Methods

In this study, we conducted two simulations to evaluate the performace of LASSO models in identifying significant features within high-dimensional datasets. The first simulation assumes sparsity in the feature space corresponding to pixels, referred to as the pixel space, while the second simulation assumes sparsity in the frequency domain, known as the frequency space. The pixel space represents the original high-dimensional domain, where each dimension corresponds to a pixel in an image. Features in this space are directly observable and may exhibit inherent correlations. In contrast, the frequency space is a transformed version of the pixel space, derived through eigen decomposition, where the data is represented in terms of its frequency components.

Let X denote a column vector representing the pixel values of a single observation, with 256 pixels in total. Its covariance matrix, Σ, is defined to have an exponential correlation structure, where $\Sigma_{ij} = -\exp(\mathrm{dist}(i, j))$. Here, $\mathrm{dist}(i, j)$ is the distance between the pixels i and j in a 16×16 grid. The matrix V contains the eigenvectors of Σ, with each column representing an eigenvector. We transform the random vector X into the frequency space by computing $X_{\text{freq}} = V^T X$. The covariance matrix of X_{freq} is then given by $\mathrm{cov}(X_{\text{freq}}) = V^T \Sigma V$, which results in a diagonal matrix.

In each simulation iteration, 1000 observations of X_{freq} are generated from a multivariate normal distribution with the covariance matrix $\mathrm{cov}\,(X_{\text{freq}})$. This process is repeated across 500 iterations. For each observation, we compute the corresponding pixel space vector X as $X = V X_{\text{freq}}$.

In the first simulation, we assume that the coefficient vector β is sparse in the pixel space. Specifically, the non-zero values in β are confined to a central 8×8 region of the image. The response variable y is drawn from a binomial distribution, with the success probability determined by $\eta = X\beta$. The non-zero coefficients in β are chosen to ensure that the probability $p = \frac{1}{1+\exp(-\eta)}$ is uniformly distributed within the interval $[0, 1]$.

In the second simulation, sparsity is assumed in the coefficient vector b within the frequency space. Here, 10% of the 256 entries in b are randomly set to non-zero values, while the rest remain zero. The response variable y is generated similarly to the first simulation, ensuring that the probability p is uniformly distributed.

For both simulations, two models are fitted: one using the covariates in the pixel space and the other using the covariates in the frequency space. Each dataset, consisting of 1000 observations with 256 features (representing 16×16 pixel images), is split into training (80%) and test (20%) sets. The regularization parameter λ is tuned using cross-validation based on the binomial deviance metric. The dataset is divided into 10 folds, with the model trained and validated iteratively across these folds while varying λ. We consider two specific λ values, 'lambda.min' which minimizes the cross-validated error, and 'lambda.1se', which represents the largest λ within one standard error of the minimum.

After selecting the optimal λ, model performance is assessed using accuracy and AUC (Area Under the Curve) metrics. Additionally, a permutation test is conducted 100 times to calculate p-values for each covariate. Across all iterations, we compute the mean and standard deviation of the metrics, as well as the percentage of significant p-values for each covariate.

Results

In Simulation 1, the distribution of the success probability p was evaluated at various β values: 0.01, 0.05, 0.1, 0.2, and 1. As shown in Figure 1, $\beta = 0.1$ yielded the most uniform distribution of p, making it the optimal choice for model fitting. Similarly, in Simulation 2, the distribution of p was assessed at various b values: 0.1, 0.2, 0.4, 0.6, 0.8, and 1. As illustrated in Figure 2, $b = 0.4$ resulted in the most uniform distribution of p.

