Sparse Modeling

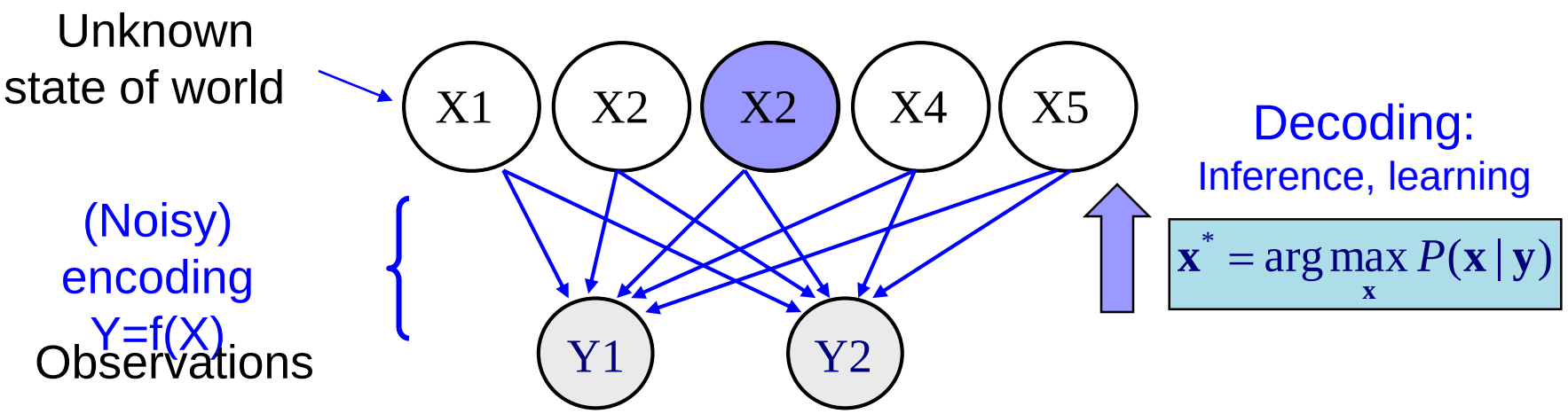
Theory, Algorithms and Applications

Irina Rish
Computational Biology Center (CBC)
IBM T.J. Watson Research Center, NY

Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

A Common Problem



Can we recover a high-dimensional X from a low-dimensional Y?

Yes, if:

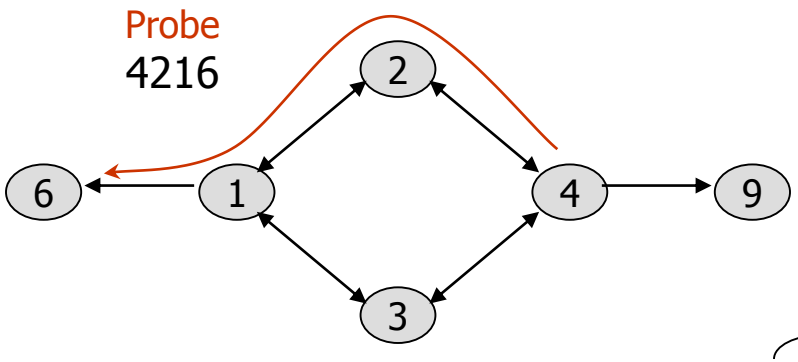
- X is structured; e.g., sparse (few $X_i \neq 0$) or compressible (few large X_i)
- encoding preserves information about X

Examples:

- **Sparse signal recovery** (compressed sensing, rare-event diagnosis)
- **Sparse model learning**

Example 1: Diagnosis in Computer Networks

(Beygelzimer, Kephart and Rish 2007)



N links (possible bottlenecks)

vector Y of end-to-end probe delays

vector X of (unknown) link delays

Routing matrix A

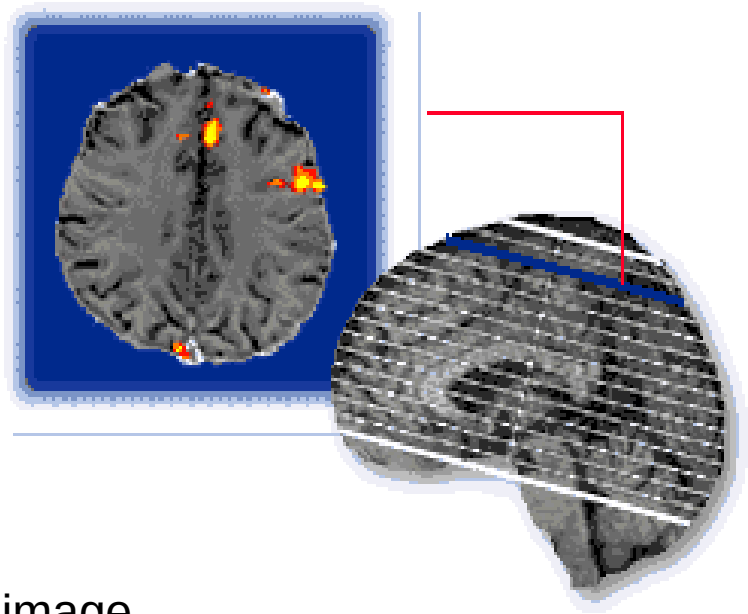
M probes

Y		X1	X2	X3	X4	X5	X6
		❶❷	❶❸	❶❻	❷❹	❸❹	❹❾
Y1	❹❷❶❻	1	0	1	1	0	0
Y2	❹❸❶❻	0	1	1	0	1	0
Y3	❸❶❷	1	1	0	0	0	0
Y4	❶❷❹❾	1	0	0	1	0	1

- **Model:** $y = Ax + \text{noise}$
- **Problem structure:** **X is nearly sparse - small number of large delays**
- **Task:** find bottlenecks (extremely slow links) using probes ($M << N$)

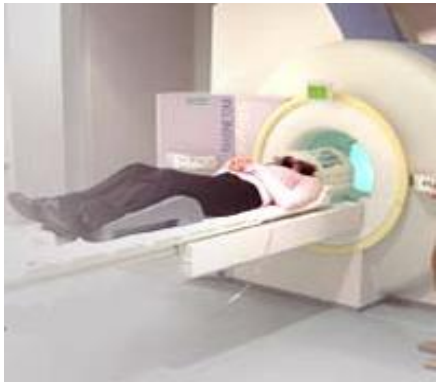
Recover sparse state (`signal') X from noisy linear observations

Example 2: Sparse Model Learning from fMRI Data



- Data: high-dimensional, small-sample
  - **10,000 - 100,000 variables** (voxels)
  - **100s of samples** (time points, or TRs)

- Task: given fMRI, predict mental states
  - emotional: angry, happy, anxious, etc.
  - cognitive: reading a sentence vs viewing an image
  - mental disorders (schizophrenia, autism, etc.)



- Issues:
  - **Overfitting:** can we learn a predictive model that generalizes well?
  - **Interpretability:** can we identify brain areas predictive of mental states?

