Principal Component Analysis (PCA)

Data Challenge Group 4

1 Introduction

In the rapidly evolving field of computer-assisted medicine, obtaining a precise three-dimensional (3D) representation of a patient's anatomy is a critical prerequisite for successful surgical planning and decision-making. However, in clinical practice, acquiring a perfect 3D model is often challenged by noisy data, the limitations of medical imaging hardware, or the availability of only partial information, such as 2D X-rays or occluded scans. This project, proposed in collaboration with TwInsight, addresses this challenge by developing a **Statistical Shape Model (SSM)** of the human femur.

1.1 Context and Problem Statement

TwInsight specializes in creating "digital twins" of patients to support orthopedic surgery planning. The central problem we address is the reconstruction of a complete, realistic anatomical shape from incomplete data. While every patient is morphologically unique, human anatomy follows a consistent structural pattern. As noted in the project guidelines, "we are all made the same (Anatomy), but we are all different (Morphology)". This duality allows us to mathematically model "what a femur looks like" by analyzing the commonalities and variations across a population.

1.2 Methodological Approach: Principal Component Analysis

To capture these anatomical variations, we utilize **Principal Component Analysis (PCA)**. A 3D mesh of a femur is composed of thousands of vertices (approximately 18,291 in our dataset), creating a high-dimensional problem that is computationally expensive to manage. PCA serves as a powerful tool for dimensionality reduction, allowing us to simplify this complexity by:

- **Defining a Mean Shape:** Calculating the average geometry of the population to serve as a reference.
- **Capturing Variations:** Identifying the principal "modes" of deformation (eigenmodes). For example, the first few modes often correspond to global changes such as the length or thickness of the bone.
- **Dimensionality Reduction:** Compressing the data from thousands of spatial coordinates into a small set of weighting coefficients, effectively reducing the order of the model from $D \approx 55,000$ to a manageable subspace (e.g., $D = 24$).

1.3 Project Goals

This report details the mathematical formulation and implementation of a Linear PCA-based Statistical Shape Model using a dataset of 24 aligned femur meshes. We aim to demonstrate how this model can identify the main modes of anatomical variation and, crucially, solve the inverse problem: reconstructing a full, realistic 3D bone shape from sparse or partial measurements.

2 Mathematical Formulation

2.1 Data Representation and High-Dimensionality

As defined in the project specifications, our dataset consists of $N = 24$ aligned femur meshes. Each mesh is composed of $M = 18,291$ vertices, where each vertex corresponds to a specific anatomical landmark shared across all subjects.

Mathematically, a single mesh i is initially represented by a matrix $\mathbf{V}_i \in \mathbb{R}^{M\times 3}$. To perform statistical analysis, we flatten this structure into a shape vector $\mathbf{x}_i \in \mathbb{R}^D$, where $D = 3M = 54,873$.

$$\mathbf{x}_i = (x_1, y_1, z_1, \ldots, x_M, y_M, z_M)^\top \tag{1}$$

2.2 Centering and Covariance

We first compute the mean shape $\bar{\mathbf{x}} \in \mathbb{R}^D$ of the population:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \tag{2}$$

We then construct the centered data matrix $\mathbf{X}_c \in \mathbb{R}^{D\times N}$ by stacking the centered vectors column-wise:

$$\mathbf{X}_c = \begin{pmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}}) \\ \cdots \\ (\mathbf{x}_N - \bar{\mathbf{x}}) \end{pmatrix} \tag{3}$$

The standard formulation of PCA requires diagonalizing the sample covariance matrix $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}_c\mathbf{X}_c^\top \in \mathbb{R}^{D\times D} \tag{4}$$

2.3 Computational Strategy: The SVD Approach

In our context, computing $\mathbf{S}$ directly is computationally intractable because $D \gg N$. The covariance matrix would contain approximately 3×10^9 entries.

To solve this "High Dimension, Low Sample Size" problem, we employ **Singular Value Decomposition (SVD)** on the data matrix $\mathbf{X}_c$. SVD decomposes the centered data matrix as:

$$\mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^\top \tag{5}$$

Where:

- $\mathbf{U} \in \mathbb{R}^{D \times N}$ contains the **Principal Components** (eigenvectors of the covariance matrix), also called "Eigen-femurs".
- $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ is a diagonal matrix of singular values σ_k, related to the eigenvalues by $\lambda_k = \frac{\sigma_k^2}{N-1}$.
- $\mathbf{W} \in \mathbb{R}^{N \times N}$ contains the right singular vectors.

