Sufficient Reductions in Regression and Classification

Efstathia Bura

Institute of Statistics and Mathematical Methods in Economics
Vienna University of Technology

Medical University of Vienna

1. Statistical Inference for Big Data
    - Regression
    - A ...simple example
    - Eigen-representation of common estimators
    - Linear SDR
    - Moment Based SDR
2. Non-linear SDR
    - Exponential Family Predictors
    - Application: Multivariate Bernoulli Predictors
    - Classification
    - Simulation Experiments
3. Future Directions
4. References

Big Data and Statistics

Challenge: how to extract **useful** and **useable** knowledge from the overwhelming amount of raw data.

This is at the core of statistical inference:

the objective of statistical methods is the reduction of data. A quantity of data. . . is to be replaced by relatively few quantities which shall adequately represent. . . the relevant information contained in the original data

Fisher (1922): "On the mathematical foundations of theoretical statistics"

Regression

- High-level objective is to model $F(Y|\mathbf{X})$: this is a difficult problem
- Standard Approach: **Linear Regression**
  - Focus on first two moments

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\mathrm{E}(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \quad \mathrm{var}(Y_i) = \sigma^2, \text{ or } \sigma^2(X_{ij})$$

  - Estimation via OLS or GLS

Regression

- High-level objective is to model $F(Y|\mathbf{X})$: this is a difficult problem
- Standard Approach: **Linear Regression**
  - Focus on first two moments

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\mathrm{E}(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \quad \mathrm{var}(Y_i) = \sigma^2, \text{ or } \sigma^2(X_{ij})$$

  - Estimation via OLS or GLS

Regression

- High-level objective is to model $F(Y|\mathbf{X})$: this is a difficult problem
- Standard Approach: **Linear Regression**
  - Focus on first two moments

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\mathrm{E}(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \quad \text{var}(Y_i) = \sigma^2, \text{ or } \sigma^2(X_{ij})$$

  - Estimation via OLS or GLS

Regression

- High-level objective is to model $F(Y|\mathbf{X})$: this is a difficult problem
- Standard Approach: **Linear Regression**
  - Focus on first two moments

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, \ldots, n$$

$$\mathrm{E}(Y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \quad \mathsf{var}(Y_i) = \sigma^2, \text{ or } \sigma^2(X_{ij})$$

  - Estimation via OLS or GLS

Linear Model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \mathsf{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n, \text{ or } \mathsf{var}(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}_{y|x}$$

If

- the **signal/information** is in the first two moments and
- the model is **general** enough, i.e. includes polynomial terms, interactions, etc

then OLS/GLS regression can be a very effective modeling/predictive tool

A ...simple example

- $\mathbf{X} \sim N_{10}(0, \mathbf{I}_{10})$, i.e. 10 independent standard normal predictors
- Generate 200 observations for each X_j, $j = 1, \ldots, 10$.
- I have generated Y from a model that I will reveal later
- Want to predict Y using X_j's: where to start?
- Let's plot the data

A ...simple example

- $\mathbf{X} \sim N_{10}(0, \mathbf{I}_{10})$, i.e. 10 independent standard normal predictors
- Generate 200 observations for each X_j, $j = 1, \ldots, 10$.
- I have generated Y from a model that I will reveal later
- Want to predict Y using X_j's: where to start?
- Let's plot the data

A ...simple example

- $\mathbf{X} \sim N_{10}(0, \mathbf{I}_{10})$, i.e. 10 independent standard normal predictors
- Generate 200 observations for each X_j, $j = 1, \ldots, 10$.
- I have generated Y from a model that I will reveal later
- Want to predict Y using X_j's: where to start?
- Let's plot the data

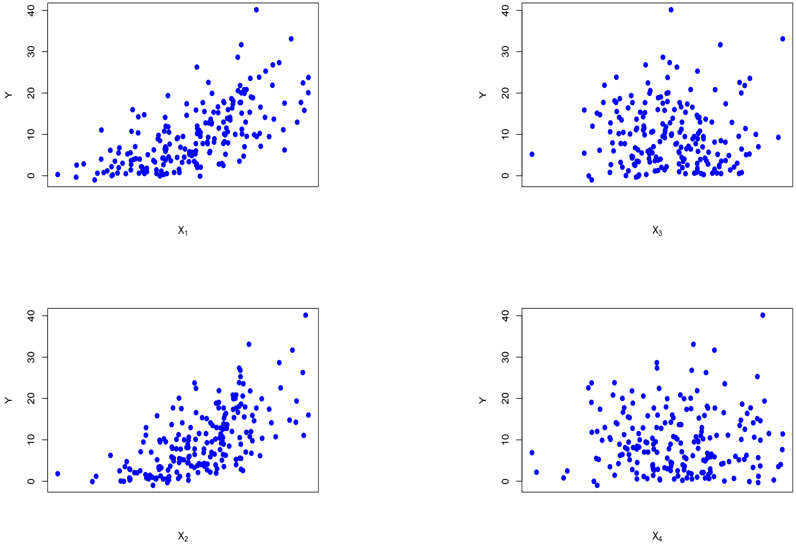
A ...simple example

- $\mathbf{X} \sim N_{10}(0, \mathbf{I}_{10})$, i.e. 10 independent standard normal predictors
- Generate 200 observations for each X_j, $j = 1, \ldots, 10$.
- I have generated Y from a model that I will reveal later
- Want to predict Y using X_j's: where to start?
- Let's plot the data

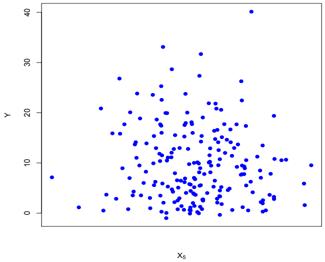
A ...simple example

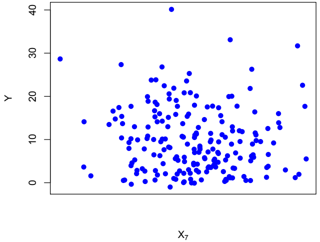
- $\mathbf{X} \sim N_{10}(0, \mathbf{I}_{10})$, i.e. 10 independent standard normal predictors
- Generate 200 observations for each X_j, $j = 1, \ldots, 10$.
- I have generated Y from a model that I will reveal later
- Want to predict Y using X_j's: where to start?
- Let's plot the data

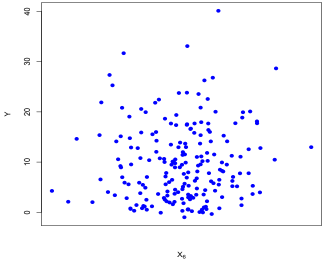
Marginal Plots

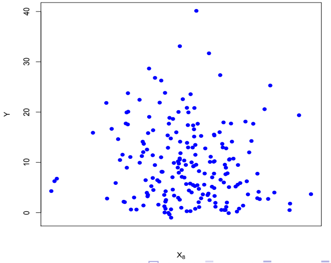


Marginal Plots









What model to fit?

- Looks like Y is a non-linear function of (at least) X_1 and X_2
- One could start from

$$
\begin{aligned}
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \\
+ \beta_6 x_3 + \beta_7 x_4 + \ldots + \beta_{13} x_{10} + \epsilon
\end{aligned}
$$

- With p moderately large modeling is challenging: very difficult to visualize how Y changes as a function of the components of $\mathbf{X}$
- Is this the final model?

What model to fit?

- Looks like Y is a non-linear function of (at least) X_1 and X_2
- One could start from

$$
\begin{aligned}
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \\
+ \beta_6 x_3 + \beta_7 x_4 + \ldots + \beta_{13} x_{10} + \epsilon
\end{aligned}
$$

- With p moderately large modeling is challenging: very difficult to visualize how Y changes as a function of the components of $\mathbf{X}$
- Is this the final model?