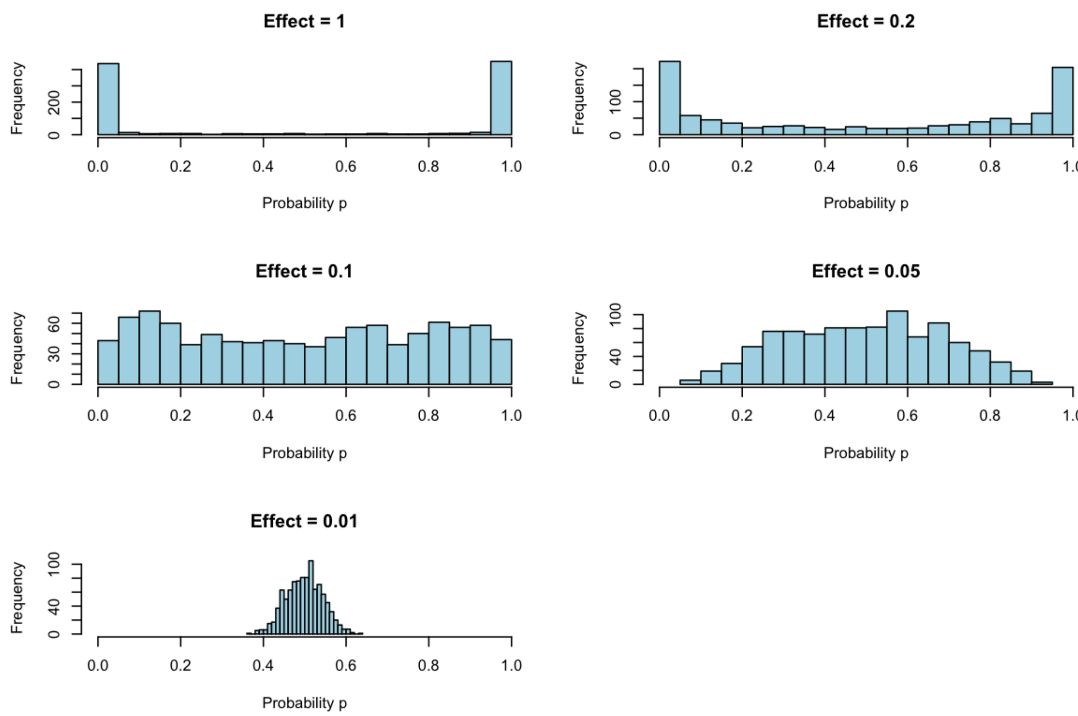

Figure 1: Distribution of success probability p at different β values in Simulation 1.

Using the selected values of β and b, the response variable y was drawn from a binomial distribution with success probabilities determined by $\eta = X\beta$ in both simulations. Figure 3 and 5 depict the group mean difference in covariate values between instances where $y = 1$ and $y = 0$ in both the pixel space and frequency space for Simulation 1 and 2, respectively. For Simulation 1, as shown in Figure 3, the heatmap in the pixel space reveals that the central region with non-zero coefficients in β corresponds to higher mean covariate values, which is consistent with the heatmap of β in Figure 4. Similarly, regions with higher or lower values in the frequency space match the corresponding values in the coefficients. A similar pattern is observed in Simulation 2, as illustrated in Figures 5 and 6, where the heatmaps for both the pixel and frequency spaces demonstrate alignment between covariate values and the corresponding non-zero coefficients.

To assess the performance of models fitted with covariates from the pixel space versus the frequency space, we evaluated the area under the curve (AUC) and prediction accuracy. LASSO models were trained using cross-validation by splitting each set of 1000 simulated observations into 80% training and 20% test sets. This process was repeated 500 times.