This method allows us to recover the principal components without ever explicitly forming the massive $D \times D$ covariance matrix.

2.4 Shape Reconstruction and Imputation

Any femur mesh $\mathbf{x}$ in the dataset can be approximated by a linear combination of the mean shape and the first K principal components (modes of variation):

$$\mathbf{x} \approx \bar{\mathbf{x}} + \sum_{k=1}^{K} b_k \mathbf{u}_k \tag{6}$$

where $\mathbf{u}_k$ are the columns of $\mathbf{U}$ corresponding to the largest singular values, and b_k are the shape parameters (weights).

Since the shape parameters b_k typically follow a normal distribution $\mathcal{N}(0, \lambda_k)$, we can not only reconstruct existing shapes but also generate new, statistically plausible femurs by sampling these parameters. This property is fundamental for the project's goal of imputing missing anatomical data from sparse inputs.

3 Implementation and Analysis of Linear PCA

In this section, we evaluate the capability of the Linear Statistical Shape Model (SSM) to capture the anatomical variability of the femur. The implementation relies on the Iterative PCA class, trained here on the complete training set to establish a baseline for reconstruction performance before addressing the missing data problem.

3.1 Dimensionality Reduction and Analysis of Variance

The primary challenge of this dataset is the "High Dimension, Low Sample Size" (HDLSS) problem, with $D \approx 55,000$ variables (vertex coordinates) but only $N = 24$ samples. To address this, we fit the PCA model using Singular Value Decomposition (SVD) on the flattened feature vectors.

3.1.1 Understanding Variance in Shape Models

In the context of Statistical Shape Modeling, "variance" quantifies the magnitude of morphological deviation from the Mean Shape. Mathematically, the total variance is the sum of the eigenvalues ($\sum \lambda_i$) of the covariance matrix.

- **Large Eigenvalues (High Variance):** Represent global, high-amplitude deformations that affect the entire bone structure (e.g., global scaling, axial twisting).
- **Small Eigenvalues (Low Variance):** Represent local, high-frequency details or noise (e.g., surface roughness).

3.1.2 Interpretation of the Explained Variance Ratio

The cumulative explained variance analysis (Figure 1) provides a quantitative measure of the model's compactness. The curve exhibits an extremely steep initial rise, which allows for two critical interpretations:

1. **High Anatomical Correlation:** The fact that a very small number of components captures the vast majority of information indicates that the 18,291 vertices are highly correlated. They do not move independently; if a vertex on the femoral shaft moves, its neighbors move in a predictable way to maintain surface continuity.
2. **Dimensionality Reduction Efficiency:** We observe that the first $K = 7$ principal components alone are sufficient to capture approximately 98% of the total variance. This implies that the essential anatomical identity of a patient can be compressed into a vector of size 7, rather than 55,000, with minimal loss of geometric fidelity.

The "knee" of the curve suggests that adding components beyond this point yields diminishing returns, as they likely begin to model noise rather than meaningful anatomical features.

3.2 Reconstruction Quality and Error Analysis

To validate the model, we projected the unseen test meshes onto the $K = 7$ principal components and then reconstructed them back into the original 3D space. This process tests the model's ability to generalize to new morphologies using a restricted set of parameters.

While the average reconstruction error provides a global performance metric, analyzing the *worst-case scenario* offers deeper insights into the model's limitations. We identified the test sample yielding the highest global Euclidean error and performed a localized analysis.

The error distribution visualization (Figure 3) is particularly instructive. By mapping the vertex-wise L2 norms onto the geometry, we observe that:

- **Low Error Regions:** The femoral shaft (diaphysis) consistently shows low reconstruction errors. This is expected, as this region has low curvature and low morphological variability across the population.