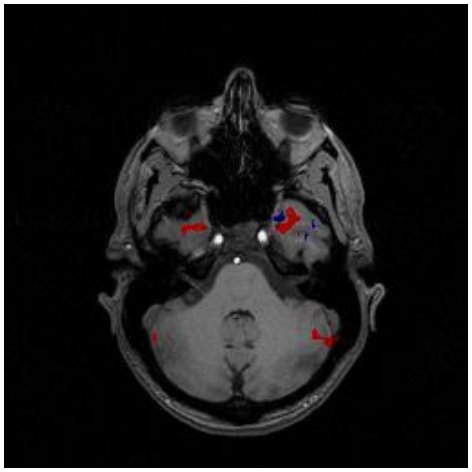
fMRI image courtesy of fMRI Research Center @ Columbia University

Sparse Statistical Models: Prediction + Interpretability

Data

x - fMRI voxels,

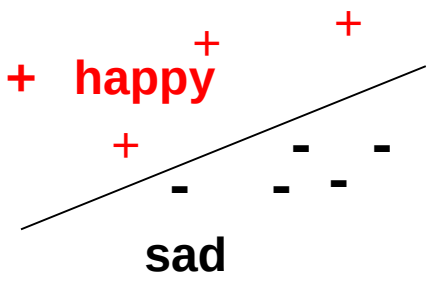
y - mental state



Small number of Predictive Variables ?

Predictive Model

y = f(**x**)



- Sparsity → variable selection → model interpretability
- Sparsity → regularization → less overfitting / better prediction

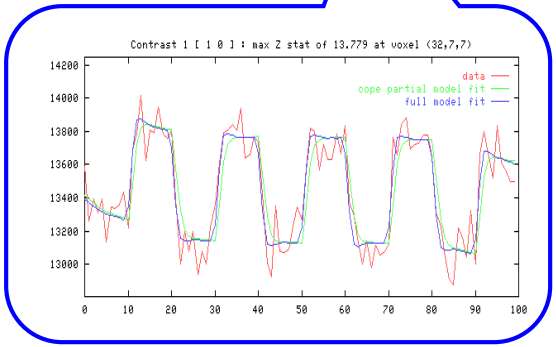
Sparse Linear Regression

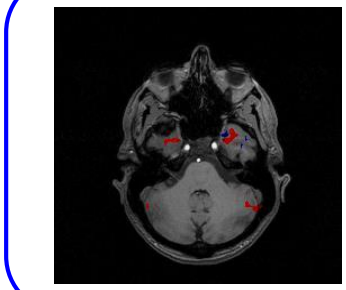
$$y = Ax + \text{noise}$$

Measurements:
mental states, behavior, tasks or stimuli

fMRI data ("encoding")
rows – samples (~500)
Columns – voxels (~30,000)

Unknown parameters ('signal')

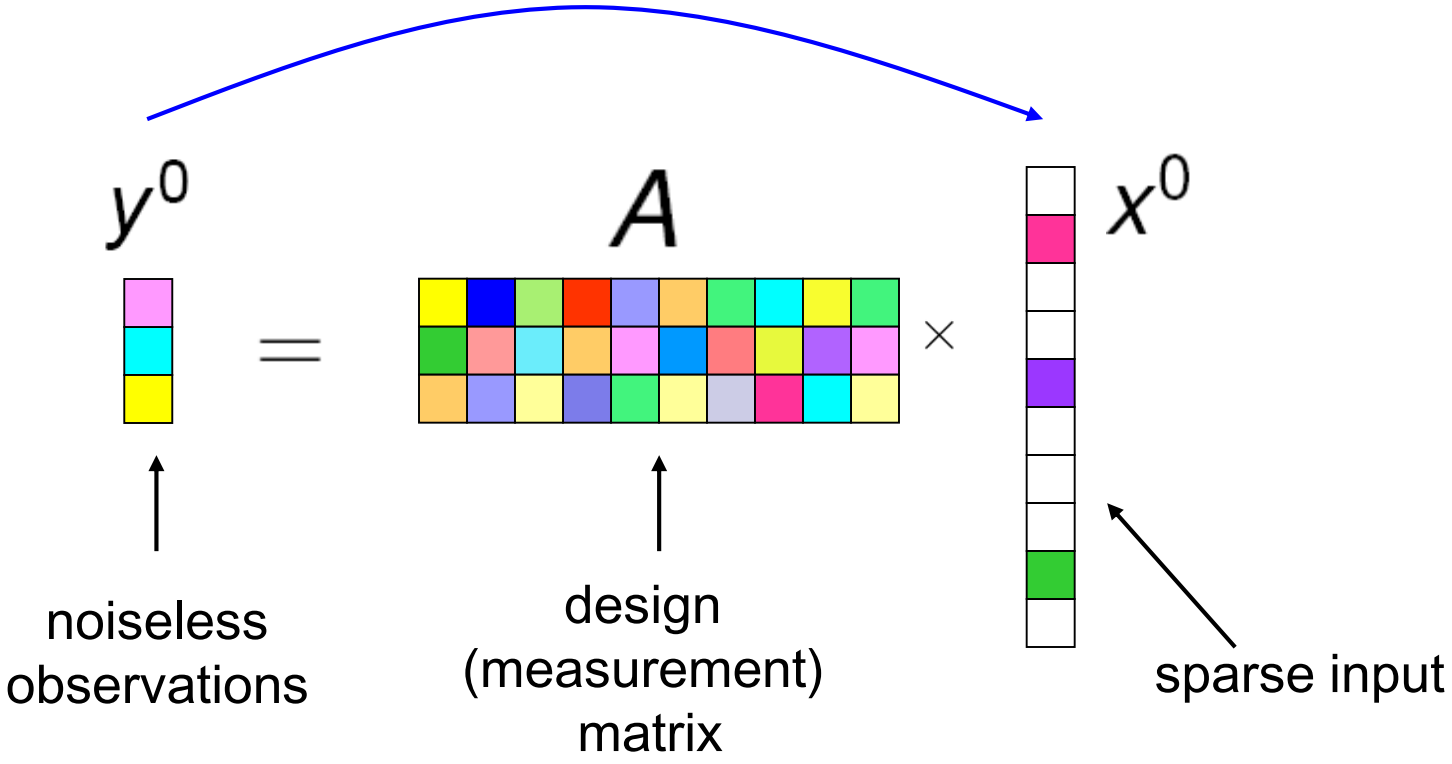




Find small number of most relevant voxels (brain areas)

