# Statistical Methods For High Dimensional Data Project Report

Maxim Nitsenko

Università degli studi di Padova, Dipartimento di Matematica, Padova, Italia
E-mail: maximnikolaevich.nitsenko@student.unipd.it
Matricola: 2129684

## 1. Introduction

This study addresses **feature selection** and **interpretability** in the context of Alzheimer' s detection using handwriting data from the DARWIN dataset. The goal is not predictive accuracy, but the identification of a **small** set of informative and interpretable variables that distinguish between healthy and diseased individuals.

A central challenge is the **high correlation** among features: 251 pairs have absolute Pearson correlation above 0.8, and 132 above 0.9. In interpretability-focused settings, it is often preferable to select entire groups of correlated variables rather than a single representative, as grouped features may reflect shared underlying mechanisms or diagnostic patterns.

To address this, we employ methods that explicitly support joint selection of correlated variables, including **elastic net** regularization, **group lasso** (with both task/feature-based and correlation-based groupings), and **group best-subset selection**. **Stability selection** is applied across all methods to ensure robustness, retaining only features with consistently high selection probability across resampled subsets.

## 2. Dataset Description

The dataset used in this study is the DARWIN (Diagnosis AlzheimeR WIth haNdwriting) dataset, designed to support early diagnosis of Alzheimer's Disease through handwriting analysis. Data were collected following a standardized experimental protocol aimed at capturing motor and cognitive alterations in handwriting dynamics typically associated with neurodegenerative conditions.

The dataset includes handwriting samples from **174 participants**: **89 patients diagnosed with Alzheimer's Disease** and **85 healthy controls**. Participants were matched across demographic variables such as age, education, gender, and occupational background to minimize confounding.

### 2.1 Tasks

Each participant performed **25 distinct handwriting tasks**, grouped into four categories:

- **Graphic tasks**: Drawing lines, circles, and retracing forms (simple motor control).
- **Copy tasks**: Copying letters, words, and paragraphs, including mirror/reverse copy tasks.

- **Memory tasks**: Recalling and writing words or names associated with images.
- **Dictation tasks**: Writing dictated content (words, phrases, numbers), engaging working memory.

Handwriting was recorded at 200 Hz using a Wacom tablet, capturing x-y coordinates, pressure, and in-air movements.

### 2.2 Features

From each handwriting task, **18 kinematic and dynamic features** were extracted, encompassing both temporal and spatial dimensions. These include:

- Total writing time, in-air time, and on-surface time.
- Speed, acceleration, and jerk for both on-surface and in-air segments.
- Pressure statistics (mean and variance).
- Pen-down event count and spatial dispersion measures.
- Generalized tremor indices derived from path irregularity.

Each participant is thus represented by $25 \times 18 = 450$ features. Given the sample size ($n = 174$) and feature dimensionality ($p = 450$), this constitutes a **high-dimensional** $p \gg n$ learning problem.

## 3. Stability Selection

Cross-validation techniques are commonly employed to select the regularization parameter $\lambda$ in Lasso regression due to their favorable predictive performance. However, such methods often result in inconsistent model selection in the context of sparse models. Specifically, while cross-validation may yield a $\lambda$ that minimizes prediction error, it does not necessarily lead to reliable identification of the true underlying variables.
To address the instability of variable selection inherent to methods like the Lasso, Meinshausen and Bühlmann (2010) introduced *Stability Selection*, a framework designed to improve the robustness of variable selection in high-dimensional settings, particularly when the number of predictors $p$ greatly exceeds the number of observations $n$.
Stability Selection combines a subsampling strategy with a base selection algorithm (e.g., Lasso or graphical Lasso) to identify variables that are consistently selected across multiple random subsets of the data. This approach mitigates the sensitivity of variable selection procedures to small changes in the data and provides a measure of selection reliability.

### Methodology

The key steps of Stability Selection are as follows:

1. **Subsampling:** Randomly draw multiple subsets of the data (typically of size $\lfloor n/2 \rfloor$) without replacement.
2. **Variable Selection:** Apply a base variable selection method (e.g., Lasso) independently to each subsample to identify active variables.

3. **Selection Probability:** For each variable, compute the empirical selection probability, i.e., the proportion of subsamples in which the variable is selected.

4. **Stable Variable Set:** Define a selection probability threshold $\pi_{\text{thr}}$ (e.g., 0.8), and retain variables whose selection probabilities exceed this threshold.

This approach yields a stable set of variables that are more likely to reflect true underlying associations, offering an interpretable and statistically grounded alternative to single-run variable selection procedures.

# 4. ELASTIC NET

When variables are highly correlated, the Lasso tends to select only one variable from each group, and does so arbitrarily. This behavior is problematic when the results are intended for hypothesis generation. In such cases, we would prefer to identify all variables that exhibit similar evidence of association with the outcome, not just a single representative.