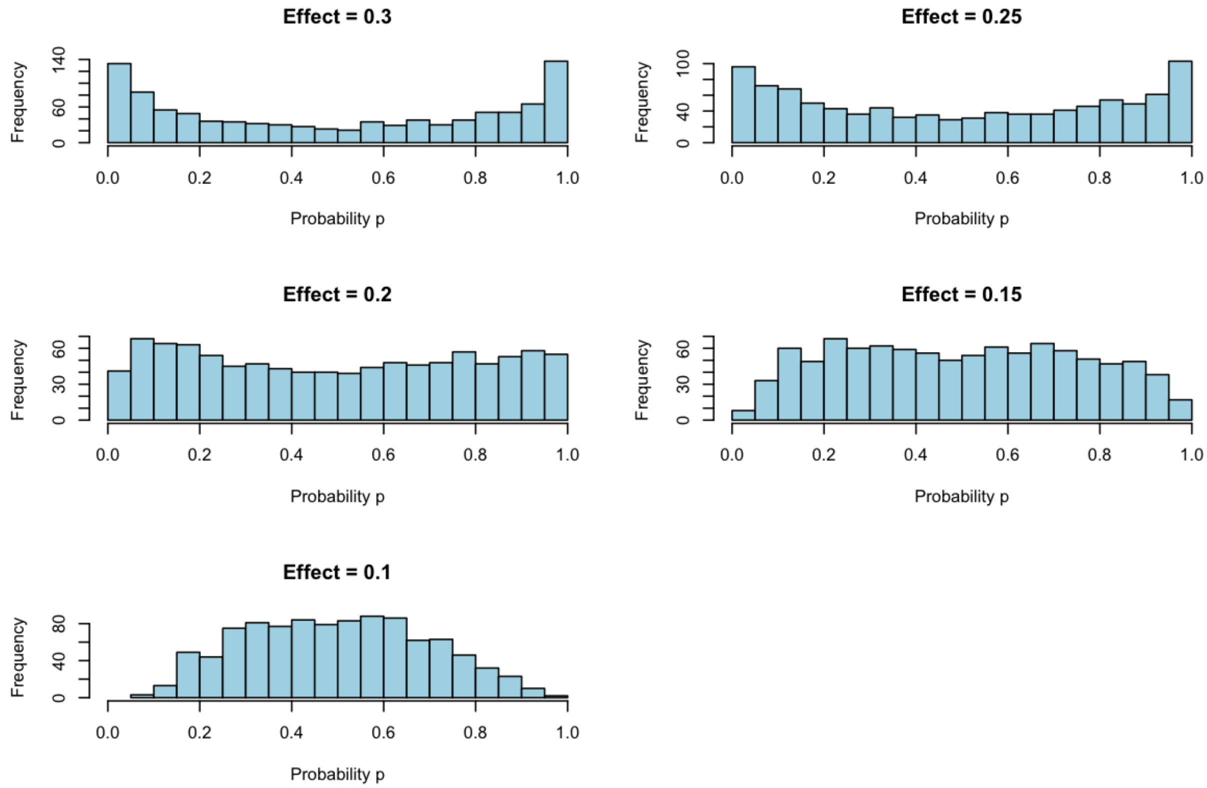

Figure 2: Distribution of success probability p at different b values in Simulation 2.

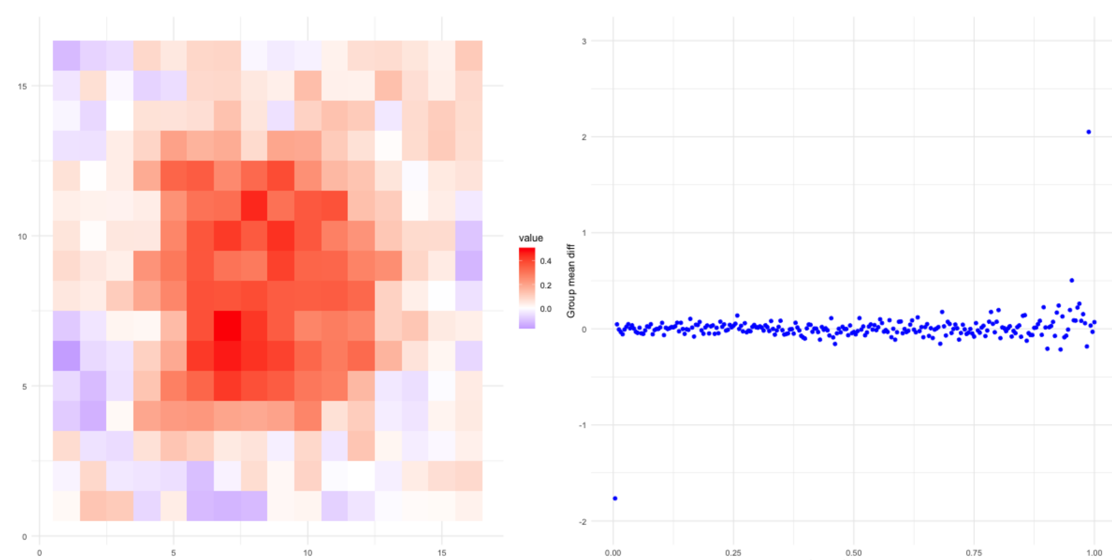

Figure 3: Group mean difference in covariate values between instances where $y = 1$ and $y = 0$ in Simulation 1, shown for both the pixel space (left) and frequency space (right).

