SDR in High Dimensional Regressions

R.D. Cook
School of Statistics
University of Minnesota

ICSA 2010

Recent work in collaboration with Liliana Forzani
Sufficient Dimension Reduction for Regression

Variables: $Y \in \mathbb{R}^1$, $X \in \mathbb{R}^p$, $(Y, X) \sim F$

Data: $(Y_i, X_i)$ iid $F$, $i = 1, \ldots, n$.

Goal: Reduce $\dim(X)$ without loss of information on $Y|X$.

Sufficient Reduction: $R : \mathbb{R}^p \rightarrow \mathbb{R}^q$, $q \leq p$,

1. $X|(Y, R(X)) \sim X|R(X)$ (Inverse regression)
2. $Y|X \sim Y|R(X)$ (Forward regression)
3. $Y \perpshuffle X|R(X)$ (Joint reduction)

Bijective trans. of $R$ are also sufficient.
**Linear reduction:** restrict $R(X) = \alpha^T X$, $\alpha \in \mathbb{R}^{p \times q}$, so $X|(Y, \alpha^T X) \sim X|\alpha^T X$. $\text{span}(\alpha)$ is called a dimension reduction subspace.

**Goal:** Exhaustive estimation of the central subspace $S_{Y|X} = \cap \text{span}(\alpha) \subseteq \mathbb{R}^p$.

**Methods:** SIR, SAVE, PHD, IHT, PFC, Contour reg., Directional reg., Likelihood-based methods, . . . , nearly all pursue estimation of $S_{Y|X}$ based on traditional asymptotic reasoning with $p$ fixed and $n \to \infty$. 
Ultimate goal is Prediction: \( X | Y \rightarrow Y | X \).

- Parametric model based on \( \hat{R} \).
- Direct inversion (Adragni & Cook ‘09)

\[
\hat{E}(Y | X_{\text{new}} = x) = \frac{\sum_{i=1}^{n} y_i \hat{f}\{R(x) | Y = y_i\}}{\sum_{i=1}^{n} \hat{f}\{R(x) | Y = y_i\}}
\]

**Issue for today**: Estimation of \( R(X) \) when \( n = o(p) \) or \( n \approx p \).
PFC Model

\[ \mathbf{X} | (Y = y) \sim \mathbf{\mu} + \mathbf{\Gamma} \mathbf{\beta} f(y) + \mathbf{N}_p(0, \Delta) \]

- \( \mathbf{\mu} \in \mathbb{R}^p, \mathbf{\Gamma} \in \mathbb{R}^{p \times d}, d < p, f(y) \in \mathbb{R}^r, d < r, \Delta > 0. \)

- \( f(y) \in \mathbb{R}^r \) known vector of basis functions, like polynomials, piecewise polynomials, Fourier series, or indicators if the response is categorical.

Graphing predictors vs. response can aid the choice.

- Advantages: Ability to handle \( d > 1 \) linear combinations, model flexibility through choice of \( f \), and avoid curse of dimensionality while perhaps taking advantage of the blessings of dimensionality.

- Normality of errors not crucial.
Sufficient Reduction $R(X)$

$$X|(Y = y) \sim \mu + \Gamma \beta f(y) + N_p(0, \Delta)$$

- $R(X) = (\Gamma^T \Delta^{-1} \Gamma)^{-1} \Gamma^T \Delta^{-1} (X - \mu) \in \mathbb{R}^d$ is a minimal sufficient reduction.

$R(X)$ holds the coordinates of the projection of $X - \mu$ onto $\text{span}(\Gamma)$ in the $\Delta^{-1}$ inner product.

$$E\{R(X)|Y = y\} = \beta f(y)$$

$$\text{var}\{R(X)|Y = y\} = (\Gamma^T \Delta^{-1} \Gamma)^{-1}$$

which is assumed to have a finite limit as $p \to \infty$. 
MLE of $R(X)$

MLE was provided by Cook and Forzani (2009),

$$\hat{R}_{MLE}(X) = (\hat{\Gamma}^T \hat{\Delta}^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}^T \hat{\Delta}^{-1} (X - \bar{X})$$

It is robust to deviations from normality and to the choice of $f$.

For example, consistency requires only that $\text{cov}(f_{\text{true}}(Y), f(Y))$ have full row rank.

