Supervised dimension reduction in large-p-small-n problems
Identifying functional elements of the human genome

Sequences of the nuclear DNA that perform a function

May represent ~5% of the human genome (Mouse Cons, *Nature* 2002; Chiaromonte et al., CSHL Symp 2004)

However, only 1-1.5% is **protein coding** (Human Cons, *Nature* 2001)

Thus, as much as 3.5-4% may harbor **functional non-coding sequences**

... among which **regulatory elements**
Data: Alignments

Mismatches (substitutions)
Gaps (small insertions and deletions)

Multiple genome alignment

CTCCCGG---CAGCCCC
CTCCCAGAGAGCTGCCCC
CT---AGAGAGCTGCCCC

Substitutions
Deletion
Insertion

MultiZ, AxtChain/AxtNet
Blanchette et al. Gen Res 2004
Kent et al. Gen Res 2003

Review of alignment algorithms

Review: Batzoglou Briefings in Bioinformatics 2005
Data: Training sets

**POSITIVE TRAINING SET:**
Alignment segments from a collection of known functional elements.

**NEGATIVE TRAINING SET:**
Alignment segments from DNA sequences believed to have no function.

For RP scores

5-way mammalian alignments of Hum/Chimp/Mus/Rat/Dog on

**Positive REG:** known human regulatory elements from literature, trimmed to smallest seq’s preserving activity; L. Elnitski, [http://bio.cse.psu.edu/mousegroup/Reg_annotations](http://bio.cse.psu.edu/mousegroup/Reg_annotations)

**Negative AR:** ancestral repeat sequences, sampled as to approximately match the size of the positive training set. **Model of neutral behavior.**
Basic classification strategy for Regulatory Potential scores

Many alignment columns \((5^m-1)\ ordered m\text{-}uples of A,C,G,T and “–”, not \(m\) gaps) Limited training data (~30,000 columns in REG, similar in AR).

Create a mapping from algn clmns to a small alphabet that retains relevant information

Characterize “word” content in each training set fitting variable order Markov models

Compute a log-odds score for (any) align segment

\[
RP = \sum_{a \text{ in segment}} \log \left( \frac{P_{\text{REG}}(s_a \mid s_{a-1}, s_{a-2} \ldots)}{P_{\text{AR}}(s_a \mid s_{a-1}, s_{a-2} \ldots)} \right) 
\]

+ if “word” content resembles REG, – if it resembles AR.

F. Chiaromonte, Sp 08
See this as a large-\( p \)-small-\( n \) regression

\[
RP(A) = \sum_{a \text{=position in } A} \log \left( \frac{p_{\text{POS}}(s_a | s_{a-1} \ldots s_{a-T})}{p_{\text{NEG}}(s_a | s_{a-1} \ldots s_{a-T})} \right) = \sum_{w=\text{word}} \lambda_w X_{A,w}
\]

\[
P_Y(w_{\text{last}} | w_{\text{previous}}) = \frac{X_{Y,w}}{\sum_{\sigma=\text{sym in alphabet}} X_{Y,(w_{\text{previous}} \sigma)}} \quad Y = \text{POS, NEG}
\]

\[
X_{Y,w} = \sum_{A=\text{member of } Y} X_{A,w}
\]

... just one approach to the investigation of the regression of \( Y \) on \( X \), where

- Response: Categorical (2 classes)
- Dim of predictor: \( p = #(\text{Alphabet})^{T+1} \) e.g. \( 6^4=1296 \) (minus 1, add to constant)
- Number of observations: \( n = 514 \) (257 in each of \( Y=\text{POS} \) and \( Y=\text{NEG} \))

Given limited availability of training data, even with very parsimonious encodings

\[
n << p
\]
Many large-$p$-small-$n$ in Genomics
Investigation of “word” content in multiple species alignments for classification (prediction of function) is not the only instance … Microarray data

... in dozens of (indep) samples

Sufficient Dimension Reduction
Cook, Wiley 1998
Chiaromonte and Cook, AIMS 2003
(Cook et al., K.C. Li et al. since early 90’s)

expression for thousands of genes…

Reduced problem

Y based reduction

select $d << n$ lin comb’s
(prior to modeling)

Cook, Wiley 1998
Chiaromonte and Cook, AIMS 2003
(Cook et al., K.C. Li et al. since early 90’s)
Sufficient Dimension Reduction: basic concepts

Identify a collection of linear combinations (a linear subspace) \textbf{sufficient} for the regression, i.e. such that

\[ Y \parallel X \mid \alpha'X, \quad S = Sp(\alpha) \]

