DIMENSION REDUCTION TECHNIQUES: A REVIEW

Efstathia Bura
Department of Statistics
The George Washington University
ebura@gwu.edu

The Relevance of Dimensionality Reduction

- Advances in data collection and storage capabilities have led to an information overload
- A system processes data in the form of a collection of real-valued vectors: speech signals, images, etc.
- Suppose the system is effective if the dimension of the vector is not *too high*
- Problem of Dimensionality Reduction appears when the data are of a higher dimension than tolerated
- Example: Stat Analysis of a Multivariate Population–interested in finding structures and/or interpreting the variables
  - Convenient to visualize the data, i.e., reducing their dimensionality to 2 or 3.
- In general, when the **intrinsic** dimensionality of the data is smaller than the actual, DR brings improved understanding of the data and their structure
  - Feature extraction
  - Representation in a different coordinate system

In Mathematical Terms

- Given the p-dimensional Random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$
- Find a lower representation $\mathbf{S} = (S_1, S_2, \ldots, S_k)^T$, $k \leq p$, with the same "information content" as $\mathbf{X}$
  - Have to select a criterion
- The $\mathbf{S}$-components are called hidden or latent variables
- Two Types of DR Problems: Linear and Non-Linear
- Concentrate on Linear DR Techniques– Result in

$$\mathbf{S}_{k\times n} = \mathbf{W}_{k\times p}\mathbf{X}_{p\times n}$$

where $\mathbf{W}_{k\times p}$ is the transformation weight matrix

$$\mathbf{X}_{p\times n} = \mathbf{A}_{p\times k}\mathbf{S}_{k\times n}$$

General Definition of DR

- Suppose we have a sample $\{\mathbf{X}\}_{i=1}^{n}$ of p-dimensional vectors lying in a data space $\mathcal{X} \subset \mathbb{R}^p$
- Fundamental Assumption for DR: the sample actually lies, at least approximately, on a manifold (linear or nonlinear) of smaller dimension than the data space.
- Goal of DR: find a representation of this manifold (a coordinate system) that will allow to project the data vectors on it and obtain a low-dimensional, compact representation of the data
- Formally, given $\{\mathbf{X}\}_{i=1}^{n} \in \mathcal{X}$, find
  - A space $\mathcal{S} \subset \mathbb{R}^k$
  - A **dimensionality reduction mapping F:**

$$\begin{aligned} F : \mathcal{X} &\to \mathcal{S} \\ \mathbf{X} &\to \mathbf{S} = F(\mathbf{X}) \end{aligned}$$

  - a smooth, nonsingular **reconstruction mapping f:**

$$\begin{aligned} f : \mathcal{S} &\to \mathcal{X} \\ \mathbf{S} &\to \mathbf{X} = f(\mathbf{S}) \end{aligned}$$

- such that

  1. $k \leq p$ is as small as possible
  2. The manifold $\mathcal{M} = f(\mathcal{S})$ approximately contains all the sample points $\{\mathbf{X}\}_{i=1}^{n}$
  3. Or, the reconstruction error of the sample,

  $$\mathrm{E_d}(\{\mathbf{X}\}_{\mathrm{i=1}}^{\mathrm{n}}) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{d}(\mathbf{X}_{\mathrm{n}}, \mathbf{X}'_{\mathrm{n}})$$

  where $\mathbf{X}'_n = f(F(\mathbf{X}_n)$ is the reconstructed vector for $\mathbf{X}_n$ and d is a suitable distance in $\mathcal{X}$

- Conditions 2 and 3 are not equivalent: 3 implies 2 but not vice versa – $F \circ f \neq 1$
  - E.g., when $\mathbf{X}$ has a distribution on $\mathcal{X}$, this is typically the case.

