Seeking Interpretable Models for High Dimensional Data

Bin Yu

Statistics Department, EECS Department
University of California, Berkeley

http://www.stat.berkeley.edu/~binyu
Characteristics of Modern Data Sets

- **Goal: efficient use of data for:**
  - Prediction
  - Interpretation (proxy: sparsity)

- **Larger number of variables:**
  - Number of variables ($p$) in data sets is large
  - Sample sizes ($n$) have not increased at same pace

- **Scientific opportunities:**
  - New findings in different scientific fields
Today’s Talk

- Understanding early visual cortex area V1 through fMRI
- Occam’s Razor
- Lasso: linear regression and Gaussian graphical models
- Discovering compressive property of V1 through shared non-linear sparsity
- Future work
Understanding visual pathway

Gallant Lab at UCB is a leading vision lab.

Da Vinci (1452-1519)  Mapping of different visual cortex areas

Polyak (1957)  Small left middle grey area: V1
Understanding visual pathway through fMRI

One goal at Gallant Lab: understand how natural images relate to fMRI signals
Mind-reading with a brain scan

Brain activity can be decoded using magnetic resonance imaging.

Kerri Smith

Scientists have developed a way of ‘decoding’ someone’s brain activity to determine what they are looking at.

“The problem is analogous to the classic ‘pick a card, any card’ magic trick,” says Jack Gallant, a neuroscientist at the University of California in Berkeley, who led the study.
Stimuli

- Natural image stimuli
Stimulus to fMRI response

- Natural image stimuli drawn randomly from a database of 11,499 images
- Experiment designed so that response from different presentations are nearly independent
- fMRI response is pre-processed and roughly Gaussian
Gabor Wavelet Pyramid

(a) Spatial frequency
- 1 cycle/FOV
- 2 cycles/FOV
- 4 cycles/FOV
- 8 cycles/FOV
- 16 cycles/FOV

(b) Orientation
- 0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°

Phase
- 0°, 90°
Features

\[ \log(|| \cdot ||^2 + 1) \rightarrow X_1 \]

\[ \log(|| \cdot ||^2 + 1) \rightarrow X_{10921} \]

Raw stimulus

Project onto Complex Gabor Wavelet Basis

Log energy of each channel

\( \mathbf{X} \) (10,921 dimensions)
"Neural" (fMRI) encoding for visual cortex V1

Predictor: \( p = 10,921 \) features of an image
Response: (preprocessed) fMRI signal at a voxel
\( n = 1750 \) samples

**Goal:** understanding human visual system
interpretable (sparse) model desired
good prediction is necessary

Minimization of an empirical loss (e.g. L2) leads to

- ill-posed computational problem, and
- bad prediction
Linear Encoding Model by Gallant Lab

Data

\[ Z_i = (X_i, Y_i), \quad i = 1, \ldots, n \]

- \( X_i \) (predictor vector) \( \in \mathbb{R}^p \)
- \( Y_i \) (response) \( \in \mathbb{R}^1 \) or \( \{-1, 1\} \)

- \( X \): p=10921 dimensions (features)
- \( Y \): fMRI signal
- \( n = 1750 \) training samples

Separate linear model for each voxel via e-L2boosting (or Lasso)

\[ Y = X\beta + \epsilon \]

Fitted model tested on 120 validation samples

- Performance measured by correlation
Modeling “history” at Gallant Lab

- Prediction on validation set is the benchmark
- Methods tried: neural nets, SVMs, e-L2boosting (Lasso)
- Among models with similar predictions, simpler (sparser) models by e-L2boosting are preferred for interpretation

This practice reflects a general trend in statistical machine learning -- moving from prediction to simpler/sparser models for interpretation, faster computation or data transmission.
Occam’s Razor

14th-century English logician and Franciscan friar, William of Ockham

Principle of Parsimony:

Entities must not be multiplied beyond necessity.
Occam’s Razor via Model Selection in Linear Regression

\[ Z_i = (X_i, Y_i), \ i = 1, ..., n \]
\[ X_i \text{ (predictor vector)} \in \mathbb{R}^p \]
\[ Y_i \text{ (response)} \in \mathbb{R}^1 \text{ or } \{-1, 1\} \]

\[ Y = X\beta + \epsilon \]

- Maximum likelihood (ML) is LS when Gaussian assumption
- There are \(2^p\) submodels
- ML goes for the largest submodel with all predictors
- Largest model often gives bad prediction for \(p\) large
Akaike (73,74) and Mallows’ Cp used estimated prediction error to choose a model:

\[ AIC(\mathcal{M}) = ||Y - X\hat{\beta}(\mathcal{M})||^2 + 2\text{dim}(\mathcal{M})\sigma^2 \]

Schwartz (1980):

\[ BIC(\mathcal{M}) = ||Y - X\hat{\beta}(\mathcal{M})||^2 + \log n \times \text{dim}(\mathcal{M})\sigma^2 \]

Both are penalized LS by $L_0$.