What model to fit?

- Looks like Y is a non-linear function of (at least) X_1 and X_2
- One could start from

$$
\begin{aligned}
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \\
+ \beta_6 x_3 + \beta_7 x_4 + \ldots + \beta_{13} x_{10} + \epsilon
\end{aligned}
$$

- With p moderately large modeling is challenging: very difficult to visualize how Y changes as a function of the components of $\mathbf{X}$
- Is this the final model?

What model to fit?

- Looks like Y is a non-linear function of (at least) X_1 and X_2
- One could start from

$$\begin{aligned} y = {} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 \\ & + \beta_6 x_3 + \beta_7 x_4 + \ldots + \beta_{13} x_{10} + \epsilon \end{aligned}$$

- With p moderately large modeling is challenging: very difficult to visualize how Y changes as a function of the components of $\mathbf{X}$
- Is this the final model?

Model Selection

- Stepwise Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- All Subset Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- Penalized Regression, e.g. LASSO

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **All are constrained by the starting model**
- Unknown effect of all the data processing (e.g. inducing collinearity?) on the validity of inference (confidence intervals, tests of hypotheses, predictions)

Model Selection

- Stepwise Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- All Subset Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- Penalized Regression, e.g. LASSO

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **All are constrained by the starting model**
- Unknown effect of all the data processing (e.g. inducing collinearity?) on the validity of inference (confidence intervals, tests of hypotheses, predictions)

Model Selection

- Stepwise Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- All Subset Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- Penalized Regression, e.g. LASSO

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **All are constrained by the starting model**
- Unknown effect of all the data processing (e.g. inducing collinearity?) on the validity of inference (confidence intervals, tests of hypotheses, predictions)

Model Selection

- Stepwise Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- All Subset Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- Penalized Regression, e.g. LASSO

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **All are constrained by the starting model**
- Unknown effect of all the data processing (e.g. inducing collinearity?) on the validity of inference (confidence intervals, tests of hypotheses, predictions)

Model Selection

- Stepwise Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- All Subset Regression with R^2, R^2_{adj}, AIC/BIC, C_p
- Penalized Regression, e.g. LASSO

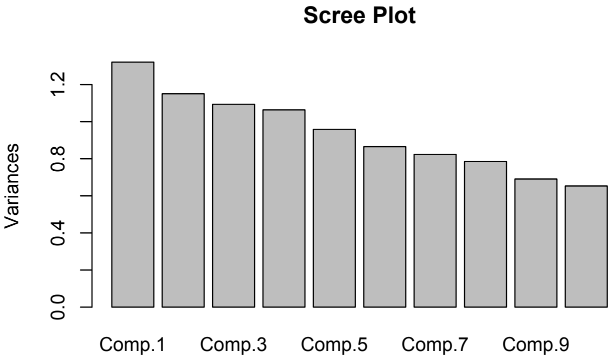
$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **All are constrained by the starting model**
- Unknown effect of all the data processing (e.g. inducing collinearity?) on the validity of inference (confidence intervals, tests of hypotheses, predictions)

Feature Extraction

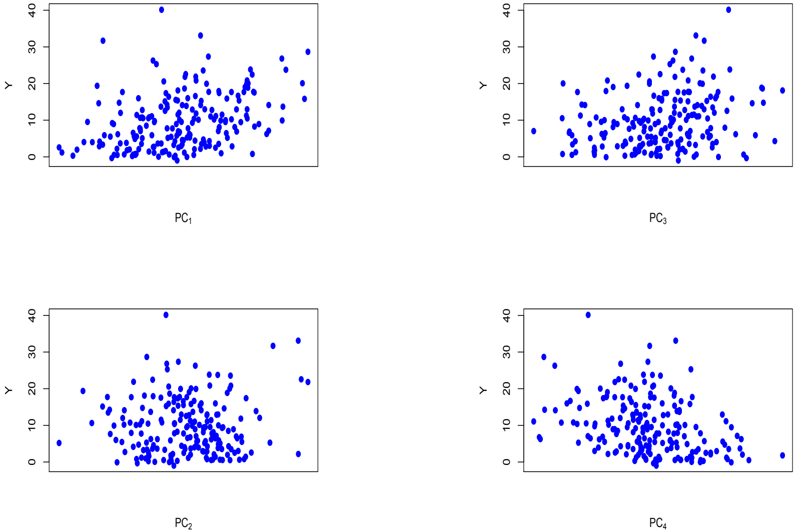
- $\mathbf{X} = (X_1, \ldots, X_p)$
- If all we needed to model Y is $\boldsymbol{\alpha}'\mathbf{X}$, i.e. a few ($< p$) linear combinations of the X's
- Then we would plot Y versus $\boldsymbol{\alpha}'\mathbf{X}$ and modeling would be much simpler
- This idea has been around for a long time: Principal Component Regression (PCR) and variants–Ridge and PLS Regression
- Let's apply PCR to our example

The scree plot



- No PC stands out

Marginal Plots for the PCs



Alternatively: Sliced Inverse Regression

Using the *dr* package in R:

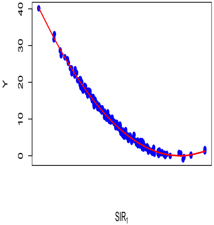
```
s1=dr(y˜x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,method="sir")
summary(s1)

Method:
sir with 14 slices, n = 200.

Large-sample Marginal Dimension Tests:
              Stat  df   p.value
0D vs >= 1D 304.73 130 4.441e-16
1D vs >= 2D 113.33 108 3.439e-01
2D vs >= 3D  78.15  88 7.647e-01
3D vs >= 4D  53.88  70 9.232e-01
```

SIR estimates the dimension to be 1!

SIR estimates the linear function needed to model Y. What next?

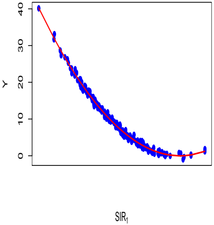


- Plot y versus the SIR predictor
- Model y as a quadratic function of this new predictor SIR1
- `summary(lm(y~SIR1+SIR1sq))`

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.24427    0.05719  161.64   <2e-16 ***
SIR1        -8.61903    0.04360 -197.69   <2e-16 ***
SIR1sq       1.95772    0.03008   65.09   <2e-16 ***

Residual standard error: 0.6485 on 197 degrees of freedom
Multiple R-squared: 0.9951,     Adjusted R-squared: 0.9951
F-statistic: 2.01e+04 on 2 and 197 DF,  p-value: < 2.2e-16
```

SIR estimates the linear function needed to model Y. What next?

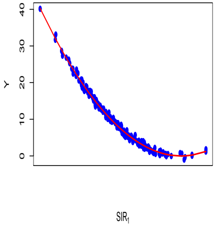


- Plot y versus the SIR predictor
- Model y as a quadratic function of this new predictor SIR1
- summary(lm(y~SIR1+SIR1sq))

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.24427    0.05719  161.64   <2e-16 ***
SIR1        -8.61903    0.04360 -197.69   <2e-16 ***
SIR1sq       1.95772    0.03008   65.09   <2e-16 ***

Residual standard error: 0.6485 on 197 degrees of freedom
Multiple R-squared: 0.9951,     Adjusted R-squared: 0.9951
F-statistic: 2.01e+04 on 2 and 197 DF,  p-value: < 2.2e-16
```

SIR estimates the linear function needed to model Y. What next?



