Dimension reduction methods to improve image classification

Xinwei Shen and Yunfei Yang

Department of Mathematics, The Hong Kong University of Science and Technology

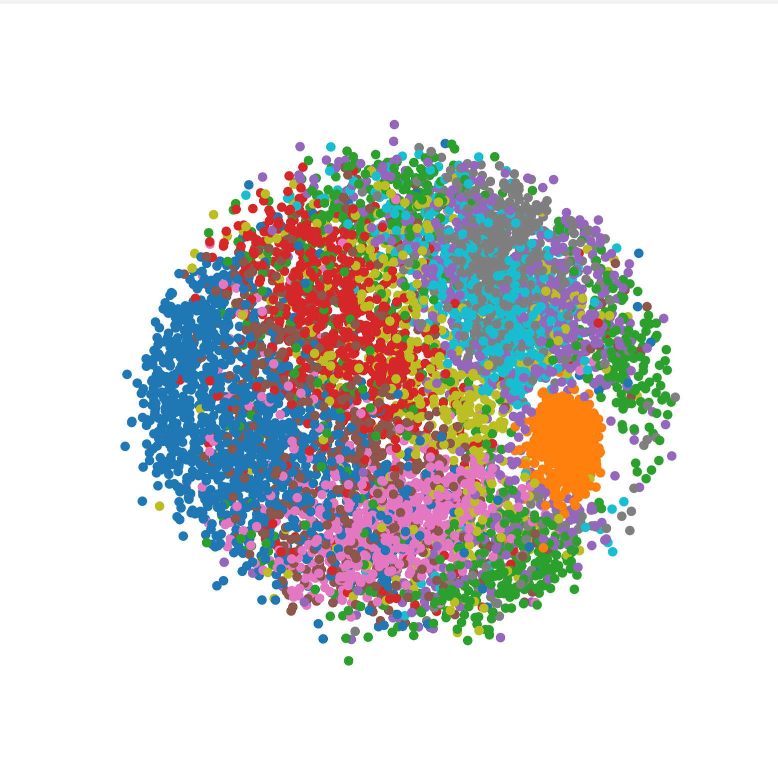
Introduction

Image classification is an important task in machine learning. In practice, high dimensionality often makes it difficult to apply classical models. For instance, it may slow down the algorithm or limit the prediction accuracy due to the redundant information within the training data. Therefore, dimensionality reduction of image features plays a crucial role in image retrieval and classification tasks. In this project, we

- explore the power of dimension reduction techniques to improve image classification accuracy more efficiently,
- and briefly analyzed the reasons behind.

Data

We analyze the Hand-written Digits dataset which contains 16×16 gray images of 10 handwritten digits. It consists of 7291 training examples, and 2007 testing examples. The following figure visualizes the training set in two dimension using multidimensional scaling (MDS), which motivates us to implement dimension reduction techniques to capture the essential information of image data.

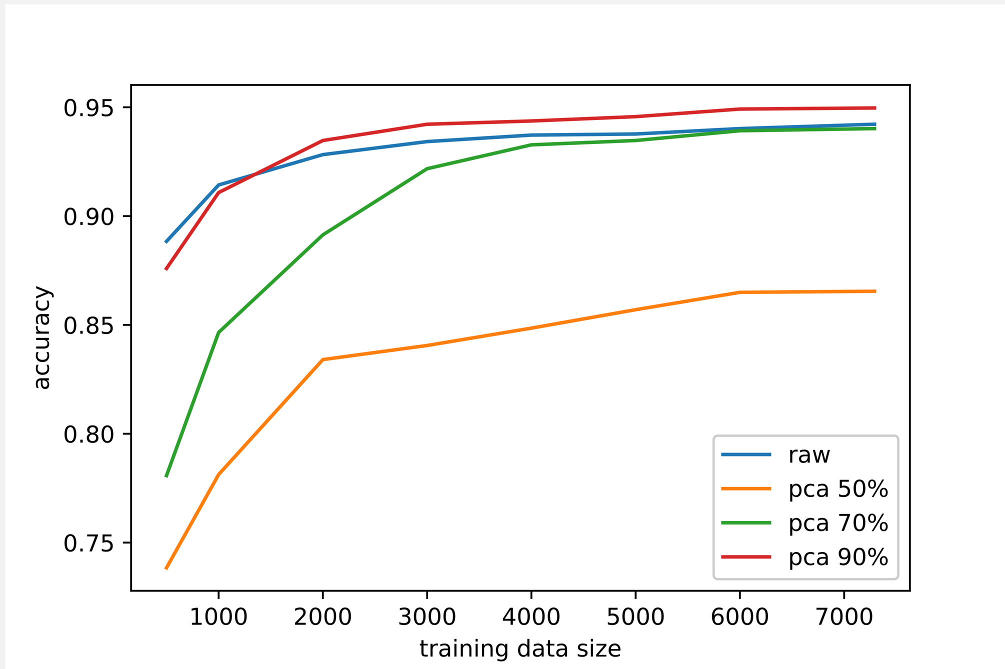


Methodology

We implement and compare both geometric dimension reduction techniques such as principal component analysis (PCA), as well as regularized estimators including Lasso, Ridge and Elastic Net. Specifically, the regularization penalty is $\lambda((1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1)$, where β is the regression coefficients, and $\alpha = 0$ for Lasso, 1 for Ridge and 0.5 for Elastic Net. We combine these methods with support-vector machine (SVM) and multinomial logistic regression (MLR), two commonly used classification models in machine learning and statistics.

