Dimension Reduction in Abundant High Dimensional Regressions

Dennis Cook

University of Minnesota

8th Purdue Symposium
June 2012

In collaboration with Liliana Forzani & Adam Rothman,
Annals of Statistics, February 2012, 353–384

Broad context

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

Data: $(Y_i, \mathbf{X}_i)$ iid, $i = 1, \ldots, n$.

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

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

span$(\boldsymbol{\alpha})$ is called a dimension reduction subspace (DRS) & $\boldsymbol{\alpha}^T\mathbf{X}$ is called a sufficient reduction.

Broad contex, cont.

Smallest reduction is characterized by

- $\mathcal{S}_{Y|\mathbf{X}} = \cap\, \mathcal{S}_{\text{DRS}}$; $R(\mathbf{X}) = \boldsymbol{\eta}^T\mathbf{X}$;
  $\text{span}(\boldsymbol{\eta}) = \mathcal{S}_{Y|\mathbf{X}} =$ central subspace.
- Can't really handle $n < p$ yet.
- Chen et al. (2010) pursue variable elimination by estimating rows of $\boldsymbol{\eta}$ to be 0, but still with $p/n \to 0$

Today's context

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

Distinctions:

- Bypass estimation of $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathbb{R}^p$ and instead estimate $R(\mathbf{X}) \in \mathbb{R}^d$ directly, with $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ 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 – $R(\mathbf{X}_{\text{new}})$ or $Y|\mathbf{X}_{\text{new}}$ – rather than variable selection.
- Use SPICE (Rothman, et al.) to estimate a critical $p \times p$ matrix of weights $\mathbf{W}$.

Tasks:

- Reductive context and $R(\mathbf{X})$
- Class of estimators $\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X})$
- Key structural assumptions
- Main results for $\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = O_p(r(n,p))$, $r \to 0$ as $n, p \to \infty$.
- Illustrations

Inverse regression

$$\mathbf{X}|(Y = y_i) \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}(y_i) + \boldsymbol{\varepsilon}_i, i = 1, \ldots, n.$$

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

Estimation

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

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

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

Weight matrix: $\widehat{\mathbf{W}} \in \mathbb{R}^{p \times p}$ is an "estimator" of $\boldsymbol{\Delta}^{-1}$ with population version $\mathbf{W}$.

Reductions :

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}) = (\widehat{\boldsymbol{\Gamma}}^T \widehat{\mathbf{W}} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T \widehat{\mathbf{W}}(\mathbf{X} - \bar{\mathbf{X}})$$

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

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

Specific estimators

Choices for $\widehat{\mathbf{W}}$: Let $\widehat{\boldsymbol{\Delta}}$ be the residual covariance matrix from the multivariate OLS fit of $\mathbf{X}$ on $\mathbf{f}$ (requires only $n > r + 4$). Then

- $\widehat{\mathbf{W}} = \mathbf{W}$, like $\mathbf{W} = \mathbf{I}_p$ or the ideal case $\mathbf{W} = \boldsymbol{\Delta}^{-1}$.
- $\widehat{\mathbf{W}} = \mathrm{diag}^{-1}(\widehat{\boldsymbol{\Delta}})$
- $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}^{-1}$, requires $n > p + r + 4$, allowing $n \asymp p$.
- $\widehat{\mathbf{W}} =$ SPICE estimator of $\boldsymbol{\Delta}^{-1}$ applied to $\widehat{\boldsymbol{\Delta}}$
- $\widehat{\mathbf{W}} =$ Moore-Penrose inverse $\widehat{\boldsymbol{\Delta}}^{-}$ of $\widehat{\boldsymbol{\Delta}}$ (simulation only).

Signal rate, $h(p)$

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

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

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

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 $\mathbf{\Delta}^{-1}$ and $\mathbf{W}$

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

1. $\|\boldsymbol{\rho}\| = O(h(p))$
2. $E(\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon}) = O(p)$ and $\mathrm{var}(\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon}) = O(p^2)$. Recall $\mathrm{var}(\boldsymbol{\varepsilon}) = \mathbf{\Delta}$.