- Plot y versus the SIR predictor
- Model y as a quadratic function of this new predictor SIR1
- `summary(lm(y~SIR1+SIR1sq))`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.24427    0.05719  161.64   <2e-16 ***
SIR1        -8.61903    0.04360 -197.69   <2e-16 ***
SIR1sq       1.95772    0.03008   65.09   <2e-16 ***

Residual standard error: 0.6485 on 197 degrees of freedom
Multiple R-squared: 0.9951,     Adjusted R-squared: 0.9951
F-statistic: 2.01e+04 on 2 and 197 DF,  p-value: < 2.2e-16
```

Summary

- Truth: $Y = (X_1 + X_2 + 3)^2 + .5N(0,1)$
- Dimension is 1: a single linear function, $X_1 + X_2$, is needed to model Y
- **SIR** identified that this is a 1-dimensional problem and estimated $\boldsymbol{\alpha}'\mathbf{X}$ that can replace $\mathbf{X}$ in the regression of Y on $\mathbf{X}$
- The **complexity of modeling** Y has been drastically reduced
- Much easier to accurately specify the model

Singular Value Decomposition

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\begin{aligned} \text{var}(\mathbf{X}) = \boldsymbol{\Sigma}_x &= \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}' \ : p \times p \\ \mathbf{V} &= (\mathbf{v}_1, \ldots, \mathbf{v}_p) \\ \boldsymbol{\Lambda} &= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_p \end{pmatrix} \end{aligned}$$

- $\text{var}(\mathbf{X}) = \boldsymbol{\Sigma}_x = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$
- $\boldsymbol{\Sigma}_x^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j' : p \times p$
- $\text{cov}(\mathbf{X}, Y) = \boldsymbol{\sigma}_{xy} : p \times 1$

Parameter Estimators

- OLS: $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\sigma}_{xy}$
- PCR: $\boldsymbol{\beta}_{PCR} = \boldsymbol{\Sigma}_x^{-1}(M)\boldsymbol{\sigma}_{xy}, \ M \leq p$
- Ridge: $\boldsymbol{\beta}_{RR} = (\boldsymbol{\Sigma}_x + \kappa\mathbf{I}_p)^{-1}\boldsymbol{\sigma}_{xy}$
- PLS: $\boldsymbol{\beta}_{PLS} = \boldsymbol{\Sigma}_x^D(u)\boldsymbol{\sigma}_{xy}$
  - $\boldsymbol{\Sigma}_x^D(u) = \mathbf{W}_u(\mathbf{W}_u'\boldsymbol{\Sigma}_x\mathbf{W}_u)^{-1}\mathbf{W}_u'$
  - $\mathbf{W}_u = (\boldsymbol{\sigma}_{xy}, \boldsymbol{\Sigma}_x\boldsymbol{\sigma}_{xy}, \ldots, \boldsymbol{\Sigma}_x^u\boldsymbol{\sigma}_{xy})$

Eigen-representation of the estimators

$$\text{var}(\mathbf{X}) = \mathbf{\Sigma}_x = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j'$$

Untargeted

- OLS: $\boldsymbol{\beta}_{OLS} = \sum_{j=1}^{p} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j' \boldsymbol{\sigma}_{xy}$
- PCR: $\boldsymbol{\beta}_{PCR} = \sum_{j=1}^{M} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j' \boldsymbol{\sigma}_{xy}$, $M \leq p$
- Ridge: $\boldsymbol{\beta}_{RR} = \sum_{j=1}^{p} \frac{1}{\lambda_j + \kappa} \mathbf{v}_j \mathbf{v}_j' \boldsymbol{\sigma}_{xy}$

Targeted

- PLS: $\boldsymbol{\beta}_{PLS} = \sum_{j=1}^{u} \frac{1}{\lambda_j} \mathbf{v}_j \mathbf{v}_j' \boldsymbol{\sigma}_{xy}$
  - Summation is only over first u eigenvalues that satisfy $\mathbf{v}_j' \boldsymbol{\sigma}_{xy} \neq 0$
  - If an eigenvalue has multiplicity $r > 1$, only one eigenvector among the r is chosen

Connection with SDR

- **PC** starts with the eigen-decomposition of var$(\mathbf{X})$ then selects directions with the maximal variance of $\mathbf{X}$
- **PLS** starts with targeted eigen-decomposition of var$(\mathbf{X})$ using the correlation of $\mathbf{X}$ with Y as ordering principle.
- **SIR** (a linear SDR method) uses the eigen-decomposition of var$(\mathrm{E}(\mathbf{X}|Y))$ with $\mathrm{E}(\mathbf{X}|Y)$ as ordering principle
- Why is that a good idea?

$$\text{var}(\mathbf{X}) = \text{var}[\mathrm{E}(\mathbf{X}|Y)] + \mathrm{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\mathrm{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\mathrm{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

$$\mathrm{var}(\mathbf{X}) = \mathrm{var}[\mathrm{E}(\mathbf{X}|Y)] + \mathrm{E}[\mathrm{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\mathrm{var}[\mathrm{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\mathrm{E}[\mathrm{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\mathrm{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

$$\text{var}(\mathbf{X}) = \text{var}[\mathrm{E}(\mathbf{X}|Y)] + \mathrm{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\mathrm{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\mathrm{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

$$\text{var}(\mathbf{X}) = \text{var}[\mathrm{E}(\mathbf{X}|Y)] + \mathrm{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\mathrm{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\mathrm{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

$$\mathrm{var}(\mathbf{X}) = \mathrm{var}[\mathrm{E}(\mathbf{X}|Y)] + \mathrm{E}[\mathrm{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\mathrm{var}[\mathrm{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\mathrm{E}[\mathrm{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\mathrm{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

$$\text{var}(\mathbf{X}) = \text{var}[\mathrm{E}(\mathbf{X}|Y)] + \mathrm{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\mathrm{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\mathrm{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Linear Sufficient Dimension Reduction

- Linear Sufficient Dimension Reduction (SDR) finds $\boldsymbol{\alpha}$ with

$$F(Y|\mathbf{X}) = F(Y|\boldsymbol{\alpha}'\mathbf{X})$$

- That is, **SDR** targets Y to find

$$\boldsymbol{\alpha}'\mathbf{X} = (\boldsymbol{\alpha}_1'\mathbf{X}, \ldots, \boldsymbol{\alpha}_d'\mathbf{X})$$

  that can replace $\mathbf{X}$ in the regression of Y on $\mathbf{X}$

- Sliced Inverse Regression (SIR) (Li 1991) is one such method

Linear Sufficient Dimension Reduction

- Linear Sufficient Dimension Reduction (SDR) finds $\boldsymbol{\alpha}$ with

$$F(Y|\mathbf{X}) = F(Y|\boldsymbol{\alpha}'\mathbf{X})$$

- That is, **SDR** targets Y to find

