



fMRI classification: Smart use of LASSO to win Ridge regression

Yongxin Taylor Xi

Ramadge Group

NIAM meeting Apr 25, 2008



Overview

- Linear regression model for binary classifier
- Regularization
 - Ridge regression (L2 norm)
 - geometry interpretation as maximizing margin
 - advantage and disadvantage
 - Lasso (L1 norm)
 - sparse solutions: ability to ignore irrelevant features
 - no good to be applied directly in voxel domain
- How to use Lasso in a sensible way?
 - Before Lasso: clustering the voxels
 - Before Lasso: dimensionality reduction such as PCA
- Experimental results
 - fMRI data, Yale face database

Linear regression model for binary classifier

- Say we have n training examples:

Data matrix X

$$X^T = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, x_i \in \mathbf{R}^p$$

Label vector Y

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Translate data to center the origin (X_c), then solve:

$$\min_b \|X_c^T b - Y\|^2$$

Note: For fMRI data with $n < p$, there exist many solutions of b such that $\|X_c^T b - Y\|^2 = 0$

Regularization (I)

- Ridge regression

Penalize the L2 norm of the solution:

$$\min_b \|X_c^T b - Y\|^2 + \lambda \|b\|_2^2$$

Intuition: choose $\min_b \|b\|_2$ such that $\|X_c^T b - Y\|^2 = 0$

Geometry:

$$x_+^T b / \|b\|_2 = 1 / \|b\|_2$$

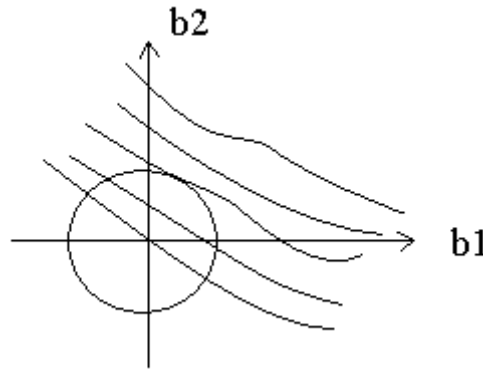
$$x_-^T b / \|b\|_2 = -1 / \|b\|_2$$

Margin: $2 / \|b\|_2$

So, maximizing margin = minimizing L2 norm

Disadvantage of ridge regression

- Solution not sparse
 - Linear combination of all features, including the noisy ones



- Why does it usually generalize pretty well?
 - The idea of maximizing margin
 - The average effect of noisy features
- Can we get rid of noisy features?

Regularization (II)

- Lasso

Penalize the L1 norm of the solution:

$$\min_b \|X_c^T b - Y\|^2 + \lambda \|b\|_1$$

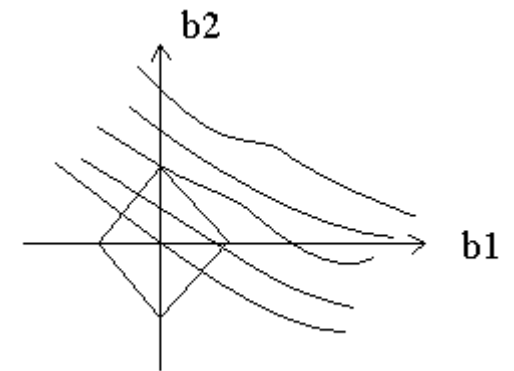
Assume $\text{rank}(X_c) = n-1$

Then b has at most $(n-1)$ nonzero entries.

- Sparse yet unstable when $n \ll p$

– $n \sim 100$

$p \sim 1,000-10,000$



How to use Lasso in a sensible way?

- Lasso is reported tackling noise and yielding good generalization in many applications. So what's going wrong here?
- Feature space huge. Single feature unreliable. Similar features greatly exist.
- Idea: clustering the features!

Way I: Cluster voxels

- Similarity measure: correlation between the voxel behavior vectors (normalized) along the time line
- Say cluster into N groups
- Obtain the exemplar of each group (linear combination of group members)
- Now we have N features. Run Lasso.

Way II: dimensionality reduction

- PCA: $p \rightarrow (n-1)$
- lossless data-driven transform
- The idea of variance: cognitive state, subject, noisy sources
- Basis criteria: maximum variance and orthogonality
- Now we have $(n-1)$ features. Run Lasso.