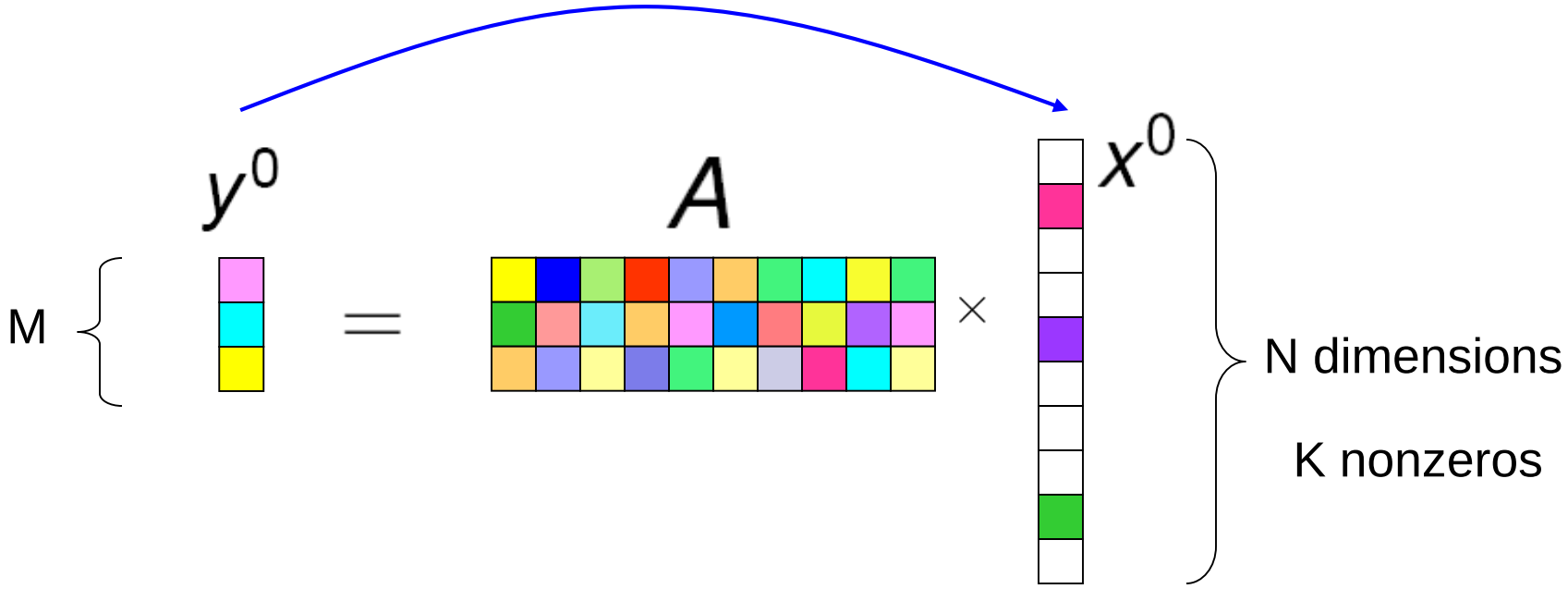
fMRI activation image and time-course courtesy of Steve Smith, FMRIB

Sparse Recovery in a Nutshell



Can we recover a sparse input efficiently from a small number of measurements?

Sparse Recovery in a Nutshell

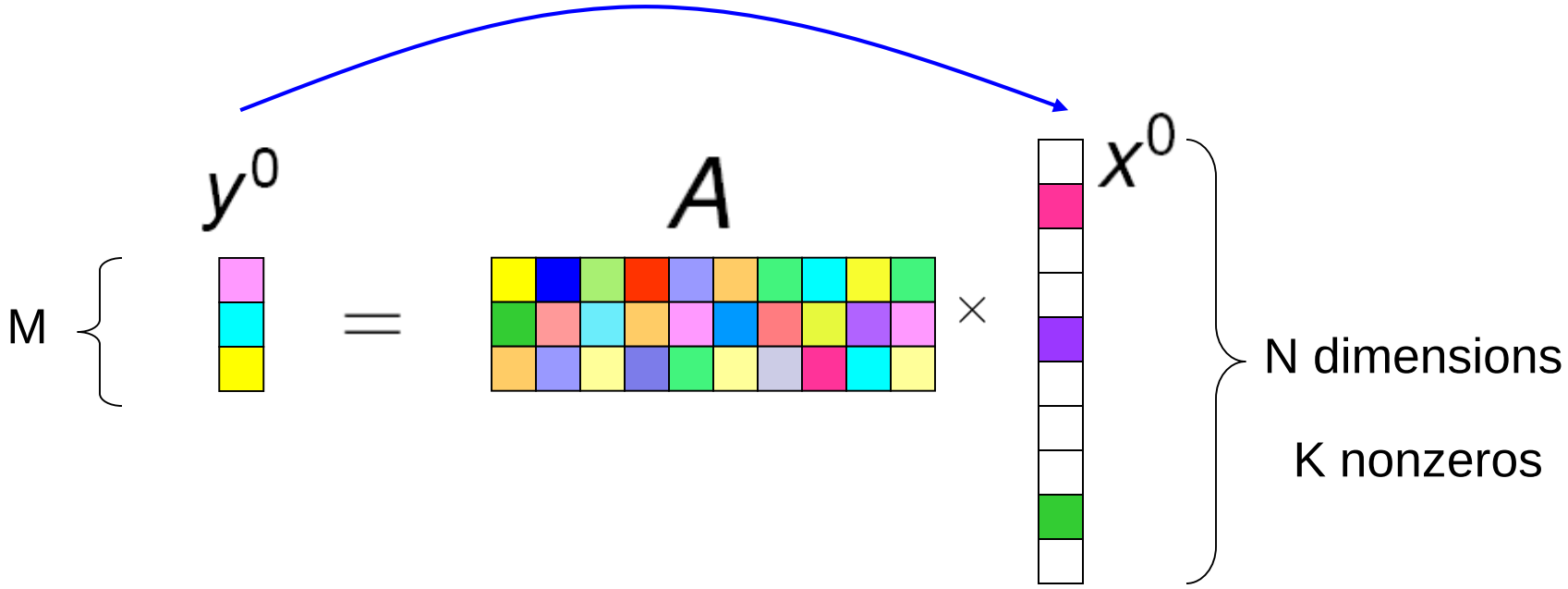


``Compressed Sensing Surprise'':

Given random A (i.i.d. Gaussian entries), x^0 can be reconstructed exactly (with high probability):

- from just $M = O(K \log(N/K))$ measurements
- efficiently - by solving convex problem $\min_x ||x||_1 \; s.t. \;\; y = Ax$
  ($\Leftrightarrow$ linear program)

