Nonlinear Dimension Reduction to Improve Predictive Accuracy in Genomic and Neuroimaging Studies

Maxime Turgeon

June 5, 2018

McGill University
Department of Epidemiology, Biostatistics, and Occupational Health

Acknowledgements

This (ongoing) work has been done under the supervision of:

- Celia Greenwood (McGill University)
- Aurélie Labbe (HEC Montréal)

CENTRE LUDMER CENTER

NEUROINFORMATIQUE & SANTÉ MENTALE | NEUROINFORMATICS & MENTAL HEALTH

Motivation

- Modern genomics and neuroimaging bring an abundance of high-dimensional, correlated measurements $\mathbf{X}$.
- We are interested in predicting a clinical outcome $\mathbf{Y}$ based on the observed covariates $\mathbf{X}$.
  - However, the collected data typically contains thousands of covariates, whereas the sample size is at most a few hundreds.
- We would also want to capture the potentially complex, nonlinear association between $\mathbf{X}$ and $\mathbf{Y}$, and between the covariates themselves.

Motivation

- With a low to medium signal-to-noise ratio, the information contained in the data should be used sparingly.
- Moreover, from a clinical perspective, we need to account for the possibility of similar clinical profiles leading to different outcomes.
  - We want **prediction**, not *classification*.

Proposed approach

This work investigates the properties of the following approach:

- Let **X** be p-dimensional and **Y** binary.
- Using nonlinear dimension reduction methods, extract K components $\hat{L}_1, \ldots, \hat{L}_K$.
- Predict Y using a logistic regression model of the form

$$\operatorname{logit}\left(\mathrm{E}\left(Y \mid \hat{L}_1, \ldots, \hat{L}_K\right)\right) = \beta_0 + \sum_{i=1}^{K} \beta_i \hat{L}_i.$$

Nonlinear dimension reduction

General principle

- In PCA and ICA, we learn a linear transformation from the latent structure to the observed variables (and back).
- On the other hand, nonlinear dimension reduction (NLDR) methods try to learn the manifold underlying the latent structure.
    - NLDR methods are *non-generative*, i.e. they do not learn the transformation.
- The main approach: preserve local structures in the data.

Multidimensional Scaling

- **Main principle**: Manifolds can be described by pairwise distances.
- Let $D = (d_{ij})$ be the matrix of pairwise distances for the observed values $\mathbf{X}_1, \dots, \mathbf{X}_n$.
- The goal is now to find $\mathbf{L}_1, \dots \mathbf{L}_n$ in a lower dimensional space such that

$$\left(\sum_{i \neq j} (d_{ij} - \|\mathbf{L}_i - \mathbf{L}_j\|)^2 \right)^{1/2}$$

  is minimized.
- The objective function can also be weighted in a such a way that preserving small distances is prioritized.

Other methods

Other methods that are considered in this work:

- Isomap;
- Laplace Eigenmaps (SE);
- kernel PCA;
- Locally Linear Embedding (LLE);
- t-distributed Stochastic Embedding (t-SNE).

All methods are implemented in the Python module `scikit-learn`.

Simulations

General framework

$$\begin{tikzcd}
X_1, \ldots, X_p \arrow[rr, dashed] & & Y \\
& L_1, \ldots, L_K \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Performance metrics

We want to measure two key properties:

1. **Calibration**: using the Brier score (*lower* is better);
2. **Discrimination**: using the AUROC (*higher* is better).

1. Swiss roll

- We first generate two uniform variables $L_1 \sim U(0, 10)$ and $L_2 \sim U(-1, 1)$.
- We then generate a binary outcome Y:

$$\operatorname{logit}\left(\mathrm{E}\left(Y \mid L_1, L_2\right)\right) = -5 + L_1 - L_2.$$

- Finally, we generate three covariates X_1, X_2, X_3:

$$(X_1, X_2, X_3) = (L_1 \cos(L_1), L_2, L_1 \sin(L_1)).$$

- We fix $n = 500$ and repeat the simulation $B = 250$ times.

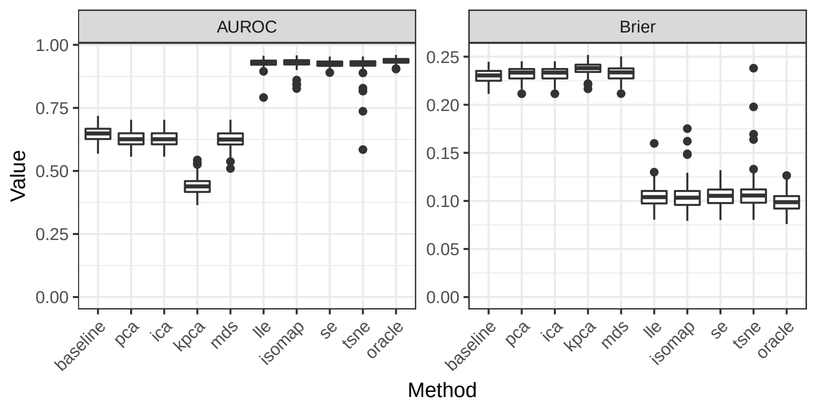
1. Swiss roll

1. Swiss roll

We compared 10 approaches:

1. **Oracle**: logistic regression with L_1, L_2 (i.e. the true model);
2. **Baseline**: logistic regression with X_1, X_2, X_3;
3. **Classical linear methods**: PCA, ICA;
4. **Manifold learning methods**: kernel PCA, Multidimensional scaling (MDS), Isomap, Locally Linear Embedding (LLE), Spectral Embedding (SE), and t-distributed Stochastic Neighbour Embedding (tSNE).

