{\boldsymbol{\beta}}_{\text{ridge}}\|}.$$

SPCA as a Regression Problem

Extension to SPCA:

- Adds an L_1 penalty (Lasso) to enforce sparsity:

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Z}_i - \mathbf{X}\beta\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1$$

- Combines Ridge (L_2) and Lasso (L_1) penalties for stability and sparsity.
- The resulting sparse loading vector $\hat{\mathbf{v}}_i$ is normalized, and the sparse principal component is given by:

$$\mathbf{X}\hat{\mathbf{v}}_i.$$

SPCA Optimization Problem

General Optimization Criterion:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^\top \mathbf{x}_i\|^2 + \sum_{j=1}^{k} \left(\lambda_1 \|\boldsymbol{\beta}_j\|_1 + \lambda_2 \|\boldsymbol{\beta}_j\|^2\right)$$

- $\mathbf{A}$: Sparse loading matrix ($p \times k$).
- $\mathbf{B}$: Coefficients of principal components ($n \times k$).

Constraints:

$$\mathbf{A}^\top \mathbf{A} = \mathbf{I}_{k \times k}$$

Role of Penalties:

- λ_1: Controls sparsity (Lasso penalty).
- λ_2: Ensures stability (Ridge penalty).

Algorithm to resolve SPCA optimization problem

Algorithm

- Fix A and solve for B using the Elastic Net.
- Fix B and solve for A using Singular Value Decomposition (SVD).
- Repeat until convergence.

SPCA in Face Recognition

Applications:

- Extracts localized features like eyes, nose, and mouth.
- Focuses on key facial regions rather than global patterns.
- Facilitates dimensionality reduction and interpretable feature extraction.

Comparison with PCA:

- PCA generates dense components, involving all features.
- SPCA introduces sparsity, isolating specific important features.

Summary of SPCA

- SPCA combines PCA with regularization techniques (Lasso, Elastic Net).
- Components are easier to interpret because only the most important variables are kept (due to Lasso penalty).
- Useful in scenarios where feature selection and sparsity are critical, like face recognition.

Introduction to Structured Sparse PCA

Overview:

- Extends Sparse PCA to incorporate prior structure in data.
- Incorporates domain knowledge to enforce sparsity patterns in principal components.
- Useful for datasets with hierarchical or group structures.

Core Idea:

- Retains sparsity while respecting predefined structures (e.g., groups of features, spatial contiguity).

Why Structured Sparse PCA?

Limitations of Sparse PCA:

- Ignores group or hierarchical relationships in features.
- May select sparse features that are not meaningful in context.

Advantages of SSPCA:

- Leverages domain knowledge to improve feature selection.
- Ensures selected features adhere to logical groupings.
- Produces interpretable principal components with real-world relevance.

Applications:

- Functional brain imaging (groups of voxels).
- Genomics (pathways or gene networks).
- Face recognition (spatially adjacent features like eyes or mouth).

Mathematical Formulation of SSPCA

Optimization Problem:

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\min_{\mathbf{A},\mathbf{B}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{A}\mathbf{B}^\top \mathbf{x}_i\|^2 + \sum_{j=1}^{k} P(\beta_j)$$

- $\mathbf{x}_i$: Input data vectors.
- $\mathbf{A}$: Sparse loading matrix ($p \times k$).
- $\mathbf{B}$: Coefficients of principal components ($n \times k$).
- $P(\beta_j)$: Structured sparsity-inducing penalty.

Structured Sparsity Penalties

Group Lasso:

$$P_{\text{group}}(\boldsymbol{\beta}) = \sum_{g \in \mathcal{G}} \lambda_g \|\boldsymbol{\beta}_g\|_2$$

- $\mathcal{G}$: Predefined groups of features.
- Ensures entire groups of features are selected or removed.

Fused Lasso:

$$P_{\text{fused}}(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$$

- Promotes sparsity and smoothness for sequential data.

Graph-based Penalty:

$$P_{\text{graph}}(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta}$$

- $\mathbf{L}$: Graph Laplacian encoding feature connectivity.
- Encourages features to be selected based on graph structure.