But we must have $p \ll n$, from traditional asymptotic arguments.
MDF Estimation (Cook & Ni, 2005)

\((\hat{\mu}, \hat{\beta}, \hat{\Gamma})\) minimize

\[
\text{trace}(X - 1_n \mu^T - F \beta^T \Gamma^T)\hat{W}^{-1}(X - 1_n \mu^T - F \beta^T \Gamma^T)^T,
\]

where the minimization is over \(\mu \in \mathbb{R}^p\), \(\Gamma \in \mathbb{R}^{p \times d}\), \(\beta \in \mathbb{R}^{d \times r}\), \(X \in \mathbb{R}^{n \times p}\) has rows \(X_i^T\) and \(F \in \mathbb{R}^{n \times r}\) has rows \(f^T(y_i)\) with \(1_n^T F = 0\).

\(\hat{W}\) is an “estimator” of \(\Delta\).

\[
\hat{R}(X) = (\hat{\Gamma}^T \hat{W}^{-1} \hat{\Gamma})^{-1} \hat{\Gamma}^T \hat{W}^{-1}(X - \bar{X})
\]

**Goal:** Characterize \(\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(?)\), as \(n, p \to \infty\).
Choices for $\hat{W}$

Let $\hat{\Delta}$ be the residual covariance matrix from the multivariate OLS fit of $X$ on $f$ (requires only $n > r + 4$). Then

- $\hat{W} = \hat{\Delta}$, requires $n > p + r + 4$.
- $\hat{W} = \text{diag}(\hat{\Delta})$.
- $\hat{W} = \text{Bickel-Levina (2008) threshold estimator of } \Delta \text{ applied to } \hat{\Delta}$.
- $\hat{W} = W$, like $W = I_p$
Signal Rate, $h(p)$

Define $1 \leq h(p) \leq p$ so that as $p \to \infty$

$$\frac{\Gamma^T W^{-1} \Gamma}{h(p)} \to G > 0,$$

where $W$ is the population version of $\hat{W}$.

**Abundant signal:** $h(p) \asymp p$

**Modest signal:** $h(p) \asymp p^{2/3}$

**Sparse signal:** $h(p) = O(1)$
Convergence Rate for $\hat{W} = \hat{\Delta}$.

Assume that $n > p + r + 4$, $\lim p/n \in [0, 1)$ and that

$$\kappa = \left( \frac{p}{hn} \right)^{1/2} \to 0$$

as $n, p \to \infty$. Then

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa)$$

- **Abundant signal**, $h \asymp p$, then $O_p(\kappa) = O_p(1/\sqrt{n})$

- **Modest signal**, say $h \asymp p^{2/3}$, then $O_p(\kappa) = O_p \left( \frac{p^{1/6}}{\sqrt{n}} \right)$

- **Sparse signal**, $h = O(1)$, then $O_p(\kappa) = O_p \left( \sqrt{p/n} \right)$
Simulations

Data generation:

\[ X \mid (Y = y) \sim \Gamma y + N_p(0, \Delta) \]

with \( d = 1, \Gamma \sim N(0, 1), Y \sim N(0, 1), \Delta = 50I_p. \)

Fitted model:

\[ X \mid (Y = y) \sim \mu + \Gamma \beta f(y) + N_p(0, \Delta) \]

with \( d = 1, \text{ a four dimensional } f(y), \text{ so } r = 4, \text{ and } \Delta > 0. \)

All results based on averages of 100 replications.
\[ \kappa = \sqrt{\frac{p}{hn}} \]

\[ \text{corr}(R, \hat{R}) \text{ vs. } n \]

\[ p/n = 1/2, \ h = p \]

\[ \text{corr}(R, \hat{R}) \text{ vs. } p \]

\[ n = 160, \ h = p \]
\[ \kappa = \sqrt{\frac{p}{hn}} \]

\[ \text{var}(Y|R(X)) \text{ vs. } p \]

\[ \text{corr}(R, \hat{R}) \text{ vs } h(p)/p \]

\( n = 160, \ p = 80 \)
• **Bet-on-Sparsity**: “Use a procedure that does well in sparse problems, since no procedure does well in dense problems.” (Friedman, et al., 2004)

• **Bet-on-Sparsity vs. Bet-on-Abundance.**

• **Screening to insure abundance**, \( h \simeq p \)

\[
\kappa = \left( \frac{p}{hn} \right)^{1/2}
\]