$$\boldsymbol{\alpha}'\mathbf{X} = (\boldsymbol{\alpha}_1'\mathbf{X}, \ldots, \boldsymbol{\alpha}_d'\mathbf{X})$$

  that can replace $\mathbf{X}$ in the regression of Y on $\mathbf{X}$

- Sliced Inverse Regression (SIR) (Li 1991) is one such method

Linear Sufficient Dimension Reduction

- Linear Sufficient Dimension Reduction (SDR) finds $\boldsymbol{\alpha}$ with

$$F(Y|\mathbf{X}) = F(Y|\boldsymbol{\alpha}'\mathbf{X})$$

- That is, **SDR** targets Y to find

$$\boldsymbol{\alpha}'\mathbf{X} = (\boldsymbol{\alpha}_1'\mathbf{X}, \ldots, \boldsymbol{\alpha}_d'\mathbf{X})$$

  that can replace $\mathbf{X}$ in the regression of Y on $\mathbf{X}$

- Sliced Inverse Regression (SIR) (Li 1991) is one such method

SIR

How to find the reduction $\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}'\mathbf{X}$ with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_d) : p \times d$, $d < p$?

- Li (1991) was the first to observe that if $\mathrm{E}(\mathbf{X}|\boldsymbol{\alpha}'\mathbf{X}) = \mathbf{A}(\boldsymbol{\alpha}'\mathbf{X})$ and $\boldsymbol{\Sigma}_x = \mathsf{var}(\mathbf{X})$, then

$$\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X}) \in \boldsymbol{\Sigma}_x \operatorname{span}(\boldsymbol{\alpha})$$

$$\boldsymbol{\Sigma}_x^{-1} \left(\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X})\right) \in \operatorname{span}(\boldsymbol{\alpha})$$

where $\operatorname{span}(\boldsymbol{\alpha})$ is the column space of $\boldsymbol{\alpha}$

How to find $\boldsymbol{\alpha}$?

Eaton 1983: If $\mathbf{Z}$ is a random vector, then $\mathbf{Z} \in \mathrm{E}(\mathbf{Z}) + \mathrm{span}(\boldsymbol{\Sigma}_z)$
Let $\mathbf{Z} = \mathrm{E}(\mathbf{X}|Y)$ to obtain,

$$\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X}) \in \mathrm{span}((\mathsf{var}(\mathrm{E}(\mathbf{X}|Y))$$

Therefore,

$$\boldsymbol{\Sigma}_x^{-1}\mathrm{span}(\mathsf{var}(\mathrm{E}(\mathbf{X}|Y))) \subset \mathrm{span}(\boldsymbol{\alpha})$$

To estimate α we need

- an estimate of Σ_x^{-1}
- an estimate of $\mathsf{var}(\mathrm{E}(\mathbf{X}|Y))$ and its rank

How to find $\boldsymbol{\alpha}$?

Eaton 1983: If $\mathbf{Z}$ is a random vector, then $\mathbf{Z} \in \mathrm{E}(\mathbf{Z}) + \mathrm{span}(\boldsymbol{\Sigma}_z)$
Let $\mathbf{Z} = \mathrm{E}(\mathbf{X}|Y)$ to obtain,

$$\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X}) \in \mathrm{span}((\mathsf{var}(\mathrm{E}(\mathbf{X}|Y))$$

Therefore,

$$\boldsymbol{\Sigma}_x^{-1}\mathrm{span}(\mathsf{var}(\mathrm{E}(\mathbf{X}|Y))) \subset \mathrm{span}(\boldsymbol{\alpha})$$

To estimate $\boldsymbol{\alpha}$ we need

- an estimate of $\boldsymbol{\Sigma}_x^{-1}$
- an estimate of $\mathsf{var}(\mathrm{E}(\mathbf{X}|Y))$ and its rank

Linear SDR in general

- How can we identify $\mathrm{span}(\boldsymbol{\alpha})$?
- General Idea: find a kernel matrix $\mathbf{M}$ so that

$$\mathcal{S}(\mathbf{M}) \subset \mathrm{span}(\boldsymbol{\alpha})$$

- SDR methods: different proposals for $\mathbf{M}$
- For example: In SIR, $\mathbf{M} = \mathrm{cov}(\mathrm{E}(\mathbf{X}|Y))$

Linear SDR in general

- How can we identify $\mathrm{span}(\boldsymbol{\alpha})$?
- General Idea: find a kernel matrix $\mathbf{M}$ so that

$$S(\mathbf{M}) \subset \mathrm{span}(\boldsymbol{\alpha})$$

- SDR methods: different proposals for $\mathbf{M}$
- For example: In SIR, $\mathbf{M} = \mathrm{cov}(\mathrm{E}(\mathbf{X}|Y))$

Linear SDR in general

- How can we identify $\mathrm{span}(\boldsymbol{\alpha})$?
- General Idea: find a kernel matrix $\mathbf{M}$ so that

$$S(\mathbf{M}) \subset \mathrm{span}(\boldsymbol{\alpha})$$

- SDR methods: different proposals for $\mathbf{M}$
- For example: In SIR, $\mathbf{M} = \mathrm{cov}(\mathrm{E}(\mathbf{X}|Y))$

Linear SDR in general

- How can we identify $\mathrm{span}(\boldsymbol{\alpha})$?
- General Idea: find a kernel matrix $\mathbf{M}$ so that

$$S(\mathbf{M}) \subset \mathrm{span}(\boldsymbol{\alpha})$$

- SDR methods: different proposals for $\mathbf{M}$
- For example: In SIR, $\mathbf{M} = \mathrm{cov}(\mathrm{E}(\mathbf{X}|Y))$

Moment Based Linear SDR

Under $F(Y|\mathbf{X}) = F(Y|\boldsymbol{\alpha}^T\mathbf{X})$

- Linearity condition: $\mathrm{E}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X})$ is linear in $\boldsymbol{\alpha}^T\mathbf{X}$

$$\boldsymbol{\Sigma}^{-1}(\mathrm{cov}(\mathbf{X}|Y)) \subseteq \mathrm{span}(\boldsymbol{\alpha})$$

- Linearity condition and constant variance $\mathrm{cov}(\mathbf{X}|\boldsymbol{\alpha}^T\mathbf{X})$

$$\mathrm{span}(\boldsymbol{\Sigma}^{-1}(\mathrm{cov}(\mathbf{X}|Y) - \boldsymbol{\Sigma})) \subseteq \mathrm{span}(\boldsymbol{\alpha})$$

Moment Based Linear SDR

- SIR, PIR, PFC, MAVE, etc: First Moment of $\mathbf{X}|Y$
- SAVE, SIRII, pHd, DR, etc: First and Second Moment $\mathbf{X}|Y$
- Most existing Linear SDR methods are based on moments of $\mathbf{X}|Y$ and very often are not exhaustive

Sufficient Dimension Reduction in General

- Let $\mathbf{R}: \mathbb{R}^p \rightarrow \mathbb{R}^d$ with $d \leq p = \dim(\mathbf{X})$, such that

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$$

- $\mathbf{R}(\mathbf{X})$ is a sufficient reduction for the regression of Y on $\mathbf{X}$: no information about Y is lost when $\mathbf{X}$ is replaced by $\mathbf{R}(\mathbf{X})$
- The reduction in the complexity of the regression is
  - driven by Y
  - not restricted by the forward regression function, no model for the response is needed
  - exhaustive

Sufficient Dimension Reduction in General

- Let $\mathbf{R}: \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p = \dim(\mathbf{X})$, such that

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$$

- $\mathbf{R}(\mathbf{X})$ is a sufficient reduction for the regression of Y on $\mathbf{X}$: no information about Y is lost when $\mathbf{X}$ is replaced by $\mathbf{R}(\mathbf{X})$
- The reduction in the complexity of the regression is
  - driven by Y
  - not restricted by the forward regression function, no model for the response is needed
  - exhaustive

Sufficient Dimension Reduction in General

