Future Research

Jordi Cortés, Daniel Fernández, Liukuan Yu, Nora Amama

XI RETREAT GRBIO
VILANOVA I LA GELTRÚ
JULY 1st, 2026

IRIS Institute for Research and Innovation in Health

GRBIO

UPC Departament d'Estadística i Investigació Operativa
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Combining Several Metrics for Synthetic Data Validation

How to apply several resemblance metrics at once?

How many resemblance metrics can be applied simultaneously without inflating Type I error?

PROBLEM

- **Many resemblance metrics** are reported in current practice.
- Each metric defines its **own hypothesis test**.
- Applying them jointly creates a **multiplicity problem**.
- **No specific recommendations** exist for synthetic data validation.

GOAL

Develop a statistically valid framework for selecting and combining resemblance metrics.

1 QUESTION 1

Which metrics will be included?

Univariate

- Kolmogorov-Smirnov
- Anderson-Darling
- Jensen-Shannon

Bivariate

- Correlation difference
- Mutual information difference

Multivariate: Classifier-based

- pMSE
- SPECKS
- PO50

Multivariate: Distribution-based

- Energy distance
- Multivariate Cramer-von Mises

Candidate families: each captures a different aspect of resemblance, so they are complementary, not redundant. We focus on a subset, shown next.

Reference: Hernandez et al. (2022). Synthetic data generation for tabular health records: a systematic review. Neurocomputing, 493, 28-45.

1 WHERE WE FOCUS

REALISTIC PLAN · EARLY OCT 2026

Where we focus: done and planned

Univariate

- **Kolmogorov–Smirnov**
- Chi-square · planned

Bivariate

- Pearson correlation-difference test · planned
- Spearman correlation-difference test · planned

Multivariate: Classifier-based

- **SPECKS**
- **pMSE**

Multivariate: Distribution-based

- Multivariate Cramér–von Mises · planned

What we will do: in bold, the metrics with results so far (KS, SPECKS and pMSE). The rest is the concrete plan we will add.

1

QUESTION 1

Why these metrics?

Univariate	**Marginal distortions one variable** at a time.
Bivariate	**Pairwise** dependence Changes.
Classifier-based	**Multivariate distinguishability** between real and synthetic data.
Distribution-based	**Multivariate discrepancy** structure.

P

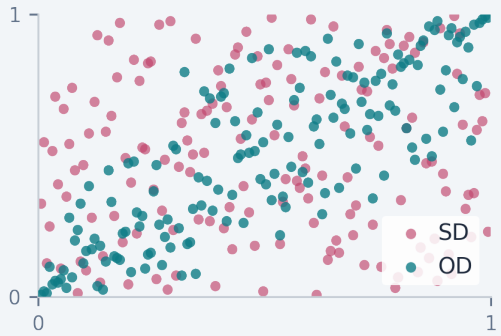
A PLANNED METRIC · DEPENDENCE / COPULA-BASED

Cramér–von Mises on copulas (planned)

THE STATISTIC

$$S_{o,s} = \frac{n_o n_s}{n_o + n_s} \int_{[0,1]^q} \left[\hat{C}_o(u) - \hat{C}_s(u)\right]^2 d\hat{C}(u)$$

Integrated squared gap between the two empirical copulas, built from ranks so the marginals are removed. **We are comparing dependencies, not the marginal distributions**



Same uniform marginals, different dependence.

1 **Build empirical copulas** → 2 **Get Cramér-von Mises statistic** → 3 **P-Value using Bootstrap**

Classifier-free: it compares the dependence structure (copula) directly, so it complements the classifier-based SPECKS and pMSE.

Rémillard & Scaillet (2009), Testing for equality between two copulas, J. Multivariate Anal. 100(3), 377–386 (empirical copulas; multiplier bootstrap; R package TwoCop).

2 QUESTION 2

How to control Type I error under multiplicity?