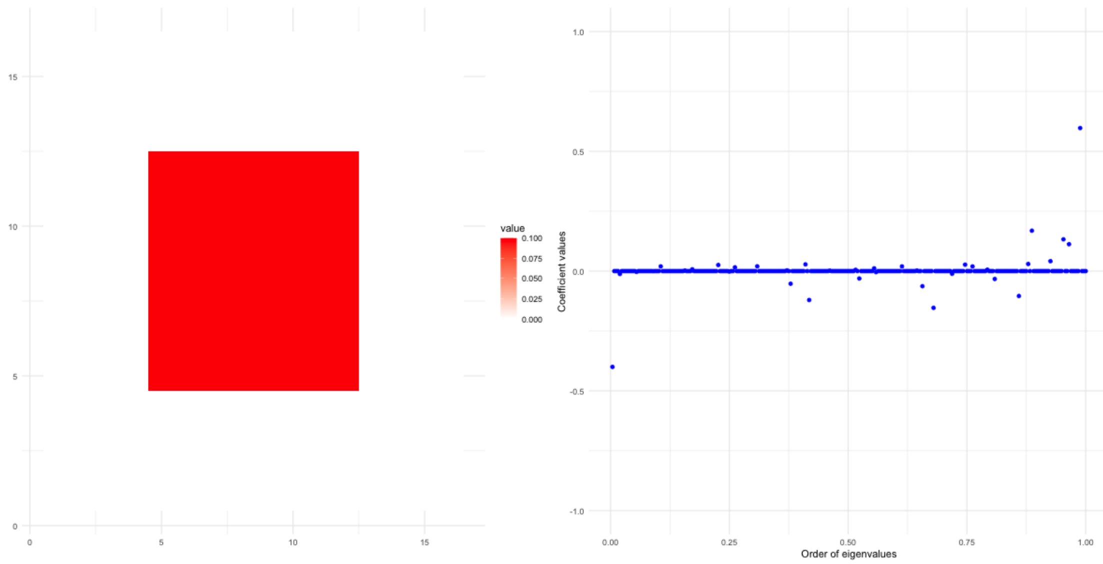

Figure 4: Actual coefficients in Simulation 1 for the pixel space (left) and frequency space (right).

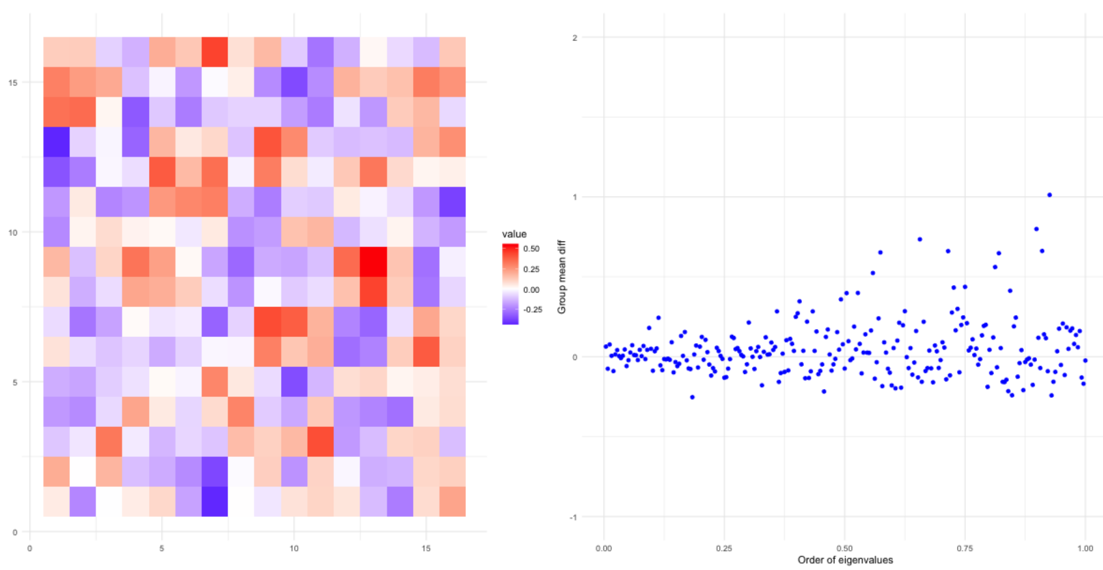

Figure 5: Group mean difference in covariate values between instances where $y = 1$ and $y = 0$ in Simulation 2.