- Let $\mathbf{R}: \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p = \dim(\mathbf{X})$, such that

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$$

- $\mathbf{R}(\mathbf{X})$ is a sufficient reduction for the regression of Y on $\mathbf{X}$: no information about Y is lost when $\mathbf{X}$ is replaced by $\mathbf{R}(\mathbf{X})$
- The reduction in the complexity of the regression is
    - driven by Y
    - not restricted by the forward regression function, no model for the response is needed
    - exhaustive

Sufficient Dimension Reduction in General

- Let $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p = \dim(\mathbf{X})$, such that

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$$

- $\mathbf{R}(\mathbf{X})$ is a sufficient reduction for the regression of Y on $\mathbf{X}$: no information about Y is lost when $\mathbf{X}$ is replaced by $\mathbf{R}(\mathbf{X})$
- The reduction in the complexity of the regression is
  - driven by Y
  - not restricted by the forward regression function, no model for the response is needed
  - exhaustive

Sufficient Dimension Reduction in General

- Let $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p = \dim(\mathbf{X})$, such that

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$$

- $\mathbf{R}(\mathbf{X})$ is a sufficient reduction for the regression of Y on $\mathbf{X}$: no information about Y is lost when $\mathbf{X}$ is replaced by $\mathbf{R}(\mathbf{X})$
- The reduction in the complexity of the regression is
  - driven by Y
  - not restricted by the forward regression function, no model for the response is needed
  - exhaustive

Sufficient Dimension Reduction in General

- Let $\mathbf{R} : \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p = \dim(\mathbf{X})$, such that

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$$

- $\mathbf{R}(\mathbf{X})$ is a sufficient reduction for the regression of Y on $\mathbf{X}$: no information about Y is lost when $\mathbf{X}$ is replaced by $\mathbf{R}(\mathbf{X})$
- The reduction in the complexity of the regression is
  - driven by Y
  - not restricted by the forward regression function, no model for the response is needed
  - exhaustive

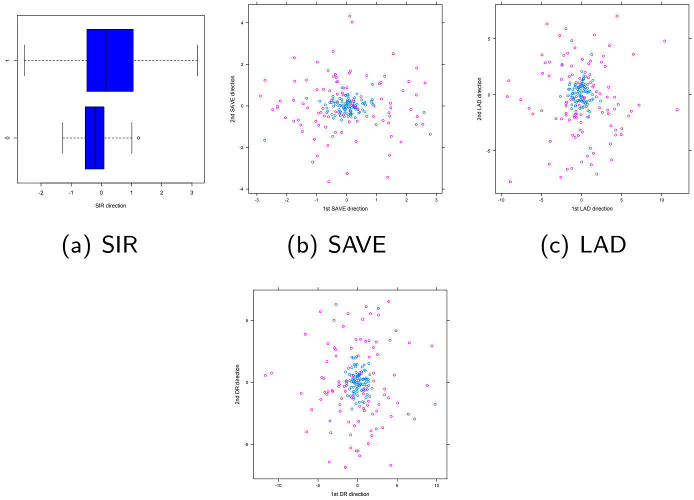
Is Moment-Based SDR all that is needed?

- Y Bernoulli with $P(Y=1)=P(Y=0)=.5$ and

$$(\mathbf{X}|Y=y)\sim N_p\left(\mathbf{0},\sigma^2(y)\mathbf{I}_p=c_y\mathbf{\Delta}\right)$$

  where $\sigma^2(0)=1$ and $\sigma^2(1)=10$.
- Let $(\mathbf{X}_1,Y_1),\dots,(\mathbf{X}_n,Y_n)$ be a sample from this model, where $n=200$ and $p=10$.
- Let's apply linear dimension reduction methods such as SIR, SAVE, DR and LAD to examine how well they perform in discriminating the two populations

Linear SDR methods



(a) SIR

(b) SAVE

(c) LAD

(d) DR

Is Moment-Based SDR good enough?

- All perform badly.
- *But* $\mathbf{X}$ does contain all discriminatory information!
- What are we missing?

Is Moment-Based SDR good enough?

- All perform badly.
- *But* $\mathbf{X}$ does contain all discriminatory information!
- What are we missing?

Is Moment-Based SDR good enough?

- All perform badly.
- *But* $\mathbf{X}$ does contain all discriminatory information!
- What are we missing?

Why Linear SDR is not enough?

- Let's consider $(Y, \mathbf{X})$ jointly normal. Then

$$Y|\mathbf{X} \sim N(\mu_y + \Sigma_{yx}\boldsymbol{\Sigma}_x^{-1}(\mathbf{X} - \boldsymbol{\mu}_x), \sigma_y^2 - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{yx}')$$

  - $\mathrm{E}(Y|\mathbf{X}) = \mu_y + \Sigma_{yx}\boldsymbol{\Sigma}_x^{-1}(\mathbf{X} - \boldsymbol{\mu}_x) = \boldsymbol{\alpha}'(\mathbf{X} - \boldsymbol{\mu}_x)$, and $\mathrm{var}(Y|\mathbf{X})$ is constant
  - Sufficient reduction is the scalar $\boldsymbol{\alpha}'\mathbf{X}$
  - Both OLS and SIR estimate the vector $\boldsymbol{\alpha}$
  - SIR always recovers the OLS predictor
- When $(Y, \mathbf{X})$ not jointly normal but $\mathbf{X}|Y \sim N_p(\boldsymbol{\mu}_Y, \boldsymbol{\Delta})$, then

$$F(Y|\mathbf{X}) = F(Y|\boldsymbol{\alpha}'\mathbf{X})$$

  and SIR (but not OLS) recovers the minimal sufficient reduction $\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}'\mathbf{X} = (\boldsymbol{\alpha}_1'\mathbf{X}, \ldots, \boldsymbol{\alpha}_d'\mathbf{X})$

Non-linear SDR

Bura and Forzani (JASA 2015) introduced Non-linear Reductions in SDR:

- When $\mathbf{X}|Y \sim N_p(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}_Y)$, then

$$\mathbf{R}(\mathbf{X}) = \left(\boldsymbol{\alpha}'(\mathbf{X}-\boldsymbol{\mu}_X), (\mathbf{X}-\boldsymbol{\mu}_X)'\boldsymbol{\Sigma}_x^{-1}(\mathbf{X}-\boldsymbol{\mu}_X)\right)$$

- The minimal sufficient reduction has a non-linear component
- Same is true for elliptically contoured (heavy tailed distns): $\mathbf{X}|Y \sim EC_p(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}, g_Y)$ (Bura and Forzani 2015)
- SIR and all other model-free SDR methods cannot recover the non-linear component and cannot be exhaustive!

Back to our simulation example

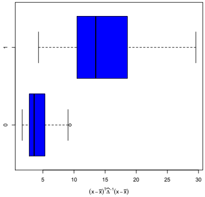
- Y Bernoulli with $P(Y=1)=P(Y=0)=.5$ and

$$(\mathbf{X}|Y=y)\sim N_p\left(\mathbf{0},\sigma^2(y)\mathbf{I}_p=c_y\boldsymbol{\Delta}\right)$$

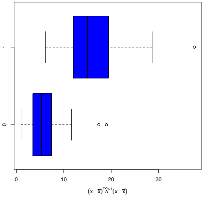
  where $\sigma^2(0)=1$ and $\sigma^2(1)=10$.

- Let $(\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)$ be a sample from this model, where $n=200$ and $p=10$.

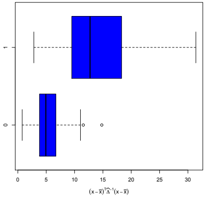
Non-linear SDR



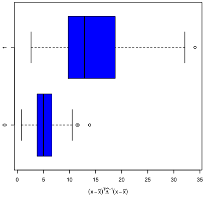
(e) Initial $\sigma^2(y), n =$ 100



(f) $8th$ iteration $\sigma^2(y), n = 100$



(g) Initial $\sigma^2(y), n =$



(h) $8th$ iteration

Motivation for Non-linear SDR

We want to find the minimal sufficient reduction $\mathbf{R}(\mathbf{X})$ so that $F(\mathbf{Y}|\mathbf{X}) = F(\mathbf{Y}|\mathbf{R}(\mathbf{X}))$

$$\textit{Fact}: \quad F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}) \quad \text{iff} \quad \mathbf{X}|(\mathbf{R}(\mathbf{X}), Y) \overset{d}{=} \mathbf{X}|\mathbf{R}(\mathbf{X})$$