Summary: Methodology should work quite well whenever \( n > p + r + 4 \) and the signal is abundant or modest. May not work well with sparse signals.
Convergence Rate for $\hat{W} = W$

\[
\rho = W^{-1/2} \Delta W^{-1/2} \\
\nu \sim N_d(0, V) \\
V = \gamma^T \rho \gamma_0 (\gamma_0^T \rho \gamma_0)^{-1} \gamma_0^T \rho \gamma
\]

where $\gamma \in \mathbb{R}^{p \times d}$ is a semi-orthogonal basis matrix for $\text{span}(W^{-1/2} \Gamma)$ and $(\gamma, \gamma_0)$ is an orthogonal matrix. Then, assuming that the diagonals of $\rho$ are bounded,

\[
\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = \frac{\nu}{\sqrt{h}} + O_p(\kappa) + O_p \left( \frac{\|\rho\|_F}{h \sqrt{n}} \right)
\]
• If $W = \Delta$ then

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa)$$

• If $\text{span}(W^{-1/2}\Gamma)$ is a reducing subspace of $\rho$ then $V = 0$ and

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa) + O_p\left(\frac{\|\rho\|_F}{h\sqrt{n}}\right)$$

• If the signal is abundant then

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = O_p\left(\frac{\|\rho\|^{1/2}}{\sqrt{p}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right)$$
Summary: The choice $W = I_p$ should give the best results in
(1) abundant regressions where (2) the diagonal elements of
$\Delta$ are bounded and (3) $\|\Delta\| = o(p)$.
Convergence Rate for $\hat{W} = \text{diag}(\hat{\Delta})$

$$\rho = \text{diag}(\Delta)^{-1/2} \Delta \text{diag}(\Delta)^{-1/2}$$

$$\nu \sim N_d(0, V)$$

$$V = \gamma^T \rho \gamma_0 (\gamma_0^T \rho \gamma_0)^{-1} \gamma_0^T \rho \gamma = 0$$

where $\gamma \in \mathbb{R}^{p \times d}$ is a semi-orthogonal basis matrix for $\text{span}(\hat{W}^{-1/2} \Gamma)$ and $(\gamma, \gamma_0)$ is an orthogonal matrix.

Abundant signal

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = \frac{\nu}{\sqrt{p}} + O_p \left( \frac{\sqrt{\log p}}{\sqrt{n}} \right)$$

$$= O_p \left( \frac{\|\rho\|^{1/2}}{\sqrt{p}} \right) + O_p \left( \frac{\sqrt{\log p}}{\sqrt{n}} \right)$$
Modest signal

If $h = p^a$, $0 < a < 1$, and $\|\rho\|_F = O(p^{(1+a)/2})$ then

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = \frac{\nu}{\sqrt{h}} + O_p(\kappa)$$

$$= O_p \left( \frac{\|\rho\|^{1/2}}{\sqrt{h}} \right) + O_p(\kappa)$$

Summary: The choice $W = \text{diag}(\hat{\Delta})$ should give good results in (1) abundant or modest regressions where (2) $\|\rho\|$ is bounded or increases slowly.

This does not reflect gains in power.
Convergence Rate for $\hat{W} = T(\hat{\Delta})$

$T$ = Bickel-Levina threshold estimator. Let $\Delta = (\delta_{ij})$, and let $\mathcal{U}(c_0(p))$ denote the class of covariance matrices with

1. bounded diagonal elements,
2. $\max_i \sum_{j=1}^{p} J(\delta_{ij} \neq 0) \leq c_0(p)$
3. $\lambda_{\min}(\Delta) \geq \varepsilon > 0$.

Recall $\kappa = (p/hn)^{1/2}$ and define

$$\omega = c_0(p) \frac{\sqrt{\log p}}{\sqrt{n}}$$

Then, assuming that $\Delta \in \mathcal{U}(c_0(p))$ we have

$$\hat{R}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa) + O_p(\omega)$$
Overarching Message

• The estimator with \( \hat{W} = \hat{\Delta}, n > p + r + 4 \), is the strongest, followed by \( \hat{W} = \text{diag}(\hat{\Delta}), n < r + 4 \).

• Any of the estimators can work well in the right situation, particularly if predictor screening is used to yield an abundant regression.

• But further study is needed for \( n < p \) regressions.