Dimension Reduction Space(s)

\[ S_{Y|X} = \bigcap_{S \text{ is a drs}} S = Sp(\beta) \]

Central Space – under mild assumptions, unique and minimal “linear core”

\[ \beta \ p \times d, \quad d = \dim(S_{Y|X}) \]

Spanning system and structural dimension

\[ Y \text{ on } W = \beta'X \]

Reduced regression

\[ \begin{align*}
Y &= \beta_1'X + \sigma \varepsilon \\
S_{Y|X} &= \text{Span}(\beta_1), d = 1 \\
Y &= g(\beta_1'X) + \sigma(\beta_2'X)\varepsilon \\
S_{Y|X} &= \text{Span}(\beta_1, \beta_2), d = 2 \\
Y &= g(\beta_1'X, \ldots \beta_d'X; \varepsilon) \\
S_{Y|X} &= \text{Span}(\beta_1 \ldots \beta_d), d
\end{align*} \]

...examples

- \textbf{any type of response} (continuous, categorical)
- extensions to cases comprising \textbf{categorical predictors} (Chiaromonte, Cook, B. Li, \textit{Annals} 2002; B. Li, Cook, Chiaromonte, \textit{Annals} 2003)
SDR: traditional methods

\[ \nu = Cov(X, Y) \]

\[ \beta = \Sigma^{-1} Cov(X, Y) \text{ OLS} \]

SLICED INVERSE REGRESSION (K.C. Li, JASA 1991)

\[ \tilde{\nu} = Var(E(X | Y)) \]

\[ p \times p \quad \text{rank}(\tilde{\nu}) = d \]

\[ \nu = (\gamma_1 \ldots \gamma_d) \]

In practice: slice if \( Y \) continuous, an LDA seed if \( Y \) categorical...
SDR: traditional methods… PROBLEM!

If \( n < p \) (or \( \sim p \)) we cannot use the inverse of the sample predictor covariance matrix!

**Diagram:**
- \( v \times p \times d \) matrix \( Sp(v) \)
- \( S_{Y|X} = \Sigma^{-1} Sp(v) \)
- \( \beta = \Sigma^{-1} v \)
- \( \hat{\beta} = \Sigma^{-1} \hat{c}_{x,y} \)
- \( \hat{\beta} = \Sigma^{-1} (\hat{y}_1 \ldots \hat{y}_d) \)

we need another way to put the seed in the right place
Enveloping the central space

\[ S_{Y|X} \subseteq M \]
\[ M = Sp(R), \quad R \quad p \times t \]

\[ S_{Y|X} = P_M^{(\Sigma)} S_{Y|X} \]
\[ = R (R'\Sigma R)^{-1} R' \Sigma Sp(\Sigma^{-1} \nu) \]
\[ = R (R'\Sigma R)^{-1} R' Sp(\nu) \]

\[ \beta = R (R'\Sigma R)^{-1} R' \nu \]

No need to invert \( \Sigma \), but only the much smaller \( t \) by \( t \) matrix \( R'\Sigma R \).

… new re-positioning
Sigma-Enveloping and iterations

\[ M = \bigoplus_{j=1}^{k} P_j^{(\Sigma)} S_{Y|X} \supseteq S_{Y|X} \]

\( P_j^{(\Sigma)} \) on jth eigenspace of \( \Sigma \)
\( k \) # of eigenspaces non-orth to \( S_{Y|X} \)

\[ R_u = (\nu, \Sigma \nu \ldots \Sigma^{u-1} \nu) \]
...span a non-decreasing sequence of subspaces

**Theorem:**
\[ \exists \widetilde{u} \leq k : \ M_u = Sp(R_u) \subset M, u < \widetilde{u} \]
\[ M_u = Sp(R_u) = M, u \geq \widetilde{u} \]

...now project
\[ S_u = P_{R_u}^{(\Sigma)} S_{Y|X} = R_u (R_u' \Sigma R_u)^{-1} R_u' Sp(\nu) \]

**Theorem:**
\[ \exists u^* \leq \widetilde{u} : \ S_u = S_{Y|X}, u \geq u^* \]
\[ \text{dim}(S_u) = R_u (R_u' \Sigma R_u)^{-1} R_u' Sp(\nu) = d \quad \forall u \]