FWER control

Family-wise error rate

- **Bonferroni**
- **Holm**
- **Hochberg**
- **Hommel**

Controls the **probability of even one false rejection** (FWER ≤ α). Safe but conservative as the panel grows.

FDR control

False discovery rate

- **Benjamini-Hochberg**
- **Benjamini-Yekutieli**

Controls the **expected proportion of false rejections** (FDR). Less stringent, so it retains more power to detect SD that truly differ.

These adjustments are standard, but **applying them to synthetic-data resemblance validation is new**. We identify which methods keep the Type I error controlled.

Chen SY, Feng Z, Yi X (2017). A general introduction to adjustment for multiple comparisons. J Thorac Dis 9(6):1725-1729.
Raab GM, Nowok B, Dibben C (2021). Assessing, visualizing and improving the utility of synthetic data. arXiv:2109.12717.

QUESTION 3

What scenarios of bad synthetic data are more detectable? Simulation framework

Data-generating mechanisms

H_0: OD and SD share the same distribution.

H_1 scenarios:

1. Mean shift
2. Variance distortion
3. Correlation distortion
4. Marginal distortion only
5. Joint-structure distortion only

Outputs

- Empirical Type I error
- FWER and FDR
- Statistical power

- Start with **simulated data**
- Then replicate on **public datasets**. Real-data tracks via Siemens Energy and US synthetic data (via Georgetown University)
- Finally, check different **generation algorithms**.

Open-source implementation

Planned package functionalities

- Compute the full panel of **resemblance metrics** from OD and SD inputs.
- Apply the chosen **multiplicity adjustment** (Bonferroni, Holm, Hochberg, Hommel, BH, BY).
- Return the union decision plus **adjusted p-values** and **interpretable values** for every metric.
- Reproducible **simulation utilities** for Type I error and statistical power.
- **Diagnostic plots** per metric and per family (e.g., statistical power curves and decision heatmaps).
- Paired **toy datasets**, real and synthetic, to try the package.

CRAN integration makes the framework citable, installable and reusable by the community.