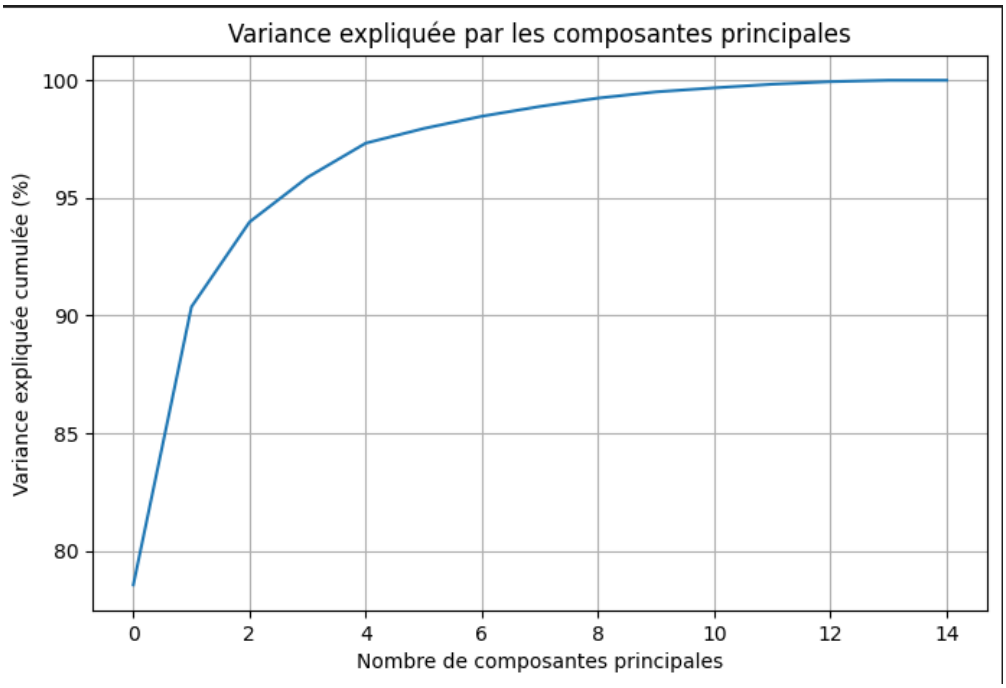

Figure 1: Cumulative explained variance vs. number of components. The rapid saturation of the curve confirms that the femur anatomy resides on a low-dimensional manifold within the high-dimensional vertex space.

Figure 2: Comparison between the original Ground Truth mesh (left) and the PCA Reconstruction (right) for the worst-performing test case.

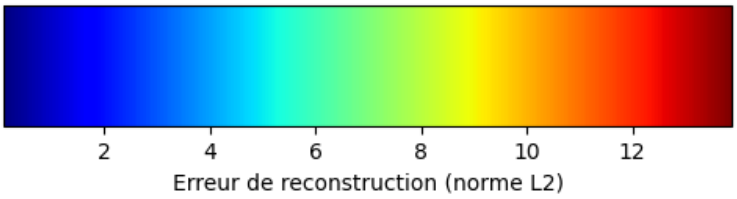

Figure 3: Vertex-wise L2 reconstruction error mapped onto the mesh. Warmer colors (red/yellow) indicate higher deviation from the ground truth.

- **High Error Regions:** The reconstruction artifacts are concentrated in areas of high anatomical complexity, such as the femoral head and the condyles. These articular surfaces exhibit subtle, non-linear variations that are difficult to capture fully with only 7 linear components.

This analysis confirms that while a low-rank linear model successfully captures the global shape, it acts as a low-pass filter, smoothing out specific high-frequency details in the most variable anatomical regions.

4 Impact of Feature Augmentation

To further improve the model's performance, we explored the technique of *Feature Augmentation*. The core idea is to enrich the input data by adding a 4^{th} dimension to the standard (x, y, z) coordinates of each vertex, potentially helping the linear PCA model capture non-linear anatomical correlations.

4.1 Part 1: Homogeneous Augmentation (Ones)

As a baseline experiment, we first augmented the data by appending a constant value of 1 to every vertex, transforming each point $\mathbf{v}_i$ from (x, y, z) to $(x, y, z, 1)$. This is akin to using homogeneous coordinates in projective geometry.

Analysis: Mathematically, the variance of a constant variable is zero. Adding a dimension with zero variance implies that it carries no information regarding the shape variability. Consequently, the Covariance Matrix simply gains a row and column of zeros (or negligible values), leaving the eigenvalues (and thus the principal components) unchanged.

As illustrated in Figure 4, the explained variance curve for this method is identical to the standard PCA. This serves as a validation of our pipeline: simply increasing dimensionality without adding information does not improve the model.

