Dimension Reduction & Prediction in Abundant High Dimensional Regressions

Dennis Cook

School of Statistics
University of Minnesota

October 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, $i = 1, \ldots, n.$

Reductions: Find the “smallest” linear reduction $R(X) = \eta^T X : \mathbb{R}^p \to \mathbb{R}^q, q \leq p,$ so that $Y \perp X | R(X).$

Goal: Prediction – $Y | X_{\text{new}}$ or $R(X_{\text{new}})$ – rather than variable selection when $n, p \to \infty$ in various alignments.

Emphasize abundant regressions, where many predictors contribute information about $Y.$ Sparsity is not ruled out, but is not required, either.
Part I: Normal linear regression, \((Y, \mathbf{X})\)

Model:

\[
Y = \mu_Y + \beta^T (\mathbf{X} - \mu_X) + \epsilon
\]

with \(\epsilon \sim N_1(0, \sigma^2)\), \(\mathbf{X} \sim N_p(\mu_X, \Sigma_X)\) and
\(R(\mathbf{X}) = \beta^T \mathbf{X}\).

Predictions: With \(\mathbf{X}_{\text{new}} \sim N_p(\mu_X, \Sigma_X)\)

\[
\hat{Y}_{\text{new}} = \bar{Y} + \hat{\beta}^T (\mathbf{X}_{\text{new}} - \bar{\mathbf{X}})
\]

where \(\hat{\beta}\) is for now a generic estimator of \(\beta\).

Approach: Compare methods using

\[
D = \hat{\beta}^T (\mathbf{X}_{\text{new}} - \bar{\mathbf{X}}) - \beta^T (\mathbf{X}_{\text{new}} - \mu_X) = O_p\{r(n, p)\}
\]

as \(n, p \to \infty\) in various alignments. We consider
\(\text{var}(D) = O\{r^2(n, p)\}\), when details permit. In all cases, calculations are w.r.t the data and \(\mathbf{X}_{\text{new}}\).
Signal rate $h(p)$

Notation: Let $\Sigma_{X|Y} = \text{var}(X|Y)$, $\sigma_{XY} = \text{cov}(X, Y)$, $\sigma_Y^2 = \text{var}(Y)$ and $R_{YX}^2 = \text{sqared pop. multiple corr. coefficient for } Y \text{ on } X$. Then

$$h(p) = \frac{\sigma_{XY}^T \Sigma_{X|Y}^{-1} \sigma_{XY}}{\sigma_Y^2} = \frac{R_{YX}^2}{1 - R_{YX}^2}$$

- If $\sigma_{XY}$ falls in a reducing subspace of $\Sigma_{X|Y}$ with bounded eigenvalues and then $h(p) \preceq \|\sigma_{XY}\|^2$.
- If as $p \to \infty$ the eigenvalues of $\text{corr}(X|Y)$ are bounded and $0 < \rho^2(X_j, Y) < 1$, then $h(p) \preceq p$.
- Terminology: A regression is abundant if $h(p) \to \infty$ and sparse otherwise. “Bounded eigenvalues” means bounded away from 0 and $\infty$. 

---

| Dimension Reduction & Prediction in Abundant High Dimensional Regressions |
Part I: Results

- **Σᵪ known**: If Σᵪ is known and \( \hat{\beta} = \Sigma^{-1} \hat{\sigma}_{XY} \) then \( \text{var}(D) \preceq p/n \).

- **OLS**: If \( n > p + 2 \), \( p/n \to [0, 1) \) and \( \hat{\beta} = \Sigma^{-1} \hat{\sigma}_{XY} \) then \( \text{var}(D) = O(\kappa^2) \) where

\[
\kappa^2 = \frac{p}{nh(p)}.
\]

Special cases: \( \text{var}(D) = O(n^{-1}) \) if \( h(p) \preceq p \) and \( \text{var}(D) = O(p/n) \) if \( h(p) \preceq 1 \).