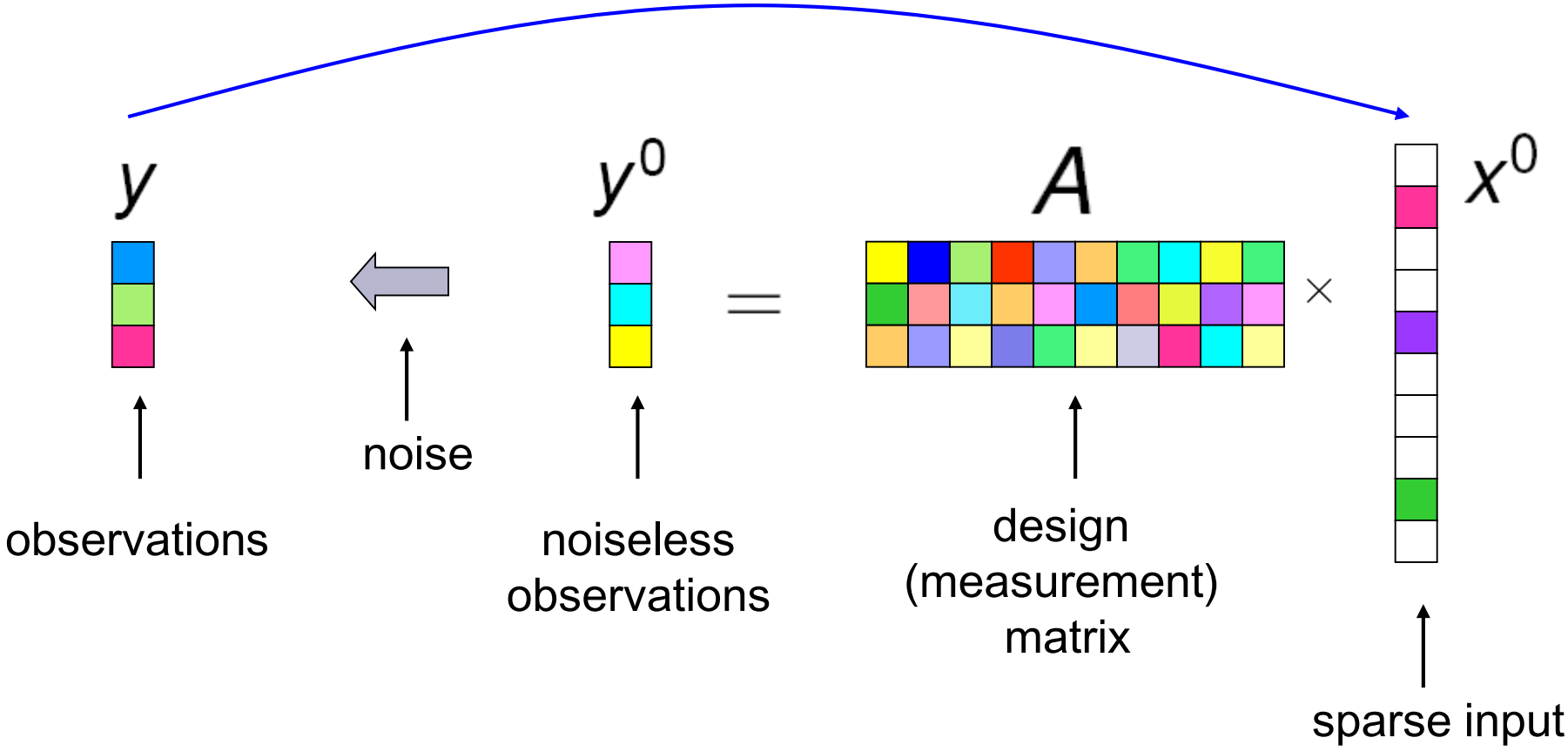
Sparse Recovery in a Nutshell



In general, if A is ``good'' (e.g., satisfies Restricted Isometry Property with a proper constant), sparse x^0 can be reconstructed with M <<N measurements by solving (linear program):

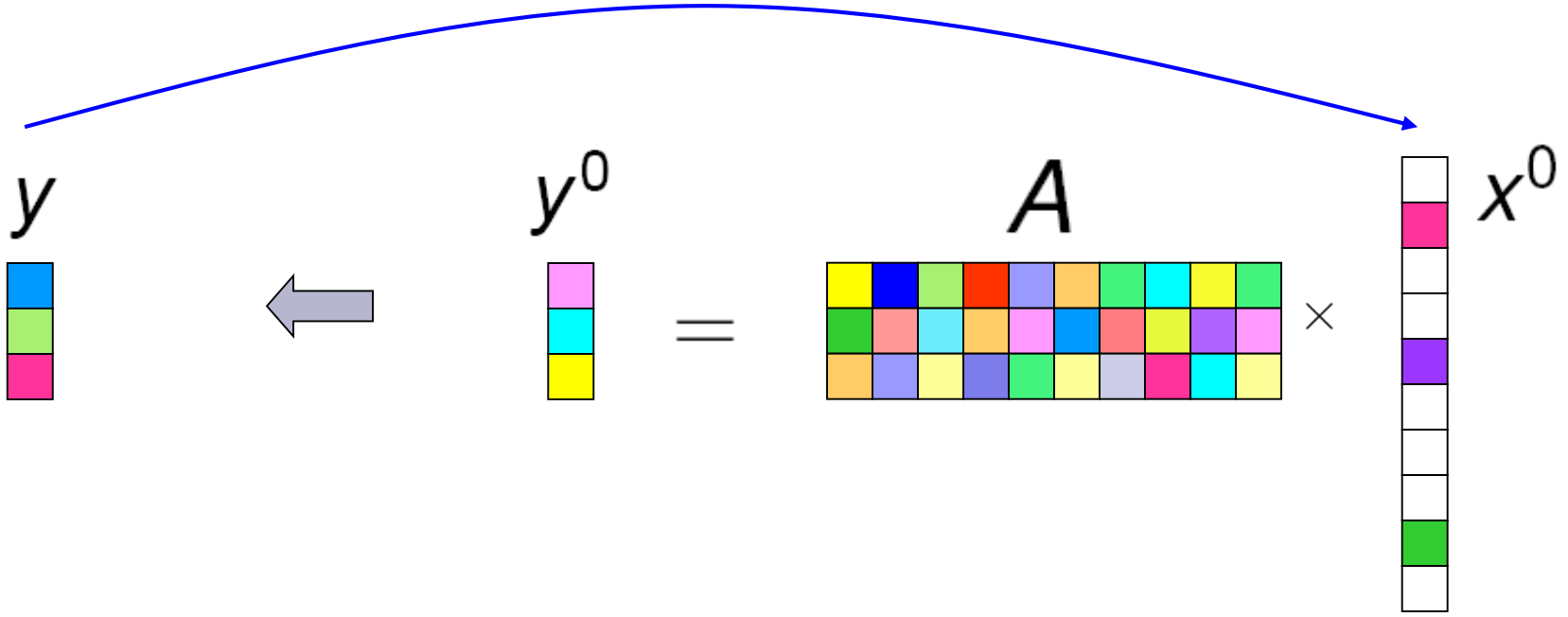
$$\min_x ||x||_1 \; s.t. \;\; y = Ax$$

Sparse Recovery in a Nutshell



And what if there is noise in observations?

Sparse Recovery in a Nutshell



Still, can reconstruct the input accurately (in l_2-sense), for A satisfying RIP; just solve a noisy version of our l_1-optimization:

$$\min_x ||x||_1 \ s.t. \ \ ||y - Ax||_2^2 \leq \epsilon$$

$$\Updownarrow$$

$$\min_x ||y - Ax||_2^2 \ s.t. \ \ ||x||_1 \leq t$$

(Basis Pursuit, aka Lasso)

Sparse Linear Regression vs Sparse Signal Recovery

- Both solve the same optimization problem
- Both share efficient algorithms and theoretical results
- However, sparse learning setting is more challenging:
  - We do not design the "design" matrix, but rather deal with the given data
  - Thus, nice matrix properties may not be satisfied (and they are hard to test on a given matrix, anyway)
  - We don't really know the ground truth (``signal") – but rather assume it is sparse (to interpret and to regularize)
- Sparse learning includes a wide range of problems beyond sparse linear regression (part 2 of this tutorial)

Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

Motivation: Variable Selection

- Filter methods:
  rank each x_i (or a small subset of X) using a ranking function $r(i)$, such as correlation or mutual information with the response y.
  Fast but suboptimal - can miss multivariate predictive patterns.

- Wrapper methods:
  rank each x_i (or a small subset of X) by its predictive accuracy, i.e., train a separate model for each x_i and evaluate its accuracy.
  Wrappers yield better predictions, but are quite expensive.

- Embedded methods:
  variable selection is *embedded* in model learning.
  (E.g., via greedy methods or certain regularization techniques).

