Dealing with Large $p$ in Relation to Data $n$ of the Same or Smaller Dimension

Peter Bickel

U.C. Berkeley

May, 2009

• In part collaborative work with Aiyou Chen, Liza Levina, Bo Li, Ya’acov Ritov and Alexandre Tsybakov.
• With assistance of Jing Lei
Outline

1. The data explosion
2. Goals and difficulties of analysis
3. The regression paradigm
   a) Prediction
   b) Variable selection
4. Pessimism of least favorable analysis as opposed to practice.
5. “Sparsity”: Models
   a) Methods in regression
   b) Covariance estimation
6. “Sparsity”: Data
   a) Regression: predictions on manifolds
   b) Dimension reduction
7. Another view of low dimensional structure
   a) Non-Gaussian signals with Gaussian noise
   b) ICA and Projection Pursuit
The Data Explosion

- Growth in amount and complexity of data in all fields exponential
- Due to:
  - Moore’s law and exponential growth in computing power, memory, etc
  - Exponential expansion of Internet
  - Exponential(?) growth of different data gathering methods (sensors, experimental techniques, telescopes, …)
The Data Explosion

DDBJ/EMBL/GenBank database growth

* Note: CON and TPA divisions are not counted in the Release statistic.

From: DDBJ(DNA Data Bank of Japan) statistics (2008)
Examples

I: Microarrays of gene expression at a given set of times and conditions say for tumor classification
   • $10^4$ genes but all but hundreds probably irrelevant

II: Atmospheric measurements
   • Average monthly measurements of (say) temperature at $10^4$ locations, $10^2$ years
   • Product of computer model or many missing values if actual
   • Distant positions have little to do with each other

III: Hand written individual postal ZIP code digits
   • $16 \times 16$ images
   • $< 10^8$ images
   • models of how people write are crude
Difficulties of Analysis

Features of many situations

- Models poor or non existent
- Some qualitative knowledge
- “Small” numbers of replicates or near replicates
Goals

I: Prediction assessment of factor interactions and importance

II: Extraction of simplifying features of data consistent with scientific understanding and with predictive value

III: Classification of digits by machine
The Regression Paradigm

- Observe \((X_i, Y_i), i = 1, \ldots, n\)
- I.I.D. \(X_i \in \mathcal{X}, Y_i \in \mathbb{R}\)
- Modelled by:

\[
Y = \eta(X) + \epsilon, \quad (X, Y) \sim \mathbb{P} \in \mathcal{P}
\]

- \(\eta(X) = \mathbb{E}[Y|X]\)
- For simplicity let \(\epsilon \sim N(0, \sigma^2)\)
Goals

• **Prediction** Construct $\hat{\eta} : \mathcal{X} \rightarrow \mathbb{R}$

  • **Regression**
    • $\mathbb{E}_P (\hat{\eta}(X) - Y)^2$ “small” for $P \in \mathcal{P}$
    • **Minimax** exhibit $\tilde{\eta}$ such that,

      $$\sup_{P} \mathbb{E}_{P} (\tilde{\eta}(X) - Y)^2 = \min_{\tilde{\eta}} \sup_{P} \mathbb{E}_{P} (\tilde{\eta}(X) - Y)^2$$

      (Least Favorable Analysis)

• **Classification**
  • $\hat{\eta} : \mathcal{X} \rightarrow \{0, 1\}$
  • $\mathbb{P}[\hat{\eta}(X) \neq Y]$ “small” for $P \in \mathcal{P}$
The Regression Paradigm

- **Linear Regression** (LR)
  - \( f_1, \ldots, f_p : \mathcal{X} \rightarrow \mathbb{R} \) given
  - \( \eta(X) = f^T(X)\beta, \ f \equiv (f_1, \ldots, f_p)^T, \ \beta \in \mathbb{R}^p \) unknown

- **NonParametric Regression** (NPR)
  - \( \eta(X) \in "A nice function space"" 
  - e.g. \( \mathcal{X} = \mathbb{R}^d, \ \mathcal{P}_s \leftrightarrow \left\{ \left| D^k\eta \right| (\cdot) \leq K < \infty, \ 1 \leq k \leq s \right\} \)
  - Sobolev spaces, Besov spaces, etc.
**Standard Solutions**

\[ LR \mathcal{P} \iff \eta(x) = f^T(x)\beta \]