Constraints in SSPCA:

Structured Sparsity:

$\boldsymbol{\beta}_j$ must respect predefined group, spatial, or graph structures.

Examples of structural constraints:

- **Group Sparsity:** Features are selected in predefined groups.
- **Smoothness:** Neighboring features have similar weights (e.g., in spatial or temporal data).
- **Graph Connectivity:** Selected features form a connected subgraph in a predefined network.

Comparison of PCA, SPCA, and SSPCA

Feature	**PCA**	**Sparse PCA**	**SSPCA**
Sparsity	No	Yes	Yes
Structure	No	No	Yes
Interpretability	Low	Medium	High
Applications	General	Feature selection	Structured data

Challenges and Takeaways of SSPCA

Challenges:

- Requires prior knowledge of feature structure.
- Computationally expensive for large datasets.
- Sensitive to hyperparameter tuning.

Key Takeaways:

- SSPCA bridges the gap between Sparse PCA and domain-specific feature selection.
- Balances sparsity, structure, and interpretability.

Application to Human-Face Recognition

Analyzing PCA and SPCA on the LFW Dataset

Objectives

Objectives:

- Evaluate the effectiveness of PCA and Sparse PCA (SPCA) for facial image analysis.
- Extract lower-dimensional representations that preserve discriminative information.

Evaluation:

- Predictive performance using a Support Vector Machine (SVM) classifier.
- Assess trade-offs in accuracy, interpretability, and computational efficiency.

Determining the Number of Principal Components

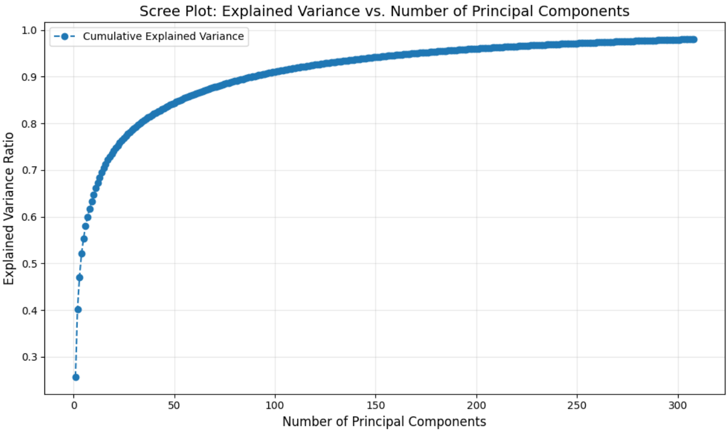


Figure: Scree Plot: Cumulative Explained Variance vs. Number of Principal Components.

Scree Plot Analysis

Two methods:

- **Elbow Point:** The curve flattens around 20–30 components, where adding further components contributes minimal additional variance.
- **Target Variance Retention:** To retain **95%** of the variance, approximately **170 components** are needed.

We choose to keep components that retain **95% of the variance** to achieve a balance between dimensionality reduction and minimal information loss.

Dimensionality Reduction with PCA

PCA Application:

- Goal: Simplify data structure and extract meaningful features.
- PCA reduces dimensions while retaining 95% of the variance.

Results:

- Original dataset shape: $1,288 \times 1,850$.
- Reduced dataset shape: $1,288 \times 171$.

Interpretation:

- Dimensionality reduction simplifies computational tasks.
- Maintains most of the dataset's variance.

Visualizing Principal Components (Eigenfaces)

Key Insights:

- Principal components (PCs) represent key patterns in the data.
- Eigenfaces: PCs for facial data, highlighting areas of variation.

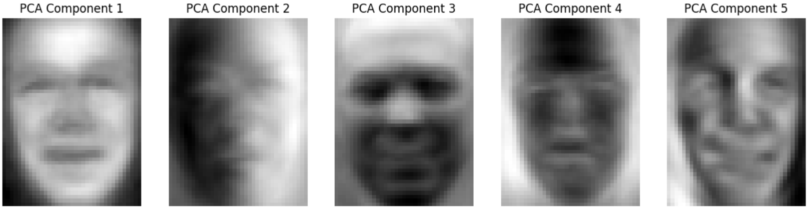


Figure: First five PCA components (eigenfaces). Lighter/darker areas correspond to regions of high/low variance.

