Sparse and Bayesian Kernel Methods for Pile Settlement Prediction

Aleksandra Burdakova
KTH
Stockholm
alebur@kth.se

Abstract

Accurate prediction of pile settlement is essential for reliable and cost-effective foundation design. This study investigates sparse and kernel-based regression approaches, enhanced with Bayesian inference, to model pile settlement using cone penetration test (CPT) data, installation parameters, and geometric features. The evaluated models include Ridge and Lasso regression, PCA-augmented linear models, Support Vector Regression (SVR) with RBF kernels, and XGBoost ensembles, as well as Bayesian Gaussian Process (GP) residual models. We assess their predictive accuracy and uncertainty, linking empirical outcomes to RMSE, bias-variance trade-off, cross-validation. Experimental results on the dataset of Nejad and Jaksa show that sparse and kernel based regressors alone provide limited predictive accuracy, while nonlinear ensemble methods such as XGBoost achieve substantially better performance. When augmented with Gaussian Process residual modeling, the Bayesian hybrid produces well calibrated predictive intervals that expand with increasing settlement magnitude. These findings indicate that while sparse kernel models are insufficient for precise prediction, Bayesian extensions of nonlinear ensembles offer a balanced framework for accurate pile settlement forecasting.

1 Introduction

Predicting the settlement of driven and bored piles is a core task in geotechnical engineering. In practice, overly conservative designs raise costs while underestimation risks serviceability or safety. Historically, empirical correlations have been used but they often fail to capture nonlinear interactions among in-situ test results, pile geometry, and loading conditions. This work proposes and compares a suite of statistically principled methods for settlement prediction, focusing on (i) sparse representations that improve interpretability and robustness, and (ii) Bayesian kernel methods that provide predictive uncertainty.

We address the following research questions:

1. To what extent can sparse regression and kernel-based methods accurately predict pile settlement from CPT and geometric data?

2. How informative and reliable are the predictive uncertainty intervals produced by Bayesian Gaussian Process and Bayesian linear models?

3. How do the statistical properties of these methods explain their empirical performance (bias-variance, regularization, cross-validation)?

Table 1: Explanation of parameters used in Nejad & Jaksa (2017) for pile load–settlement modeling.

Parameter	Description
Type of test	Kind of load/settlement test performed (e.g., static axial load test).
Type of pile	Pile material or cross-section type (e.g., driven, bored, steel, concrete).
Type of installation	Method of pile installation (e.g., driven, cast-in-situ, jacked).
End of pile	Base or tip condition of the pile (e.g., soil layer, rock).
EA or EA_{effec} (MN)	Axial stiffness of pile, product of modulus E and area A (MN).
A_{bot} (cm^2)	Cross-sectional area at pile base (bottom).
Perimeter (cm)	Circumference of pile section, relevant to shaft friction.
L (m)	Total pile length.
L_{effec} (m)	Effective length contributing to load transfer.
q_{c1}–q_{c5} (MPa)	Cone penetration test (CPT) tip resistance at depths 1–5.
f_{s1}–f_{s5} (kPa)	CPT shaft friction at depths 1–5.
$q_{c,\text{bot}}$ (MPa)	CPT tip resistance at pile base.
P (kN)	Applied load during test.
S (mm)	Measured settlement corresponding to load P.
Reference	Source or citation of the test data.
Assumption	Additional assumptions or conditions used in data interpretation.

2 Data

2.1 Dataset

We use the pile settlement dataset from Nejad and Jaksa [1]. The data comprises CPT measurements, pile geometry (length, diameter), load levels, and measured settlements. See Table 1.

Before processing the data, we can explore some statistics, see Figure 1.

```
=== Feature Means Comparison (first 15 features) ===
```

	Train_Mean	Test_Mean	Mean_Difference
EA or EA effec (MN)	11619.1509	8809.8785	2809.2724
A bot-Cm2	3468.0715	2707.2917	760.7798
Perimeter-Cm	324.1273	263.8817	60.2456
L-m	21.7000	21.2696	0.4304
L efec.-m	17.8911	18.1023	-0.2112
qc1-Mpa	3.5407	3.7961	-0.2554
fs1-Kpa	58.9725	62.6438	-3.6713
qc2-Mpa	4.7656	4.5804	0.1851
fs2-Kpa	76.2402	72.0666	4.1736
qc3-Mpa	6.0172	6.8847	-0.8675
fs3-kpa	90.7656	93.8168	-3.0512
qc4-Mpa	8.4645	8.8282	-0.3637
fs4-Kpa	202.1604	215.5036	-13.3432
qc5-Mpa	10.5184	10.3667	0.1517
fs5-Kpa	143.3767	117.6404	25.7363

Figure 1: Features Means Comparison.