- **Least Squares**
  - \( \hat{\beta}_{p \times 1} = [F^TF]^{-1}F^T\mathbf{y}, F_{n \times p} \equiv ||f_j(X_i)|| \)
  - \( \hat{\mathbf{Y}} = F[F^TF]^{-1}F^T\mathbf{y} \), predicted values
  - \( \hat{\eta}(x) = \hat{\beta}^T f(x) \)
The Worst Case

- $p = \infty$.
- Eg: $\eta(X) = \mathbb{E}(Y|X) \in \mathcal{P}_s$
  \[ \mathcal{P}_s = \{ \eta : |D^s \eta| \leq M \} \]
- Solution: Regularize
- Approximate $\eta(X)$ by $\sum_{j=1}^{K} \beta_j f_j(X)$
  \[ K \ll p, n, \text{ and regress as if approximation is model.} \]
Standard Solutions

• **Classification**: 2 Classes (NPC)
  - $Y = 0$ or $1$
  - $\eta(X) = \mathbb{P}(Y = 1|X)$

• **Classification**: 2 Classes (LC)
  - $\mathbb{P} \leftrightarrow \lambda(x) \equiv \log \left( \frac{\mathbb{P}[Y = 1|x]}{\mathbb{P}[Y = 0|x]} \right) = f^T(x)\beta$
  - Logistic Regression

\[
\hat{\lambda}(x) = f^T(x)\hat{\beta} \text{ where } \hat{\beta} \equiv \text{MLE}
\]
The “Overfitting” Problem (LR/LC)

\[ \mathbb{E} \left| \hat{Y} - F^T \beta \right|^2 = \frac{\sigma^2 p}{n}, \]

where \( F^T \beta \) is the best predictor of \( Y \) if \( \beta \) is known.

Note:

- If \( p > n \), LS has no unique solution for \( \beta \).
- If \( \frac{p}{n} \to \infty \), \( \hat{Y} \) worthless.
The Least Favorable Bottom Line for Prediction

\[ \triangle(s,d) \equiv \min_{\hat{\eta}} \max_{\mathcal{P}_s} \mathbb{E}_{\mathcal{P}} (\hat{\eta}(X) - \eta(X))^2 \approx n^{-\frac{2s}{2s+d}} \]

for

\[ \mathcal{P}_s = \{ \eta : |D^s\eta| \leq M \}. \]

(i) If \( s \to \infty \), behavior independent of \( d \), \( n^{-1} = \) parametric case

(ii) For moderate \( s \), even small \( d \) kill you!

(iii) For instance, \( s = 2, d = 12, n = 10^4 \Rightarrow \triangle(s,d) = .1 \)
Some Empirical Evidence

<table>
<thead>
<tr>
<th></th>
<th>d</th>
<th>classes</th>
<th>n</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zip code digits</strong></td>
<td>256</td>
<td>10</td>
<td>10,000</td>
<td>&lt; 0.025</td>
</tr>
<tr>
<td><strong>Microarrays</strong></td>
<td>3–4,000</td>
<td>2–3</td>
<td>70–100</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>Spam</strong></td>
<td>57</td>
<td>2</td>
<td>4,600</td>
<td>&lt; 0.07</td>
</tr>
</tbody>
</table>

Is there no curse of dimensionality? Why is theory no guide to practice?

(1) Hastie, Tibshirani, Friedman (2001) Elements of Statistical Learning (obtained from other sources)
(2) Dudoit, Fridlyand, Speed (2002) JASA
**Explanation I**

- “Sparse models”
- $\eta(\cdot)$ much more special than apparent.
- For suitable $f_1, f_2, \ldots$

$$\eta = \sum_{j=1}^{\infty} \beta_j f_j.$$

- All but $K, K \ll n, p$ of $\beta_k = 0.$
“Sparse Model”

A. Each $f_j$ depends only on one of $X_1, \ldots, X_p$

$$\eta = \sum_{j=1}^{p} g_j(X_j).$$

B. Each $f_j$ function of $\{X_k : k \in S, |S| \ll n, p\}$.

$$\eta = \mathbb{E}(Y|\{X_k : k \in S\}).$$

Two big problems

(i) Choice of $\{f_j\}$.