Generating Synthetic Data Using AA/ADA

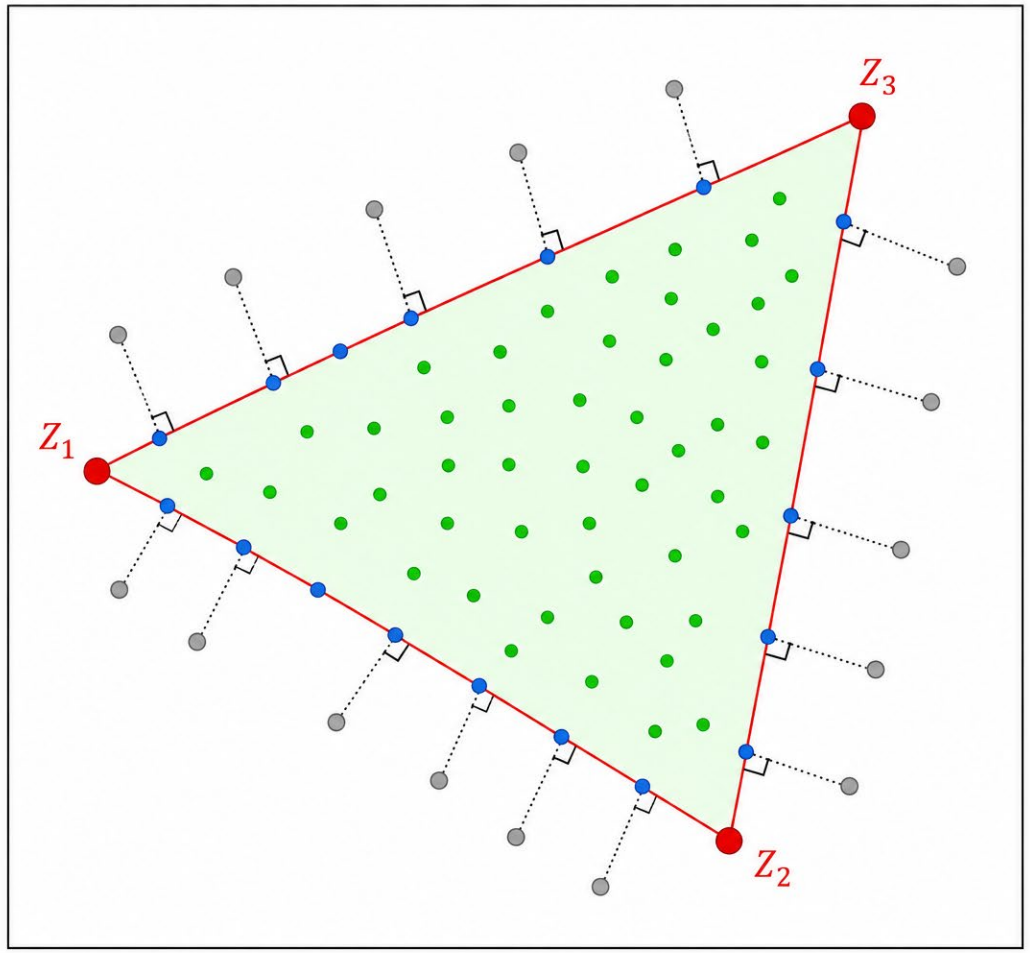
1 QUESTION 1

Can residual augmentation improve the generator?

! Current limitation

- Pure AA-Dirichlet generation uses **$\tilde{X} = \tilde{A}Z$**.
  - Synthetic records stay **within the prototype hull**.
  - Increasing k improves coverage, **but local OD may remain outside**.
- **Higher rejection rate of**

$H_0: F_p^{OD}(p) = F_p^{SD}(p)$.



Archetypes
Z_1, Z_2, Z_3

Simplex (convex hull) spanned by the archetypes

Real points inside the simplex (perfectly reconstructed)

Synthetic points generated on the simplex boundary

Real points outside the simplex (not perfectly reconstructed)

Residuals (E)
Perpendicular distance from each real point to its projection on the simplex boundary

Residual for point i:

$$E_i = \| x_i^{real} - x_i^{synthetic} \|_2$$

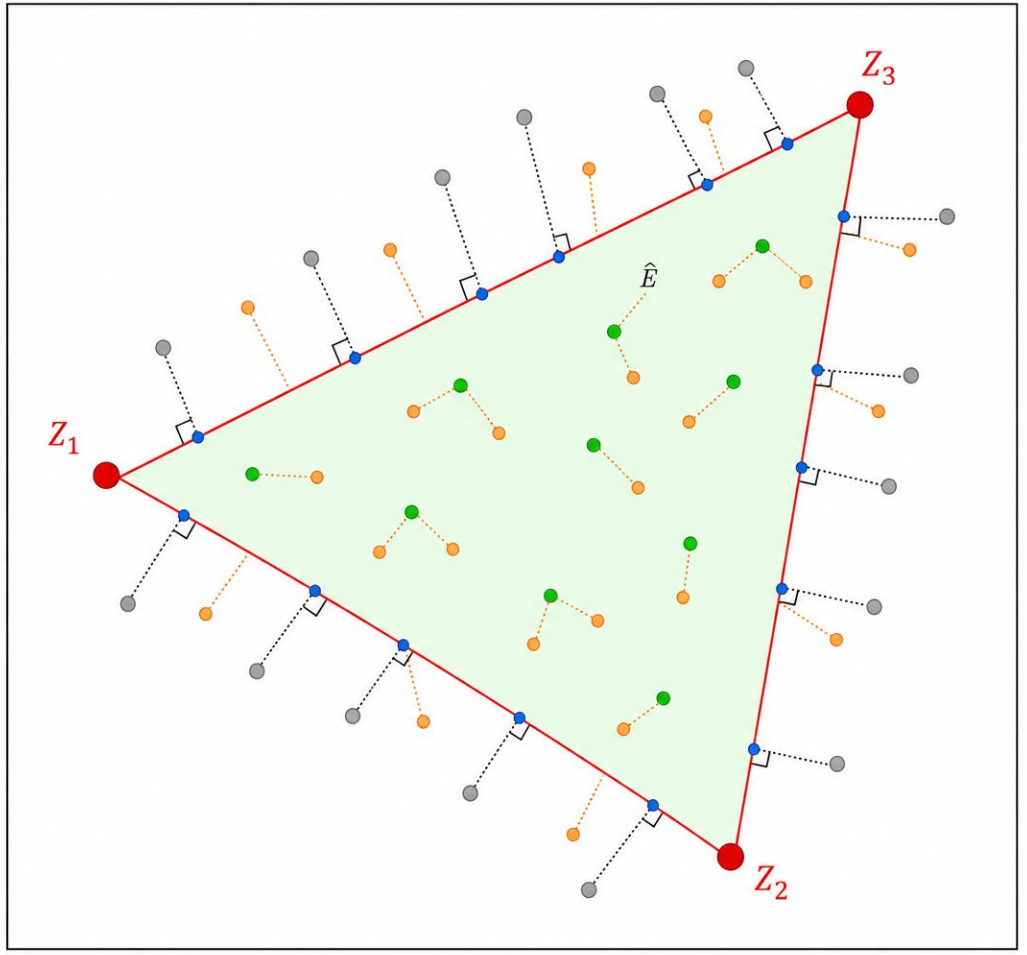
Euclidean distance between the real point and its projection.