- **Moore-Penrose inverse**: If \( n < p \), the eigenvalues of \( \Sigmaᵪ \) are bounded as \( p \to \infty \), and \( \hat{\beta} = \hat{\Sigma}^{-1} \hat{\sigma}_{XY} \), then \( \text{var}(D) \preceq 1 \) if either \( n/p \to [0, 1) \) or if \( p - n \) is constant and \( h(p) \preceq p \).
Predicting near the data: If $X_{\text{new}} \mid (\tilde{X}, \tilde{\Sigma}_X) \sim N_p(\tilde{X}, \tilde{\Sigma}_X)$ and $\hat{\beta} = \tilde{\Sigma}_X \hat{\sigma}_{XY}$, then

$$\text{var}(D) \asymp \frac{\min(n, p)}{nh(p)} + \frac{1}{n}.$$  

Special cases: If $n > p$ then $\text{var}(D) \asymp \kappa^2$. If $n < p$ then $\text{var}(D) \asymp h^{-1}(p) + n^{-1}$.

**$\Sigma_{X\mid Y}$ known:** Recall $\Sigma_X = \Sigma_{X\mid Y} + \sigma_{XY} \sigma_{XY}^T / \sigma_Y^2$. If $\Sigma_{X\mid Y}$ is known and $\hat{\beta} = \left( \Sigma_{X\mid Y} + \hat{\sigma}_{XY} \hat{\sigma}_{XY}^T / \hat{\sigma}_Y^2 \right)^{-1} \hat{\sigma}_{XY}$ then

$$D = O_p(\kappa^2) + O_p \left( \frac{\kappa}{\sqrt{h(p)}} \right) + O_p \left( \frac{1}{\sqrt{n}} \right)$$

Special case: If $p > n$ and $h(p) \asymp p$ then $D = O_p(n^{-1/2})$.  

Dimension Reduction & Prediction in Abundant High Dimensional Regressions
\[ \hat{\beta} = \left( \Sigma_{X|Y}^{-1} - \Sigma_{X|Y}^{-1}(\hat{\sigma}_Y^2 + \hat{\sigma}_{XY}^T \Sigma_{X|Y}^{-1} \hat{\sigma}_{XY}) \Sigma_{X|Y}^{-1} \right) \hat{\sigma}_{XY} \]

and if \( \|\Sigma_{X|Y}^{1/2}(\Sigma_{X|Y}^{-1} - \Sigma_{X|Y}^{-1})\Sigma_{X|Y}^{1/2}\| = O_p(\omega) \), then add \( O_p(\omega) \) to the previous order of \( D \).

Qualitative conclusion from part I: Incorporating knowledge of or constraints on \( \Sigma_{X|Y}^{-1} \) can lead to better estimators that dealing with \( \Sigma_X \) or \( \Sigma_X^{-1} \).
Part II: Inverse regression

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

- \( \mu \in \mathbb{R}^p, \Gamma \in \mathbb{R}^{p \times d}, \beta \in \mathbb{R}^{d \times r}, d < p \& r; d, r \) fixed.
- \( E(\varepsilon_i) = 0, \text{var}(\varepsilon_i) = \Sigma_{X|Y} > 0, \varepsilon \perp Y. \)
- \( f(y) \in \mathbb{R}^r \) 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 \( \text{rank}\{\text{cov}(f(Y), g(Y))\} = r. \)
- \( R(X) = (\Gamma^T\Sigma_{X|Y}^{-1}\Gamma)^{-1}\Gamma^T\Sigma_{X|Y}^{-1}(X - \mu_X) \in \mathbb{R}^d. \)
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 $\Sigma_{X|Y}^{-1}$ with population version $W$.