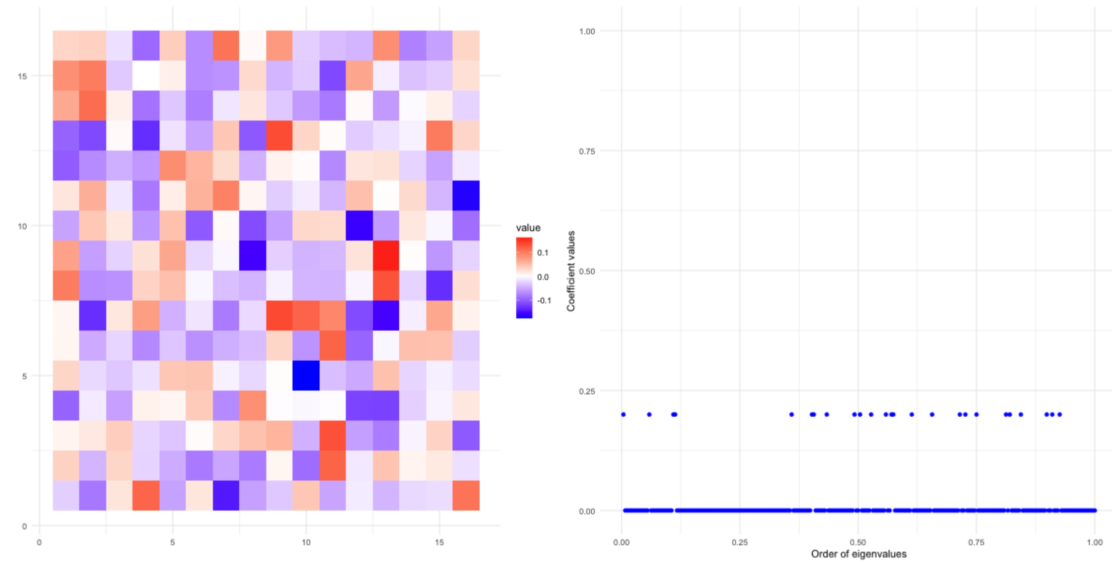

Figure 6: Actual coefficients in Simulation 2 for the pixel space (left) and frequency space (right).

Table 1 presents the average AUCs and accuracies over the 500 iterations. Regardless of whether sparsity is assumed in the pixel space (Simulation 1) or the frequency space (Simulation 2), models fitted in the frequency space consistently outperformed those fitted in the pixel space. Specifically, in Simulation 1, using 'lambda.min' as the regularization value, models fitted with covariates from the pixel space achieved an AUC of 0.803 (SE = 0.031) and an accuracy of 72.5% (SE = 0.032). In contrast, models fitted with covariates from the frequency space achieved a slightly higher AUC of 0.827 (SE = 0.029) and a higher accuracy of 74.6% (SE = 0.031). A similar trend was observed in Simulation 2, with models fitted in the frequency space demonstrating superior performance regardless of the regularization parameter used.

Table 1: Comparison of AUC and accuracy between models fitted in the pixel space and frequency space across 500 iterations for Simulation 1 and Simulation 2.

Simulation	**Model in Pixel Space**		**Model in Frequency Space**	
	AUC (SE)	**Accuracy (SE)**	**AUC (SE)**	**Accuracy (SE)**
Simulation 1				
lambda.min	0.803 (0.031)	0.725 (0.032)	0.827 (0.029)	0.746 (0.031)
lambda.1se	0.800 (0.031)	0.723 (0.032)	0.826 (0.029)	0.745 (0.031)
Simulation 2				
lambda.min	0.756 (0.036)	0.687 (0.035)	0.816 (0.032)	0.735 (0.033)
lambda.1se	0.734 (0.040)	0.669 (0.036)	0.816 (0.032)	0.734 (0.034)

The mean estimated coefficients in both the pixel space and frequency space were calculated for Simulation 1 and Simulation 2. Figure 7 displays the mean estimated β values. The left column shows the estimates when models were fitted using `lambda.min`, while the right column corresponds to models fitted using `lambda.1se`. The top row presents the results for Simulation 1, and the bottom row for Simulation 2. When comparing these estimated β values to the actual coefficients shown in Figures 4 and 6, it is evident that the estimated values closely align with the true coefficients.