A Main Result

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = \boldsymbol{\nu} + O_p(\kappa) + O_p(\psi) + O_p(\omega).$$

- $\boldsymbol{\nu} = R_{\mathbf{W}}(\varepsilon_{\text{new}}) - R(\varepsilon_{\text{new}})$, which does not depend on n
  - $E(\boldsymbol{\nu}) = 0$ & $\text{var}(\boldsymbol{\nu})$ is bounded as $p \to \infty$
  - $\text{var}(\boldsymbol{\nu}) \to 0$ as $p \to \infty$ if $\|\boldsymbol{\rho}\| = o(h(p))$
    - $\|\boldsymbol{\rho}\| = o(p)$ in abundant regressions
    - No help in sparse regressions
  - $\text{var}(\boldsymbol{\nu}) = 0$ if $\text{span}(\mathbf{W}^{1/2}\boldsymbol{\Gamma})$ reduces $\boldsymbol{\rho}$. Holds trivially if $\mathbf{W} = \boldsymbol{\Delta}^{-1}$ so $\boldsymbol{\rho} = \mathbf{I}_p$.

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = \boldsymbol{\nu} + O_p(\kappa) + O_p(\psi) + O_p(\omega).$$

- $\kappa \longrightarrow 0$ as $n, p \longrightarrow \infty$:

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

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 $\widehat{\mathbf{W}} = \boldsymbol{\Delta}^{-1}$ then $\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = O_p(\kappa)$. κ^{-1} is the oracle rate.
4. If $n > p + r + 4$, $\varepsilon \sim N(0, \boldsymbol{\Delta})$ & $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}^{-1}$, then $\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = O_p(\kappa)$. (Allows $n \asymp p$.)

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = \boldsymbol{\nu} + O_p(\kappa) + O_p(\psi) + O_p(\omega).$$

- $\psi(n, p, \boldsymbol{\rho})$:

$$\psi(n, p, \boldsymbol{\rho}) = \frac{\|\boldsymbol{\rho}\|_F}{h(p)\sqrt{n}}$$

- $\omega(n, p)$: Define $\mathbf{S} = \mathbf{W}^{-1/2}(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{W}^{-1/2}$.
    - $\|\mathbf{S}\| = O_p(\omega)$.
    - $\|E(\mathbf{S}^2)\| = O(\omega^2)$.
- If the regression is abundant and $\|\boldsymbol{\rho}\| = O(1)$, then

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = O_p(n^{-1/2}) + O_p(\omega)$$

$\widehat{\mathbf{W}}$ = SPICE estimator of $\boldsymbol{\Delta}^{-1}$ based on $\widehat{\boldsymbol{\Delta}}$

Assume that (A) the eigenvalues of $\boldsymbol{\Delta}$ are bounded as $p \to \infty$, (B) the errors are sub-Gaussian, (C) the SPICE tuning parameter $\asymp (\frac{\log p}{n})^{1/2}$.

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

Then for SPICE

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

and

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = O_p(n^{-1/2}) + O_p(\omega)$$

If s is bounded and the regression is abundant then

$$\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}}) - R(\mathbf{X}_{\text{new}}) = O_p(n^{-1/2}\log^{1/2} p)$$

Simulations

Data generation:

$$\mathbf{X}|(Y = y) \sim \boldsymbol{\Gamma} y + N_p(0, \boldsymbol{\Delta})$$

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

Fitted model:

$$\mathbf{X}|(Y = y) \sim \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}(y) + \boldsymbol{\varepsilon}$$

with $d = 1$, $\mathbf{f}(y) = (y, y^2, y^3, y^4)^T$, so $r = 4$,
All results based on averages over 200 replications of the correlation between $\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{X}_{\text{new}})$ and $R(\mathbf{X}_{\text{new}})$ based on 100 $\mathbf{X}_{\text{new}}$ samples.