Model Selection as Regularized Optimization

Regularization constrains the model space to avoid overfitting:

$$\min_{\beta} L(Z, \beta) \quad s.t.\ R(\beta) \leq t$$

$$\Updownarrow$$

$$\min_{\beta} L(Z, \beta) + \lambda R(\beta)$$

- $Z = \{Z^1, ..., Z^n\}$ - data (e.g., $Z^i = (X_{(i,:)}, y_i)$)
- β - vector of model parameters
- $L(\cdot)$ - loss function (e.g., model's error on the data)
- $R(\cdot)$ - regularization penalty (e.g., model's complexity)
- λ - regularization parameter

Bayesian Interpretation: MAP Estimation

- Loss: negative log-likelihood
- Regularization: negative log-prior on model parameters
- Learning: maximum a posteriori (MAP) probability estimation

$$\arg\max_{\beta} \log P(Z|\beta)P(\beta|\lambda)$$

$$\Updownarrow$$

$$\arg\min_{\beta} -\log P(Z|\beta) - \log P(\beta|\lambda)$$

$$\Updownarrow$$

$$\arg\min_{\beta} L(Z,\beta) + R(\beta,\lambda)$$

Best Subset Selection

- find best subset of M predictors, i.e.

$$\min_{\beta} L(Z, \beta) \quad s.t. \ ||\beta||_0 \leq M$$

where l_0-norm $||\beta||_0$ is the number of nonzeros $|\{i|\beta_i \neq 0\}|$

- NP-hard problem!

- various approximations (mainly greedy):

forward stepwise regression $\Leftrightarrow$ Orthogonal Matching Pursuit (Mallat and Zhang, 1993)
stagewise OMP (StOMP) (Donoho et al., 2006)
regularized OMP (ROMP) (Needell and Vershynin, 2009)
subspace pursuits (Dai and Milenkovic, 2008)
CoSaMP (Needell and Tropp, 2008)
SAMP(Do et al., 2008)
GraDeS (Gradient Descent with Sparsification) (Garg and Khandekar, 2009), etc. etc.

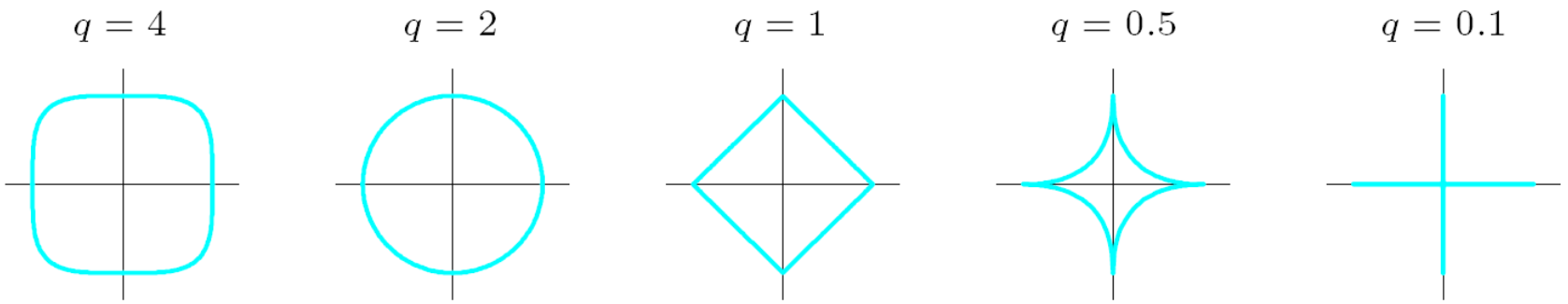
see more at http://dsp.rice.edu/cs (Compressive Sensing Resources)

- Alternative approach:

l_1-norm relaxations of l_0 (or, more generally, l_q-norms, $0 < q \leq 1$)

What is special about l_1-norm? Sparsity + Computational Efficiency

l_q-norm constraints for different values of q



Convexity $\Rightarrow$ efficient optimization methods

Sparsity $\Rightarrow$ variable selection

- $q < 1$: convexity, but no sparsity (no "sharp edges")
- $q > 1$: sparsity (sharp edges), but no convexity
- $q = 1$: sparsity and convexity

Image courtesy of [Hastie, Friedman and Tibshirani, 2009]

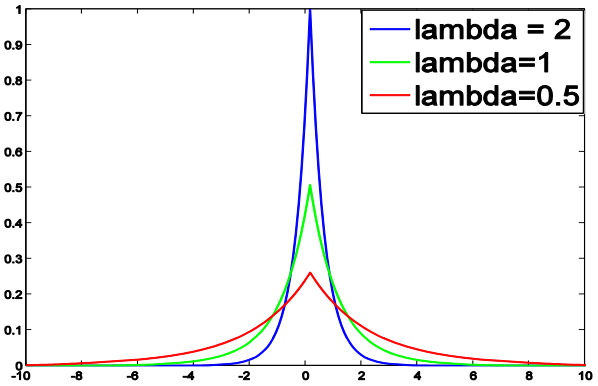
LASSO: Least Absolute Shrinkage and Selection Operator

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

- First proposed by (Tibshirani, 1996)
- Known as Basis Pursuit (Chen et al., 1999) in signal processing
- Bayesian view: MAP estimation with:

  - independent Gaussian observations $y_i \sim e^{-\frac{1}{2}(y - X^i\beta)^2}$ and

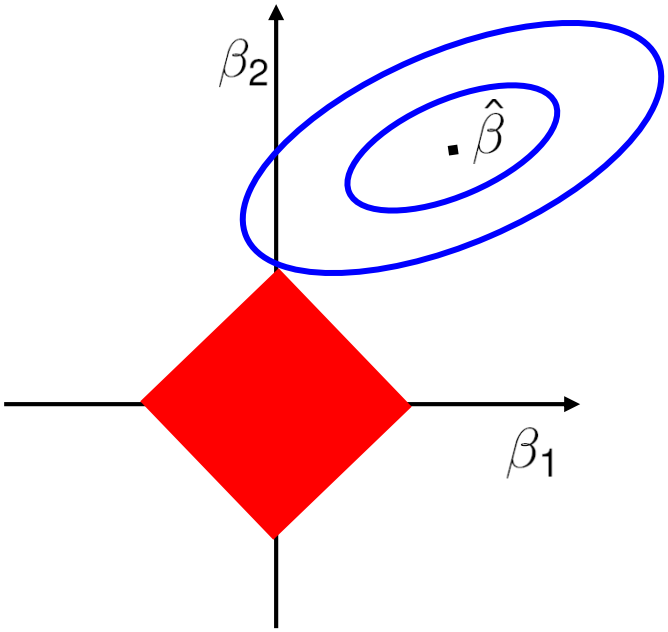
  - independent Laplace parameters $\beta_j \sim e^{-\lambda|\beta_j|}$



- Laplace prior enforces solution sparsity $\Longleftrightarrow$ variable selection