Figure 8 presents the mean estimated b values plotted against the ordered eigenvalues. The eigenvalues are ranked from smallest to largest, with the smallest assigned an order of 1. To standardize the scale, the orders are then divided by the total number of eigenvalues, resulting in values between 0 and 1 on the x-axis. In Simulation 1, the number of non-zero coefficient estimates closely matches the true b values shown in Figure 4. These non-zero coefficients are primarily concentrated among the largest eigenvalues, indicating that the model correctly identifies the most significant components. In contrast, Simulation 2 shows that non-zero coefficient estimates are more uniformly distributed along the x-axis.

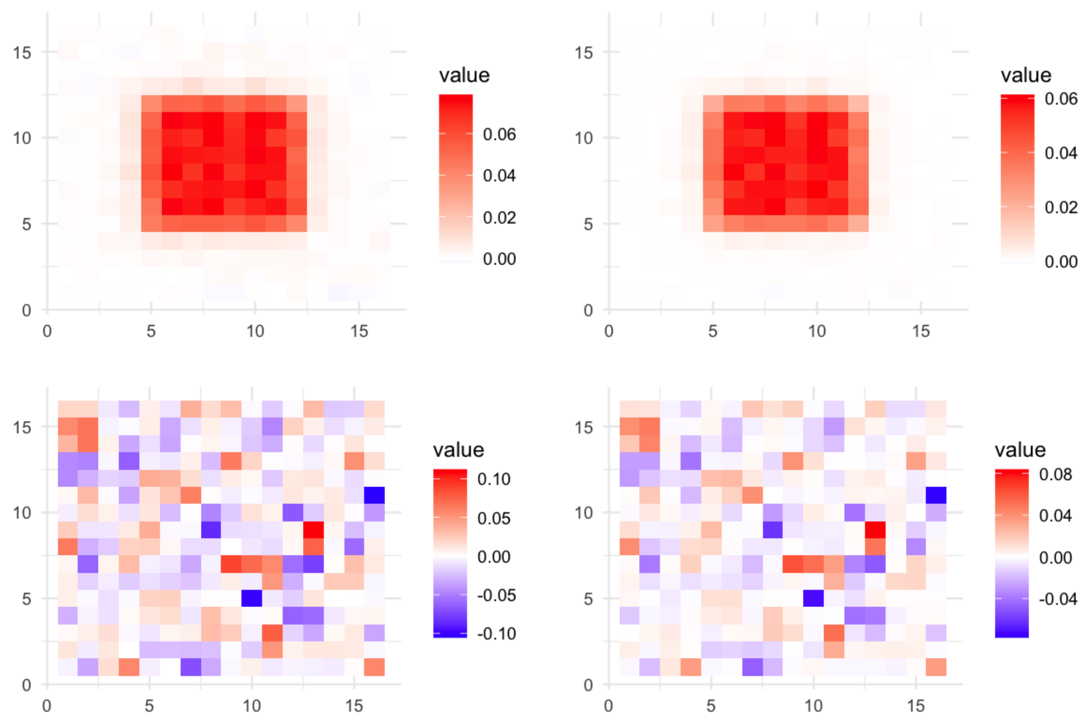

Figure 7: Mean estimated β values across simulations, with models fitted using `lambda.min` (left) and `lambda.1se` (right). The top row shows results for Simulation 1, while the bottom row shows results for Simulation 2.

Figure 9 shows the percentage of significant elements in β when fitting models in the pixel space for Simulation 1 (left) and Simulation 2 (right). A permutation test was performed, where the original coefficient estimates were compared to those obtained after permuting the covariates 100 times. The p-value was calculated as the proportion of times the absolute value of the original estimates exceeded those in the permutations. The results indicate that, regardless of whether sparsity is assumed in the pixel space or frequency space, significant p-values provide little information about the actual non-zero coefficients. Figure 10 shows the