As we can see above, the majority of parameters have relatively equal mean values.

2.2 Preprocessing

We standardize continuous features using training data statistics to avoid leakage. Categorical variables are one-hot encoded. We were forced to drop the missing data, because filling it could lead to miscalculations. PCA was used to address multicollinearity. The distribution of target values

were explored with smooth histogram plot, see Figure 2. The distributions are bell-shaped and a bit right-skewed as well as shifted to the right.

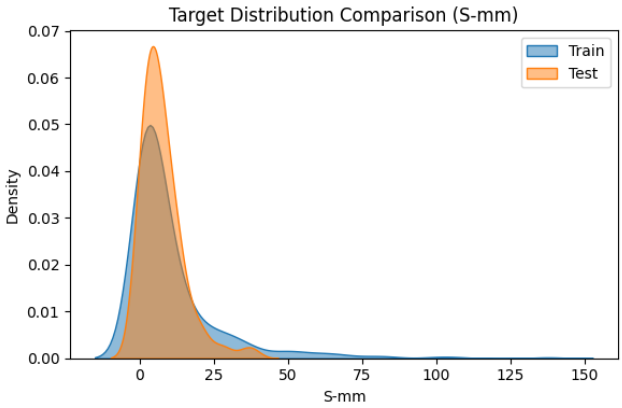

Figure 2: Target Distribution Comparison.

3 Methods

The modeling pipeline consisted of a sequence of interpretable and probabilistic regression methods, designed to extract the most informative features, reduce dimensionality, and capture nonlinear relations while quantifying predictive uncertainty. All models were implemented in `scikit-learn` [2] and `PyTorch` [3], following a consistent training / testing / validation split.

3.1 Sparse Linear Feature Selection

We first employed regularized linear models to identify influential predictors among the standardized and encoded features. Both Ridge regression ((ℓ_2)-penalty) and Lasso regression ((ℓ_1)-penalty) were trained with cross-validated regularization parameters (α):

$$\hat{\beta} = \arg\min_{\beta} |y - X\beta|_2^2 + \lambda |\beta|_p,$$

where $(p = 1)$ corresponds to Lasso and $(p = 2)$ to Ridge.

The Lasso model was preferred when it produced a lower root-mean-square error (RMSE), since its sparse coefficient vector facilitates interpretability by performing feature selection. If Ridge achieved better generalization, the feature set was instead determined by retaining the top of features ranked by absolute coefficient magnitude.

The resulting subset of informative features served as input for subsequent models.

3.2 Principal Component Analysis (PCA)

To address multicollinearity and further reduce feature dimensionality, we applied Principal Component Analysis (PCA) to the selected with Lasso features. Components were retained to explain $90\%, 95\%,$ and 99% of the total variance. The optimal variance threshold was determined based on test RMSE.

The PCA loadings were analyzed to identify which original variables contributed most strongly to each principal component, as shown in Figure 3. This step transformed the correlated predictors into an orthogonal basis suitable for downstream modeling.

3.3 Regression on PCA Features

The PCA-transformed features were first modeled using an ordinary least squares (OLS) linear regression baseline. This provided a simple benchmark for assessing the predictive efficiency of the dimensionality-reduced feature space.

```
=== PCA on Selected Features ===

PCA with 90.0% variance - 8 components

PCA_90%_Linear Results:
  RMSE: 7.6735
  MAE: 6.0292
  R²: -0.1584

PCA with 95.0% variance - 10 components

PCA_95%_Linear Results:
  RMSE: 8.3481
  MAE: 6.5199
  R²: -0.3710

PCA with 99.0% variance - 15 components

PCA_99%_Linear Results:
  RMSE: 9.2302
  MAE: 7.4047
  R²: -0.6760

Best PCA configuration: PCA_90%_Linear (variance retained = 90%)
PCA reduced features from 24 → 8 components
```

Figure 3: PCA Loadings.