Way II (continue)

- Solve:
$$\min_b \|X_c^T P b - Y\|^2$$
$$s.t. \|b\|_1 \leq r$$
- P is PCA transform matrix
 - size: $p \times (n-1)$
- b is solution vector
 - size: $(n-1) \times 1$
- Lasso asks for sparsity of the modes

Experiments: data

- Data: monkeydog data (no time series)
- 392 examples
 - = 7 categories x 8 rounds x 7 subjects
- $p = 2048$ voxels from IT cortex
- 7 category: woman face, man face, monkey face, dog face, house, chair, and shoe. Solution is 7 vectors b_1, \dots, b_7 .
- 2 category: face v. s. object
- Goal: linear classification with good prediction

Train for each subject

- PCA+LASSO

$$\min_b \|X_c^T P b - Y\|^2$$

$$s.t. \# \text{ nonzeros } (b) \leq m \leq n-1$$

56 examples / subject

Use 55 to train

Test on the rest

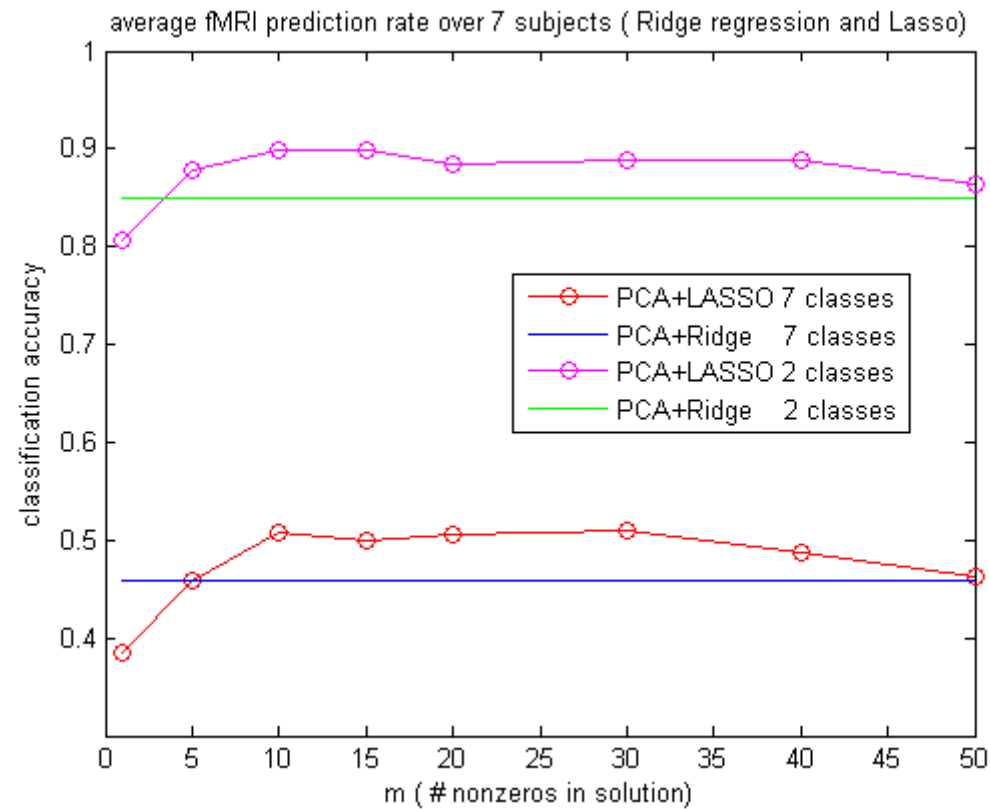
- $n=55$; P : 2048×54 ; b : 54×1
- Choose m to be 1, 5, 10, 20, 30, 40, 50