(ii) Choice of $k$ with $\beta_k \neq 0$. 
A Solution for (ii)

- **LASSO (Tibshirani):**
  \[ \hat{\beta} = \text{arg min} \left\{ \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{\infty} \beta_j f_j(X_i))^2 + \lambda \sum_{j=1}^{\infty} |\beta_j| \right\}. \]

- **Dantzig selector (Candes and Tao (2007) AS)**
  \[ \min |\beta|_1, \quad \text{for } |F^T(Y - F\beta)|_\infty \leq \lambda. \]

- **B., Ritov & Tsybakov (2008):** LASSO $\asymp$ Dantzig asymptotically.
Selection of Important Variables

- Collinearity problem.
- Suitable versions of LASSO find sparsest stable set (Meinshausen & Bühlmann (2009)).
- But is this what is wanted?
Sparse Models

Covariance Estimation

\(X_1, \ldots, X_n\) i.i.d. \(\mathbb{E}(X_1) = \mu, \ \text{Var}(X_1) \equiv \mathbb{E}(X_1 - \mu)(X_1 - \mu)^T \equiv \Sigma\)

- Principal component analysis (PCA)
- Linear or quadratic discriminant analysis (LDA/QDA)
- Inferring independence and conditional independence (graphical models)
- Implicit estimation of linear regression, \(\Sigma_{XX}^{-1}\Sigma_{XY}\), \(X\): regressors, \(Y\): response

Covariance itself is usually not the end goal:
- PCA requires estimation of the eigenstructure
- LDA/QDA, regression and conditional independence require the inverse
Pathologies of empirical covariance matrix

Estimate $\Sigma$ by $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (X_i - \bar{X})^T$

- MLE, for Gaussian unbiased (almost), well-behaved (and well studied) for fixed $d, n \to \infty$. But noisy if $d$ is large.
- Singular if $d > n$, so $\hat{\Sigma}^{-1}$ is not uniquely defined.
- Computational issues with $\hat{\Sigma}^{-1}$ for large $d$. (for Moore-Penrose)
- LDA completely breaks down if $d/n \to \infty$
- Eigenstructure inconsistent for i.i.d. components model as soon as $d/n \to c > 0$.
- Regression breaks down even for prediction
The Basic Problem

• Unspecified $\Sigma$ not sparse even if $\mu$ known

• Sparse models: many possibilities

  Simplest:

  • Permutation invariant sparsity

    A) Each row of $\Sigma$ sparse or sparsely approximable in operator norm

      e.g. If $S_i = \{j : \sigma_{ij} \neq 0\} \ |S_i| \leq s$ for all $i$.

    B) Each row of $\Sigma^{-1}$ sparsely approximable.

  • $A$) roughly implies $B$) if $\lambda_{\text{min}}(\Sigma) \geq \delta > 0$, where

    $\lambda_{\text{min}} \equiv \min(\text{eigenvalue}), \lambda_{\text{max}} \equiv \max(\text{eigenvalue})$

  • Interpretation in Gaussian case

    A) $X_i \perp X_j$ for all $j \in S^c_i$

    B) $X_i \perp X_j | X_k : k \neq j, j \in (S^i)^c$, where $S^i \equiv \{j : \sigma^{ij} \neq 0\}$
Regularized Estimates

If \( M \equiv ||m_{ij}||, T_t(M) \equiv ||m_{ij}1(|m_{ij}| \geq t)|| \)

- Not positive definite in general
- Permutation invariant

A) \( \hat{\Sigma}_t \equiv T_t(\hat{\Sigma}) \)

B) \( \tilde{\Sigma}_\lambda^{-1} \equiv \arg \min \left\{ \text{trace} \left( \hat{\Sigma}M \right) - \log(\det(M)) + \lambda \sum_{i \neq j} |\sigma^{ij}| \right\}, \) 
  \( t, \lambda \) chosen by cross validation
Regularized Estimates