4.2 Part 2: Radial Augmentation (Barycenter)

In the second approach, we augmented the data by adding the **Euclidean distance to the barycenter** as the 4^{th} dimension. For a vertex $v_i = (x_i, y_i, z_i)$ and the mesh barycenter B, the new feature vector is (x_i, y_i, z_i, d_i), where $d_i = ||v_i - B||_2$.

4.2.1 Variance Analysis

Unlike the constant augmentation, the distance to the barycenter varies significantly across the population, encoding information about the "volume" or "inflation" of the bone.

Figure 5 clearly shows the Barycenter curve "above" the others. This higher Area Under the Curve (AUC) implies that for the same number of components K, the Barycenter-

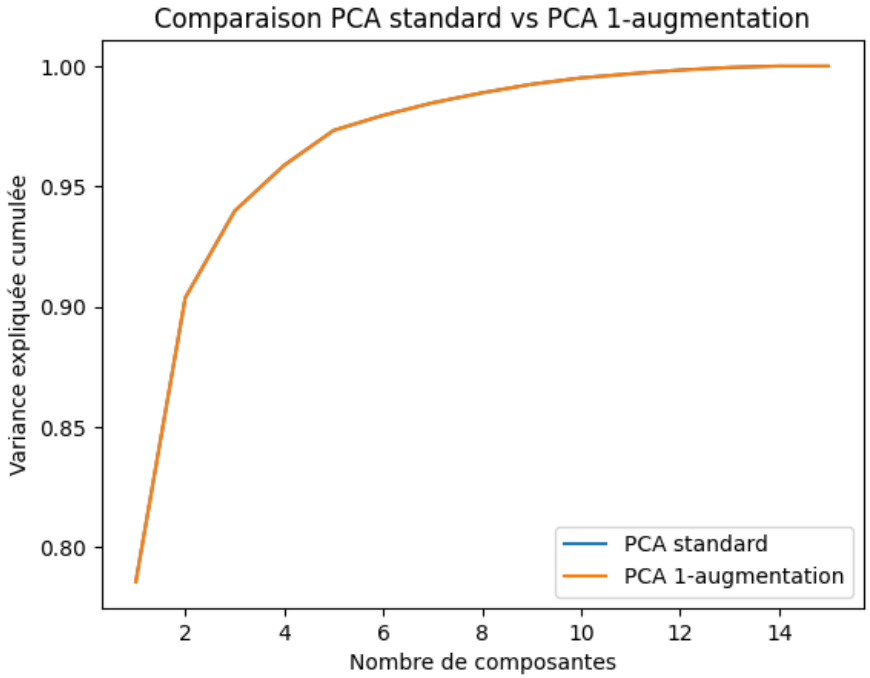

Figure 4: Comparison of Cumulative Explained Variance: Standard PCA vs. PCA with Ones Augmentation. The curves perfectly overlap, confirming that adding a constant feature adds no predictive power to the model.

augmented model captures more shape variability.

4.2.2 Heatmap Comparison and Interpretation

To understand *where* this improvement comes from, we compared the reconstruction error heatmaps of the Standard PCA versus the Barycenter PCA.

The visual improvement is most striking in the Femoral Head (the "ball" of the femur), as seen in Figure 7.

- **The Limitation of Linear PCA:** The femoral head is roughly spherical. A change in its size (e.g., expansion) is a non-linear transformation in Cartesian coordinates ($x^2 + y^2 + z^2 = r^2$). A standard Linear PCA struggles to approximate this curvature using only linear combinations of x, y, z vectors, resulting in higher errors (red zones) on spherical surfaces.

- **The Barycenter Advantage:** By explicitly providing the *radius* (d) as a feature, we "linearize" the problem. An expansion of the femoral head corresponds to a simple linear increase in this 4^{th} dimension. The PCA can easily learn to scale this parameter, effectively capturing the "inflation/deflation" mode of the joint without needing complex combinations of spatial coordinates. This allows the model to reconstruct the spherical curvature with much higher fidelity.