Experiment 1

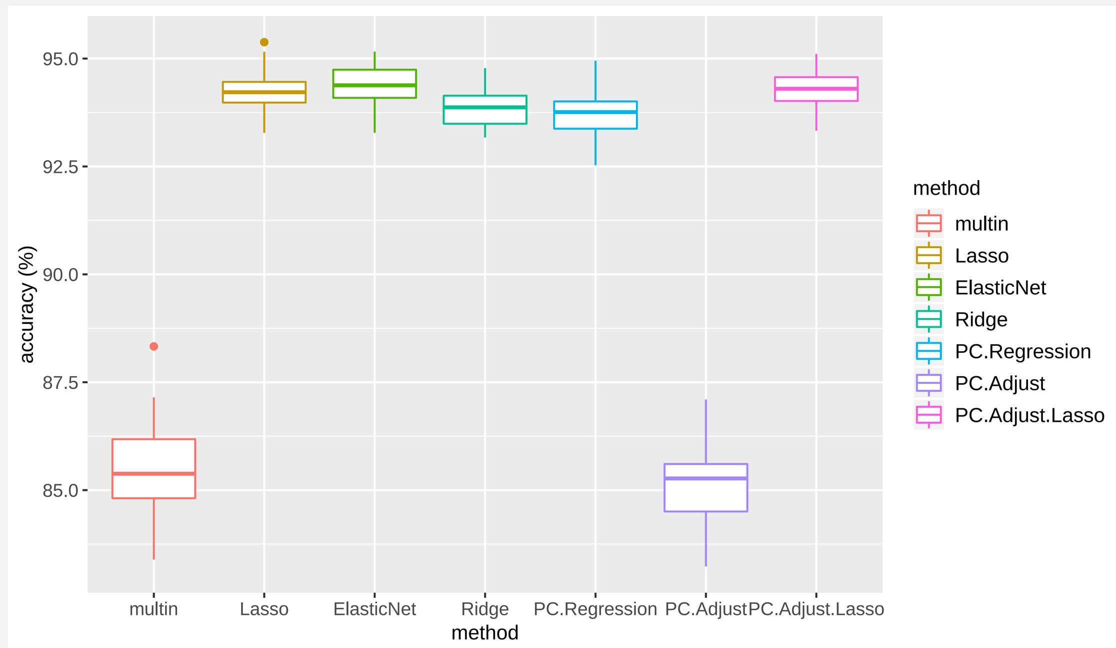
We first explore the effect of PCA under different dimension to sample size ratios. Specifically, we apply PCA to training sets with various sizes and select 7, 17, 55 principal components (PCs) such that the percentage of variance explained is greater than 50%, 70% and 90% respectively. Then, we use these PCs as new features to implement SVM.



- When the sample size becomes large enough, pca 70% and 90% have similar or even better performance than the original data, suggesting that PCA can well capture the essential pattern of the data.
- Meanwhile, using a few PCs as features significantly reduces the training time, leading to more efficient training.
- We analyze that the reason why PC SVM behaves less satisfying with a small sample size is that in high dimensional settings, sample mean and sample covariance obtain higher risk, and PCA may fail to capture the signals.

Experiment 2

We then consider biased estimators by regularization, Lasso, Ridge and Elastic Net of which the theories are relatively well established for MLR. We randomly split data into training and testing sets as 80% : 20% and implement the standard MLE and the three biased estimators. The tuning parameter is chosen based on 10-fold cross-validation. Repeating this procedure for 40 times, we create the boxplots for prediction accuracy of different methods.



- All the regularized estimators generalize better than the standard model. This implies that the underlying structure of the regression model is sparse, which coincides with the intuition that only a few pixels matter to distinguish the digits.
- Ridge behaves worse than others since it only shrinks the coefficients but not to zero.
- PC regression (with first 50 PCs) performs well, consistent with experiments 1.

Experiment 3

In many high-dimensional cases, there exists some latent confounding variables affecting both predictors and response, leading to a small but dense perturbation on the sparse structure. We use the first 5 PCs to represent the confounding factors and regress them out from both predictors and response in hope of removing the hidden confounding effects. The figure in Exp. 2 shows

- without penalty (PC.Adjust), the model makes no difference from the standard MLR;
- with Lasso penalty (PC.Adjust.Lasso), it even slightly outperforms standard Lasso, which is potentially because this deconfounding adjustment improves the irrepresentable condition for Lasso to achieve variable selection consistency.

Conclusion and Future Work

For both geometric dimension reduction techniques and regularized estimators, we conduct experiments to demonstrate how they represent signals from data, leading to more effective and efficient classification performance. We also analyze the potential theoretical reason supporting the results.

In the future, we plan to explore more in high dimensional setting, i.e., when dimension exceeds sample size. Aimed at the curse of dimensionality, we will try more methods like James-Stein estimate and non-convex penalty.

References

[1] Robert Tibshirani, Martin Wainwright and Trevor Hastie. *Statistical learning with sparsity: the lasso and generalizations.* Chapman and Hall/CRC, 2015.

[2] Domagoj Ćevid and Peter Bühlmann and Nicolai Meinshausen. Spectral Deconfounding and Perturbed Sparse Linear Models. *arXiv e-prints*, *arXiv:1811.05352*, Nov 2018.