In Gaussian case

\[ \| \hat{\Sigma}_t - \Sigma \| = O(p \sqrt{\frac{\log d}{n}}), \quad t \asymp A \sqrt{\frac{\log d}{n}}. \]

\[ \| \cdot \| \equiv \text{operator norm} \Rightarrow \text{rate holds for inverses, eigen structures} \]

\[ \| \hat{\Sigma}_\lambda - \Sigma \|_F = O(p \sqrt{\frac{\log d}{n}}), \quad t \asymp A \sqrt{\frac{\log d}{n}}, \]

where \( \| M \|_F^2 \equiv \sum_{i,j} m_{ij}^2 \) : Frobenius norm

Frobenius \geq \text{operator}
**Example 1**

- \( \mathbf{X}_k = \{X(i, j) : i = \text{latitude}, \ j = \text{longitude}\} \)
- \( \mathbf{X}_{n \times d} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T \)
- \( \mathbb{E}(\mathbf{X}) = \mathbf{0} \)
- \( \Sigma = \text{Var}(\mathbf{X}) = \mathbb{E}\left(\mathbf{XX}^T\right) = \sum_{j=1}^{d} \lambda_j \mathbf{e}_j \mathbf{e}_j^T \)
- \( \mathbf{e}_1, \ldots, \mathbf{e}_d : \text{Principal components.} \)
- \( \lambda_1 > \cdots > \lambda_d : \text{Eigenvalues.} \)
- **Goal**: Estimate, interpret \( \mathbf{e}_j, j = 1, \ldots, K \)
  such that \( \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{d} \lambda_j} \) large.
Figures

EOF pattern #1 (field)

EOF PATTERN #1 (CLIM.PACT)
**Figures**

**EOF pattern #2 (Thresholding)**
Figures

EOF pattern #1 (thresholding)
Explanation II

**Sparse Data (SD)**

Regression:

$X$ lives on (or close) to $\mathbb{M}$ an $m$-dimensional Riemannian manifold $X = T_{d \times 1}(U), U \in O \subset \mathbb{R}^m, m << d, T$ 1-1, smooth!

or $|X - T(U)| \leq \varepsilon$ for some $U$.

**Note:** Special case

If $\mathbb{M}$ is a hyperplane this leads to principal components regression.
**Sparse Data and/or Sparse Models**

**SD: Sparse Data**

- Example: Classification 2 classes
  
  \[ p_j(x) = \text{density of } X \text{ given } Y = y \]
  
  \[ \frac{p_1(x)}{p_0} = \frac{q_1(u)}{q_0} \]
Sparse Data and/or Sparse Models

Consequence of SD

Let \( p(y|x) \equiv \) density of \( Y|X = x \) with respect to \( \mu \) a function on \( \mathbb{R}^{d+1} \).

If \( \varepsilon = 0 \),

\[ p(y|x) = q(y|u), \; u = T^{-1}(x) \]

where \( q \) is conditional of \( Y|U = u \) with respect to \( \mu T^{-1} \)

SUGGESTS: \( \hat{p}(y|x) \approx \hat{q}(y|u) \)

Estimation of regression function on \( \mathbb{R}^{d+1} \) is in fact estimation on \( \mathbb{R}^{m+1} \).
Sparse Data and/or Sparse Models

Consequence of SD

If $\hat{p}_1, \hat{p}_0$ are kernel estimates of $p_1, p_0$ with bandwidth $h$ based on $X_{10}, \ldots, X_{n0}, X_{11}, \ldots, X_{n1}$

$$\frac{\hat{p}_1}{\hat{p}_0}(x) = \frac{\hat{P}_1[|X - x| \leq h]}{\hat{P}_0[|X - x| \leq h]}$$

and if $M$ is flat,

$$\approx \frac{\# \{i : |U_{i0} - T^{-1}(x)| \leq h\}}{\# \{i : |U_{i1} - T^{-1}(x)| \leq h\}}$$

for $nh \to \infty$, $h \to 0$

where $U_{ij} = T^{-1}(X_{ij})$, $j = 0, 1$.