1 QUESTION 1

Can residual augmentation improve the generator?

+ Planned residual step

- Calculate residuals: $E = X - \tilde{A}Z$.
- Generate residual noise, e.g., $\hat{E} \sim N(\bar{E}, S_E)$.
- Final synthetic data: $\tilde{X} = \tilde{A}Z + \hat{E}$.



Archetypes
Z_1, Z_2, Z_3

Simplex (convex hull) spanned by the archetypes

Real points inside the simplex (perfectly reconstructed)

Synthetic points generated on the simplex boundary

Perturbed points (synthetic with additional noise)

Generated noise (from the corresponding green or blue point)

Real points outside the simplex (not perfectly reconstructed)

Residuals (E)
Perpendicular distance from each real point to its projection on the simplex boundary

Generated residual (noise) for point i:

$$\hat{E}_i = \| x_i^{perturbed} - x_i^{target} \|_2$$

$$\hat{E}_i \sim \mathcal{N}(\bar{E}, S_E^2)$$

with $\bar{E} = \text{mean}(E)$ and $S_E = \text{sd}(E)$

2

QUESTION 2

Does the AA simulation framework also work for ADA?

! Why ADA next?

- ADA **uses real observations** as archetypoids.
- Stronger **interpretability** than AA:
  - Every synthetic point has a percentage of any Archetypoid.
- Drawback: **Computational cost**.

Simulation plan

- Reuse the **18 scenarios**: n = 30 or 100; p = 3, 5, 10; ρ = 0, 0.3, 0.5.
- Fit **ADA** across k and compute RSS/PC paths.
- Use **PC* = 0.05** as the initial stopping rule.
- Evaluate **rejection rate** and **coverage**.

Key comparison: AA-Dirichlet vs ADA-Dirichlet · same scenarios, same validation protocol, different prototype constraint.

Penalized ADA (PADA)

Next Steps

1 Sensitivity Analysis in λ

How the penalty **parameter λ** affects the selection of archetypoids.

2 Simulation Study

Simulation experiments to evaluate the method under **known ground truth conditions**.

3 Radial Approach

Alternative ways of incorporating representativeness via a radial approach for selecting archetypoids.