The Curse of Dimensionality (Bellman, 1961)

- In the absence of simplifying assumptions, sample size required to estimate a function of several variables grows exponentially with the number of variables
- *Empty Space Phenomenon:* high-dimensional spaces are inherently sparse
  - one-dimensional standard normal: 68% of the mass is contained in [-1,1]
  - 10-dimensional standard normal, the same hypersphere contains only 0.02% of the mass

Supervised and Unsupervised Learning

- **Unsupervised Learning**: $\mathbf{X}_n$ comprise all the data
  - Principal Component Analysis: finds a few orthogonal linear combinations of the $\mathbf{X}$-components with the largest variance
  - Factor Analysis and Principal Factor Analysis: estimates unknown common factors
  - Projection Pursuit: given a projection index that defines the ”interestingness” of a direction, PP finds directions maximizing the index
  - Independent Component Analysis: finds linear projections that are as nearly statistically independent as possible
  - Multidimensional Scaling: finds a k-dimensional representation of $\mathbf{X}$ so that the distances among the points in the new space reflect the proximities in the data
  - Neural Nets

Supervised Learning: Prediction

- a response variable or vector is available
- try to reduce the dimension of $\mathbf{X}$ *after* imposing a specific structure on the regression curve $\mathrm{E}(\mathbf{Y}|\mathbf{X})$
  - *additive, generalized additive and projection pursuit models* (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1990), ACE (Breiman and Friedman, 1985), MARS (Friedman, 1991), *partially linear or spline models* (Green and Silverman, 1994), *single- and multi-index models* along with different fitting methods such as *average derivatives* (Härdle, 1990; Newey and Stoker, 1989; Samarov, 1993), *interaction splines*, and CART (Breiman, Friedman, Olshen and Stone, 1984)
  - A concise discussion of the above and additional nonparametric modelling techniques can be found in Fan and Gijbels (1996)

Pre-Modelling Approach

- Reduce the regressor dimension prior to assuming any model for the regression relationship
  - Partial Least Squares: SVD on $\mathbf{X}Y$
    - extensively used in chemometrics
    - emphasis on predicting the response and not on understanding the underlying relationship of the variables
  - "Global" methods – involving a spectral decomposition of an appropriate matrix
  - "Local" methods – Structure Adaptive Approaches

Global Methods for Dimension Reduction

- $\mathbf{Y}$ is a $q \times 1$ response vector
- $\mathbf{X}$ is a $p \times 1$ predictor vector

Sufficient dimension reduction in regression focuses on finding $k \leq p$ linear combinations

$$\boldsymbol{\eta}_1^T \mathbf{X}, \ldots, \boldsymbol{\eta}_k^T \mathbf{X}$$

that can replace $\mathbf{X}$

- without loss of information
- without requiring restrictive conditions on $\mathbf{Y}|\mathbf{X}$

If $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k) \, : p \times k$ matrix, the previous statement is expressed by

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\eta}^T \mathbf{X}$$

- **Goal:** Estimate the smallest subspace spanned by $\boldsymbol{\eta}$.

Definitions:

- $S(\boldsymbol{\eta})$, the subspace spanned by the columns of $\boldsymbol{\eta}$, is called a dimension reduction subspace.
- The **Central Subspace** $S_{\mathbf{Y}|\mathbf{X}} = \cap S(\boldsymbol{\eta})$ is the smallest dimension-reduction subspace which provides the greatest dimension reduction in the predictor vector.
- The dimension of the Central Subspace is the **Structural Dimension** of the regression
- Problem: estimate (1) the dimension of $S_{\mathbf{Y}|\mathbf{X}}$ and (2) basis elements of $S_{\mathbf{Y}|\mathbf{X}}$

- The estimation is based on finding a kernel matrix $\mathbf{M}$ so that $S(\mathbf{M}) \subset S_{\mathbf{Y}|\mathbf{X}}$

Estimation Methods

Two main approaches

- First moment methods
  - SIR and variations (Li, 1991): $\mathbf{M} = \mathrm{Cov}(\mathrm{E}(\mathbf{X}|\mathbf{Y}))$
  - polynomial inverse regression (Bura and Cook, 2001): $\mathbf{M} = \mathrm{E}(\mathbf{X}|\mathbf{Y})$
- Second moment methods
  - pHd (Li, 1991): $\mathbf{M} = \mathrm{E}((\mathbf{Y} - \mathrm{E}(\mathbf{Y}))\mathbf{X}\mathbf{X}^{\mathrm{T}})$
  - SAVE (Cook and Weisberg, 1991): $\mathbf{M} = \mathrm{E}(\mathrm{Cov}(\mathbf{X}) - \mathrm{Cov}(\mathbf{X}|\mathbf{Y})^2$
  - SIRII (Li, 1991): $\mathbf{M} = \mathrm{E}(\mathrm{Cov}(\mathbf{X}|\mathbf{Y}) - \mathrm{E}(\mathrm{Cov}(\mathbf{X}|\mathbf{Y})))^2$.

