# *Sufficient Dimension Reduction using Support Vector Machine and it's variants*

Andreas Artemiou
School of Mathematics, Cardiff University

@AG DANK/BCS Meeting 2013

*SDR*

*PSVM*

*Real Data*

*Current Research and other problems*

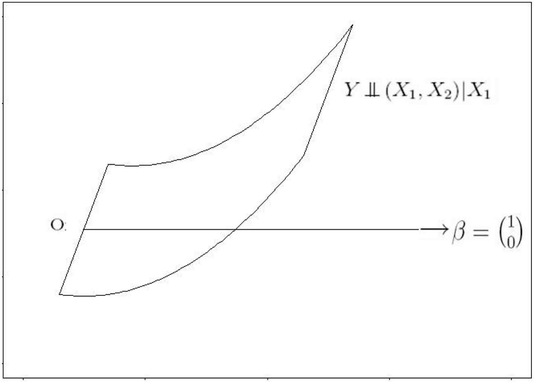
# Sufficient Dimension Reduction

- Objective: Attempt to identify a small number of directions that can replace a $p$ dimensional predictor vector $\mathbf{X}$ without loss of information on the conditional distribution of $Y|\mathbf{X}$.
- In other words, our objective is to estimate $\boldsymbol{\beta}$ under:

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

  where $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$, $d \leq p$.
- If $d < p$ dimension reduction is achieved.
- See for example Li K. C. (1991), (1992), Cook R. D. (1994), (1996), (1998), Xia et al (2002), Li, B. and Wang, S. (2007).

# *Example*



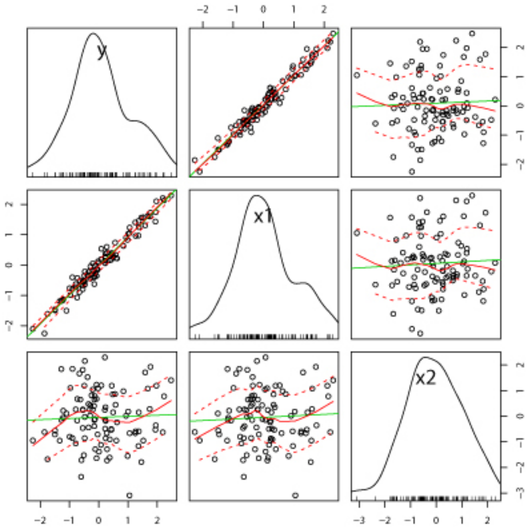
# *Sufficient Dimension Reduction*

- The space spanned by the column vectors of $\beta$ is called **Dimension Reduction Subspace** (DRS) and it is denoted with $\mathcal{S}(\beta)$.
- If the intersection of all DRSs is a DRS itself we call it the **Central Dimension Reduction Subspace** (CDRS) which is denoted with $\mathcal{S}_{Y|\mathbf{X}}$ and it has the smallest dimension among all possible DRSs.
- CDRS doesn't always exist, but if it exists is unique. (Cook 1998)

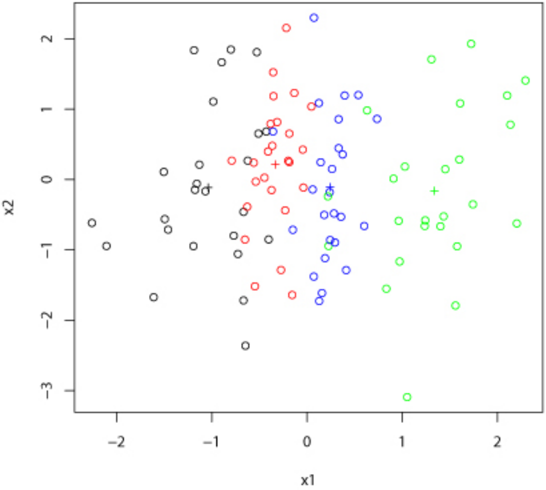
# *Example on SIR*

- This is a toy example with 100 data from bivariate standard normal.
- $Y = X_1 + \varepsilon$
- $\varepsilon \sim N(0, 0.2^2)$
- We have 4 slices from 25 points each.