This is the uniform kernel density ratio estimate based on $\{U_{ij}\}$

NB: $h$ needs to be related to $m$ Not $d$.

Choice: By V fold cross validation and/or estimation of local dimension.

∴ Difficulty generated by $m$ not $d$. 
Choice of Tuning Parameters

- $h$: here.
- $\lambda$: LASSO and Dantzig.
- $t$: Thresholding in covariance matrix estimation.

Stable Method: $V$-fold cross validation.

See

- L. Gy’orfi et al. (2002)
Dimension Estimation

• Previous discussion suggests manifold knowledge not needed for regression but matters for clustering

• Black box: $n$ points in $\mathbb{R}^d$ in $\rightarrow n$ points in $\mathbb{R}^m$ out, with $m < d$. In a smooth way

• Intuition: Locally
  • everything is linear
  • geodesic distances between neighbors are preserved

• Major embedding algorithms
  • Locally Linear Embedding (Roweis and Saul (2000))
  • Isomap (Tenenbaum et al. (2000))
  • Laplacian Eigenmaps (Belkin and Niyogi (2002))
  • Hessian Eigenmaps (Donoho and Grimes (2003))

• Not clear how to determine $m$

• Plausible that $m = m(x)$ (local dimension)
Local Dimension Estimation Methods

- Methods based on nearest $k$ neighbours

If $X_1, \ldots, X_n$ are an i.i.d. sample from a density $f(x)$ in $\mathbb{R}^m$, then

$$\frac{k}{n} \approx f(x)V(d)T_k(x)^d$$

- $V(d)$ is the volume of the unit sphere in $\mathbb{R}^d$,
- $T_k(x)$ is the Euclidean distance from $x$ to its $k$-th nearest neighbor.

I: Estimate dimension by regressing $\log T_k$ on $\log k$ (Pettis et al. (1979))
II: A Maximum Likelihood Estimator of Local Dimension

**Idea:** fix a point $x$, assume $f(x) \approx \text{const.}$ in a small sphere of radius $r = \frac{1}{\nu}, \nu = o(n^{-1/d})$ and treat the observations as a homogeneous Poisson process.

- $X_i = g(Y_i) \in \mathbb{R}^d, Y_i$ are sampled from an unknown density $f$ on $\mathbb{R}^m$, with unknown $m \leq d$; $g$ is a smooth manifold mapping (locally).
- At $x$, approximate the binomial process $\{N(t), 0 \leq t \leq r\}$

$$N(t) = \sum_{i=1}^{n} 1\{||X_i - x|| \leq t\}$$

by a Poisson process with rate $\lambda(t) = f(x)V(m)mt^{m-1}$ and log-likelihood (letting $f(x) = e^{\theta}$)

$$L(m, \theta) = \int_0^R \log \lambda(t)dN(t) - \int_0^R \lambda(t)dt.$$
A Maximum Likelihood Estimator of Intrinsic Dimension

- Nice exponential family; MLE for $m$

\[
\hat{m}_R(x) = \left[ \frac{1}{N(R,x)} \sum_{j=1}^{N(R,x)} \log \frac{R}{T_j(x)} \right]^{-1}
\]

- More convenient in practice: fix $k$ NN

\[
\hat{m}_k(x) = \left[ \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{R_k(x)}{T_j(x)} \right]^{-1}
\]

- Unless local or cluster estimates are desired, average over points

\[
\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(X_i), \quad \hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{m}^k
\]

- The MLE of $\theta = \log f(x)$ can be used to estimate entropy.
A Little Theory

(1) If $m$ is fixed, $k \to \infty$, $\frac{k}{n} \to 0$, $f(x) > 0$ and continuous at $x$, then

$$\sqrt{k} (\hat{m}_k(x) - m) \Rightarrow N(0, m^2)$$

(2)