Insights from PCA Components

Description of the First Five Components:

- **Component 1:** Captures global brightness and symmetry across the face.
- **Component 2:** Highlights asymmetry, such as lighting differences between left/right sides.
- **Component 3:** Focuses on shadows and variations in the mouth and chin regions.
- **Component 4:** Emphasizes variations in the eye and brow regions.
- **Component 5:** Reflects changes in cheeks and mouth (e.g., facial expressions).

Observations and Limitations of PCA

Insights:

- Reduced dimensionality (171 PCs) retains 95% of variance.
- PCA components (eigenfaces) highlight key variations in facial features.

Limitations:

- Lack of sparsity: All features (pixels) contribute to each PC, making them less interpretable.
- Difficulty associating PCs with specific regions or patterns in the face.

Sparse PCA: Addressing PCA Limitations

Motivation:

- PCA produces dense principal components (PCs) that are difficult to interpret.
- Sparse PCA introduces sparsity, isolating specific features to enhance interpretability.

Sparse PCA Applied to the LFW Dataset:

- Reduced dimensionality from $1,850$ features to 171 sparse components.
- Maintains meaningful variance while highlighting localized regions.

Visualizing Sparse PCA Components

Key Insights:

- Unlike PCA, Sparse PCA components highlight specific localized regions.
- Enhances interpretability by focusing on distinct facial features (e.g., eyes, mouth).

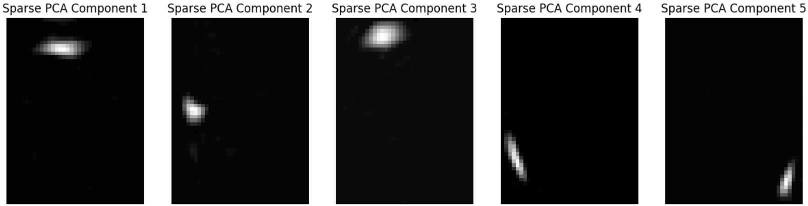


Figure: First five Sparse PCA components. Each component captures localized facial features.

Interpreting Sparse PCA Components

Descriptions of the First Five Components:

- **Component 1:** Highlights a localized region, likely corresponding to an eye or part of the forehead.
- **Component 2:** Focuses on a separate facial feature, such as the other eye or a shadowed region.
- **Component 3:** Captures a distinct upper facial feature, potentially related to the eyes or forehead.
- **Component 4:** Highlights the lower face, possibly the mouth or jawline.
- **Component 5:** Isolates another localized region, potentially the cheeks.

Comparison Between PCA and Sparse PCA

Key Differences:

- **Component Structure:**
  - PCA: Dense, global patterns involving all pixels.
  - Sparse PCA: Sparse, localized features (e.g., eyes, nose, mouth).
- **Interpretability:**
  - PCA components are harder to relate to specific facial features.
  - Sparse PCA components focus on distinct regions, improving interpretability.

MiniBatch Sparse PCA

MiniBatch Sparse PCA: A variant of Sparse PCA for efficient dimensionality reduction with sparse components.

- **Why Use It?** Efficient for large datasets; reduces memory and computation.
- **How It Works:** Processes data in **mini-batches** to balance speed and accuracy.

Key Parameters:

- `n_components`: Number of components.
- `alpha`: Controls sparsity.
- `batch_size`: Size of mini-batches.
- `max_iter`: Iterations for convergence.

Visualizing MiniBatch Sparse PCA Components

Key Insights:

- MiniBatch Sparse PCA components capture localized and sparse facial features.
- Sparsity enhances interpretability by isolating meaningful regions of variation.

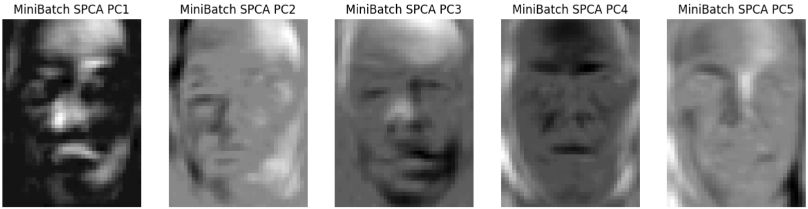


Figure: First five MiniBatch Sparse PCA components. Each component highlights specific facial features.