# *Example on SIR*



# *Example on SIR*



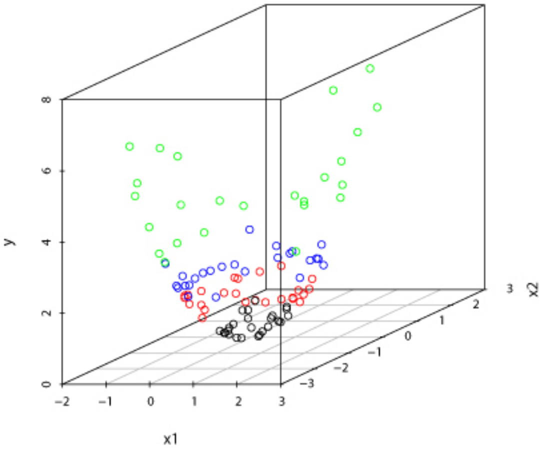
# *Algorithm of SIR by Li (1991)*

- Standardize the data $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - E(\mathbf{X}))$
- Slice the response variable into $H$ slices.
- In each slice find the mean of the data points on the hyperplane defined by the standardized predictors $\mathbf{Z}$, $\hat{m}_i, i = 1, \ldots, H$
- Build a candidate matrix $\hat{M} = \sum_{i=1}^{H} \hat{m}_i \hat{m}_i^{\top}$
- Perform eigenvalue decomposition to find the $d$ eigenvectors corresponding to the $d$ largest eigenvalues, $\hat{\eta}_1, \ldots, \hat{\eta}_d$.
- Use $\hat{\beta}_i = \mathbf{\Sigma}^{-1/2}\hat{\eta}_i$
- We need to assume the linearity condition, that is $E(\mathbf{X}|\boldsymbol{\beta}^{\top}\mathbf{X})$ is a linear function of $\boldsymbol{\beta}^{\top}\mathbf{X}$.

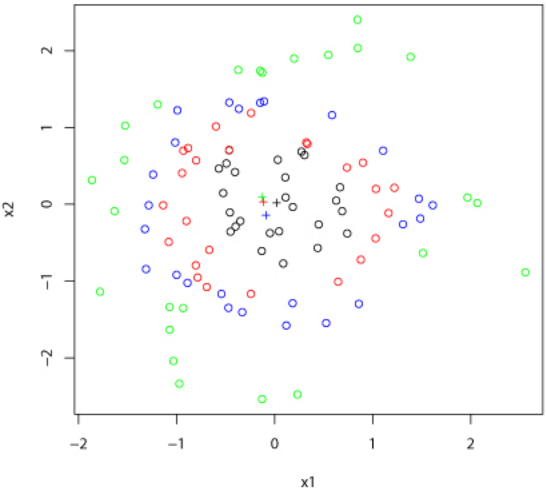
# *Example that SIR fails*

- This is a toy example with 100 data from bivariate standard normal.
- $Y = X_1^2 + X_2^2 + \varepsilon$. This creates the bowl dataset.
- $\varepsilon \sim N(0, 0.01^2)$
- We have 4 slices from 25 points each.

# *Example that SIR fails*



# *Example that SIR fails*



# *Constant Conditional Variance (CCV)*

- SAVE in addition to the LCM assumption, it also requires the CCV assumptions which is stated as: the $\mathrm{var}(\mathbf{X}|\beta^{\mathsf{T}}\mathbf{X})$ is non-random.
- LCM and CCV together was shown to be equivalent to assuming the predictors are normally distributed.
- principal Hessian direction (pHd) by Li (1992) and Cook (1998) and Directional Regression by Li and Wang (2007) are two other methods that were proposed and need both the LCM and the CCV.