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{m}_k(X_i) - m \right) \Rightarrow N(0, \sigma^2(m))$$

at least for $\frac{k^{3/2}}{n} \to \infty$

(3) If $m \to \infty$ and $n \to \infty$, (1),(2) hold only if $\forall \varepsilon > 0$

$$\frac{\varepsilon^m n \pi^{m/2}}{\Gamma \left( \frac{m}{2} + 1 \right)} \rightarrow \infty$$

$$\Rightarrow \frac{\log n}{m} - \frac{\log m}{2} > 0$$
## Dimension of Selected Data Sets

**Zip code digits\(^{(1)}\)**

<table>
<thead>
<tr>
<th>Digit</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>9</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>(\hat{k})</td>
<td>52</td>
<td>40</td>
<td>143</td>
<td>76</td>
<td>112</td>
<td>69</td>
<td>79</td>
<td>75</td>
<td>59</td>
<td>39</td>
</tr>
</tbody>
</table>

(1) Hastie, Tibshirani, Friedman (2001) *Elements of Statistical Learning*(obtained from other sources)

\[ n = 1000 - 3000 \text{ per digit.} \]
Another View of Sparsity

Friedman & Tukey (1974, IEEE Computing Transactions)
Huber (1983) AS

If $p$ is large, almost all univariate projections of $X$ are Gaussian.
A Different Generalization (Kawanabe, Blanchard & Spokoiny et. al. 2006, JMLR)

\[ X_{p \times 1} = S + \Delta, \quad S \in \mathcal{V}, \quad \mathcal{V} \text{ a linear subspace}, \dim(\mathcal{V}) = r < p \]

\[ \Delta \sim \mathcal{N}(0, \Gamma), \quad S \perp \Delta. \]

- An equivalent formulation:
  \[ X_{p \times 1} = S + \epsilon \]
  \[ P[S \in \mathcal{V}] = 1, \epsilon \text{ Gaussian and } \epsilon \in \mathcal{V}^T, \text{ where} \]
  \[ \mathcal{V} = [v_1, \ldots, v_r] \]
  \[ \mathcal{V} \text{ unknown.} \]


- Can identify projection operator \( \Pi_{\mathcal{V}} : \mathbb{R}^p \to \mathcal{V}. \)
Noisy ICA

- \( \mathbf{X} = A\mathbf{\epsilon} + \Delta, \quad \Delta \sim \mathcal{N}(0, \Gamma) \perp \mathbf{\epsilon}. \)
- \( \mathbf{\epsilon} \) non-Gaussian \( \rightarrow \) Independent components.
- If \( \text{Rank}(A) < p \)

**Conjecture:**

Can estimate \( A \) efficiently.

**Note:** If \( \Delta = 0, \text{Rank}(A) = p \), efficient estimation possible (Chen and B (AS 2006)).
Conjecture (Special case $r = 1$, general case follows)


Let $P_v$ be the true distribution of

$$
\frac{v^T(X - \mu)}{\sigma_v},
$$

where $\mu = \mathbb{E}X$ and $\sigma_v^2 = \text{Var}(v^T X)$ and let $p_v$ be the corresponding density.

Let $\hat{P}_v$ be the empirical distribution of

$$
\frac{v^T(X_i - \bar{X})}{\hat{\sigma}_v}, \quad i = 1, \ldots, n,
$$

where $\hat{\sigma}_v^2$ is the empirical variance of $v^T X_i$. 
Let \( \hat{p}_v \) be an estimate of \( p_v \) such that \( \int \log \hat{p}_v d\hat{P}_v \) is an efficient estimator of \( \int \log p_v dP_v \).

Let

\[
\hat{V} = \arg \max_{|v|=1} \int \log \frac{\hat{p}_v}{\varphi} d\hat{P}_v,
\]

where \( \varphi \) is the standard Gaussian density. Then \( \hat{V} \) is efficient.

Conjecture is supported by theorems 14.1 and 14.2 of Hyvarinen et. al.
Review

• Least favorable analysis:
  One can do nothing for $p$ large compared to $n$.

• “Sparsity” in:
  a) models
  b) Data representation
  as key to doing well.

• Examples:
  regression
  covariance estimation
  projection pursuit

• Partial validation: dimension estimation.