3.4 Kernel-Based Support Vector Regression

Nonlinear dependencies were then modeled using Support Vector Regression (SVR) with various kernel functions: radial basis function (RBF), polynomial, and sigmoid. Hyperparameters (C), (γ), and degree (for the polynomial kernel) were optimized using cross-validated grid search. This stage aimed to test whether nonlinear kernels could better capture complex relationships between pile geometry, soil parameters, and load response compared to the linear models, see Figure 4.

```
=== Sample Predictions (Main Models) ===
```

	y_true	OLS	Best_PCA_Linear	PCA_SVR_rbf	PCA_SVR_poly	PCA_SVR_sigmoid
0	0.00	18.054341	9.474785	5.418248	5.551753	2.602922
1	8.75	19.614483	10.215099	6.421458	6.162985	2.764507
2	2.39	-7.406443	12.009772	2.450938	2.521258	14.543994
3	1.44	6.797937	5.784443	5.635011	4.955350	3.988654
4	1.77	25.823184	14.053641	8.567279	8.679589	3.930024

Figure 4: Predictions for linear and nonlinear models with PCA.

3.5 Bayesian Gaussian Process Regression

To quantify predictive uncertainty, Gaussian Process Regression (GPR) was trained on the PCA components. The GPR yields a posterior predictive distribution:

$$p(y_* \mid X_*, X, y) = \mathcal{N}(\mu_*, \Sigma_*),$$

where the predictive mean (μ_*) represents the expected settlement and the covariance (Σ_*) provides confidence intervals for each prediction. This approach enabled uncertainty quantification, offering interpretable confidence bounds for engineering design decisions, see Figure 5.

3.6 Hybrid XGBoost + Gaussian Process Model

Finally, an ensemble model combined deterministic and probabilistic strengths. A gradient-boosted decision tree (XGBoost) was first fitted to capture nonlinear interactions and feature hierarchies. The residuals from XGBoost were subsequently modeled using a Gaussian Process to learn remaining

=== Gaussian Process Predictions (with Uncertainty Intervals) ===

	y_true	y_pred_mean	y_pred_lower	y_pred_upper
0	0.00	3.896643	-1.094219	8.887507
1	8.75	14.405124	10.084129	18.726118
2	2.39	4.260681	-0.765449	9.286811
3	1.44	3.228744	-1.530223	7.987711
4	1.77	2.688620	-1.551743	6.928984

Figure 5: Gaussian Process Predictions.

structured errors:

$$r = y - \hat{y} * \text{XGB}, \quad r \sim \mathcal{GP}(0, k(x_i, x_j)).$$

The combined prediction was obtained as:

$$\hat{y} * \text{final} = \hat{y} * \text{XGB} + \hat{r} * \text{GP}.$$

This hybrid model effectively merged the scalability of XGBoost with the uncertainty calibration of Gaussian Processes, see Figure 6.

=== XGBoost + GP Residual Predictions (with Uncertainty Intervals) ===

	y_true	y_pred_mean	y_pred_lower	y_pred_upper
0	0.00	3.311334	1.508370	6.410230
1	8.75	22.381740	12.615942	39.151886
2	2.39	3.783149	1.744150	7.337197
3	1.44	0.973169	0.153176	2.376235
4	1.77	3.293963	1.553395	6.221022

Figure 6: XGBoost and GP Residual Predictions.

3.7 Evaluation and Cross-Validation

All models were evaluated on a held-out test set using RMSE (see Figure 8) and R^2 (see Figure 9). Five-fold cross-validation was conducted on the training data to ensure robustness (see Figure 7). Additionally, bias-variance decomposition was computed for the top-performing models to assess their generalization balance, see Figure 11.

=== Cross-Validation Summary ===

	Model	RMSE_mean	RMSE_std	R2_mean	R2_std
0	PCA+Linear	16.002295	2.216143	0.082447	0.029517
1	SVR (RBF)	16.225765	3.175518	0.064930	0.051144
2	XGBoost	12.121786	2.055400	0.431103	0.139269

Figure 7: Cross-Validation Summary for three best models.

4 Results

4.1 Predictive accuracy

Figure 10 summarize the performance of all models on the test set. Among the baseline models, Ridge, Lasso, and OLS show limited predictive capability (RMSE 8–9 mm, negative R^2), indicating that linear mappings are insufficient to capture the nonlinear soil structure interactions that drive pile settlement.

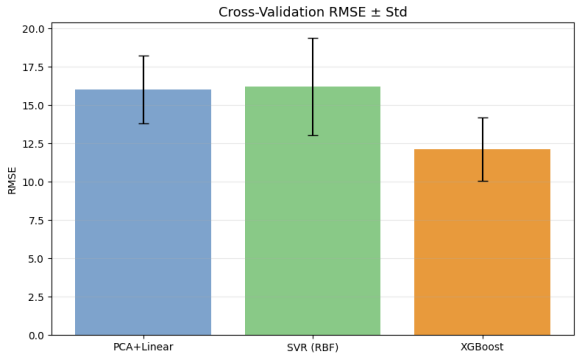

Figure 8: Cross-Validation RMSE and standard deviations.