Rissanen’s Minimum Description Length (MDL) principle gives rise to many different different criteria. The two-part code leads to BIC.
Model Selection for image-fMRI problem

For the linear encoding model, the number of submodels

\[ 2^{11000} \approx 2 \times 10^{3000} \]

Combinatorial search: too expensive and often not necessary

A recent alternative:

continuous embedding into a convex optimization problem through L1 penalized LS (Lasso) -- a third generation computational method in statistics or machine learning.
The $L_1$ penalty is defined for coefficients $\beta$:

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$$

**Used initially with $L_2$ loss:**

- Signal processing: Basis Pursuit (Chen & Donoho, 1994)
- Statistics: Non-Negative Garrote (Breiman, 1995)
- Statistics: LASSO (Tibshirani, 1996)

$$\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \|\beta\|_1 \right\}.$$ 

**Properties of Lasso**

- Sparsity (variable selection) and regularization
- Convexity (convex relaxation of $L_0$-penalty)
Lasso: computation and evaluation

The “right” tuning parameter unknown so “path” is needed (discretized or continuous)

Initially: quadratic program for each a grid on \( \lambda \). QP is called for each \( \lambda \).


Theoretical studies: much work recently on Lasso in terms of L2 prediction error L2 error of parameter model selection consistency
Model Selection Consistency of Lasso

Set-up:

Linear regression model

\[ Y = X \cdot \beta + \epsilon \]

\[ p \times 1 \quad n \times p \quad p \times 1 \quad n \times 1 \]

\( n \) observations and \( p \) predictors

\[ X = (X_1, X_2), \quad X_1 : n \times s, \quad X_2 : n \times (p - s), \]
\[ \beta = (\beta_1, \ldots, \beta_s, 0, 0, \ldots 0)' \in \mathbb{R}^p \]

Assume

\( (A): \quad X'X/n \to C > 0, \quad \max_i(||X_i||^2)/n \to 0, \quad \lambda_n = o(n). \)

Knight and Fu (2000) showed L2 estimation consistency under (A).
Model Selection Consistency of Lasso

- \( p \) small, \( n \) large (Zhao and Y, 2006), assume (A) and
  \[ \lambda_n = n^{1/2+c}, \ c \in (0, 1/2) \]

Then roughly* Irrepresentable condition (1 by \( p-s \) matrix)

\[ |\text{sign}((\beta_1, \ldots, \beta_s))(X'_1 X_1)^{-1} X'_1 X_2| < 1 \]

* Some ambiguity when equality holds.

- Related work: Tropp(06), Meinshausen and Buhlmann (06), Zou (06), Wainwright (06)

Population version

\[ |\text{sign}((\beta_1, \ldots, \beta_s))\Sigma_{SS}^{-1}\Sigma_{SS^c}^{-1}| < 1, \]
\[ S = \{1, \ldots, s\}, \ S^c = \{s + 1, \ldots, p\} \]
Irrepresentable condition (s=2, p=3): geometry

\[
\Sigma = \begin{pmatrix}
1 & 0 & r \\
0 & 1 & r \\
r & r & 1 \\
\end{pmatrix}, \quad |r(\text{sign}(\beta_1) + \text{sign}(\beta_2))| < 1
\]

\[\beta_1 = 1, \beta_2 = 1, \beta_3 = 0\]

\[\beta_1 = 1, \beta_2 = -1, \beta_3 = 0\]

- \(r=0.4\)
- \(r=0.6\)
Model Selection Consistency of Lasso

- Consistency holds also for \( s \) and \( p \) growing with \( n \), assume

  irrepsentable condition
  bounds on max and min eigenvalues of design matrix
  smallest nonzero coefficient bounded away from zero.

Gaussian noise (Wainwright, 06):

\[
    n \sim s \log p, \quad \text{or} \quad p \sim \exp(n/s)
\]

Finite 2k-th moment noise (Zhao&Y,06):

\[
    n > p^{c/k}, \quad \text{or} \quad p < n^{c'k}
\]
Consistency of Lasso for Model Selection

- **Interpretation of Condition** – Regressing the irrelevant predictors on the relevant predictors. If the $L_1$ norm of regression coefficients

\[
\| (X_1'X_1)^{-1}X_1'X_2 \|_1 < 1
\]

- **Larger** than 1, Lasso **can not** distinguish the irrelevant predictor from the relevant predictors for some parameter values.
- **Smaller** than 1, Lasso **can** distinguish the irrelevant predictor from the relevant predictors.

