Introduction

This document details an analytical study on the USPS handwritten digit dataset, specifically focusing on the digits 5 and 8. Comprising 1098 images, each represented by 256 pixels, the dataset encapsulates the grayscale intensities of handwritten digits, with pixel values ranging from -1 (white) to 1 (black).

The main objective of this study is to employ LASSO regression to predict the handwritten digits based on their pixel values.

Data Preprocessing

To align with William's preprocessing steps, pixel values were adjusted and scaled to fall between 0 and 1. Boundary values (0 or 1) were moved one-tenth towards the closest value. Subsequently, the logit function was applied, expanding the pixel value range from 0-1 to $\mathbb{R}$. Each image was then standardized to have a mean of 0 and a standard deviation of 1.

Spatial Filtering

Define Y as the 1098×256 matrix of observed pixel values, and C as a 256×256 matrix representing the spatial structure of the pixels. Let M be the centering matrix, transforming C into a matrix with zero mean. The eigendecomposition of MCM is denoted as $Q_C \Lambda_C Q_C^t$, where Q_C contains the eigenvectors, and Λ_C contains the eigenvalues. Through the operation $Y_C = Y * Q_C$, we project the original pixel values into the spatial frequency domain.

Three types of similarity matrices (C) were constructed. The adjacency matrix identified immediate spatial relationships by computing the Euclidean distance between pixels and marking them as adjacent (with a value of 1) if they were within a specified threshold, set to 2.

The Matern correlation matrix modeled the spatial correlations using the Matern covariance function, with range (phi) and smoothness (nu) parameters optimized via a grid search based on semivariance analysis.

The empirical correlation matrix was established by calculating the Pearson correlation coefficients between the logit-transformed pixel values across all images, setting the diagonal entries to zero to avoid self-correlation.

Model Fitting

LASSO regression was utilized to predict the handwritten digits, transforming the response variable into binary (0 for digit 5 and 1 for digit 8). The data were divided into 80% for training and 20% for testing. The optimal regularization parameter, λ, was determined through cross-validation on the training set. The λ yielding the lowest cross-validation error was selected to fit the model, with performance evaluated on the test set using mean squared error (MSE).

Various covariate matrices were employed to fit the models: the original data (scaled 0-1), the logit-transformed data, the logit-transformed data projected onto each spatial frequency space, and the logit-transformed data projected onto the top 10 eigenvectors of each spatial frequency space.

Results

Model	**MSE** ($\times 10^{-2}$)
Original (sacled to 0-1)	4.51
Logit	4.02
Adjacency Matrix Projected	4.31
Matern Matrix Projected	4.44
Empirical Matrix Projected	4.21

Table 1: MSE for data projected onto different spatial frequency spaces

Model	**MSE** ($\times 10^{-2}$)
Adjacency Matrix, Top 10 Eigenvectors	7.12
Matern Matrix, Top 10 Eigenvectors	7.37
Empirical Matrix, Top 10 Eigenvectors	5.50

Table 2: MSE for data projected onto the top 10 eigenvectors of each spatial frequency space

Table 1 shows the model using the logit-transformed values outperforms other projections with the lowest MSE, suggesting it's the most accurate. The Empirical Matrix method exhibits the highest robustness in model performance across Tables 1 and 2. It maintains a lower MSE increase when reducing dimensions to the top 10 eigenvectors, which indicates that this approach retains more information and predictive accuracy during dimensionality reduction than the other methods compared.