\( d \)-dim projection bounces around until it settles on the CS
\[
\hat{R}_u = (\hat{\nu}, \hat{\Sigma} \hat{\nu} \ldots \hat{\Sigma}^{u-1} \hat{\nu}) \quad u = 1, 2, \ldots \\
\hat{\beta}_u = \hat{R}_u (\hat{R}_u' \hat{\Sigma} \hat{R}_u)^{-1} \hat{R}_u' \hat{\nu} \rightarrow \hat{\beta}
\]

works like Partial Least Squares when \( \nu = \text{Cov}(X,Y) \)

Logic similar to Iterative Hessian Transformation method (Cook and B. Li, Annals 2004)

**How do we select \( u^* \)?**

\[
\Delta_u = \beta_{u+1} - \beta_u \\
F_u = \text{trace}(\Delta_u' \Sigma \Delta_u) \quad \text{"dislocation"}
\]

**Theorem:**

\[
S_u = S_{Y|X} (\beta_u = \beta) \quad \text{iff} \quad F_t = 0, \forall t \geq u
\]

\[
f_u = n \text{ trace } (\hat{\Delta}_u' \hat{\Sigma} \hat{\Delta}_u)
\]

pragmatically benchmark with a “large sample” chi-square approximation
Back to the discrimination of regulatory elements

- Response: Categorical in 2 classes, $Y=1$ (POS) and $Y=0$ (NEG)
- Dim of predictor: $p = 6^4 = 1296$ (frequencies of encoded words of length 4)
- Number of observations: $n = 514$ (257 in each of $Y=1$ and 0)

\[ \hat{\nu} = \frac{1}{\|\bar{x}_1 - \bar{x}_0\|} (\bar{x}_1 - \bar{x}_0). \]

Linear Discriminant Analysis seed

Choosing $u^*$:

<table>
<thead>
<tr>
<th>$u$</th>
<th>1 to 2</th>
<th>2 to 3</th>
<th>3 to 4</th>
<th>4 to 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_u$</td>
<td>209706.95</td>
<td>107480.93</td>
<td>68478.26</td>
<td>58031.57</td>
</tr>
<tr>
<td>$\tilde{f}_u$</td>
<td>52.19</td>
<td>14.77</td>
<td>6.97</td>
<td>3.97</td>
</tr>
<tr>
<td>$g$</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>p-values</td>
<td>$2.734 \times 10^{-11}$</td>
<td>0.02215</td>
<td>0.2230</td>
<td>0.5543</td>
</tr>
</tbody>
</table>

... non-significant dislocation when passing from $u=3$ to 4, and thereafter

For this and many applications the right $u^*$ is small; bound above by $k = \text{number of } \Sigma\text{-eigenspaces non-orthogonal to } S_{n|x}$ (more later...)
Figure 1: Smoothed histograms of $w = \hat{B}_u'x$ for $y = 0, 1$ in subsequent iterations ($u = 1, 2, 3, 10$). The separation increases up to $u = 3$ and stabilizes thereafter.
Simulations: **quantitative** response, \( d=1 \)

\[ Y \mid X \sim Y \mid \beta'X \sim g(\beta'X) + \tau \varepsilon , \quad \varepsilon \perp X, \varepsilon \sim N(0, I_p) \]

\[ X \sim N(\mu, \Sigma). \quad \mu = 0_p \text{ and } \Sigma = \text{Diag}(\lambda) \]

**Predictor dimension:**

\[ p_{(1)} = 10, \; p_{(2)} = 500, \; p_{(3)} = 98 \]

**Regression function:**

\[ g_{(1)} = \text{linear}, \; g_{(2)} = \text{linear + 0.1cubic} \]

**Predictor coefficients:**

\[ \beta_{(1)} = \frac{1}{\sqrt{p}} 1_p \]

\[ \beta_{(2)} = \frac{1}{\sqrt{0.4p}} (1 \ldots 1, 0 \ldots 0, 1 \ldots 1)' \]

\[ \beta_{(3)} = \frac{1}{\sqrt{2}} (1, 0 \ldots 0, 1)' \]

all active, 40% active, 2 active

**Predictor covariance eigenvalues:**

\[ \lambda_{(1)} = \frac{2}{3} (2 \ldots 2, 1 \ldots 1)' \]

\[ \lambda_{(2)} = \frac{2}{(p+1)} (p, (p-1) \ldots 1)' \]

\( q=2 \) eigenspaces, multiplicity \( p/2 \)

\( q=p \) eigenspaces, multiplicity 1

\( \ldots \) etc.

Several simulation scenarios, each sampled 100 times with a sample size \( n=100 \)

F. Chiaromonte, Sp 08
Compare performance of

- **Iterative estimator**
- **Full OLS**; standard SDR when \( n > p \).
- **OLS of first \( j (=3,10) \) Principal Components**; common approach when \( p > n \).
  