- Think of $\mathbf{Y}$ as a parameter and consider the distribution $\mathbf{X}|\mathbf{Y}$
- Find the sufficient "statistic" $\mathbf{R}(\mathbf{X})$ for the "parameter" $\mathbf{Y}$
- $\mathbf{R}(\mathbf{X})$ is the function of $\mathbf{X}$ you need to characterize $\mathbf{Y}$: If $\mathbf{R}(\mathbf{X})$ is a sufficient statistic for the inverse regression $\mathbf{X}|Y$ then it is a sufficient reduction for the forward regression $Y|\mathbf{X}$.

Motivation for Non-linear SDR

We want to find the minimal sufficient reduction $\mathbf{R}(\mathbf{X})$ so that $F(\mathbf{Y}|\mathbf{X}) = F(\mathbf{Y}|\mathbf{R}(\mathbf{X}))$

$$\textit{Fact}: \quad F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}) \quad \text{iff} \quad \mathbf{X}|(\mathbf{R}(\mathbf{X}), Y) \overset{d}{=} \mathbf{X}|\mathbf{R}(\mathbf{X})$$

- Think of $\mathbf{Y}$ as a parameter and consider the distribution $\mathbf{X}|\mathbf{Y}$
- Find the sufficient "statistic" $\mathbf{R}(\mathbf{X})$ for the "parameter" $\mathbf{Y}$
- $\mathbf{R}(\mathbf{X})$ is the function of $\mathbf{X}$ you need to characterize $\mathbf{Y}$: If $\mathbf{R}(\mathbf{X})$ is a sufficient statistic for the inverse regression $\mathbf{X}|Y$ then it is a sufficient reduction for the forward regression $Y|\mathbf{X}$.

Motivation for Non-linear SDR

We want to find the minimal sufficient reduction $\mathbf{R}(\mathbf{X})$ so that $F(\mathbf{Y}|\mathbf{X}) = F(\mathbf{Y}|\mathbf{R}(\mathbf{X}))$

$$\textit{Fact}: \quad F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}) \quad \text{iff} \quad \mathbf{X}|(\mathbf{R}(\mathbf{X}), Y) \overset{d}{=} \mathbf{X}|\mathbf{R}(\mathbf{X})$$

- Think of $\mathbf{Y}$ as a parameter and consider the distribution $\mathbf{X}|\mathbf{Y}$
- Find the sufficient “statistic” $\mathbf{R}(\mathbf{X})$ for the “parameter” $\mathbf{Y}$
- $\mathbf{R}(\mathbf{X})$ is the function of $\mathbf{X}$ you need to characterize $\mathbf{Y}$: If $\mathbf{R}(\mathbf{X})$ is a sufficient statistic for the inverse regression $\mathbf{X}|Y$ then it is a sufficient reduction for the forward regression $Y|\mathbf{X}$.

Model-based SDR

- Assuming a model for $\mathbf{X}|Y$ removes the need for other assumptions
  - Reductions are exhaustive and not necessarily linear
  - We can derive their functional forms and how to estimate them
  - Different from RKHS-based methods (KSIR, GSIR, KDR, KCCA, etc)
- **Tool:** Classic Sufficient Statistics Theory, e.g. Fisher's factorization theorem–if

$$f(\mathbf{x}|y) = h(\mathbf{x})g(\mathbf{T}(\mathbf{x}), y)$$

then $\mathbf{T}(\mathbf{x})$ is sufficient for Y

$\mathbf{X}|Y$ in the Exponential Family

Regressions with all continuous, all categorical, or mixtures of continuous and categorical predictors

- Natural family of distributions for $\mathbf{X}|Y$ is the exponential:

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x})$$

- Multivariate normal, gamma, exponential, Bernoulli, Poisson, Dirichlet, multinomial, etc.
- The natural "parameters": $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yk})^T$, $k \geq p$.
- $\mathbf{T}(\mathbf{x})$ is the *minimal sufficient statistic* for Y

$\mathbf{X}|Y$ in the Exponential Family

Regressions with all continuous, all categorical, or mixtures of continuous and categorical predictors

- Natural family of distributions for $\mathbf{X}|Y$ is the exponential:

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x})$$

- Multivariate normal, gamma, exponential, Bernoulli, Poisson, Dirichlet, multinomial, etc.
- The natural "parameters": $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yk})^T$, $k \geq p$.
- $\mathbf{T}(\mathbf{x})$ is the *minimal sufficient statistic* for Y

$\mathbf{X}|Y$ in the Exponential Family

Regressions with all continuous, all categorical, or mixtures of continuous and categorical predictors

- Natural family of distributions for $\mathbf{X}|Y$ is the exponential:

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x})$$

- Multivariate normal, gamma, exponential, Bernoulli, Poisson, Dirichlet, multinomial, etc.
- The natural "parameters": $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yk})^T$, $k \geq p$.
- $\mathbf{T}(\mathbf{x})$ is the *minimal sufficient statistic* for Y

$\mathbf{X}|Y$ in the Exponential Family

Regressions with all continuous, all categorical, or mixtures of continuous and categorical predictors

- Natural family of distributions for $\mathbf{X}|Y$ is the exponential:

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x})$$

- Multivariate normal, gamma, exponential, Bernoulli, Poisson, Dirichlet, multinomial, etc.
- The natural "parameters": $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yk})^T$, $k \geq p$.
- $\mathbf{T}(\mathbf{x})$ is the *minimal sufficient statistic* for Y

$\mathbf{X}|Y$ in the Exponential Family

- **Minimal Sufficient Reduction:**

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{X}) - \mathrm{E}(\mathbf{T}(\mathbf{X}))$$

with

$$\boldsymbol{\alpha} = \mathrm{span}\{(\boldsymbol{\eta}_Y - \mathrm{E}_Y(\boldsymbol{\eta}_Y) = (\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}), Y \in \mathcal{S}_Y\}$$

- To estimate $\mathbf{R}(\mathbf{X})$ we need to estimate the natural *parameters* $\boldsymbol{\eta}_Y$

(Bura, Duarte and Forzani 2015, to appear in JASA)

Estimation via Generalized Linear Models

$\mathbf{D} : k \times r$, and $\mathbf{f}_Y \in \mathbb{R}^r$ known functions of Y, so that

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}})$$
$$\operatorname{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}) = \operatorname{span}(\mathbf{D})$$

- Obtain estimates of $\bar{\boldsymbol{\eta}}$ and $\mathbf{D}$ via Iterative Reweighted Least Squares (IRLS estimates are MLEs)
- $\widehat{d}$, the estimate of the rank d of $\mathbf{D}$, is obtained with asymptotic tests or BIC/AIC
- The first $\widehat{d}$ eigenvectors of $\widehat{\mathbf{D}}$, $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d$, yield the MLE of