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

$$R(X) = (\Gamma^T \Sigma_{X|Y}^{-1} \Gamma)^{-1} \Gamma^T \Sigma_{X|Y}^{-1} (X - \mu)$$

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

| Dimension Reduction & Prediction in Abundant High Dimensional Regressions |
Specific estimators

Choices for $\hat{W}$: Let $\hat{\Sigma}_{X|Y}$ 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 $W = \Sigma_{X|Y}^{-1}$.
- $\hat{W} = \text{diag}^{-1}(\hat{\Sigma}_{X|Y})$
- $\hat{W} = \hat{\Sigma}_{X|Y}^{-1}$, requires $n > p + r + 4$, allowing $n \approx p$.
- $\hat{W} = \text{SPICE estimator of } \Sigma_{X|Y}^{-1}$ applied to $\hat{\Sigma}_{X|Y}$
- $\hat{W} = \text{Moore-Penrose inverse } \hat{\Sigma}_{X|Y}^{-1}$ of $\hat{\Sigma}_{X|Y}$ (simulation only).
Signal rate, $h(p)$

As an extension of the previous signal rate, we define $h(p)$ 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 still the pop. $\hat{W}$.

**Abundant signal:** $h(p) \to \infty$

**Sparse signal:** $h(p) \asymp 1$
Agreement between $\Sigma_{X|Y}^{-1}$ and $W$

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

1. $\| \rho \| = O(h(p))$
2. $E(\varepsilon^T W \varepsilon) = O(p)$ and $\text{var}(\varepsilon^T W \varepsilon) = O(p^2)$. 
**A Main Result**

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

- \(\nu = R_W(\varepsilon_{\text{new}}) - R(\varepsilon_{\text{new}})\), 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 \(\|\rho\| = o(h(p))\)
- \(\text{var}(\nu) = 0\) if \(\text{span}(W^{1/2}\Gamma)\) reduces \(\rho\). Holds trivially if \(W = \Sigma_{X|Y}^{-1}\) 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 \( n, p \to \infty \) (recall \( \kappa^2 = p/\{nh(p)\} \)):
  1. \( \kappa = 1/\sqrt{n} \) when \( h(p) \approx p \).
  2. \( \kappa = \sqrt{p/n} \) when \( h(p) \approx 1 \).
  3. If \( \hat{W} = \Sigma_{X|Y}^{-1} \) then \( \hat{R}_W(X_{\text{new}}) - R(X_{\text{new}}) = O_p(\kappa) \). \( \kappa^{-1} \) is the oracle rate.
  4. If \( n > p + r + 4, \varepsilon \sim N(0, \Sigma_{X|Y}) \) & \( \hat{W} = \Sigma_{X|Y}^{-1} \), then 

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

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

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

- If \( h(p) \asymp p \) and \( ||\rho|| = O(1) \), then
  \[
  \hat{R}_{\hat{W}}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(n^{-1/2}) + O_p(\omega)
  \]
SPICE

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

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

Then for SPICE

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

and

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

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

$$\hat{R}_{\hat{w}}(X_{\text{new}}) - R(X_{\text{new}}) = O_p(n^{-1/2} \log^{1/2} p)$$
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:**

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

\[
\omega_i(x) = \frac{\hat{g}(R(x)|Y_i)}{\sum_{i=1}^{n} \hat{g}(R(x)|Y_i)}
\]

\[
\hat{g} = \exp \left\{ -2^{-1} [\hat{R}_W(x) - \hat{\beta}f(y_i)]^T \hat{\Gamma}^T \hat{W} \hat{\Gamma} [\hat{R}_W(x) - \hat{\beta}f(y_i)] \right\}.
\]
Dimension Reduction & Prediction in Abundant High Dimensional Regressions
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.
$f(y): f(y) = (y, y^2, \text{Ind}(\text{beef}))^T$
Introduction

Part I: Normal reg.

Part II: Inverse reg.

Estimation

Key structure

Main results

Spectroscopy

Conclusions

\[ \hat{R}_\Delta^{-1} \]

\[ \hat{R}_{\text{spice}} \]

\[ \hat{R}_{\text{diag}} \]

\[ \hat{R}_I \]

Dimension Reduction & Prediction in Abundant High Dimensional Regressions
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 $n > p + r + 4$, $\hat{\Sigma}_{X|Y}^{-1}$ 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, without necessarily requiring $n$ be larger than the reduced $p$. 