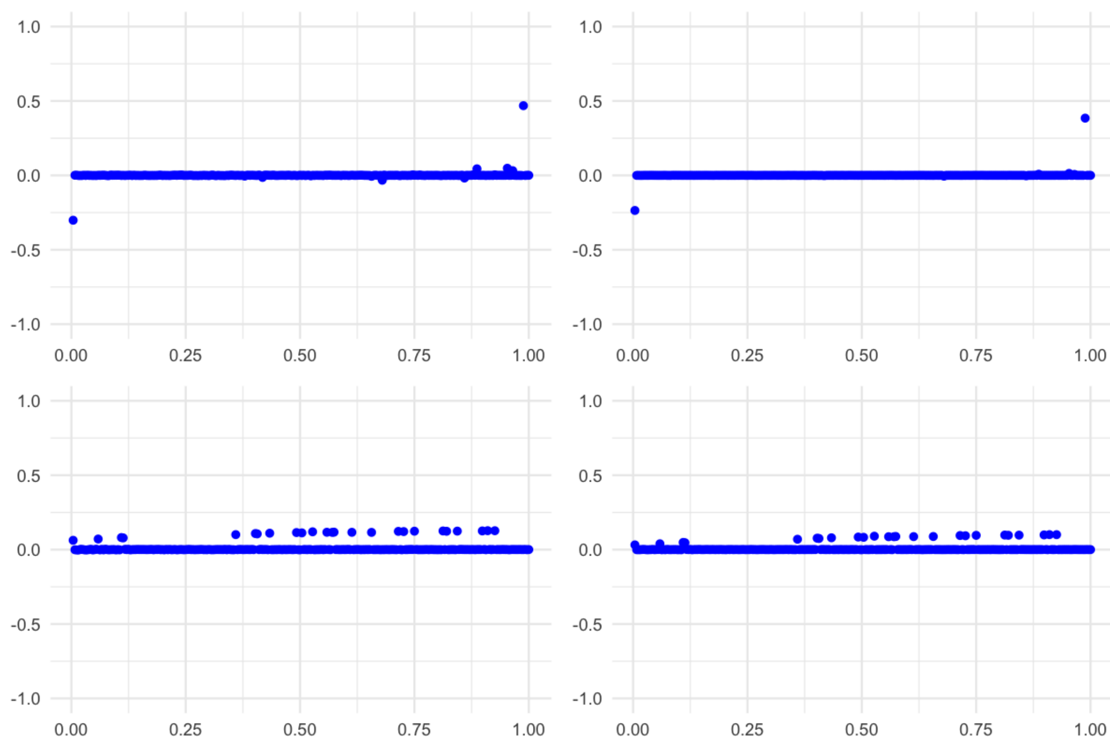

Figure 8: Mean estimated b values across simulations, plotted against ordered eigenvalues. Models fitted using `lambda.min` are on the left and models fitted with `lambda.1se` on the right. The top row shows results for Simulation 1, while the bottom row shows results for Simulation 2.

corresponding percentage of significant p-values of b across both simulations, plotted against the order of eigenvalues. No clear pattern emerged, indicating that evaluating significance based on p-values may not be effective for identifying meaningful pixels or frequencies.

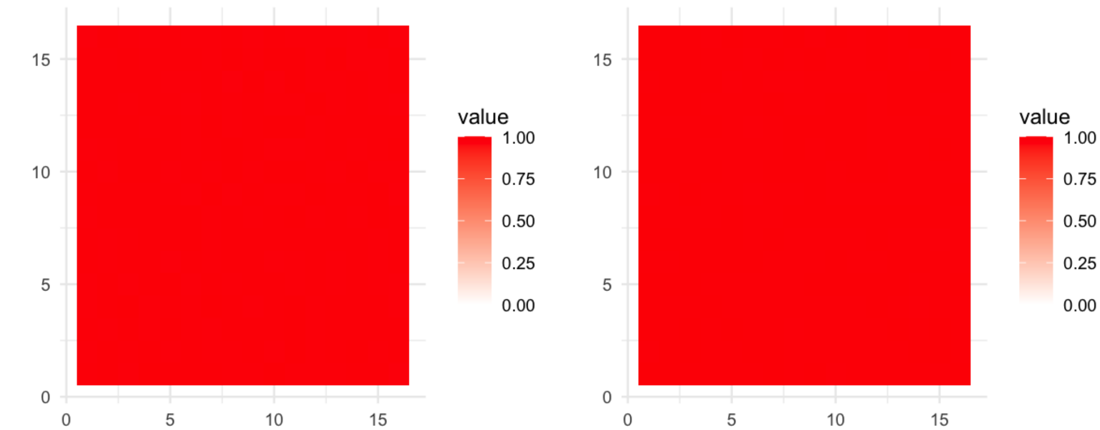

Figure 9: Percentage of significant p-values for elements of β when fitting models in the pixel space in Simulation 1 (left) and Simulation 2 (right).