# Some cases are not that simple..

- Let $\mathbf{X} \sim N(0, I_6)$, $p = \dim \mathbf{X} = 6$ and $\epsilon \sim N(0,1)$, $\epsilon \perp\!\!\!\perp \mathbf{X}$
- $Y = X_1 X_2 + X_3 X_4 + X_5 X_6 + \epsilon$
- In this case we cannot achieve any dimension reduction since $d = 6$ under the conditional model $Y \perp\!\!\!\perp \mathbf{X} | \beta^\top \mathbf{X}$.

# *Nonlinear Sufficient Dimension Reduction*

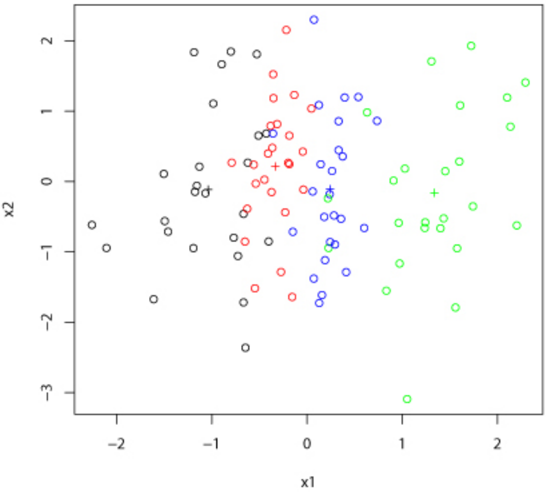
- We can achieve dimension reduction through nonlinear feature extraction under the conditional model

$$Y \perp\!\!\!\perp \mathbf{X} | \phi(\mathbf{X})$$

where $\phi : \mathbb{R}^p \to \mathbb{R}^d$ can be either linear or nonlinear function of the predictors.

- So in the previous example we will need only one direction to describe the above dimension reduction space, that is $d = 1$.
- See for example, Wu (2008) and Fukumizu, Bach and Jordan (2009) and Li, Artemiou and Li (2011).

# *Revisiting our toy example* $Y = X_1 + \varepsilon$



# *Objective function*

- For the soft margin SVM to construct the hyperplane we minimize:

$$\text{minimize} \quad \boldsymbol{\psi}^{\intercal}\boldsymbol{\psi} + \frac{\lambda}{n}\sum_{i=1}^{n}\xi_i \;\; \text{among } (\boldsymbol{\psi}, t, \boldsymbol{\xi}) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^n$$

$$\text{subject to} \quad \xi_i \geq 0, \;\; Y_i[\psi^{\intercal}(\mathbf{X}_i - \bar{\mathbf{X}}) - t] \geq 1 - \xi_i, \quad i = 1, \ldots, n.$$

- $Y_i$ is a binary variable with values $-1$ or 1, to indicate in which population the point belongs to.
- $\lambda$ is a constant called misclassification penalty or cost

# *Objective function*

- Fixing $(\psi, t)$ the above translates into minimizing:

$$\text{minimize} \quad \psi^\top \psi + \frac{\lambda}{n} \sum_{i=1}^{n} \left(1 - Y_i[\psi^\top(\mathbf{X}_i - \bar{\mathbf{X}}) - t]\right)^+$$

  among $(\psi, t) \in \mathbb{R}^p \times \mathbb{R}$.
- $a^+ = \max\{0, a\}$.
- This can be written in the population level as:

$$\psi^\top \psi + \lambda E[1 - Y(\psi^\top(\mathbf{X} - E\mathbf{X}) - t)]^+.$$

# Modifications to create PSVM

- The objective function in the previous slide is not the best in our case.
- We minimize the following objective function for PSVM:

$$\psi^\mathsf{T}\boldsymbol{\Sigma}\psi + \lambda E[1 - \tilde{Y}(\psi^\mathsf{T}(\mathbf{X} - E\mathbf{X}) - t)]^+.$$

- $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$ is added to the first term.
- We use $\tilde{Y} = I(Y \leq q) - I(Y > q)$ where $q$ is the boundary between slices.