- PCA+Ridge

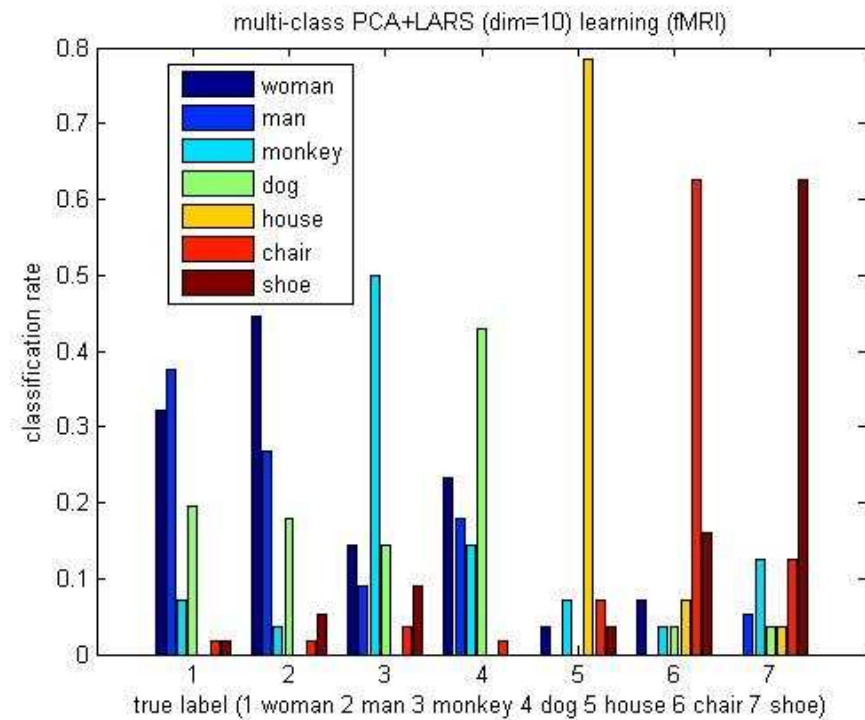
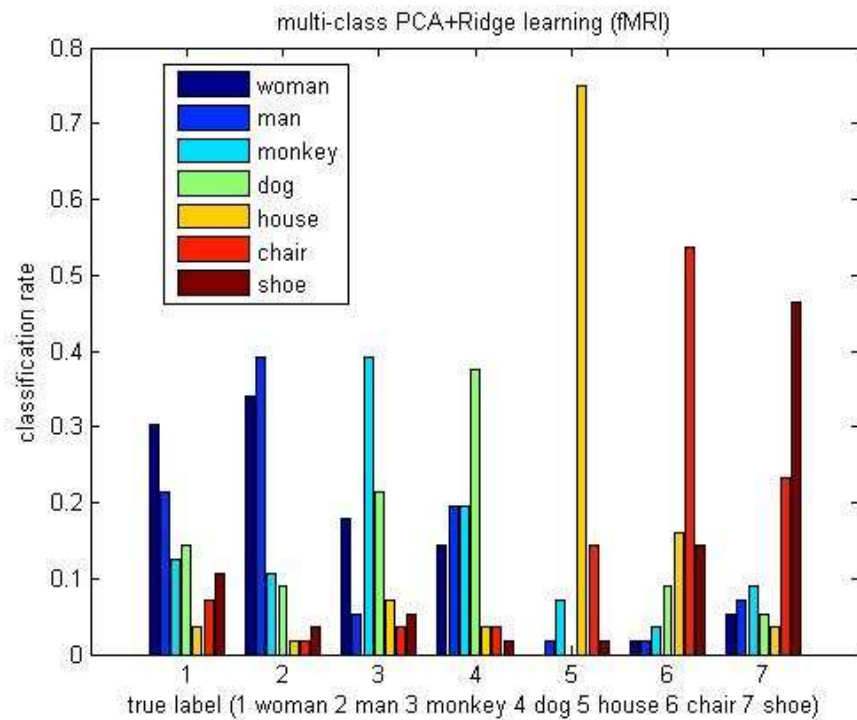
$$\min_b \|X_c^T P b - Y\|^2 + \lambda \|b\|_2^2$$

- $\# \text{ nonzeros } (b)=54$
- $\text{Lamda}=0.1, 1, 5, 10, 15, 20$

Leave-one-out prediction result



Diffusion maps



Advantage

- Help exclude noisy features
- Offer interpretations of modes/clusters
- Run fast using LARS (speed: on the same order of Ridge regression)

Looking ahead:

Clustering+LASSO may work even better!

Interpretation of modes:

FUN EXPERIMENTS WITH FACES

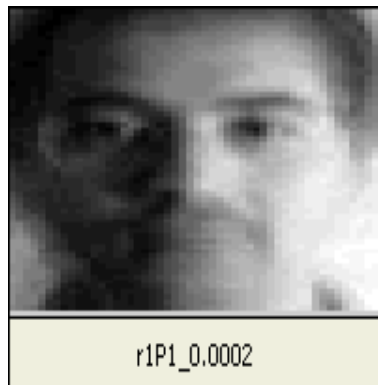
Yale Face Database

- 15 people, 11 images each
- center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.



Lighting classification

—light from left or right?