Insights from MiniBatch Sparse PCA Components

Description of the First Five Components:

- **Component 1:** Captures the overall face structure, focusing on eyes and mouth.
- **Component 2:** Highlights contours around the nose and mouth regions.
- **Component 3:** Emphasizes the lower face, particularly the mouth and jawline.
- **Component 4:** Focuses on the upper face, especially the eyes and forehead.
- **Component 5:** Balances mid-face features, including the nose and cheeks.

Prediction Model: Overview

Objective:

- Compare PCA, Sparse PCA, and MiniBatch SPCA by evaluating their performance in a classification task.

Methodology:

- Support Vector Machine (SVM) was chosen as the classifier.
- Suitable for high-dimensional data like facial feature vectors.

Support Vector Machine (SVM): Key Concepts

How SVM Works:

- Finds hyperplanes that separates data points from different classes.
- Maximizes the margin (distance to the nearest data points).

Kernel Trick:

- Maps data into a higher-dimensional space to handle non-linear separations.
- Common kernels:
  - **Linear**: For linearly separable data.
  - **RBF (Radial Basis Function)**: Effective for non-linear relationships.
  - **Polynomial**: Handles more complex boundaries.

SVM Hyperparameter Optimization

Two hyperparameters to tune:

- **C:** Controls the trade-off between maximizing the margin and minimizing classification errors.
- **Gamma:** Determines the influence of a single data point.

Optimization Strategy:

- Grid search over hyperparameter values:
  - $C \in \{1, 5, 10, 20\}$.
  - $\gamma \in \{\text{scale}, 0.1, 0.01, 0.001\}$.
- 5-fold cross-validation to select the best combination.

Prediction Results: Accuracy Comparison

Results After Grid Search:

- **No PCA:** 0.845 ($C = 5$, $\gamma =$ scale).
- **PCA:** 0.845 ($C = 5$, $\gamma =$ scale).
- **Sparse PCA:** 0.841 ($C = 10$, $\gamma =$ scale).
- **MiniBatch SPCA: 0.853** ($C = 5$, $\gamma =$ scale).

Analysis of Results

Why MiniBatch SPCA Performs Best:

- Balances sparsity with computational efficiency.
- Preserves more discriminative structure in the data.

Why PCA and Sparse PCA Are Similar:

- PCA captures global patterns effectively, which are sufficient for classification.
- Sparse PCA focuses on localized features but may lose some discriminative information if sparsity is overly enforced.

Conclusion

Project Summary:

- Applied PCA and Sparse PCA for dimensionality reduction on the **LFW dataset**.
- **PCA:**
  - Reduced dataset to 171 components, retaining 95% of the variance.
  - Captures global patterns but lacks interpretability.
- **Sparse PCA:**
  - Introduced sparsity, improving interpretability by highlighting localized features.
  - Slight reduction in classification accuracy compared to PCA.
- **MiniBatch Sparse PCA:**
  - Achieved the best accuracy (0.853).
  - Balanced sparsity with computational efficiency.

To Go Further: Potential Improvements

Future Explorations:

- **Applying SSPCA:**
  - Structured Sparse Principal Component Analysis (SSPCA) introduces additional structure to the sparsity pattern.
  - Captures meaningful relationships between features.
  - Compare SSPCA's predictive accuracy with SPCA and MiniBatch SPCA.
- **Experimenting with Different Variance Thresholds:**
  - Current PCA retains 95% of variance.
  - Explore thresholds like 90%, 85%, and 99% to observe impact on predictive accuracy.
  - **Trade-off:** Lower thresholds may remove noise, while higher thresholds preserve more information.

Bibliography I

- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265-286. https://doi.org/10.1198/106186006X113430.
- Jenatton, R., Obozinski, G., & Bach, F. (2010). Structured sparse principal component analysis. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 366-373. http://proceedings.mlr.press/v9/jenatton10a/jenatton10a.pdf.