*Theorem*

*Suppose $E(\mathbf{X}|\beta^\mathsf{T}\mathbf{X})$ is a linear function of $\beta^\mathsf{T}\mathbf{X}$. If $(\psi^*, t^*)$ minimizes the objective function above among all $(\psi, t) \in \mathbb{R}^p \times \mathbb{R}$, then $\psi^* \in \mathcal{S}_{Y|\mathbf{X}}$.*

# *Nonlinear PSVM*

- We minimize the following objective function for PSVM:

$$\langle \psi, \Sigma\psi \rangle_{\mathcal{H}} + \lambda E[1 - \tilde{Y}(\psi(\mathbf{X}) - E\psi(\mathbf{X}) - t)]^{+}.$$

- $\mathcal{H}$ is a Hilbert space of functions of $\mathbf{X}$
- $\Sigma$ is a bounded adjoint operator induced by the bilinear form $b(f_1, f_2) = \mathrm{cov}[f_1(\mathbf{X}), f_2(\mathbf{X})]$, ($b$ is defined from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$) (or in simple words, $\Sigma$ is the covariance operator).

# Nonlinear PSVM

*Theorem*

*Suppose the mapping* $\mathcal{H} \to L_2(P_{\mathbf{X}}), \quad f \mapsto f$ *is continuous and*

1. $\mathcal{H}$ *is a dense subset of* $L_2(P_{\mathbf{X}})$*,*
2. $Y \perp\!\!\!\perp \mathbf{X} | \phi(\mathbf{X})$*.*

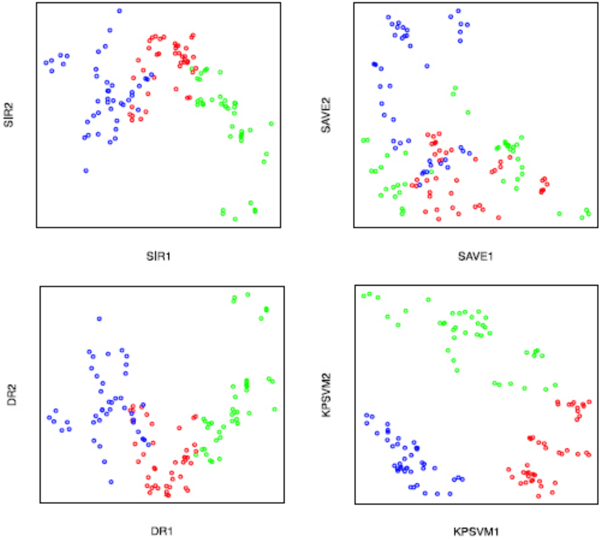
*If* $(\psi^*, t^*)$ *minimizes* $\langle \psi, \Sigma\psi \rangle_{\mathcal{H}} + \lambda E[1 - \tilde{Y}(\psi(\mathbf{X}) - E\psi(\mathbf{X}) - t)]^+$ *among all* $(\psi, t) \in \mathcal{H} \times \mathbb{R}$*,* $\psi^*(\mathbf{X})$ *is unbiased.*

- No linearity condition needed, while previous work on nonlinear dimension reduction assumed linearity (i.e. Wu 2008).

# *Vowel data*

- Vowel data from UCI repository.
- Differentiate 3 vowels from head (red), heed (green) and hud (blue)
- Training: 144 cases
- Testing: 126 cases
- Use training data to find the sufficient directions and plot the testing data on these directions.