In contrast, Ridge regression (based on the $\ell_2$ penalty) tends to retain all correlated variables, distributing the effect among them. Elastic Net combines both penalties to balance these behaviors. It introduces a mixing parameter $\alpha \in [0, 1]$, where $\alpha = 1$ corresponds to Lasso and $\alpha = 0$ to Ridge. This allows the model to benefit from both sparsity and grouping effects.

## 4.1 $\alpha$ and $\lambda$ Selection

To optimize the trade-off between sparsity and group retention, a custom metric $\phi$ is introduced. Let $\mathcal{X}$ denote the complete set of all variables in the dataset, and let $S$ denote the set of variables selected by the elastic net with parameters $\alpha$ and $\lambda$. For each variable $X \in S$, define

$$\pi_X := \frac{\sum\limits_{\substack{X_i \in S \\ X_i \neq X}} |\rho(X, X_i)|}{\sum\limits_{\substack{X_i \in \mathcal{X} \\ X_i \neq X}} |\rho(X, X_i)|}$$

as the proportion of total correlation for $X$ that is retained within the selected set. The average proportion of included correlation is then given by

$$\pi := \frac{1}{|S|} \sum_{X_i \in S} \pi_{X_i}.$$

The objective is to maximize $\pi$ to favor the inclusion of highly correlated variables, while also encouraging sparsity. Define the sparsity metric as

$$\sigma := 1 - \frac{|S|}{|\mathcal{X}|},$$

which measures the proportion of variables excluded from the model. The objective is to select $(\alpha, \lambda)$ that maximizes the joint criterion

$$\phi := \pi \cdot \sigma.$$

$\phi$ is computed over a grid of $(\alpha, \lambda)$ values using the elastic net, and the top configurations maximizing $\phi$ are identified. To assess the robustness of variable selection, a stability

analysis is performed: for each of the top 10 configurations, the dataset is repeatedly subsampled and the elastic net model is fitted with the corresponding $(\alpha, \lambda)$. For each feature, its selection frequency across 1000 subsamples is computed. Finally, variables that are selected in at least 90% of these subsamples under at least one of the top configurations are retained. This yields a stable and interpretable feature set.

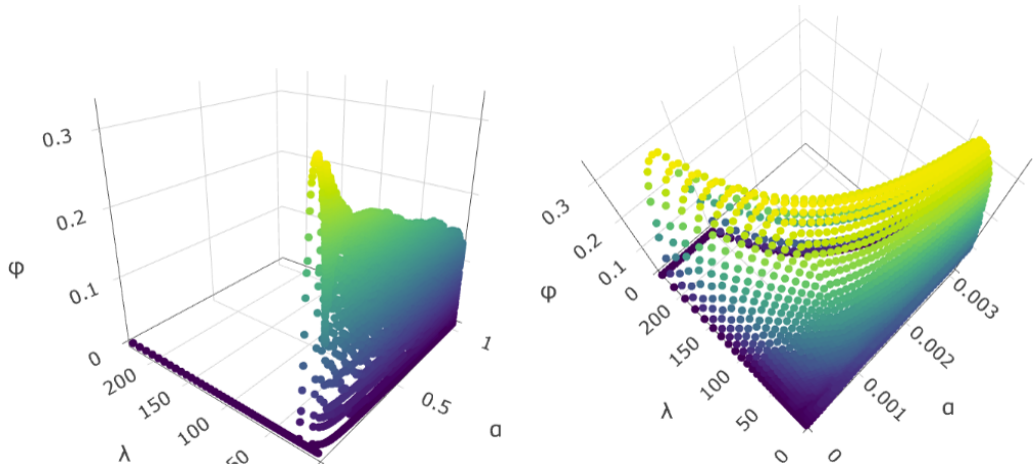


Figure 1: $\phi$ evaluated over the $(\alpha, \lambda)$ grid (left) and zoomed view near the optimum (right).

### 4.2 Results

The elastic net procedure with stability selection identified a set of stable features for Alzheimer's detection from handwriting data. Features from all tasks except 1, 4, 14, 19, and 20 were selected, indicating that most tasks contributed informative features. Commonly selected features included temporal variables (`paper_time`, `total_time`, `air_time`), movement characteristics (`disp_index`, `mean_jerk`, `gmrt`), and pen pressure (`pressure_mean`), reflecting important variations in handwriting patterns relevant to Alzheimer's detection.