4 R Package

R package implementing the proposed methodology

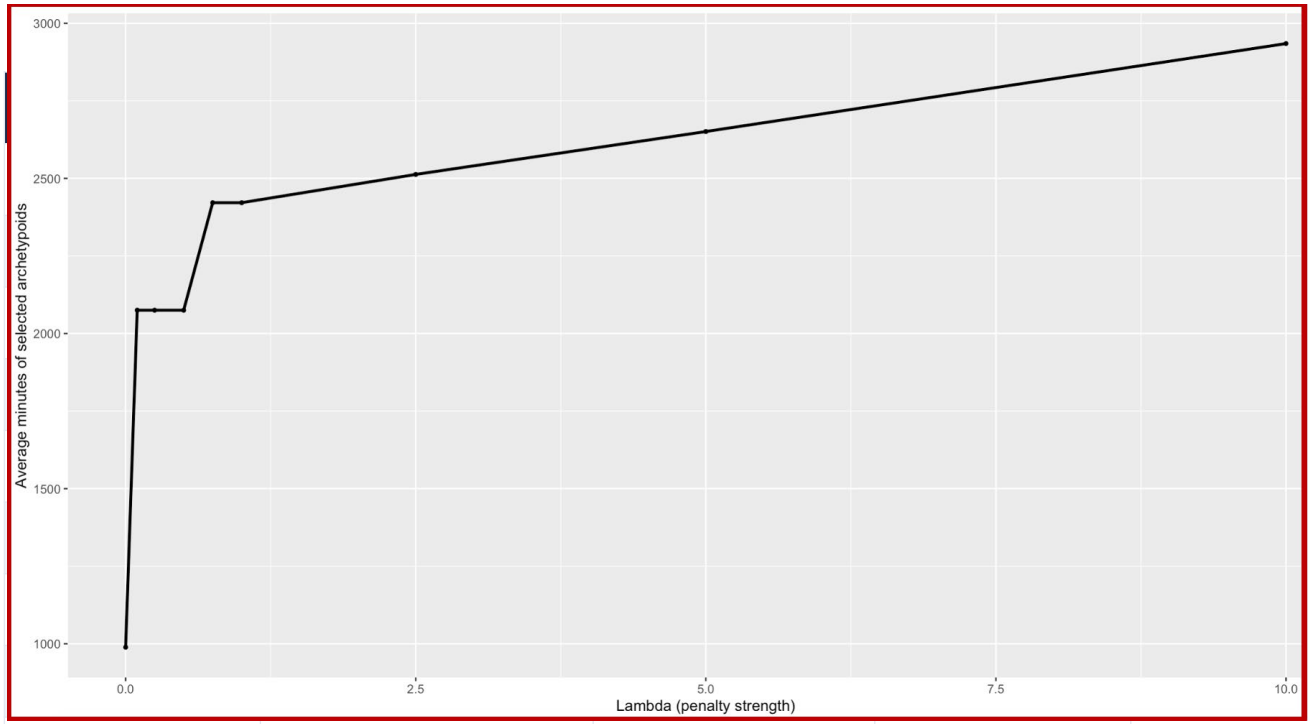
Influence of the Penalty (λ)

Case-study
1. NBA players
2. We want representativeness through minutes (M) played

For fixed k = 3, compute the **average minutes** of selected archetypoids across a range of λ values.

1. **Avg. minutes ↑** → more representative players.
2. **RSS ↑** → less extreme archetypoids.
3. **Ratio (min/RSS) ↑** → better balance?

Results for varying λ (k = 3)

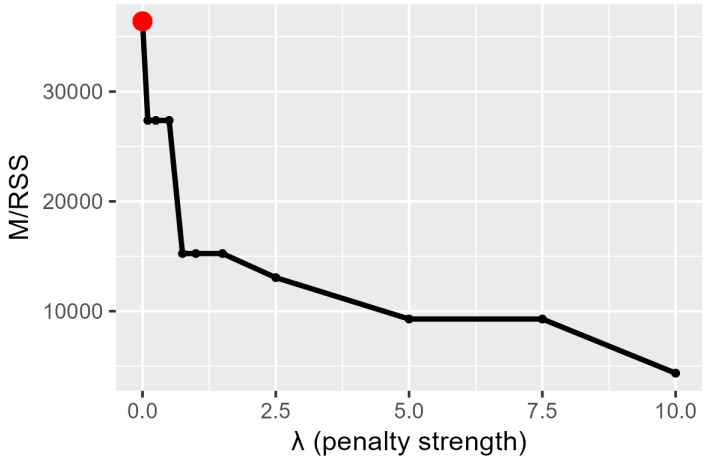


Large λ favors **high-minute players** at the cost of a higher RSS.