Equivalent Constrained Formulation: A Geometric View

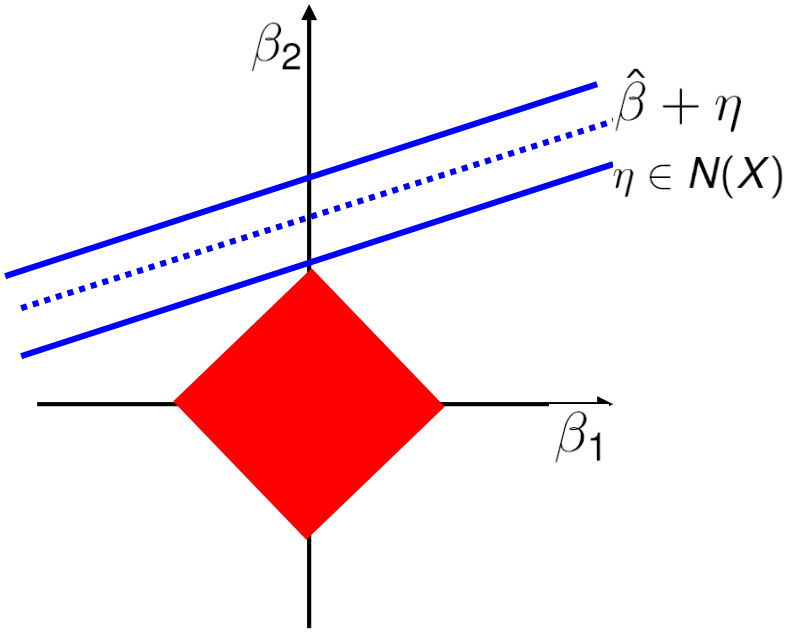
$$\min_{\beta} ||y - X\beta||_2^2 \quad s.t. \ ||\beta||_1 \leq t$$



$p \leq n$

unique OLS solution

$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2$



$p > n$

multiple OLS solutions $\hat{\beta} + \eta$:

$\forall \eta \in N(X)$ (null-space), $y = X(\hat{\beta} + \eta)$

Algorithms

- Standard quadratic programming methods: too slow

- Least Angle Regression (LARS) (Efron et al., 2004):

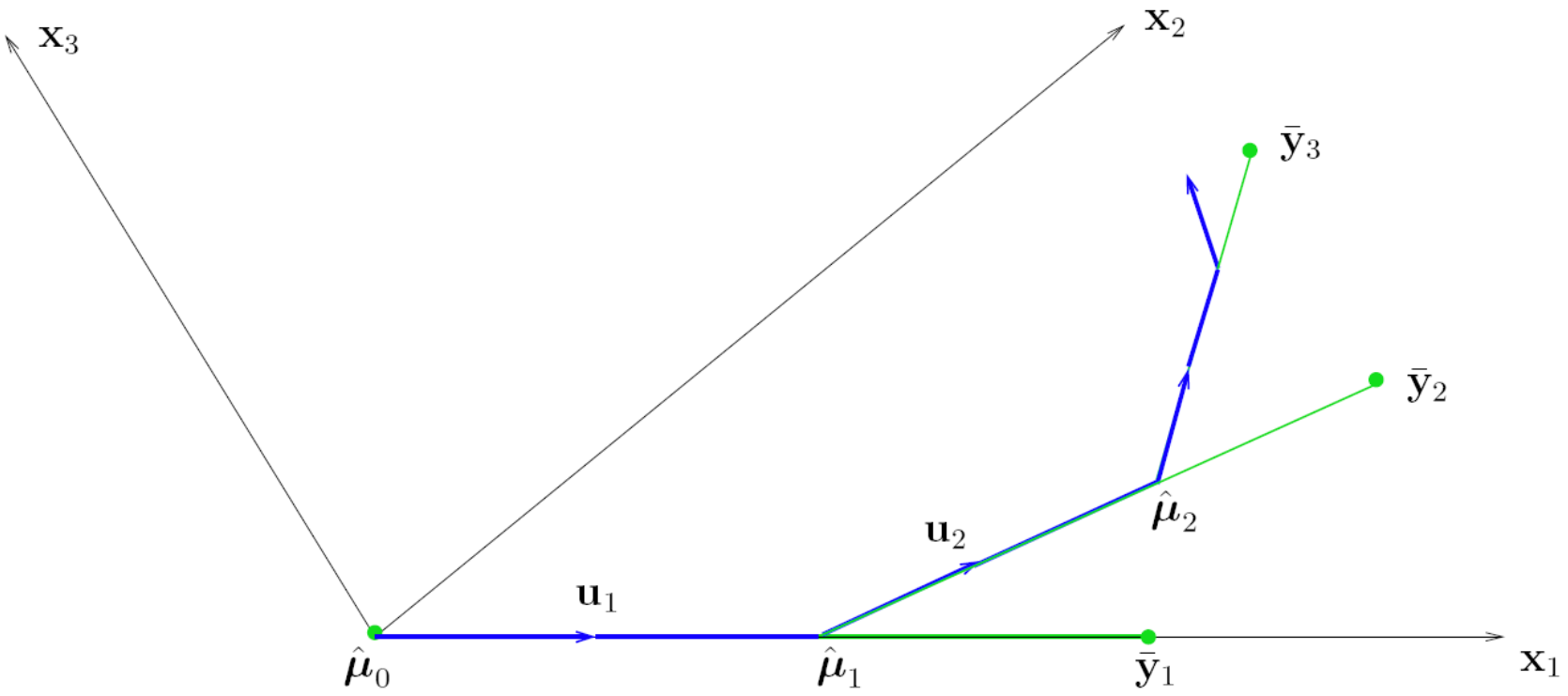
  much faster; moreover, produces the entire solution path (all solutions for all values of the regularization parameter λ) at the cost of a single least-squares fit. Similar to homotopy (continuation) method of (Osborne et al., 2000b).

- Coordinate descent (Fu, 1998), (Daubechies et al., 2004), (Friedman et al., 2007a), (Wu and Lange, 2008):

  for fixed λ, optimizes each parameter at a time; using warm-starts, it can compute the solutions on a grid of λ values faster than LARS (however, the full path is NOT computed)

- Many other methods, including generalizations to other losses; various software packages, e.g., see http://dsp.rice.edu/cs

Geometric View of LARS



At step k, LARS estimate $\mu_{\mathbf{k}}$ moves towards the current OLS estimate $\bar{\mathbf{y}}_k$ in the direction $\mathbf{u_k}$ equiangular among the current predictors.

The direction changes before reaching $\bar{\mathbf{y}}_k$ when a new variable enters the active set.

Image courtesy of [Hastie, 2007]

Predictive Performance

Three scenarios (Tibshirani, 1996):

	Best Subset	Ridge	Lasso
a few large β_i	best	worst	2nd
medium number of moderate β_i	worst	2nd	best
large number of small β_i	worst	best	2nd

Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

Model Selection Consistency of LASSO

- Let X_S be the columns of the nonzero variables in the true model (support), and let X_{S^C} be the remaining columns (complement)

- (Strong) Irrepresentability condition for model selection (Zhao and Yu, 2006a; Yuan and Lin, 2007b; Zou, 2006; Wainwright, 2009b)

$$||(X_S^T X_S)^{-1} X_S^T X_{S^C}||_\infty \leq 1 - \epsilon, \quad \text{for some } 0 < \epsilon \leq 1$$

  states that the least-squares regression coefficients (i.e., correlations) for the non-essential variables (X_{S^C} columns) on support variables in X_S must not be large.