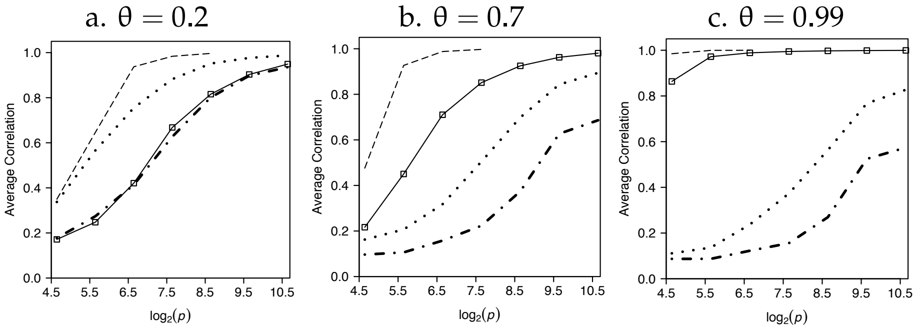


Figure: $n = p/2$

$\widehat{\mathbf{W}} = \text{SPICE}$, – – – – –

$\widehat{\mathbf{W}} = \text{Moore-Penrose inverse of } \widehat{\boldsymbol{\Delta}}$, ———

$\widehat{\mathbf{W}} = \text{diag}^{-1}\widehat{\boldsymbol{\Delta}}$, ·········

$\widehat{\mathbf{W}} = \mathbf{I}_p$ —·—·—·—

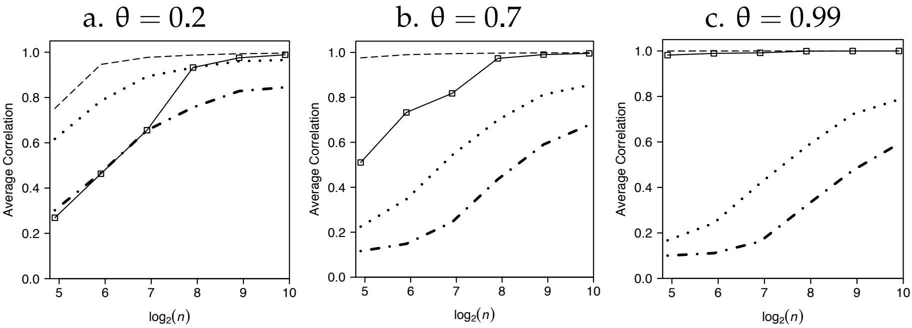


Figure: $p = 100$

$\widehat{\mathbf{W}} = \text{SPICE}$, — — — — —

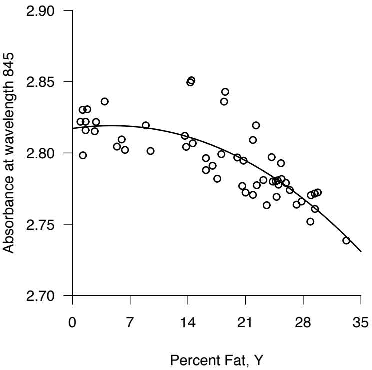
$\widehat{\mathbf{W}} = $ Moore-Penrose inverse of $\widehat{\boldsymbol{\Delta}}$, ————

$\widehat{\mathbf{W}} = \text{diag}^{-1}\widehat{\boldsymbol{\Delta}}$, · · · · · · · · ·