1. Swiss roll–Results



2. Random quadratic forms

- We first generate K latent variables $L_1, \ldots, L_K$.
- All p covariates are generated as *random quadratic forms* of the latent variables.

1. Select a random subset $L_1, \ldots, L_k$ of the K latent variables.
    - E.g. L_1 and L_4.
2. Form all possible quadratic combinations of the selected variables.
    - E.g. L_1^2, $L_1 L_4$, L_4^2.
3. Sample coefficients from standard normal and sum all terms.
    - E.g. $X_i = -0.5L_1^2 - 0.1L_1L_4 + 0.7L_4^2$.

2. Random quadratic forms

- The association between Y and $L_1, \ldots, L_5$ is defined via

$$\operatorname{logit}\left(\mathrm{E}\left(Y \mid L_1, \ldots, L_5\right)\right) = \sum_{i=1}^{5} \beta_i L_i,$$

  where

$$\beta_i = \frac{(-1)^i 2}{\sqrt{5}}.$$

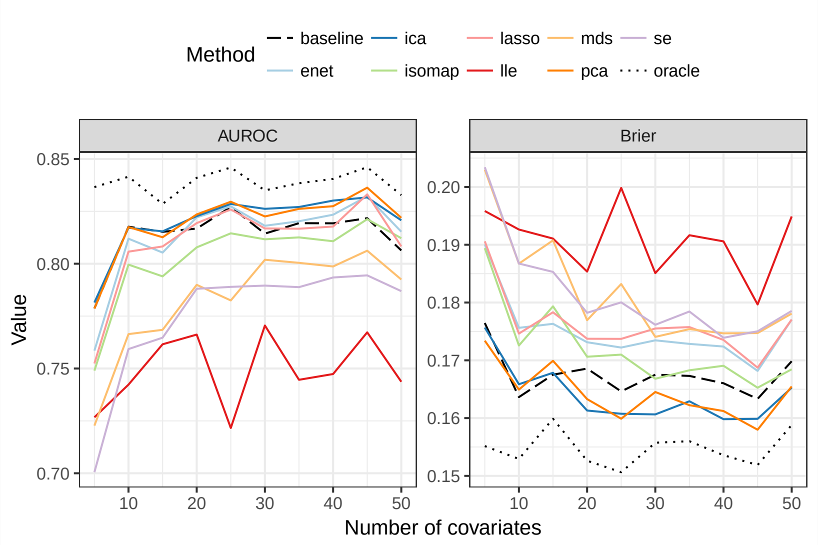
- The sample size varies as $n = 100, 150, 250,$ 300.
- The distribution of the covariates:
  - Standard normal;
  - Folded standard normal;
  - Exponential with mean 1.
- The simulation was repeated $B = 50$ times.

2. Random quadratic forms

We compared 12 approaches:

1. **Oracle**: logistic regression with only the first five covariates (i.e. the true model);
2. **Baseline**: logistic regression with all p variables;
3. **Lasso regression** using all p variables;
4. **Elastic-net regression** using all p variables;
5. **Classical methods and nonlinear extensions**: PCA, ICA, kernel PCA, and Multidimensional scaling (MDS);
6. **Manifold learning methods**: Isomap, Locally Linear Embedding (LLE), Spectral Embedding (SE), and t-distributed Stochastic Neighbour Embedding (tSNE).

2. Random quadratic forms–Results



Discussion

Summary

- The Swiss roll example shows that manifold learning methods recover the latent structure, which leads to good predictive performance.
- The random quadratic form example shows that highly complex models can lead to *worse* performance that classical PCR.
- NLDR methods have known limitations:
    - Trouble with manifolds with non-trivial homology (holes and self-intersections)
    - Sensitive to choice of neighbourhoods.
- **Where is the boundary between both regimes?**

Theoretical results

- Whitney's and Nash's embedding theorems guarantee that any (smooth or Riemannian) manifold can be embedded without intersections in a Euclidean space of high enough dimension.
- *Johnson-Lindenstrauss lemma*: We can project high-dimensional data points and preserve distances if dimension of lower space is high enough.

Final remarks

- Where does nature fit in all this? What kind of latent structures may underlie neuroimaging or genomic data?
- **Future Work**: Find low dimensional example with low performance, and high-dimensional example with good performance.
    - The latter implies finding a way to generate a high-dimensional structure with no self-intersection.

Questions or comments?

For more information and updates, visit `maxturgeon.ca`.