$$\widehat{\mathbf{R}}(\mathbf{X}) = \widehat{\alpha}'(\mathbf{T}(\mathbf{X}) - \bar{\mathbf{T}}(\mathbf{X}))$$

- Regress Y on $\widehat{\mathbf{R}}(\mathbf{X})$

Estimation via Generalized Linear Models

$\mathbf{D} : k \times r$, and $\mathbf{f}_Y \in \mathbb{R}^r$ known functions of Y, so that

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}})$$
$$\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}) = \mathrm{span}(\mathbf{D})$$

- Obtain estimates of $\bar{\boldsymbol{\eta}}$ and $\mathbf{D}$ via Iterative Reweighted Least Squares (IRLS estimates are MLEs)
- $\widehat{d}$, the estimate of the rank d of $\mathbf{D}$, is obtained with asymptotic tests or BIC/AIC
- The first $\widehat{d}$ eigenvectors of $\widehat{\mathbf{D}}$, $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d$, yield the MLE of

$$\widehat{\mathbf{R}}(\mathbf{X}) = \widehat{\alpha}'(\mathbf{T}(\mathbf{X}) - \bar{\mathbf{T}}(\mathbf{X}))$$

- Regress Y on $\widehat{\mathbf{R}}(\mathbf{X})$

Estimation via Generalized Linear Models

$\mathbf{D} : k \times r$, and $\mathbf{f}_Y \in \mathbb{R}^r$ known functions of Y, so that

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}})$$
$$\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}) = \mathrm{span}(\mathbf{D})$$

- Obtain estimates of $\bar{\boldsymbol{\eta}}$ and $\mathbf{D}$ via Iterative Reweighted Least Squares (IRLS estimates are MLEs)
- $\widehat{d}$, the estimate of the rank d of $\mathbf{D}$, is obtained with asymptotic tests or BIC/AIC
- The first $\widehat{d}$ eigenvectors of $\widehat{\mathbf{D}}$, $\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_d$, yield the MLE of

$$\widehat{\mathbf{R}}(\mathbf{X}) = \widehat{\boldsymbol{\alpha}}'(\mathbf{T}(\mathbf{X}) - \bar{\mathbf{T}}(\mathbf{X}))$$

- Regress Y on $\widehat{\mathbf{R}}(\mathbf{X})$

Estimation via Generalized Linear Models

$\mathbf{D} : k \times r$, and $\mathbf{f}_Y \in \mathbb{R}^r$ known functions of Y, so that

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}})$$
$$\operatorname{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}) = \operatorname{span}(\mathbf{D})$$

- Obtain estimates of $\bar{\boldsymbol{\eta}}$ and $\mathbf{D}$ via Iterative Reweighted Least Squares (IRLS estimates are MLEs)
- $\widehat{d}$, the estimate of the rank d of $\mathbf{D}$, is obtained with asymptotic tests or BIC/AIC
- The first $\widehat{d}$ eigenvectors of $\widehat{\mathbf{D}}$, $\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_d$, yield the MLE of

$$\widehat{\mathbf{R}}(\mathbf{X}) = \widehat{\boldsymbol{\alpha}}'(\mathbf{T}(\mathbf{X}) - \bar{\mathbf{T}}(\mathbf{X}))$$

- Regress Y on $\widehat{\mathbf{R}}(\mathbf{X})$

Application: Multivariate Bernoulli Predictors

- In many data sets, predictors of a target variable are binary or categorical
- Examples include gene association studies, image processing, natural language processing, social networks, spatial statistics
- The multivariate Bernoulli distribution models potentially dependent binary variables; a member of the exponential family
- Regressions/classifications with multivariate Bernoulli predictors are extensively used in machine learning/data mining

Ising Model

- The Ising model is an undirected graphical model that allows up to pairwise interaction effects and it has been extensively used to model multivariate binary data
- The Ising probability function belongs to the exponential family with natural parameter vector $\boldsymbol{\eta}_y = (\theta_{11}(y), \theta_{22}(y), \dots, \theta_{pp}(y), \theta_{12}(y), \dots, \theta_{p-1,p}(y))^T$, and sufficient statistic

$$\mathbf{T}(\mathbf{X}) = (X_1, \dots, X_p, X_1X_2, \dots, X_1X_p, \dots, X_{p-1}X_p)^T$$

both with $p + p(p-1)/2$ elements

Ising Distribution

Binary inverse predictors $X_1|Y, \ldots, X_p|Y$, with $X_j|Y \in \{1, 0\}, 1 \le j \le p$, with density function,

$$p(x_1, \ldots, x_p|Y = y) = \frac{1}{Z(\boldsymbol{\Theta}(y))} \exp\left(\sum_{j=1}^{p} \theta_{jj}(y)x_j + \sum_{1 \le j \le j' \le p} \theta_{jj'}(y)x_j x_{j'}\right)$$

where $\boldsymbol{\Theta}(y) = (\theta_{jj'}(y))_{p \times p}$ is a symmetric matrix specifying the network structure. The partition function $Z(\theta(y))$ ensures that the density function is proper

Natural Parameters

- $\theta_{jj}(y)$: corresponds to the main effect for variable $X_j|Y$
- $\theta_{jj'}(y)$: corresponds to the interaction effect between variables $X_j|Y$ and $X_{j'}|Y$

$$\theta_{jj'}(y) = \log \frac{P(X_j = 1, X_{j'} = 1|\mathbf{X}_{-j,-j'}, Y)P(X_j = 0, X_{j'} = 0|\mathbf{X}_{-j,-j'}, Y)}{P(X_j = 1, X_{j'} = 0|\mathbf{X}_{-j,-j'}, Y)P(X_j = 0, X_{j'} = 1|\mathbf{X}_{-j,-j'}, Y)}$$

- X_j and $X_{j'}$ are conditionally independent given Y and all other $\mathbf{X}$-variables if and only if $\theta_{jj'}(y) = 0$

Classification: Zoo data

- 101 animals classified into 7 categories: amphibian, bird, fish, insect, invertebrate, mammal, and reptile
- 15 binary predictors were measured on each animal: hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic and cat-size
- Objective: predict animal category

Estimation via multivariate logistic regression

- $\boldsymbol{\eta}_y = (\theta_{11}(y), \theta_{22}(y), \ldots, \theta_{pp}(y), \theta_{12}(y), \ldots, \theta_{p-1,p}(y))^T$
- Define

$$f_{yk} = I(y = k) - \frac{n_k}{n} \quad \text{for } k = 1, \ldots, r,$$

$r = 7 - 1 = 6$, where 7 is the number of distinct values of the response

- With n samples on Y and $\mathbf{X}$, the multivariate GLM is

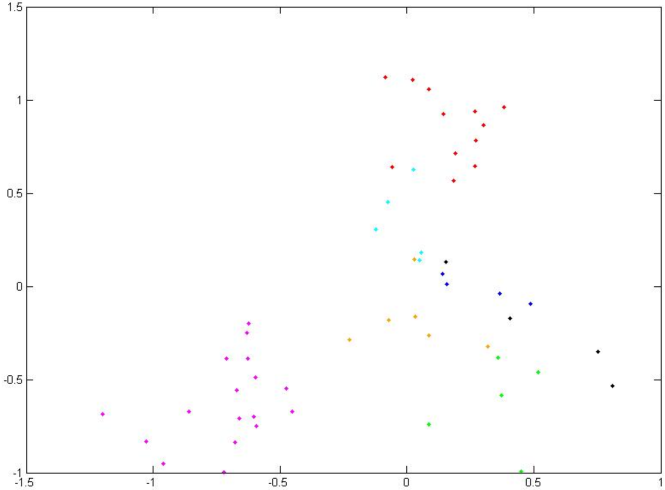
$$\boldsymbol{\eta}_n = \mathbf{F}_n \widetilde{\mathbf{D}}$$

  - $\boldsymbol{\eta}_n = (\theta_{i(jj')})$:$n \times (p + p(p-1)/2)$ random matrix
  - $\mathbf{F}_n = (f_{il}) = (f_{y_i,l})$: $n \times (r+1)$ fixed matrix
  - $\widetilde{\mathbf{D}} = (c_{lj})$, the $(r+1) \times (p + p(p-1)/2$ matrix of coefficients

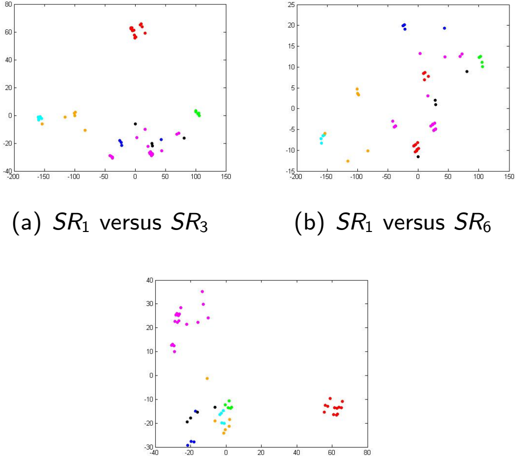
Estimation

- Impossible to fit the full pairwise dependence Ising model: requires the estimation of $7 \times 120 = 840$ parameters with 101 observations