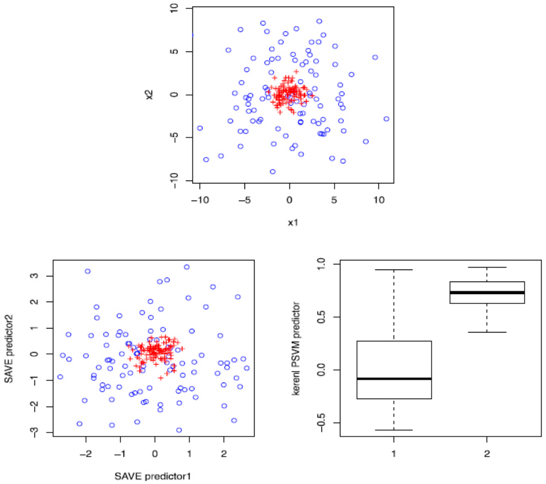
# Vowel data - pictures



# *Is this a classification method?*

- Not necessarily.
- Linear and nonlinear SDR has its own power in reducing, discriminating, visualizing and interpreting high dimensional data.
- Assume $Y \sim \mathrm{Bernoulli}(p = 1/2)$ and $\mathrm{var}(\mathbf{X}|Y = 0) = \mathrm{diag}(1, 1, 0, \ldots, 0)$ and $\mathrm{var}(\mathbf{X}|Y = 1) = \mathrm{diag}(10, 10, 0, \ldots, 0)$.
- $n = 200$ and $p = 10$.

# *Example - Comparison*



# *Define the dimension of CDRS*

- One of the most important aspects of dimension reduction is to determine how big should be the CDRS.
- There are two different approaches that were proposed.
- Sequential tests, BIC type criterion

## Modifications to create PSVM

- Objective function for PSVM:

$$\psi^{\mathsf{T}}\mathbf{\Sigma}\psi + \lambda E[1 - \tilde{Y}(\psi^{\mathsf{T}}(\mathbf{X} - E\mathbf{X}) - t)]^{+}.$$

- First part strictly convex but second is not which implies non-unique $t$. (Burges and Crisp (1999))
- Asymptotics depend on the value of $t$, i.e. Hessian matrix of $\boldsymbol{\theta} = (\psi, t)$:

$$2\mathrm{diag}(\mathbf{\Sigma}, 0) + \lambda \sum_{\tilde{y}=-1,1} P(\tilde{Y} = \tilde{y}) f_{\psi^{\mathsf{T}}\mathbf{X}|\tilde{Y}}(k|\tilde{y}) E(\mathbf{X}^{*}\mathbf{X}^{*\mathsf{T}}|\psi^{\mathsf{T}}\mathbf{X} = k).$$

# *Using Lq PSVM (Artemiou and Dong (almost ready to be submitted))*

- PSVM minimizes:

$$\psi^\top \mathbf{\Sigma} \psi + \lambda E[1 - \tilde{Y}(\psi^\top (\mathbf{X} - E\mathbf{X}) - t)]^+.$$

- We can use L2 SVM Abe (2002):

$$\psi^\top \mathbf{\Sigma} \psi + \lambda E[[(1 - \tilde{Y}(\psi^\top (\mathbf{X} - E\mathbf{X}) - t))]^+]^2.$$

- It gives strictly convex optimization problem.

# *Summary*

- Using SVM (and its variants) we create a class of techniques which have:
  - A unique framework for linear and nonlinear dimension reduction
  - Dimension Reduction with no matrix inversion in the linear case (pending the use of appropriate software)
  - Dimension reduction with no assumptions on the marginal distribution of $\mathbf{X}$ in the nonlinear case
  - We have new insights on the asymptotic properties of this class of methods.

## *Very short list of references*

- It is a long list so please do not hesitate to contact me if you need something... !!!


- Artemiou, A. and Shu, M. (to appear). A cost based reweighed method for PSVM.
- Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wile
- Li, B., Artemiou, A. and Li L. (2011). Principal Support Vector Machine for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics*, **39**, 3182–3210
- Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–342.
- Wu, H. M. (2008). Kernel sliced inverse regression with applications on classification. *Journal of Computational and Graphical Statistics*, **17**, 590–610.

# *Thank you*