Dimension Reduction in Abundant High Dimensional Regressions

Dennis Cook
University of Minnesota

8th Purdue Symposium
June 2012

In collaboration with Liliana Forzani & Adam Rothman,
Broad context

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

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

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

Reductions: Pursue \( R(X) = \alpha^T X : \mathbb{R}^p \to \mathbb{R}^q, q \leq p, \) so that \( Y \perp X|R(X). \)

\( \text{span}(\alpha) \) is called a dimension reduction subspace (DRS) & \( \alpha^T X \) is called a sufficient reduction.
Broad contex, cont.

**Smallest reduction** is characterized by

- $S_{Y|X} = \cap S_{DRS}$; $R(X) = \eta^T X$;
  $\text{span}(\eta) = S_{Y|X} = \text{central subspace}$.

- Can’t really handle $n < p$ yet.

- Chen et al. (2010) pursue variable elimination by estimating rows of $\eta$ to be 0, but still with $p/n \rightarrow 0$.
Today’s context

Estimation of $R(X) = \eta^T X$ when $n, p \to \infty$ with $n = o(p)$ or $n \asymp p$ or $p = o(n)$, where still $\text{span}(\eta) = S_{Y|X}$.

Distinctions:

- Bypass estimation of $S_{Y|X} \subseteq \mathbb{R}^p$ and instead estimate $R(X) \in \mathbb{R}^d$ directly, with $d = \dim(S_{Y|X})$ fixed.
- Emphasize **abundant** regressions, where many predictors contribute information about $Y$.
  - Food Science
  - Chemometrics
  - Biomedical Engineering

Sparsity is not ruled out, but is not required, either.
Today’s context, cont.

- Pursue prediction – $R(X_{\text{new}})$ or $Y | X_{\text{new}}$ – rather than variable selection.
- Use SPICE (Rothman, et al.) to estimate a critical $p \times p$ matrix of weights $W$.

Tasks:

- Reductive context and $R(X)$
- Class of estimators $\hat{R}_W(X)$
- Key structural assumptions
- Main results for $\hat{R}_W(X_{\text{new}}) - R(X_{\text{new}}) = O_p(r(n, p)), r \rightarrow 0$ as $n, p \rightarrow \infty$.
- Illustrations
Inverse regression

\[
X|Y = y_i \sim \mu + \Gamma \beta f(y_i) + \epsilon_i, \ i = 1, \ldots, n.
\]

- \( \mu \in \mathbb{R}^p, \Gamma \in \mathbb{R}^{p \times d}, \beta \in \mathbb{R}^{d \times r}, d < p \& r; d, r \) fixed.
- \( E(\epsilon_i) = 0, \ \text{var}(\epsilon_i) = \Delta > 0, \ \epsilon \perp Y. \)
- \( R(X) = (\Gamma^T \Delta^{-1} \Gamma)^{-1} \Gamma^T \Delta^{-1} (X - \mu) \in \mathbb{R}^d. \)
- \( f(y) \in \mathbb{R}^r \) known vector of basis functions, like piecewise polynomials or indicators if the response is categorical. Can replace \( f \) with an approximation \( g \) without affecting the results if \( \text{rank}\{\text{cov}(f(Y), g(Y))\} = r. \)
Estimation

Let $X \in \mathbb{R}^{n \times p}$ have rows $X_i^T$ and $F \in \mathbb{R}^{n \times r}$ have rows $f^T(y_i)$ with $1_n^T F = 0$. Then choose $(\hat{\mu}, \hat{\beta}, \hat{\Gamma})$ to minimize the Frobenius norm

$$
\|(X - 1_n \mu^T - F \beta^T \Gamma^T) \hat{W}^{1/2}\|_F
$$

over $\mu \in \mathbb{R}^p$, $\Gamma \in \mathbb{R}^{p \times d}$, $\beta \in \mathbb{R}^{d \times r}$.

**Weight matrix:** $\hat{W} \in \mathbb{R}^{p \times p}$ is an “estimator” of $\Delta^{-1}$ with population version $W$.

**Reductions:**

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

$$
R(X) = (\Gamma^T \Delta^{-1} \Gamma)^{-1} \Gamma^T \Delta^{-1}(X - \mu)
$$

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

Dennis Cook | Dimension Reduction in Abundant High Dimensional Regressions
Specific estimators

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

- $\hat{W} = W$, like $W = I_p$ or the ideal case $W = \Delta^{-1}$.
- $\hat{W} = \text{diag}^{-1}(\hat{\Delta})$
- $\hat{W} = \hat{\Delta}^{-1}$, requires $n > p + r + 4$, allowing $n \asymp p$.
- $\hat{W} = \text{SPICE estimator of } \Delta^{-1} \text{ applied to } \hat{\Delta}$
- $\hat{W} = \text{Moore-Penrose inverse } \hat{\Delta}^{-}$ of $\hat{\Delta}$ (simulation only).
Signal rate, $h(p)$

Assume there exists $h(p) = O(p)$ so that as $p \to \infty$

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

where $\Gamma \in \mathbb{R}^{p \times d}$, $G \in \mathbb{R}^{d \times d}$, and $W \in \mathbb{R}^{p \times p}$ is the pop. $\hat{W}$.