Expression classification

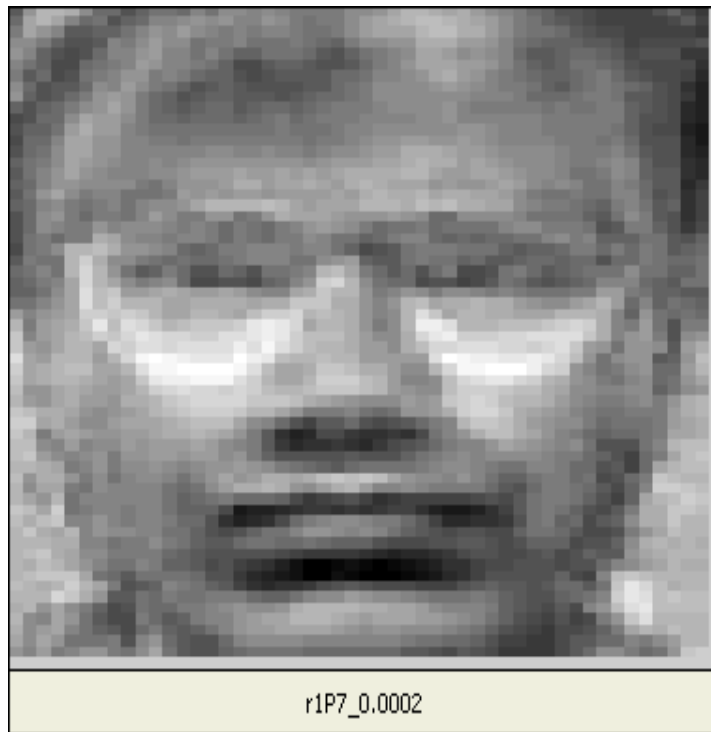
—is this person happy or not?



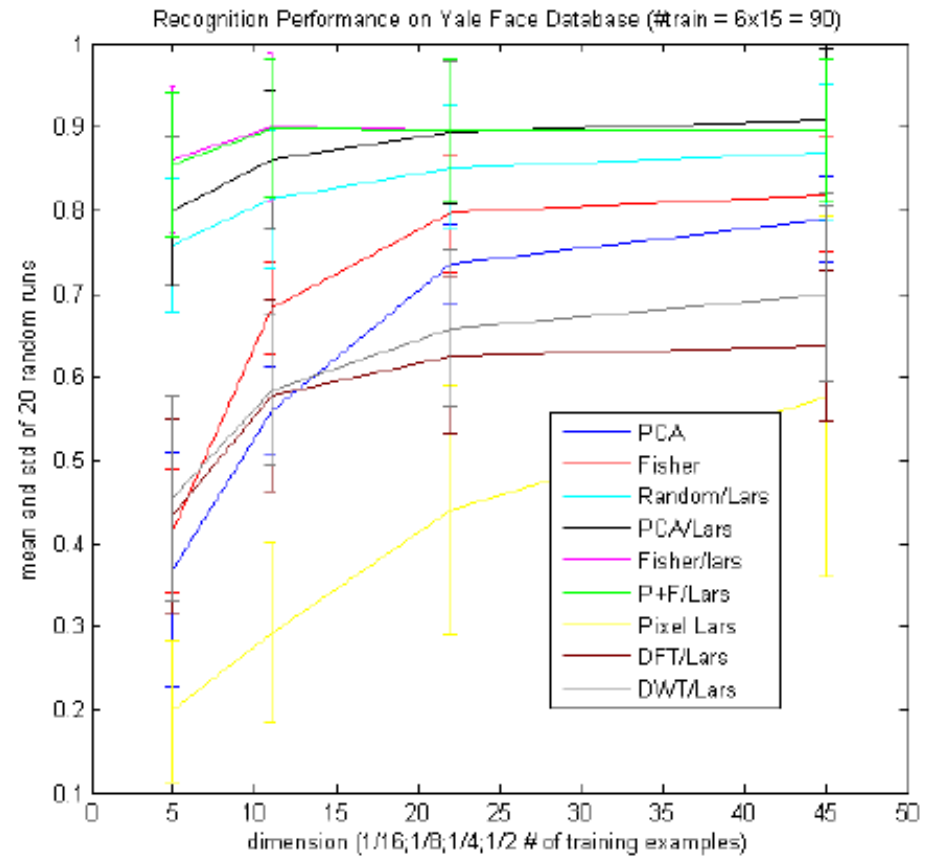
Fun continues...

Eye-glasses detection

—wearing glasses or not?



Face recognition rate (15 category)



Thank You

Questions?