• Sliced Average Variance Estimation (SAVE) is possibly the most exhaustive:

- it gains information from both the inverse mean function and the differences of the inverse covariances.
- Schott (1993) essentially showed: SAVE= SIR $\oplus$ SIRII

Continuous Response - Continuous predictors

Let $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$. Assume there exists a $k \times d$ matrix $\boldsymbol{\eta}$ such that

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \boldsymbol{\eta}^T \mathbf{Z}$$

with $d << k$.

- $S(\boldsymbol{\eta})$ is the range space of $\boldsymbol{\eta}$: a dimension reduction subspace
- $S_{\mathbf{Y}|\mathbf{Z}}$: Central Subspace in the $\mathbf{Z}$-scale
- $S_{\mathbf{Y}|\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1/2} S_{\mathbf{Y}|\mathbf{Z}}$
- The estimation of the central subspace in the two scales yields equivalent results

- All first and second moment methods can be used to estimate directions in $S_{Y|\mathbf{Z}}$

- The Inverse Mean Subspace:

$$S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} = \mathrm{span}\, \mathrm{E}(\mathbf{Z}|\mathbf{Y})$$

- The Inverse Covariance Subspace:

$$S_{\mathrm{Var}(\mathbf{Z}|\mathbf{Y})} = \mathrm{span}\{\mathbf{I} - \mathrm{Var}(\mathbf{Z}|\mathbf{Y})\}^2$$

- Two Important Conditions:

1. Linearity Condition: $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z})$ is linear in $\boldsymbol{\eta}^T\mathbf{Z}$
2. Constant Variance Condition: $\mathrm{Var}(\mathbf{Z}|\boldsymbol{\eta}^T\mathbf{Z})$ is constant

- Linearity Condition yields:

$$S_{\mathrm{E}(\mathbf{Z}|\mathbf{Y})} \subseteq S_{\mathbf{Y}|\mathbf{Z}}$$

- The Linearity and Constant Variance Conditions yield:

$$S_{\mathrm{Var}(\mathbf{Z}|\mathbf{Y})} \subseteq S_{\mathbf{Y}|\mathbf{Z}}$$

- Under both conditions, any weighted average of $\mathrm{E}(\mathbf{Z}|\mathbf{Y})$, $\mathrm{E}(\mathbf{Z}|\mathbf{Y})\mathrm{E}(\mathbf{Z}|\mathbf{Y})^{\mathrm{T}}$ or $\mathbf{I} - \mathrm{Var}(\mathbf{Z}|\mathbf{Y})$ span a subspace of $S_{\mathbf{Y}|\mathbf{Z}}$

- Both conditions refer to the marginal distribution of the predictors
- They are satisfied when $\mathbf{Z}$ is normal but normality is not necessary
- Ellipticity of the regressor vector guarantees the linearity condition
- They can be empirically checked by considering the scatterplot matrix of the predictors.
- The linearity of $\mathrm{E}(\mathbf{Z}|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z})$ can be ascertained if the scatterplots look roughly linear or random, and the homogeneity of the variance holds if there are no apparent fluctuations in data density
- Only substantial departures from both conditions are problematic: Transformations of the regressors

Estimation of the Structural Dimension

- Let $d = \dim S_{\mathbf{Y}|\mathbf{Z}}$
- Let $\hat{\mathbf{M}}$ be a sample estimate of the kernel matrix $\mathbf{M}$
- Let $(\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_p)^T$ be the eigenvectors corresponding to the eigenvalues (or singular values) $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p$ of $\hat{\mathbf{M}}$
- Let $\hat{\Lambda}_k \sim n \sum_{j=k+1}^{p} \hat{\lambda}_j$ be a test statistic
- $\hat{\Lambda}_d$ has an asymptotic weighted chi-squared distribution
- $\hat{d} = k$ if $\hat{\Lambda}_{k-1}$ is large whereas $\hat{\Lambda}_k$ is small

Marginal moment based dimension reduction

k-th Moment Dimension Reduction Subspaces

- Let $M^{(k)}(Y|\mathbf{X})$ denote the kth centered moment of $F(Y|\mathbf{Y})$
- $M^{(1)}(Y|\mathbf{X}) = \mathrm{E(Y}|\mathbf{X})$
- $M^{(2)}(Y|\mathbf{X}) = \mathrm{Var}(Y|\mathbf{X})$