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

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

**Near Sparse signal:** $h(p) = o(p^{1/3})$

**Sparse signal:** $h(p) = O(1)$
Agreement between $\Delta^{-1}$ and $W$

Define $\rho = W^{1/2}\Delta W^{1/2} \in \mathbb{R}^{p \times p}$. $\rho = I_p$ if $W = \Delta^{-1}$. Let $\| \cdot \|$ denote the spectral norm. Then we assume

1. $\| \rho \| = O(h(p))$

2. $E(\epsilon^T W \epsilon) = O(p)$ and $\text{var}(\epsilon^T W \epsilon) = O(p^2)$.

Recall $\text{var}(\epsilon) = \Delta$. 

Dennis Cook | Dimension Reduction in Abundant High Dimensional Regressions
A Main Result

\[ \hat{R}_W(X_{\text{new}}) - R(X_{\text{new}}) = \nu + O_p(\kappa) + O_p(\psi) + O_p(\omega). \]

- \( \nu = R_W(\epsilon_{\text{new}}) - R(\epsilon_{\text{new}}) \), which does not depend on \( n \)
  - \( E(\nu) = 0 \) & \( \text{var}(\nu) \) is bounded as \( p \to \infty \)
  - \( \text{var}(\nu) \to 0 \) as \( p \to \infty \) if \( \|\rho\| = o(h(p)) \)
    - \( \|\rho\| = o(p) \) in abundant regressions
    - No help in sparse regressions
  - \( \text{var}(\nu) = 0 \) if \( \text{span}(W^{1/2}\Gamma) \) reduces \( \rho \). Holds trivially if \( W = \Delta^{-1} \) so \( \rho = I_p \).
\[ \hat{R}_W(X_{\text{new}}) - R(X_{\text{new}}) = \nu + O_p(\kappa) + O_p(\psi) + O_p(\omega). \]

- \( \kappa \to 0 \) as \( n, p \to \infty \):

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

1. \( \kappa = 1/\sqrt{n} \) in abundant regressions, \( h(p) \asymp p \).
2. \( \kappa = \sqrt{p/n} \) in sparse regressions, \( h(p) = O(1) \).
3. If \( \hat{W} = \Delta^{-1} \) then \( \hat{R}_W(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa) \). \( \kappa^{-1} \) is the oracle rate.
4. If \( n > p + r + 4 \), \( \varepsilon \sim N(0, \Delta) \) & \( \hat{W} = \hat{\Delta}^{-1} \), then \( \hat{R}_W(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa) \). (Allows \( n \asymp p \).)
\[ \hat{R}_{\tilde{W}}(X_{\text{new}}) - R(X_{\text{new}}) = \nu + O_p(\kappa) + O_p(\psi) + O_p(\omega). \]

- \(\psi(n, p, \rho)\):
  \[ \psi(n, p, \rho) = \frac{\|\rho\|_F}{h(p) \sqrt{n}} \]

- \(\omega(n, p)\): Define \( S = W^{-1/2}(\tilde{W} - W)W^{-1/2} \).
  - \(\|S\| = O_p(\omega)\).
  - \(\|E(S^2)\| = O(\omega^2)\).

- If the regression is abundant and \(\|\rho\| = O(1)\), then
  \[ \hat{R}_{\tilde{W}}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(n^{-1/2}) + O_p(\omega) \]
$\hat{W} = \text{SPICE estimator of } \Delta^{-1} \text{ based on } \hat{\Delta}$

Assume that (A) the eigenvalues of $\Delta$ are bounded as $p \to \infty$, (B) the errors are sub-Gaussian, (C) the SPICE tuning parameter

$$\omega \approx \left( \frac{\log p}{n} \right)^{1/2}.$$

Let $s = s(p)$ be the total number of non-zero off diagonal elements of $\Delta^{-1}$.