- Relaxing the consistency conditions via Lasso modifications:

- bootstrap Lasso (BOLASSO) Bach (2008a) and stability-selection (Meinshausen and Buehlmann, 2008) use bootstrap approach: learn multiple Lasso models on data subsets, and then include the intersection of nonzeros (Bach, 2008a) or only frequent-enough nonzeros (Meinshausen and Buehlmann, 2008). This gets rid of "unstable" variables and improves the model-selection consistency and stability to the choice of λ parameter.

Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

Beyond LASSO

$$Loss(\mathbf{x}) + \lambda ||\mathbf{x}||_1$$

Other likelihoods (loss functions)

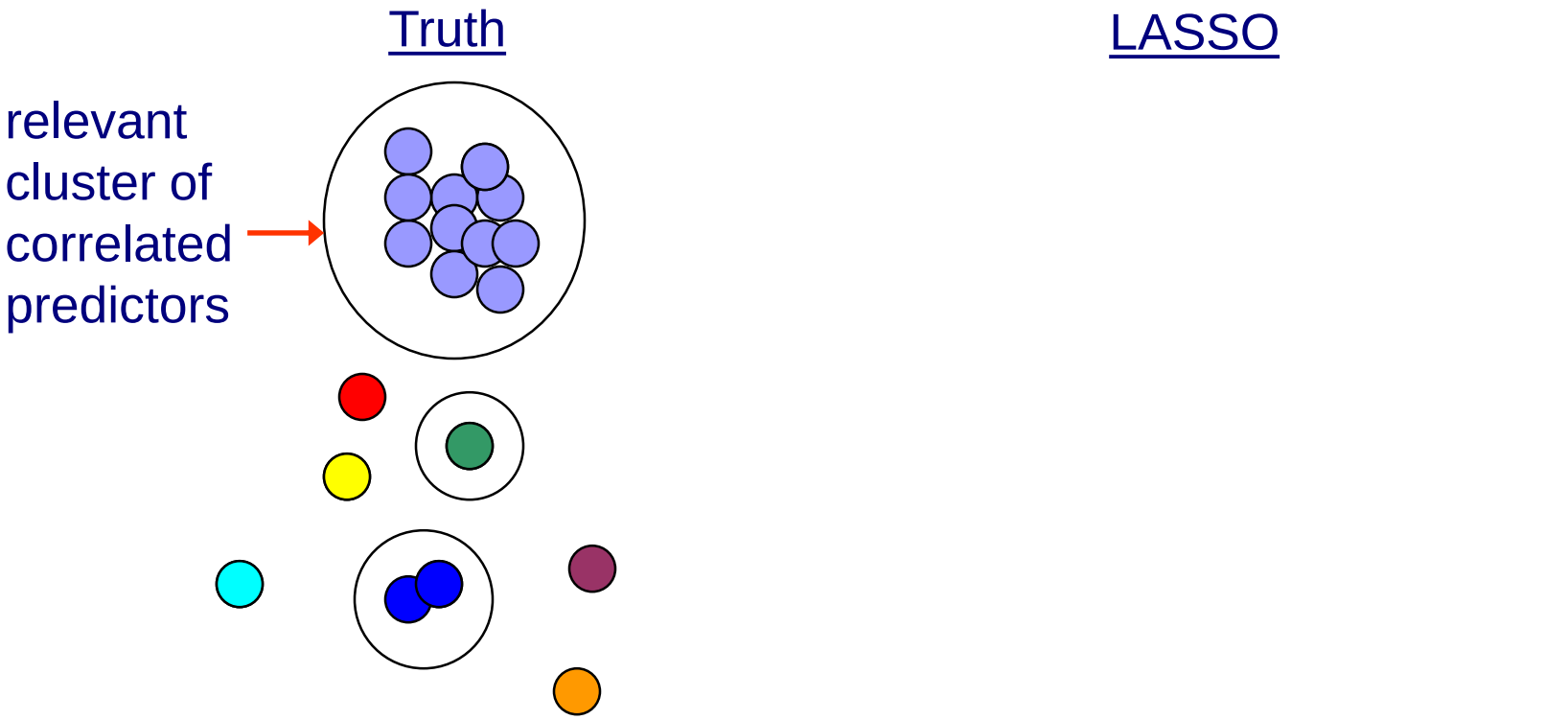
- Generalized Linear Models (exponential family noise)
- Multivariate Gaussians (Gaussian MRFs)

Adding structure beyond sparsity

- Elastic Net
- Fused Lasso
- Block l_1-l_q norms:
  - group Lasso
  - simultaneous Lasso

Some Limitations of LASSO

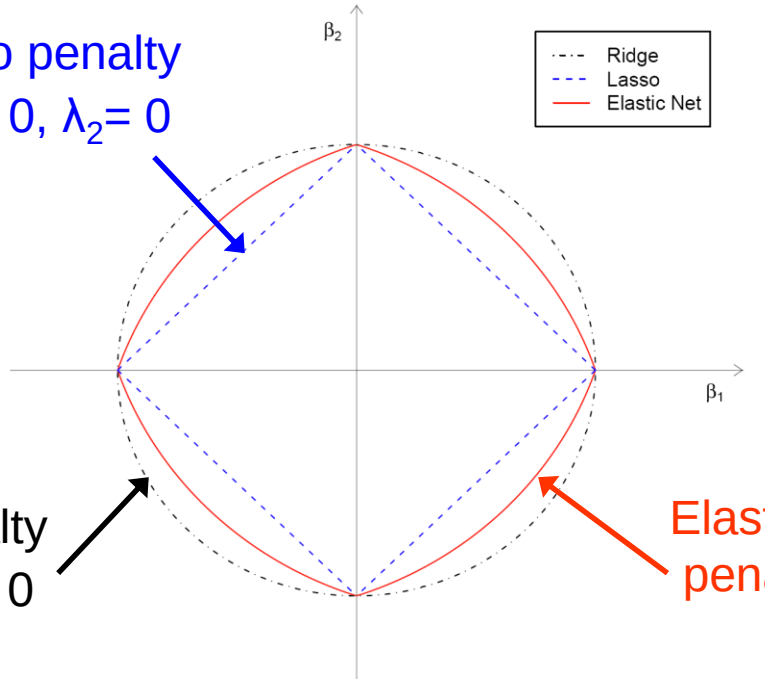
- selects at most n variables when $p > n$ (Osborne et al., 2000)
  (but what if more predictors are relevant?)
- does not group correlated variables (Zou and Hastie, 2005):
  - even if $X_i = X_j$, has many solutions with $\beta_i \neq \beta_j$
  - tends to select one variable out of a group of correlated ones



Elastic Net (Zou and Hastie, 2005)

$$\hat{\beta} = \arg\min_{\beta} ||y - X\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2$$

lasso penalty
$\lambda_1 > 0$, $\lambda_2 = 0$



ridge penalty
$\lambda_1 = 0$, $\lambda_2 > 0$

Elastic Net penalty

Elastic Net penalty:

$$\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1,$$

where $\alpha = \dfrac{\lambda_2}{\lambda_2 + \lambda_1}$

- l_1 keeps singularities at vertices $\Rightarrow$ sparsity
- l_2 enforces strictly convex edges $\Rightarrow$ grouping effect
- l_2 removes the limitation on the number of selected variables