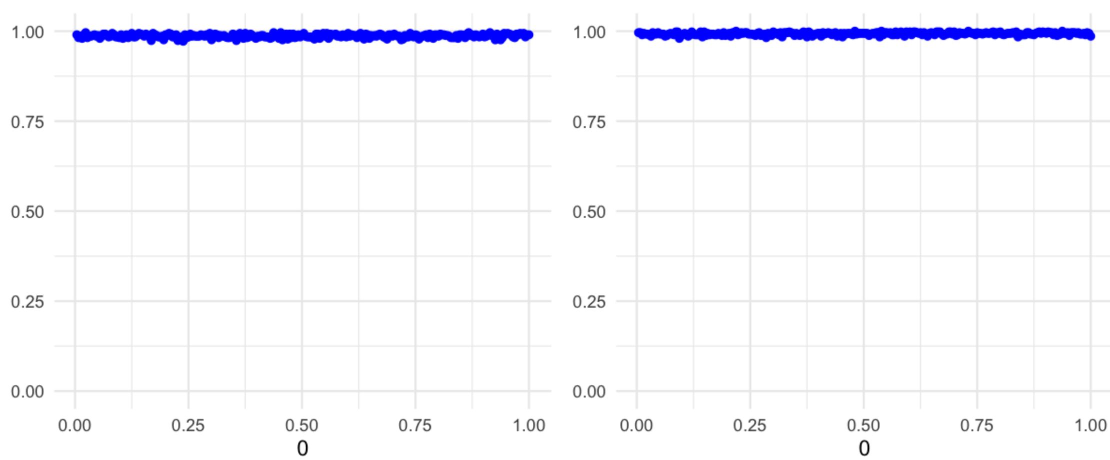

Figure 10: Percentage of significant p-values for elements of b across ordered eigenvalues in both simulations.

Figure 11 presents the frequencies associated with the top three eigenvalues, which represent the dominant patterns in the pixel space. The frequency associated with the smallest eigenvalue is also shown, highlighting the least significant variance.

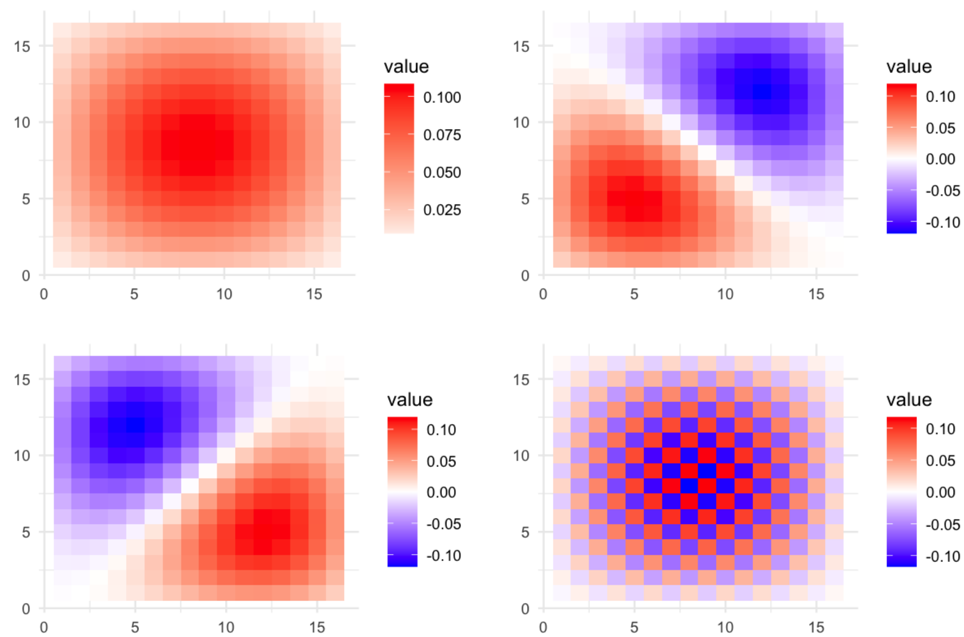

Figure 11: Frequencies associated with the top three eigenvalues (top row and bottom left) and the frequency associated with the smallest eigenvalue (bottom right), highlighting the primary and least significant patterns in the pixel space.