How to choose λ?

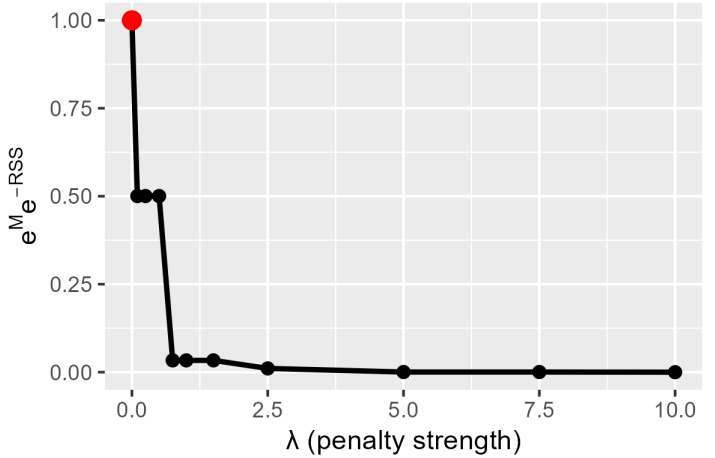
1 **Minutes-to-RSS Ratio (M/RSS).** Choose the value of λ according to:

$$\lambda_1^* = \text{argmax}_\lambda \frac{\bar{M}(\lambda)}{RSS(\lambda)}$$



2 **Exponential Trade-off Criterion.** Choose the value of λ according to:

$$\lambda_2^* = \text{argmax}_\lambda \left[exp\left(-\frac{RSS(\lambda)}{RSS(0)} \right) exp\left(\frac{\bar{M}(\lambda)}{\bar{M}(0)} \right) \right]$$

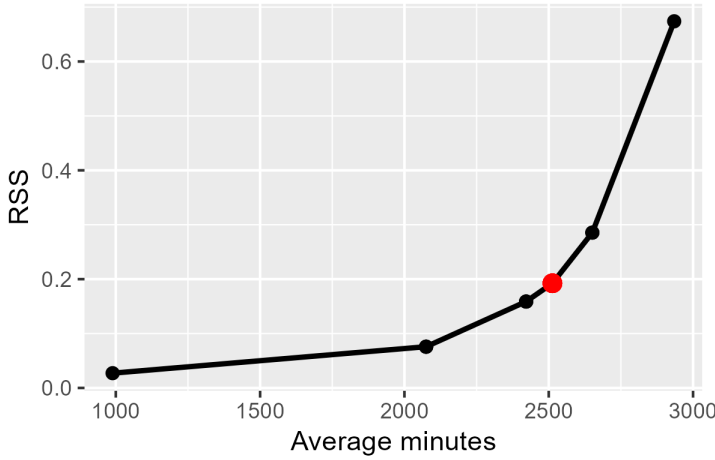


3 **Pareto Knee (Elbow) Criterion.** For each value of λ we have the pair:

$$(RSS(\lambda), M(\lambda))$$

Choose the value of λ where:

- further increases produce only small gains in representativeness, but
- substantially increase RSS.