- We assume that the Ising model is sparse; that is, that some natural parameters $\theta_{jj'}(y)$ are zero using the l_1 penalties of Cheng et al. (2012).
- After screening for sparsity, 66 of the 120 terms of $\mathbf{T}(\mathbf{X})$ were retained
- Hair, airborne, predatory, venomous and domestic main effects; rest are interactions

SR_1 vs SR_2 under Dependence – EF-DR



Sufficient Plots under Independence (Cook and Li 2009)



(a) SR_1 versus SR_3

(b) SR_1 versus SR_6

(c) SR_3 versus SR_5

Results

- Our EF-DR Method (Dependence Model)
  - We plot the first two sufficient reductions: $\widehat{\boldsymbol{\alpha}}_1'(\mathbf{T}(\mathbf{X}) - \bar{T}(\mathbf{X}))$ versus $\widehat{\boldsymbol{\alpha}}_2'(\mathbf{T}(\mathbf{X}) - \bar{T}(\mathbf{X}))$
  - Dimension is 2: all colors are separated by simple closed curves
  - Perfect in-sample classification
- In contrast, the competing independence model requires 6 reductions to classify the animals.
- We also applied KDR (a RKHS method) with Gaussian kernel to the Zoo data. Several KDR directions are needed to separate the seven classes.

Classification Accuracy

	EF-DR	KDR
$d = 2$		
LDA	0.139	0.297
dQDA	0.158	0.257
$d = 3$		
LDA	0.119	0.158
dQDA	0.109	0.139

Table: LDA and dQDA misclassification errors

Simulation

- $Y \sim N(0, 0.5)$
- Given Y, $\mathbf{X} = (X_1, \ldots, X_p)^T$ is Bernoulli with pairwise correlation structure among contiguous pairs as follows, (X_1, X_2), $(X_3, X_4), \ldots, (X_{p-1}, X_p)$, and all other interactions of all orders are zero.
- Therefore, $\boldsymbol{\eta}_Y = (\theta_{11}, \theta_{22}, \ldots, \theta_{pp}, \theta_{12}, \theta_{34}, \ldots, \theta_{(p-1)p})$.
- $\mathbf{T}(\mathbf{X}) = (X_1, X_2, \ldots, X_p, X_1X_2, X_3X_4, \ldots, X_{p-1}X_p)$.
- The natural parameter $\boldsymbol{\eta}_Y$ is generated as

  $$\boldsymbol{\eta}_Y = \mathbf{A}C^T(\mathbf{f}_Y - \mathrm{E}\mathbf{f}_Y)$$

  where $C = 1$, $\mathbf{f}_Y = Y$.
- For $p = 4$, $\mathbf{A} = (1, 1, 1, 1, 10, 10)^T/\sqrt{204}$, and for $p > 4$, we set $A_5 = \ldots = A_p = 0$, so that, for all p, the minimal sufficient reduction is $[(X_1 - \mathrm{E}(X_1)) + (X_2 - \mathrm{E}(X_2)) + (X_3 - \mathrm{E}(X_3)) + (X_4 - \mathrm{E}(X_4)) + 10(X_1X_2 - \mathrm{E}(X_1X_2)) + 10(X_3X_4 - \mathrm{E}(X_3X_4))]/\sqrt{204}$.

- Three models:
  (a) assuming the true correlation structure;
  (b) assuming the $\mathbf{X}$ components are independent given Y, which is Cook and Li's (2009) approach; and
  (c) assuming that all pairwise interactions are present as in the full Ising model.
- Accuracy: the maximum angle between the true subspace spanned by $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}$, $Y \in \mathcal{S}_Y$, and the estimated one.

	$n = 100$	$n = 200$	$n = 300$	$n = 500$	N
$p = 4$					
(a)	35.12 (16.65)	26.29 (12.14)	21.94 (8.50)	17.49 (8.09)	100
(b)	43.69 (16.30)	39.29 (12.52)	37.64 (10.49)	37.19 (7.33)	100
(c)	45.81 (15.47)	41.12 (13.78)	30.95 (13.62)	29.56 (12.24)	100
$p = 6$					
(a)	40.42 (12.10)	34.82 (13.15)	28.76 (16.96)	23.31 (13.24)	50
(b)	64.24 (15.84)	63.16 (23.91)	58.08 (25.59)	55.42 (30.12)	50
(c)	51.06 (14.56)	42.51 (13.25)	38.21 (13.78)	36.80 (12.89)	50
$p = 10$					
(a)	50.38 (10.78)	40.43 (10.27)	33.97 (10.47)	28.84 (9.16)	50
(b)	75.23 (10.86)	72.16 (11.93)	73.14 (12.11)	74.45 (9.90)	50
(c)	57.12 (13.41)	55.63 (12.71)	50.22 (12.03)	47.01 (10.82)	50

Table: Mean angles and their standard deviations in parentheses between the true and estimated subspaces

Features of EF-DR

The EF-DR sufficient reductions

(1) are exhaustive,

(2) are linear functions of the sufficient statistics, which can be both linear and nonlinear functions of the predictors,

(3) have explicit functional forms,

(4) their estimates are MLEs and hence efficient.

EF-DR also applies when the response and the predictors have a joint exponential family distribution and the response is a vector

Future Directions

- Functional SDR
    - predictors and/or response are curves
- SDR in Macro-forecasting

References

- Bura, E. and Cook, R. D. (2001a). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society* B **63**, 393-410.
- Bura, E. and Cook, R. D. (2001b). Extending SIR: The Weighted Chi-Square Test, *Journal of the American Statistical Association* **96**, 996–1003.
- Bura, E. and Pfeiffer, R. (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statistics and Probability Letters*, **58**, 2275-2280.
- Bura, E. and Yang, J. (2011). Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach. *Journal of Multivariate Analysis*, **102**, 130-142.
- Bura, E. and Forzani, J. (2015). Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, **110** (509), 420-434.
- Bura, E., Duarte, S. L. and Forzani, J. (2016). Sufficient reductions in regressions with inverse predictors in the exponential family. *Journal of the American Statistical Association*, **111**, 1-17.
- Cheng, J., Levina, E., Wang, P., and Zhu, J. (2012). Sparse Ising Models with Covariates, preprint (http://arxiv.org/pdf/1209.6342v1.pdf).

- Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression, *Statistical Science*, 22, 1-26.
- Cook, R.D. and Forzani, L. (2008). Principal Fitted Components for Dimension Reduction in Regression. *Statistical Science*, **23**, 485-501.
- Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004). Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces, *Journal of Machine Learning Research*, **5**, 73-99.
- Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–342.
- Pfeiffer, R., Forzani, L. and Bura, E. (2012). Sufficient Dimension Reduction for Longitudinally Measured Predictors. *Statistics in Medicine, Special Issue: Biomarker Working Group: Issues in the Design and Analysis of Epidemiological Studies with Biomarkers*, **31(22)**, 2414-2427.
- Tomassi, D., Forzani, L., Bura, E. and Pfeiffer, R. (2016). Sufficient reductions when the predictors have detection limits. To appear in *Biometrics.*
- Yee, T.W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models, *Statistical Modelling*, **3**, 15-41.