**Normalization/Standardization Strategy for AlignedUMAP**

**Data Structure:** Data is oriented in long-format. Columns designate features and rows indicate session summaries. The AlignedUMAP technique enables the visualization of individual and group trajectories across a temporal measurement (in our case, sessions). Depending on the features of interest used to generate the AlignedUMAP, standardization of the feature measurements is crucial to avoid skew in feature influence within embedding space.

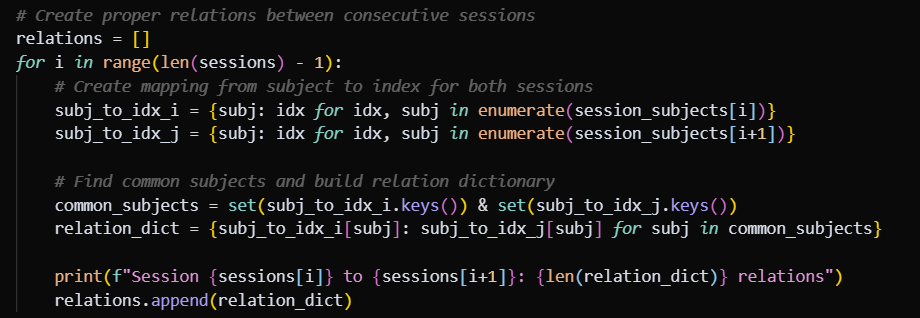
**I want to ensure we are using the correct standardization technique.**

**Perplexity Inquiry:** “I have temporal data for individual subjects across 15 sessions. I have multiple features for each session for each subject. I am using AlignedUMAP to map trajectories of individual subjects across sessions. I used StandardScalar normalization for each feature because they are measured on different scales. I am unsure whether standardization should be implemented session by session, standardize to the first session, or standardize across all sessions. I am interested in looking at treatment differences across sessions of the subjects that can be grouped by treatment.”

**Result:**

* Standardize each session separately
  + What it does: For each session, you standardize features across all subjects (mean=0, std=1, for each feature in each session).
  + Pros: Removes session-specific mean and variance, highlighting within-session differences.
  + Cons: Destroys absolute changes over time. You lose the ability to track how feature values evolve across sessions, since each session is “reset” to zero mean and unit variance.
* Standardize to the First Session
  + What it does: Use the mean and STD from the first session to standardize all subsequent sessions.
  + Pros: Preserves the interpretations of change relative to baseline (first session). Good if you want to track deviations from baseline across time.
  + Cons: Assumes the first session is a meaningful reference for all subjects and treatments. If baseline difference exists between groups, this could introduce bias.
* Standardize Across All Sessions
  + What it does: Compute mean and STD across all sessions and subjects for each feature and use these to standardize the entire dataset.
  + Pros: Preserves relative differences and temporal trajectories. Ensures comparability across sessions and treatments.
  + Cons: If treatment groups are very different at baseline, this could mask those differences somewhat.
* **Recommended Best Practice:** Standardize across all sessions and all subjects for each feature.
  + Ensures that all features are on the same scale, allowing AlignedUMAP to learn meaningful trajectories.
  + It preserves the temporal structure and allows you to see how subjects move through feature space over time, which is essential for trajectory analysis.
  + Avoids artificially removing temporal trends or introducing bias from a particular session.

Setting up AlignedUMAP:

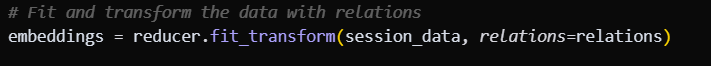


* This code snippet creates the relationships between consecutive sessions by:
  + Creating dictionaries mapping subjects to their indices in each session
  + Finding common subjects between consecutive sessions
  + Creating a relation dictionary that maps indices between sessions for the same subjects
  + **This step is crucial for AlignedUMAP as it helps maintain the temporal relationship between sessions to map trajectories.**

A screen shot of a computer

AI-generated content may be incorrect.

* This configures the AlignedUMAP algorithm with:
  + 2D embeddings (with session as a third dimension)
  + 15 nearest neighbors for local structure preservation
  + Minimum distance of 0.1 between points
  + Euclidean distance metric
  + Fixed random state for reproducibility



* This creates the actual embeddings by:
  + Fitting the AlignedUMAP model to the session data.
  + Using the relations to maintain temporal consistency
  + Producing a 2D embedding for each subject in each session.

A computer screen shot of text

AI-generated content may be incorrect.

* This creates a mapping dictionary that:
  + Links each (subject, session) pair to its corresponding position in the embedding array.
  + Makes it easier to look up specific subjects’ embeddings later.

A graph of a diagram

AI-generated content may be incorrect.Standardization (StandardScalar) across sessions.

Standardization (StandardScalar) within session.

A graph of a graph with red and blue dots

AI-generated content may be incorrect.

Clusters across sessions

A diagram of a cluster of blue dots

AI-generated content may be incorrect.

Clusters within session.

A graph of blue dots

AI-generated content may be incorrect.