Using correlation between \( \hat{\beta}'x \) and \( \beta'x \)

\[
|c(\hat{\beta}; \beta)| = \left| \hat{\beta}'\Sigma\beta \right| / \left\{ (\hat{\beta}'\Sigma\hat{\beta})^{1/2} (\beta'\Sigma\beta)^{1/2} \right\}
\]

averaged over simulation runs
Results

Table 2: Average absolute correlations between $\hat{\beta}'x$ and $\beta'x$.

<table>
<thead>
<tr>
<th></th>
<th>$p = 10, n = 100$</th>
<th></th>
<th>$p = 500, n = 100$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Iterative</td>
<td>PC(3)</td>
<td>PC(10) = OLS</td>
<td>Iterative</td>
</tr>
<tr>
<td>linear, all predictors, $q = 2$</td>
<td>0.9945</td>
<td>0.6413</td>
<td>0.9961</td>
<td>0.3153</td>
</tr>
<tr>
<td>linear, 40% predictors (top/bottom 20%), $q = 2$</td>
<td>0.9939</td>
<td>0.6692</td>
<td>0.9957</td>
<td>0.2733</td>
</tr>
<tr>
<td>linear, 2 predictors (top/bottom 1), $q = p$</td>
<td>0.9844</td>
<td>0.9107</td>
<td>0.9959</td>
<td>0.3294</td>
</tr>
<tr>
<td>linear+0.1cubic, all predictors, $q = 2$</td>
<td>0.9938</td>
<td>0.6741</td>
<td>0.9961</td>
<td>0.2783</td>
</tr>
<tr>
<td>linear+0.1cubic, 40% predictors (top/bottom 20%), $q = 2$</td>
<td>0.9935</td>
<td>0.6942</td>
<td>0.9960</td>
<td>0.2915</td>
</tr>
<tr>
<td>linear+0.1cubic, 2 predictors (top/bottom 1), $q = p$</td>
<td>0.9839</td>
<td>0.9096</td>
<td>0.9961</td>
<td>0.3887</td>
</tr>
</tbody>
</table>

sd's in 0.002-0.010 range

Table 3: Average absolute correlations between $\hat{\beta}'x$ and $\beta'x$ ($p = 98, n = 100$).

<table>
<thead>
<tr>
<th></th>
<th>Iterative</th>
<th>PC(3)</th>
<th>PC(10)</th>
<th>PC(98) = OLS</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9705</td>
<td>0.3317</td>
<td>0.5637</td>
<td>0.9589</td>
<td>0.1249</td>
</tr>
</tbody>
</table>

more simulation results...
The iterative estimator does very well, also when $p >> n$.

It does better (wrt OLS-type) the larger is $p$ relative to $n$, and already better than full OLS for $p \sim n$.

It does better (wrt OLS-type) the more sub-spherical structure in the predictor covariance – i.e. the fewer eigenspaces.
Our simulation results show that our iterative estimator performs very well in large-$p$-small-$n$ scenarios, and on par with standard estimators requiring inversion of $\hat{\Sigma}$ when $n > p$. We also have evidence that our estimator may offer an edge over standard estimators when the sample size is not much larger than the predictor dimension, and that its performance is best when the predictor covariance has a small number of eigenspaces. This last fact suggests that powers of $\hat{\Sigma}$ may allow us to effectively exploit “structure” in the predictor variability for the purpose of estimating the central space of a regression.

(Cook, B. Li and Chiaromonte, 2007)
By guiding our search for the CS, \( \Sigma \)-enveloping lets us exploit the variability structure of the predictors to overcome “shortage” of information (\( n < p \) or \( \sim p \)).

more effective the more sub-spherical structure it can grab on to:

\[
\Sigma = \Sigma_o + \sigma^2 I \quad \text{rank}(\Sigma_o) = p_o \text{ is small}
\]

Training data for discrimination of regulatory elements, word frequencies indeed covary in a highly structured way because words share sub-words:

Figure 2: Scree plot of the first forty eigenvalues of \( \hat{\Sigma} \) in the regulatory elements application (solid circles). Superimposed are the first forty eigenvalues of the sample covariance for simulated normal data of the same dimension (\( p = 1296 \)) and size (\( n = 514 \)), for which the underlying \( \Sigma \) has \( q = 4 \) eigenspaces (triangles)—see main text for details. All eigenvalues are multiplied by 100.