If

$$Y \perp\!\!\!\perp \{M^{(1)}(Y|\mathbf{X}), ..., M^{(k)}(Y|\mathbf{X})\}|\boldsymbol{\eta}^T\mathbf{X},$$

then $\mathcal{S}(\boldsymbol{\eta})$ is called a k-th moment dimension reduction subspace (DRS) for the regression of Y on $\mathbf{X}$

- **The *central k-th moment subspace (CKMS)*** $\mathcal{S}^{(k)}_{Y|\mathbf{X}}$ is the intersection of all k-th moment DRSs

$$\mathcal{S}^{(1)}_{Y|\mathbf{X}} \subseteq \ldots \subseteq \mathcal{S}^{(k)}_{Y|\mathbf{X}}$$

The First Moment of $F(Y|\mathbf{X})$

- The regression function $\mathrm{E(Y|\mathbf{X})}$
  - Central Mean Subspace: $S_{\mathrm{E(Y|\mathbf{X})}} \subset S_{Y|\mathbf{X}}$
  - $S_{\mathrm{E(Y|\mathbf{X})}} = S_{Y|\mathbf{X}}$ when $Y \perp\!\!\!\perp \mathbf{X}|\mathrm{E(Y|\mathbf{X})}$, e.g.,

$$Y = f(\boldsymbol{\beta}^T\mathbf{X}) + \epsilon, \quad \epsilon \perp\!\!\!\perp \mathbf{X}$$

  - Inference for the Central Mean Subspace $S_{\mathrm{E(Y|\mathbf{X})}}$ in Cook and Li (2002)
  - Investigated OLS, SIR, pHd, SAVE and proposed alternatives

In general, marginal methods like Cov_k (Yin and Cook 2002) and pHd (Li 1992)

- use moment estimates of moments of functions of the response and predictor vectors avoiding nonparametric estimation
- they can be easily extended to multivariate response regressions, e.g., *multi* Cov_k (Yin and Bura, 2003)

Categorical Response – Continuous Predictors

- Both SIR and SAVE can be used to estimate directions in $S_{Y|\mathbf{Z}}$
- Cook and Lee (1999): Binary Response
- Cook and Critchley (2000) and Cook and Yin (2001): Generalization and Discriminant Analysis
- SIR is equivalent to Linear Discriminant Analysis in the sense that they both estimate the same discriminant linear combinations of the predictors when they are normal.
  - Disadvantage: LDA and SIR find at most $C-1$ linear combinations for discrimination. In binary regression, **at most 1**.
- SAVE is equivalent to Quadratic Discriminant Analysis when predictors are normal

Continuous Response – Mixed Type Predictors

- Projections of Categorical Variables are not well defined
- *Partial Dimension Reduction:* One Categorical Predictor W and continuous random vector $\mathbf{X}$
- Chiaromonte, Cook and Li (2001) and Li, Cook and Chiaromonte (2003) developed *Partial Sliced Inverse Regression* for the subpopulations in $(Y, \mathbf{X})$ defined by the W-categories
- Problematic when there are many categorical predictors, if not all – e.g. epidemiological studies

Discussion and Limitations

- At least the linearity condition has to be satisfied for any of these methods to apply
- First moment methods are sensitive to linear trends in dependence of Y on $\mathbf{X}$
  - $Y = (\beta^T \mathbf{X})^2 + \epsilon$, with $\mathbf{X} \sim N(0, \mathbf{I}_p)$
  - SIR will estimate 0 and miss β
- All estimators are $\sqrt{n}$-consistent
- **But** $\hat{S}_{\mathbf{Y}|\mathbf{X}}$ is not an exhaustive estimate for $S_{\mathbf{Y}|\mathbf{X}}$
- The tests for dimension are sequential. Not much is known about their power or "optimality".
- What is the structural dimension of non-linear manifolds? A measure of complexity and hence non-discrete?
- Other?

Local Methods: concentrate on local features

- Multi-index NP-regression modelling
- not quite pre-modelling: NP-estimation of the link function along with the index space
- virtually no assumptions on $\mathbf{X}$
- slower than $\sqrt{n}$-convergence
- computationally intensive
- Xia, Tong, Li and Zhu (2002) and Hristache, Juditsky, Polzehl, Spokoiny (2001)