- **Sufficient Conditions (Verifiable)**
  - Constant correlation
  \[
  \text{corr}(X_i, X_j) = \rho_0 \in [0, 1)
  \]
  - Power decay correlation
  \[
  \text{corr}(X_i, X_i) = \rho^{|i-j|}
  \]
  - Bounded correlation*
  \[
  |\text{corr}(X_i, X_j)| < \frac{1}{2s - 1}
  \]
L1 penalized log Gaussian Likelihood

Given \( n \) iid observations of \( X \) with

\[
X = (X_1, \ldots, X_p) \in \mathbb{R}^p, \text{ mean zero, } \Sigma^* = EX'X, \Theta^* := \Sigma^{-1}
\]

Banerjee, El Ghaoui, d’Aspremont (08):

\[
\hat{\Theta} := \arg\max_{\Theta > 0} \left[ -\log \text{lik}_{\text{Gaussian}}(\Theta) + \lambda_n \|\Theta\|_{1,\text{off}} \right]
\]

by a block descent algorithm.
Ravikumar, Wainwright, Raskutti, Yu (08) gives sufficient conditions for model selection consistency.

Hessian:

\[ \Gamma^* := \nabla^2_{\Theta}[-\log \det(\Theta)] \bigg|_{\Theta = \Theta^*} = \Theta^{*-1} \otimes \Theta^{*-1} \]

\[ K_{\Gamma^*} = \|\|(\Gamma^*_{SS})^{-1}\|\|_{\infty} \quad K_{\Sigma^*} = \|\|\Sigma^*\|\|_{\infty} = \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\Sigma_{ij}^*| \]

\[ \theta_{\min}^* := \min_{i,j} |\Theta_{ij}| \]

Define “model complexity”:

\[ K := \max\{K_{\Sigma^*}^2 K_{\Gamma^*}, K_{\Sigma^*}^4 K_{\Gamma^*}^2, K_{\Sigma^*} K_{\Gamma^*}/\theta_{\min}^*\} \]
Assume the irrepresentable condition below holds

\[ \|\| (\Gamma^*_{SS})^{-1} \Gamma^*_{SS^c} \|\|_{\infty} \leq (1 - \alpha) \in [0, 1) \]

1. \( X \) sub-Gaussian with parameter \( \sigma \) and effective sample size \( n / \log p > f_1(K, \alpha, \sigma) \)

Or

2. \( X \) has 4\( m \)-th moment, \( \tau > 2 \)

\[ n/p^{\tau/m} > f_2(K, \alpha, m) \]

Then with high probability as \( n \) tends to infinity, the correct model is chosen.
Success prob’s dependence on n and p (Gaussian)

- Edge covariances as \( \Sigma_{ij}^* = 0.1 \). Each point is an average over 100 trials.
- Curves stack up in second plot, so that \((n/\log p)\) controls model selection.
Success prob’s dependence on “model complexity” $K$ and $n$

- Curves from left to right have increasing values of $K$.
- Models with larger $K$ thus require more samples $n$ for same probability of success.
Back to image-fMRI problem: Linear sparse encoding model on complex “cells”

Gallant Lab’s approach:

- Separate linear model for each voxel
  - \[ Y = Xb + e \]
- Model fitting via e-L2boosting and stopping by CV
  - \( X: p=10921 \) dimensions (features or complex “cells”)
  - \( n = 1750 \) training samples
- Fitted model tested on 120 validation samples (not used in fitting)
  
**Performance measured by correlation \((cc)\)**
Adding nonlinearity via Sparse Additive Models

- Additive Models (Hastie and Tibshirani, 1990):

\[ Y_i = \sum_{j=1}^{p} f_j (X_{ij}) + \varepsilon_i, \quad i = 1, \ldots, n \]

- Sparse: \( f_j \equiv 0 \) for most \( j \)

- High dimensional: \( p \gg n \)

SpAM (Sparse Additive Models)

Sparse Additive Models (SpAM)  
(Ravikumar, Lafferty, Liu and Wasserman, 07)

One Attempt at Sparsity in High Dimensions

Optimization: minimize \[ \mathbb{E} \left( Y - \sum_j \beta_j f_j(X_j) \right)^2 \]
subject to \[ \| \beta \|_1 \leq L_n \]
\[ \mathbb{E}(f_j^2) = 1, \quad \mathbb{E}(f_j) = 0 \]

Problem: Not convex

Equivalent Convex Formulation:

Optimization: minimize \[ \mathbb{E} \left( Y - \sum_j f_j(X_j) \right)^2 \]
subject to \[ \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2)} \leq L_n \]
\[ \mathbb{E}(f_j) = 0 \]
Sparse Backfitting

**Input:** Data \((X_i, Y_i)\), regularization parameter \(\lambda\).

**Iterate until convergence:**

For each \(j = 1, \ldots, p\):

- **Compute residual:** \(R_j = Y - \sum_{k \neq j} \hat{f}_k(X_k)\)

- **Estimate projection** \(f_j = \mathbb{E}(R_j \mid X_j)\), smooth: \(\hat{f}_j = S_j R_j\)

- **Estimate norm:** \(s_j = \sqrt{\mathbb{E}[f_j]^2}\)

- **Soft-threshold:** \(\hat{f}_j \leftarrow \left[1 - \frac{\lambda}{\hat{s}_j}\right]^+ \hat{f}_j\)

**Output:** Estimator \(\hat{f}(X_i) = \sum_j \hat{f}_j(X_{ij})\).
Simple Cell and Complex Cell Models for V1 Neurons

- Image input
  - Gabor wavelet
  - Non-negative rectification
  - Output

- Image input
  - Gabor wavelet quadrature pair
  - Squaring
  - Fixed nonlinearity
  - Output
Pooled-Complex “Cell” Model

Allows more flexible fitting, Pooling: aggregated neural responses in fMRI?
SpAM V1 Model (Ravikumar et al, NIPS08)

Connections and components in dashed region are to be estimated, under the assumption that many of them are null.
SpAM V1 encoding model (1331 voxels from V1)

For each voxel,

- Start with 10921+ complex cells (features)
- **Pre-selection** of 100 complex cells via correlation
- Fitting of **SpAM to 100 complex cells** with AIC stopping
- **Pooling** of SpAM fitted complex cells according to location and frequency to form pooled complex cells
- Fitting **SpAM to 100 complex cells and pooled complex cells** with AIC stopping
Prediction performance ($R^2$)

Median improvement 12%.
Spatial RFs and Tuning Curves
Nonlinearities

Compressive effect (finite energy supply)

Common nonlinearity for each voxel?
Shared Nonlinearity

- Shared V-SpAM model (Ravikumar et al, 08):

\[ Y = \sum_{j=1}^{p} \alpha_j f(X_j) + \varepsilon \]

where \( J = \{ \alpha_j \neq 0 \} \) is small or sparse.

\[
\text{minimize}_{\{f \in L_2[a,b], \alpha\}} \quad E \left( Y - \sum_j \alpha_j f(X_j) \right)^2 \\
\text{subject to} \quad \sum_j |\alpha_j| \leq L_n \\
\sum_j E \left( f^2(X_j) \right) \leq K_n
\]
Nonlinearities vs Shared-Nonlinearity

Compressive effect (finite energy supply)

Common nonlinearity for each voxel?
Shared-nonlinearity vs linearity: $R^2$ prediction

Median improvement 16%
Shared-nonlinearity vs nonlinearity: $R^2$ prediction

Median improvement 4.9%
Shared-nonlinearity vs. nonlinearity: sparsity

Recall: 100 original features pre-selected.

On average:

V-SpAM: 70 predictors (original and pooled)

Shared V-SpAM: sparser (13 predictors) and better prediction
On-going: summarizing shared nonlinearity (compressive effect)
On-going: mapping voxels on cortex V1 area
Summary

- L1 penalized minimization model selection consistency irreppresentable condition came from KKT and L1 penalty

  Effective sample size n/logp? Not always. Depending on the tail of relevant distribution.

  Model complexity parameters matter.

- Understanding fMRI signals in V1

  Discovered shared nonlinear compressive effect of fMRI response

  Supporting evidence: improved prediction with sparser models biologically sensible
Future work

- **V1 voxels:**
  - *improved decoding?* (to decode images seen by the subject using fMRI signals)
  - *strength borrowing?* across voxels for linear and nonlinear models.

- **Higher vision area voxels:** *benchmark model does not work well.*
  - *hope: improved encoding models* (hence making decoding possible)
    - via nonlinearity/interaction and borrowing strength

  **V4 challenge:** *how to build feedback* into the modeling.

- **Better image representation**
Acknowledgements

Co-authors:
P. Zhao and N. Meinshausen (Lasso)
P. Ravikumar, M. Wainwright, G. Raskutti (Graphical model)
P. Ravikumar, V. Vu, K. Kay, T. Naleris, J. Gallant
  (V-SpAM and shared V-SpAM)

Funding:
NSF-IGMS Grant (06-07 in Gallant Lab)
Guggenheim Fellowship (06)
NSF DMS Grant
ARO Grant