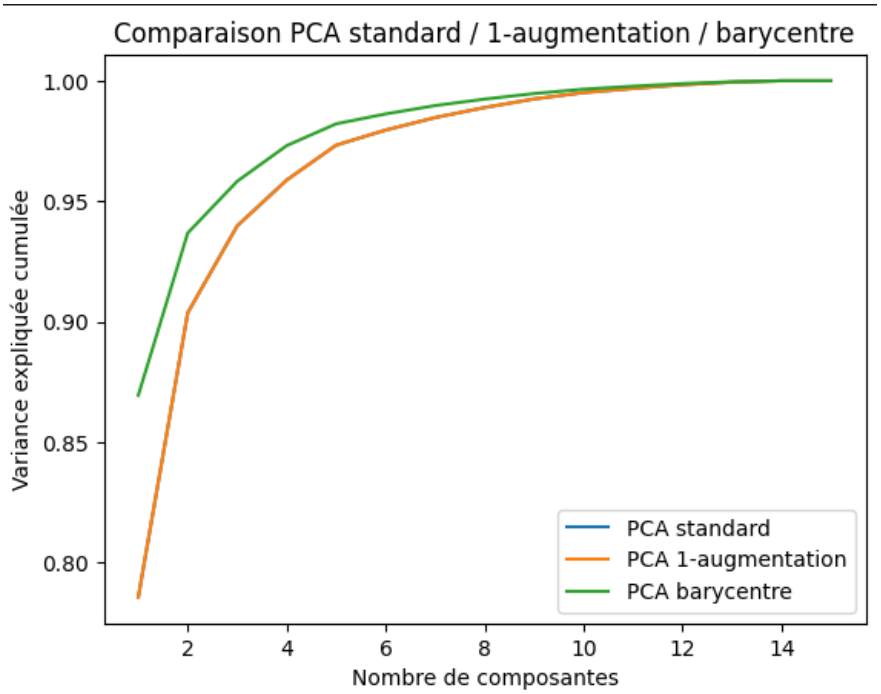

Figure 5: Cumulative Explained Variance of the three methods. The Barycenter curve (Green) rises significantly faster than the Standard PCA (Blue) and Ones (Orange), indicating a more efficient compression of anatomical information.

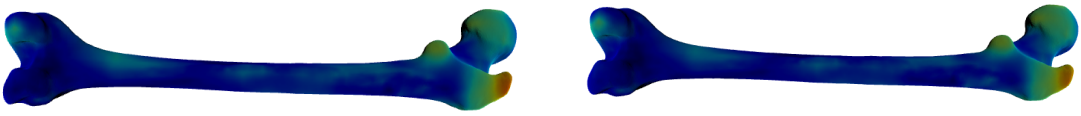

Figure 6: Vertex-wise reconstruction error. Left: Standard PCA. Right: Barycenter Augmented PCA. The global error reduction is visible, particularly in the extremities.

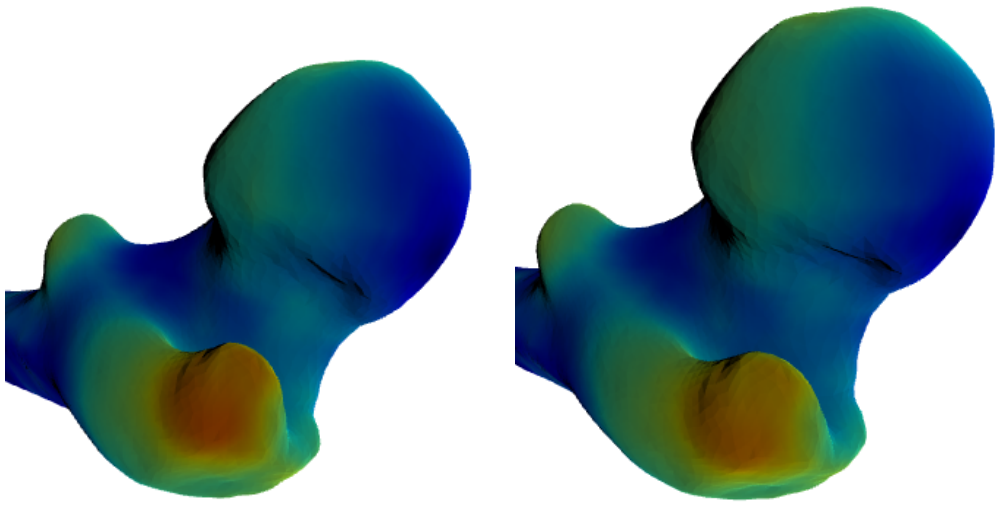

Figure 7: Zoom on the Femoral Head. The Standard PCA (Left) shows a distinct "red" error hotspot. The Barycenter PCA (Right) shows a much lighter color, indicating significantly improved reconstruction in this spherical region.