## 5. GROUP LASSO

Group Lasso extends standard Lasso for datasets where predictors are naturally organized into groups. By applying an $\ell_2$ norm penalty to each group, it encourages group-level sparsity, either selecting all variables in a group or excluding the entire group simultaneously. To alleviate statistical variation and obtain a more precise estimate of truly important tasks, this was coupled with resampling. Stability selection was used to enhance the reliability of group selection by mitigating variability inherent in high-dimensional settings and ensuring that only robustly selected groups are retained.

### 5.1 Results

Group Lasso was applied to the DARWIN dataset under three distinct group structures:

1. **Task-Based Groups**: Variables grouped by handwriting task identity (25 groups).

- **Stable Tasks**: Tasks 3, 7, 8, 9, 15, 17 and 19 were consistently selected, suggesting that these specific writing activities are more sensitive to cognitive and motor impairments associated with Alzheimer's Disease.

2. **Feature-Based Groups**: Variables grouped by feature type across all tasks (18 groups).

   - **Stable Features**: The most frequently selected features included `air_time`, `disp_index`, `max_x_extension`, and `paper_time`. These metrics capture key aspects of motor control and handwriting variability that appear to differentiate patients from controls.

3. **Correlation-Derived Groups**: Features grouped into connected components of a graph constructed from pairwise correlations thresholded at $|\rho| > 0.8$ (260 groups).

   - **Stable Components**: Three components were consistently selected: one containing temporal features `air_time` and `total_time` from task 15, another containing `num_of_pendown` from task 19, and a third containing `air_time` and `total_time` from task 24, indicating that timing coordination in specific tasks and pen-lifting behaviors are key discriminative features.

# 6. GROUP BEST SUBSET SELECTION

Group Best Subset Selection (GBSS) is an extension of traditional best subset selection that operates at the group level rather than the individual feature level. This method is particularly effective in high-dimensional settings, where evaluating all feature subsets is computationally prohibitive. By operating on predefined feature groupings, GBSS reduces the search space and enhances interpretability.

The procedure consists of the following steps:

1. **Group Partitioning**: Features are partitioned into groups. Two grouping structures are considered: task-based and feature-based, as previously described.

2. **Subset Evaluation**: For a given group size parameter $s$, all combinations of $s$ groups are considered. For each combination, a logistic regression model is fit and the model deviance is recorded.

3. **Cross-Validation**: To select the optimal number of groups $s$, $k$-fold cross-validation is performed. For each value of $s$, the average misclassification error across folds is computed, and the value $s^*$ minimizing this error is retained.

4. **Resampling-based Group Selection**: After selecting the optimal subset size $s^*$ via cross-validation, the GBSS procedure is applied repeatedly to random subsamples of the data. For each subsample, the selected group combination of size $s^*$ is recorded. The frequency of each unique group subset is then computed, allowing identification of the most consistently selected group combinations across different data splits.

### 6.1 Results

- **Task-based GBSS**: The optimal subset size was found to be 2, with the most stable task pairs being (2, 25), (8, 25), (2, 19), and (2, 21). This suggests that tasks 2, 8, 19, 21, and 25 are particularly important for discrimination.
- **Feature-based GBSS**: The optimal subset size was 1, with the most frequently selected features being `total_time`, `air_time`, and `num_of_pendown`. These features align with findings from other methods, reinforcing their relevance.

## 7. Conclusion

This study identified a small set of interpretable handwriting features that consistently distinguish between Alzheimer's patients and healthy controls. Through the application of multiple feature selection methods combined with stability selection, several key findings emerged that provide insights into the motor and cognitive manifestations of Alzheimer's disease in handwriting tasks.
Despite employing fundamentally different selection strategies, all methods converged on a core set of features related to temporal dynamics:

### 7.1 Task-Level Findings

**Tasks 19** (copying the fields of a postal order), **15** (copying the word "bottiglia" in reverse), and **8** (writing letter "l" four times in cursive) emerged as the most discriminative, each being selected by 3 out of 4 methods. This consistency across different selection approaches suggests these specific handwriting activities are particularly sensitive to the motor and cognitive impairments associated with Alzheimer's Disease.

### 7.2 Feature-Level Insights

- **Time-based metrics**: Total time taken to complete the task (`total_time`), time in air (`air_time`), and time writing on paper (`paper_time`), emerged as the most robust discriminators across all approaches, suggesting that motor execution speed and timing control are fundamental markers of disease progression.
- **Spatial dispersion measures**: Dispersion index which measures how the handwritten trails are dispersed in the entire piece of paper (`disp_index`), appeared in both Elastic Net and Group Lasso feature-based analysis, indicating that movement irregularity is a relevant indicator of neurodegeneration.
- **Pen control features**: Total number of pendowns recorded during the execution of the task (`num_of_pendown`) were identified in both Group Lasso component-based analysis and GBSS feature-based analysis, reflecting difficulties in fine motor coordination.