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

Table 1: Model performance for data projected onto different spatial frequency spaces

Model	Accuracy		AUC	
	min	1se	min	1se
Original (scaled to 0-1)	0.945	0.950	0.996	0.995
Logit	0.959	0.959	0.998	0.996
Adjacency Matrix Projected	0.964	0.973	0.996	0.995
Matern Matrix Projected	0.964	0.964	0.995	0.995
Empirical Matrix Projected	0.959	0.968	0.997	0.998

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

Model	Accuracy		AUC	
	min	1se	min	1se
Adjacency, Top 10 Eigenvectors	0.932	0.914	0.986	0.983
Matern, Top 10 Eigenvectors	0.927	0.914	0.986	0.984
Empirical, Top 10 Eigenvectors	0.964	0.959	0.992	0.995

In Table 1, the models exhibit comparable performance in terms of AUC, with marginal differences between them. Regarding accuracy, the projected models demonstrate equivalent or slightly improved results compared to the logit model, whether we consider the minimum lambda or the minimum lambda plus one standard error. In Table 2, the Empirical model outperforms the other two models in all metrics.