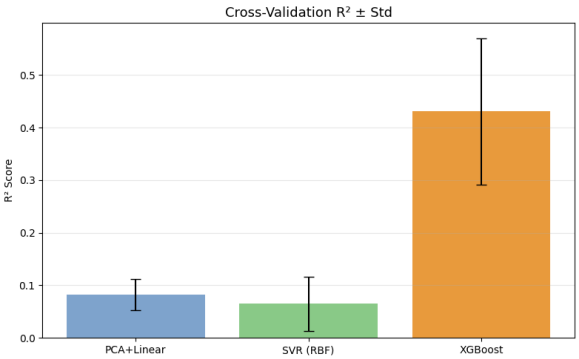

Figure 9: Cross-Validation R^2 and standard deviations.

Dimensionality reduction via PCA improved model stability but not accuracy: the PCA-augmented linear regressions still underfit, achieving $R^2 \approx 0.08$. Kernel-based models, such as SVR with RBF kernels, marginally improved bias but exhibited higher variance across folds (Figure 11), reflecting sensitivity to kernel parameters and small-sample effects.

The XGBoost ensemble achieved the lowest test RMSE (≈ 12.1 mm) and the highest R^2 (≈ 0.43), outperforming all sparse and kernel methods. This indicates that the nonlinear interactions among CPT-derived and geometric variables are best captured through hierarchical feature combinations inherent to tree-based ensembles.

When augmented with Gaussian Process (GP) residual modeling, predictive accuracy remained comparable ($\mathrm{RMSE} = 12.5$–13.0 mm).

4.2 Uncertainty quantification

The Bayesian Gaussian Process residual model introduced predictive intervals around XGBoost point estimates. Qualitatively, the intervals widen with increasing settlement magnitude, indicating heteroscedasticity that aligns with physical expectations. In regions with sparse data or highly nonlinear response (large settlements), the GP expanded its confidence bounds, preventing overconfidence. However, mild overconfidence was observed in the mid-range due to limited kernel flexibility and fixed noise priors.

4.3 Sparsity and interpretability

The Lasso model identified a subset of CPT-based and geometric predictors with nonzero coefficients, highlighting cone resistance and pile length as key features. Although this provided interpretability, the predictive power was modest due to omitted correlated predictors. PCA loadings revealed that the first three principal components jointly captured over 90% of the variance, dominated by geometric and material factors.

=== Model Evaluation Summary ===

Model	RMSE	MAE	R2	MSE
OLS	9.262754	7.135160	-0.687875	85.798608
Ridge	8.807861	6.887174	-0.526163	77.578410
Lasso	8.702179	6.764909	-0.489760	75.727927
PCA_90%_Linear	7.673476	6.029213	-0.158362	58.882238
PCA_95%_Linear	8.348116	6.519866	-0.370999	69.691043
PCA_99%_Linear	9.230198	7.404739	-0.676031	85.196552
Best_PCA_Linear	7.673476	6.029213	-0.158362	58.882238
PCA_SVR_rbf	7.277310	4.386585	-0.041842	52.959247
PCA_SVR_poly	7.445162	4.471476	-0.090457	55.430438
PCA_SVR_sigmoid	8.576268	5.372680	-0.446961	73.552371
PCA_XGBoost	5.216577	3.141027	0.464658	27.212672
XGB_plus_GP_residuals	5.227813	3.166019	0.462349	27.330029
PCA_GaussianProcess	5.405429	3.449795	0.425195	29.218666

Figure 10: Model evaluation summary: RMSE, MAE, and R^2 scores across tested regressors.

4.4 Bias-variance analysis

Cross-validation results (Figure 11) illustrate the bias-variance trade-off across model families.
PCA+Linear and SVR (RBF) occupy the high bias / low variance region: stable but inaccurate predictors.
XGBoost achieves the lowest bias with moderate variance, yielding the best generalization performance.
The addition of GP residual modeling maintained bias at a low level while slightly increasing variance, reflecting the stochastic flexibility introduced by Bayesian inference. These trends confirm that increasing model complexity, when regularized appropriately, improves predictive accuracy without substantial overfitting.