Then for SPICE

$$\omega = \left( \frac{(s + 1) \log p}{n} \right)^{1/2}$$

and

$$\hat{R}_W(x_{\text{new}}) - R(x_{\text{new}}) = O_p(n^{-1/2}) + O_p(\omega)$$

If $s$ is bounded and the regression is abundant then

$$\hat{R}_W(x_{\text{new}}) - R(x_{\text{new}}) = O_p(n^{-1/2} \log^{1/2} p)$$

Dennis Cook | Dimension Reduction in Abundant High Dimensional Regressions
Simulations

Data generation:

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

with \( d = 1, \Gamma \sim N(0, 1), Y \sim N(0, 1) \) and \( \Delta = D^{1/2} \Theta D^{1/2} \) where \( \text{diag}(D) \sim U(1, 101) \), \( \Theta = (1 - \theta)I_p + \theta 1_p 1_p^T \).

Fitted model:

\[ X|Y = y \sim \mu + \Gamma \beta f(y) + \varepsilon \]

with \( d = 1, f(y) = (y, y^2, y^3, y^4)^T \), so \( r = 4 \),

All results based on averages over 200 replications of the correlation between \( \hat{R}_W(X_{\text{new}}) \) and \( R(X_{\text{new}}) \) based on 100 \( X_{\text{new}} \) samples.
### Figure: $n = p/2$

- $\hat{W} = \text{SPICE}$, 
  - 
- $\hat{W} = \text{Moore-Penrose inverse of } \hat{\Delta}$, 
  - 
- $\hat{W} = \text{diag}^{-1} \hat{\Delta}$, 
  - 
- $\hat{W} = I_p$

#### Key Points

- **a.** $\theta = 0.2$
- **b.** $\theta = 0.7$
- **c.** $\theta = 0.99$

### Simulation Results

<table>
<thead>
<tr>
<th>$\log_2(p)$</th>
<th>4.5</th>
<th>5.5</th>
<th>6.5</th>
<th>7.5</th>
<th>8.5</th>
<th>9.5</th>
<th>10.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Correlation</td>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
<td></td>
</tr>
</tbody>
</table>

---

**Dennis Cook | Dimension Reduction in Abundant High Dimensional Regressions**
**Figure: **$p = 100$

$\hat{W} = \text{SPICE}$, ---

$\hat{W} = \text{Moore-Penrose inverse of } \hat{\Delta}$, -----

$\hat{W} = \text{diag}^{-1}\hat{\Delta}$, ········

$\hat{W} = I_p$ ————
Spectroscopy: Pork

**Goal:** Predict the percentage of fat $Y$ in a pork sample.

**Data:** $n = 54$ samples of pork. Predictors are absorbance spectra measured at $p = 100$ wavelengths.

$f(y): f(y) = (y, y^2, y^3)^T$ based on graphical evaluation:

![Graph showing absorbance at wavelength 845 vs. percent fat Y](image-url)
Dimension $d$: Adapting a permutation test (Cook and Yin 2001) we inferred $d = 1$.

Prediction:

$$\hat{E}\{Y|X = x\} = \sum_{i=1}^{n} w_i(x) Y_i$$

$$w_i(x) = \frac{\hat{g}(R(x)|Y_i)}{\sum_{i=1}^{n} \hat{g}(R(x)|Y_i)}$$

$$\hat{g} = \exp \left\{-2^{-1} [\hat{R}_{\hat{W}}(x) - \hat{\beta}_f(y_i)]^T \hat{\Gamma}^T \hat{W} \hat{\Gamma} [\hat{R}_{\hat{W}}(x) - \hat{\beta}_f(y_i)] \right\}.$$
a. $\hat{R}_\Delta$

b. $\hat{R}_{\text{spice}}$

c. $\hat{R}_{\text{diag}}$

d. $\hat{R}_I$
Spectroscopy: Pork and Beef

**Goal:** Predict the percentage of fat $Y$.

**Data:** $n = 103$ samples of pork or beef. Predictors are absorbance spectra measured at $p = 95$ wavelengths.

$f(y): f(y) = (y, y^2, \text{Ind}(\text{beef}))^T$
a. \( \hat{R}_\Delta^{-1} \)

b. \( \hat{R}_{\text{spice}} \)

c. \( \hat{R}_{\text{diag}} \)

d. \( \hat{R}_I \)
Some conclusions

- The notion of abundance can be important, depending on the application.
- Any of the estimators can work well in abundant or near-abundant regressions. Generally,
  - When \( n > p + r + 4 \), \( \hat{\Delta}^{-1} \) and SPICE seem the best.
  - When \( n < p + r + 4 \), SPICE is so far the overall winner, but has computational problems with large \( p \) or large conditional predictor correlations. More work on Moore-Penrose inverse and other possibilities needed.
- Screening methods can be developed to insure abundance or near-abundance.