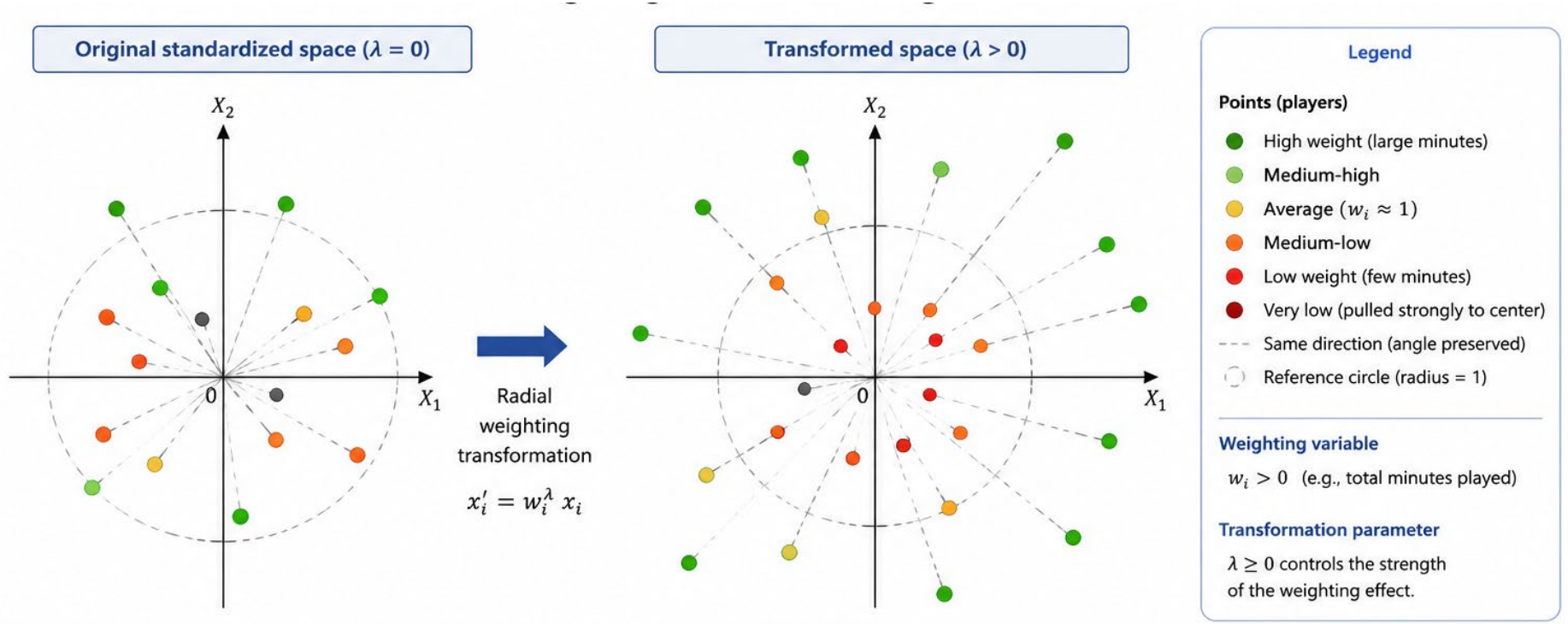


Radial Approach - RADA (Geometric intuition)

- It requires previous **standardization**

In a different way of PADA, here:

- We **keep the original optimization** problem but…
- We **change the initial geometry** of the points.



Geometric effect

- All points move along the same radial line.
- Points with $w_i > 1$ are pushed away from the origin (more prominent).
- Points with $w_i < 1$ are pulled toward the origin (less prominent).
- Points with $w_i = 1$ remain unchanged.

Key properties

- Radial (isotropic)
- Preserves angles between points
- Smooth and continuous
- Preserves overall structure

Interpretation

Players with more minutes (higher w_i) become more "extreme" (further from the center) and thus more likely to be selected as archetypoids.

Players with few minutes are shrunk toward the center and are less likely to be selected.

When $\lambda = 0$, no transformation is applied. *As* λ, increases, the differences induced by the weights are amplified.

Future Research

XI RETREAT GRBIO
VILANOVA I LA GELTRÚ
JULY 1st, 2026