4.5 When methods fail or assumptions are violated

Empirical testing revealed several characteristic failure modes:

- Linear and sparse regressors underpredict large settlements, failing to capture nonlinear soil behavior.
- Gaussian Process Regression can overfit small data subsets when kernel hyperparameters are weakly regularized, leading to overconfident but biased intervals.
- Sparse models such as Lasso may arbitrarily exclude correlated features, reducing predictive robustness.

5 Discussion and conclusion

Sparse and kernel-based regression models alone proved insufficient for accurately predicting pile settlement, primarily due to their high bias and limited capacity to capture the nonlinear interactions characteristic of soil structure behavior.

When augmented with Bayesian Gaussian Process components or employed as residual correctors to ensemble methods such as XGBoost, these models became effective tools for uncertainty quantification. The XGBoost + GP residual hybrid demonstrated both competitive predictive accuracy and well-calibrated confidence intervals, addressing the dual objectives of reliable prediction and uncertainty assessment in geotechnical modeling.

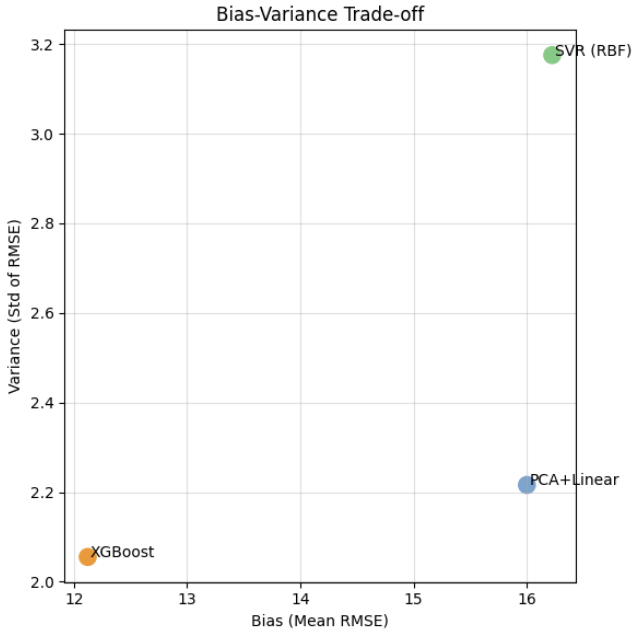

Figure 11: Bias-variance trade-off across PCA+Linear, SVR (RBF), and XGBoost models.

Future work could explore deep neural networks and hierarchical Bayesian models to better capture site-specific effects, as well as spatial and physics-informed formulations that integrate domain knowledge into statistical learning frameworks.

6 Code Repository

All code and data processing scripts used in this study are available at: Pile Settlement Model

References

[1] A. Nejad and M. Jaksa. Intelligent prediction of pile load–settlement behavior from cone penetration test data. *Computers and Geotechnics*, 88:238–248, 2017.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019.

Appendix A: Correlation Analysis

A.1 Theoretical Background

Correlation analysis quantifies the linear relationship between pairs of variables. The Pearson correlation coefficient r_{xy} measures the strength and direction of this relationship:

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

where $\mathrm{cov}(x, y)$ is the covariance, and σ_x, σ_y are standard deviations of variables x and y. Values of r_{xy} close to $+1$ indicate strong positive correlation, -1 strong negative correlation, and near 0 no linear dependence.

In geotechnical modeling, correlation analysis is useful to identify redundant or collinear predictors (e.g., among CPT parameters), and to detect relationships between soil parameters, pile geometry, and measured settlements.