$\widehat{\mathbf{W}} = \mathbf{I}_p$ —·—·—·—

Spectroscopy: Pork

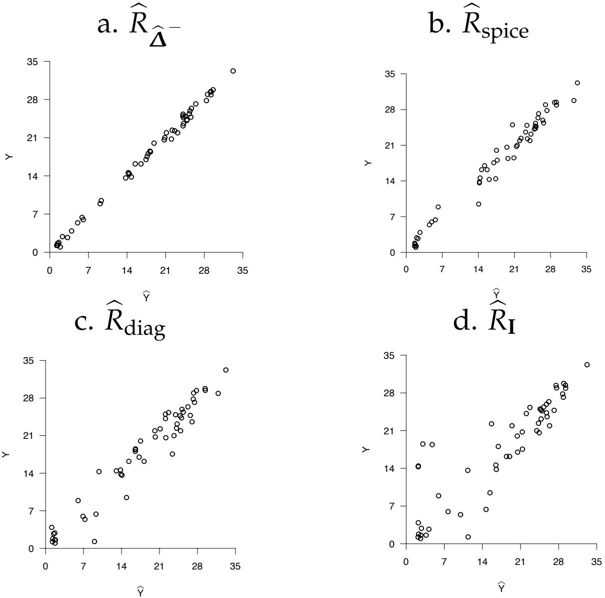
Goal: Predict the percentage of fat Y in a pork sample.
Data: $n = 54$ samples of pork. Predictors are absorbance spectra measured at $p = 100$ wavelengths.
$f(y)$: $f(y) = (y, y^2, y^3)^T$ based on graphical evaluation:



Dimension d: Adapting a permutation test (Cook and Yin 2001) we inferred $d = 1$.

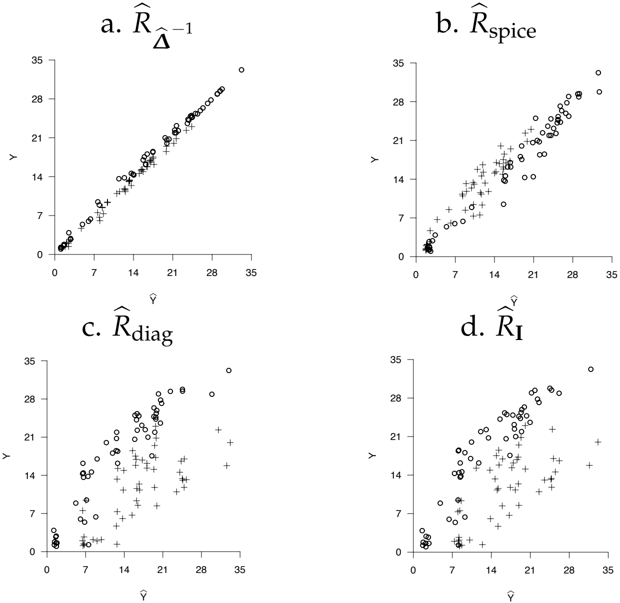
Prediction:

$$\begin{aligned}
\widehat{E}\{Y|\mathbf{X}=\mathbf{x}\} &= \sum_{i=1}^{n} w_i(\mathbf{x})Y_i \\
w_i(\mathbf{x}) &= \frac{\hat{g}(R(\mathbf{x})|Y_i)}{\sum_{i=1}^{n}\hat{g}(R(\mathbf{x})|Y_i)} \\
\hat{g} &= \exp\left\{-2^{-1}[\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{x}) - \widehat{\boldsymbol{\beta}}\mathbf{f}(y_i)]^T\widehat{\boldsymbol{\Gamma}}^T\widehat{\mathbf{W}}\widehat{\boldsymbol{\Gamma}}[\widehat{R}_{\widehat{\mathbf{W}}}(\mathbf{x}) - \widehat{\boldsymbol{\beta}}\mathbf{f}(y_i)]\right\}.
\end{aligned}$$


a. $\widehat{R}_{\widehat{\boldsymbol{\Delta}}^-}$
b. $\widehat{R}_{\text{spice}}$
c. $\widehat{R}_{\text{diag}}$
d. $\widehat{R}_{\mathbf{I}}$
Y
$\widehat{Y}$
0
7
14
21
28
35

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.
$f(y)$: $f(y) = (y, y^2, \text{Ind}(beef))^T$


a. $\widehat{R}_{\widehat{\boldsymbol{\Delta}}^{-1}}$
b. $\widehat{R}_{\text{spice}}$
c. $\widehat{R}_{\text{diag}}$
d. $\widehat{R}_{\mathbf{I}}$
Y
$\widehat{Y}$
0
7
14
21
28
35

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 $n > p + r + 4$, $\widehat{\boldsymbol{\Delta}}^{-1}$ 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.