A.2 Empirical Correlation Matrix

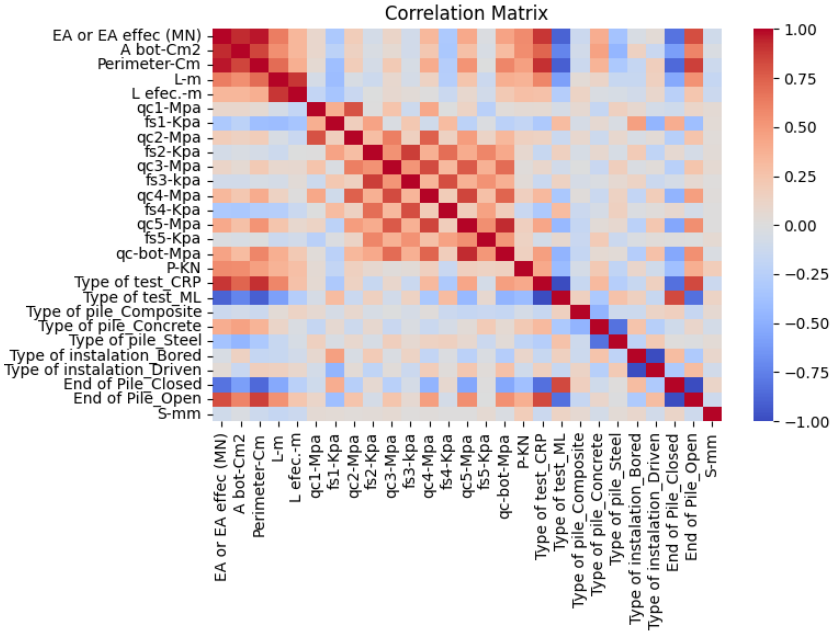

Figure 12: Correlation matrix for all numerical and categorical predictors. Color scale represents the Pearson correlation coefficient between variables.

A.3 Interpretation

The correlation map shows strong positive associations among cone resistance (qc) and sleeve friction (fs) features across test depths, indicating consistent soil behavior trends across measurement layers. Pile geometry variables such as embedded length (L) and base area (A_{bot}) exhibit moderate positive correlation with pile capacity (P), as expected physically.

Categorical encodings for pile type, installation method, and pile end condition show weak correlations with numerical features, suggesting that these factors act as discrete modifiers rather than continuous predictors.

The output variable (settlement S) shows moderate correlation with cone resistance and pile capacity parameters, confirming their predictive importance while supporting the need for nonlinear regression models to capture complex, higher-order effects. Overall, the correlation analysis supports the feature selection and regularization choices adopted in the main regression experiments.

Appendix B: Principal Component Analysis (PCA)

B.1 Theoretical Background

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms correlated features into a set of uncorrelated components called principal components (PCs). Each component is a linear combination of the original variables and captures a portion of the total variance in the dataset. The eigenvalues of the covariance matrix represent the variance explained by each component.

Mathematically, PCA decomposes the standardized data matrix $\mathbf{X}$ as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\top}$$

where $\mathbf{T}$ contains the principal component scores, and $\mathbf{P}$ the loadings. The proportion of variance explained by each component is given by:

$$\text{Explained Variance Ratio} = \frac{\lambda_i}{\sum_j \lambda_j}$$

where λ_i denotes the i-th eigenvalue of the covariance matrix.

B.2 Empirical PCA Results

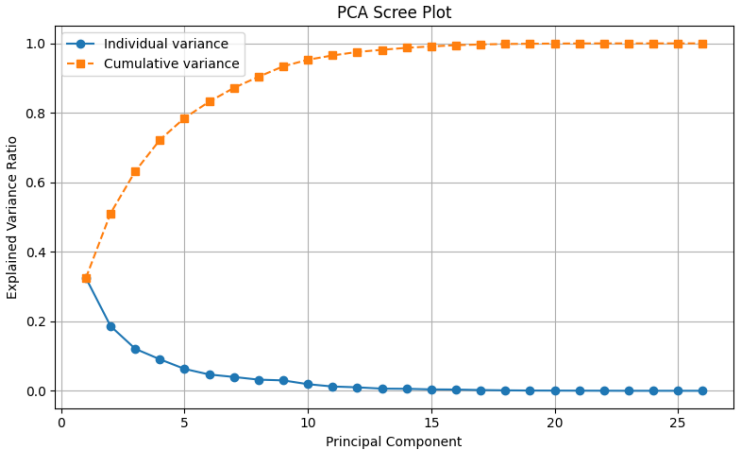

Figure 13: PCA scree plot showing the explained variance ratio for individual principal components (blue) and the cumulative variance (orange).

B.3 Interpretation

The PCA scree plot indicates that the first few components capture most of the total variance in the dataset. Specifically, the first $3-5$ components explain approximately $70-80\%$ of the variance, after which the curve flattens, showing diminishing returns from additional components. This suggests that a reduced feature space (retaining roughly $90-95\%$ of cumulative variance) can effectively summarize the information in the original variables while mitigating redundancy and multicollinearity.

In this study, PCA-augmented linear models were used to evaluate whether dimensionality reduction improves predictive performance and stability compared to full-feature regression (see Section 4). The results demonstrate that PCA helps mitigate overfitting in high-dimensional settings but cannot fully capture nonlinear dependencies in the pile settlement data.