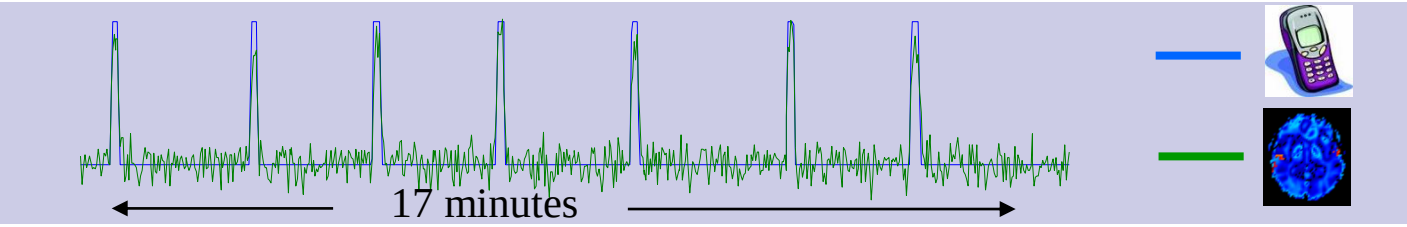
NOTE: to eliminate "double-shrinkage", Elastic Net computes a re-scaled version $(1+\lambda_2)\hat{\beta}$ of the above *naive EN* estimate $\hat{\beta}$

Example: Application to fMRI Analysis

Pittsburgh Brain Activity Interpretation Competition (PBAIC-07):

- subjects playing a videogame in a scanner
- 24 continuous response variables, e.g.
  - Annoyance
  - Sadness
  - Anxiety
  - Dog
  - Faces
  - Instructions
  - Correct hits

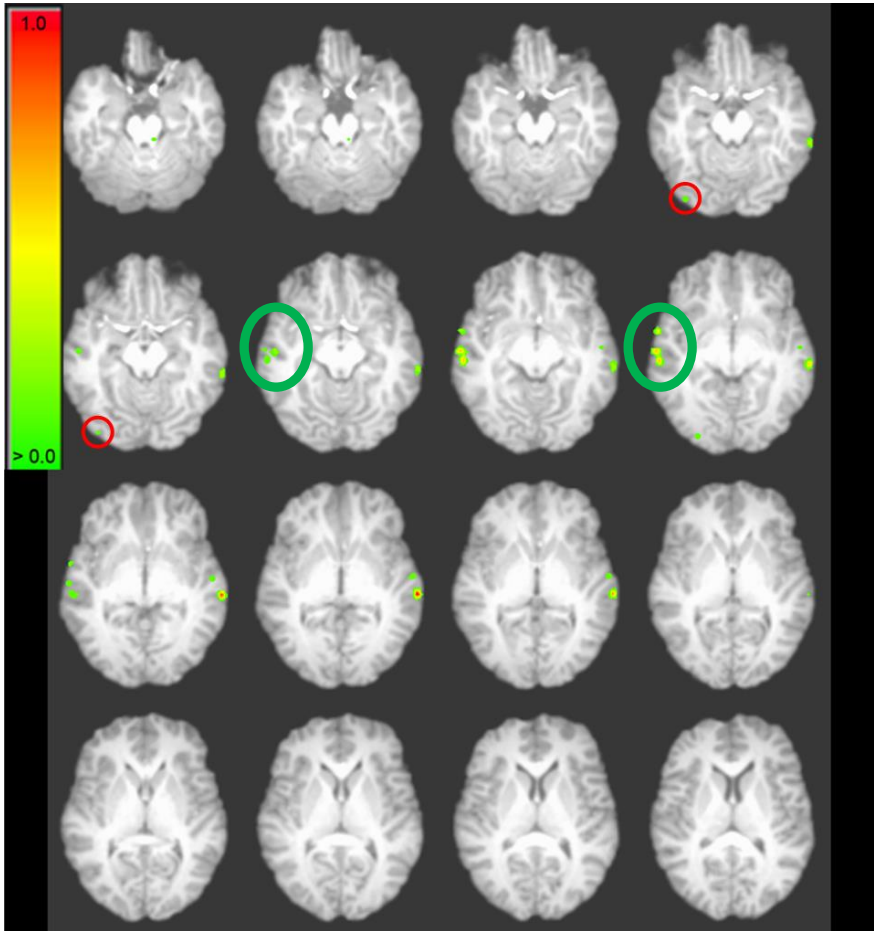
Goal: predict responses from fMRI data



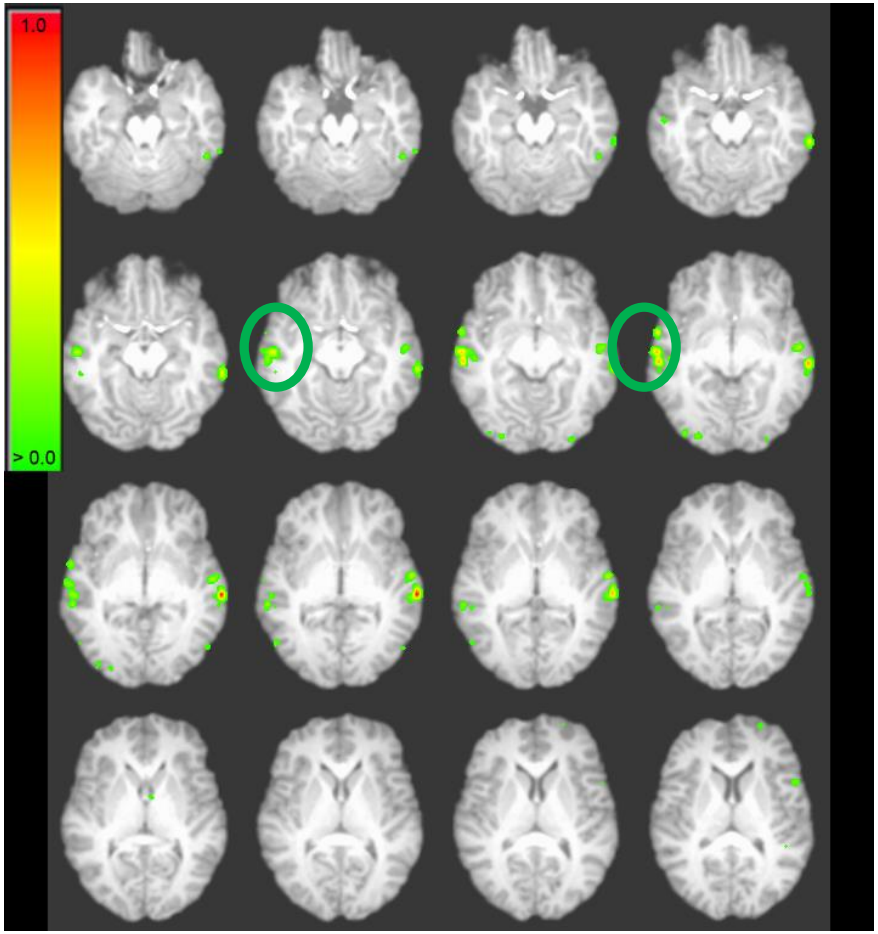
Grouping Effect on PBAIC data

(Carroll, Cecchi, Rish, Garg, Rao 2009)

Predicting 'Instructions' (auditory stimulus)



Small grouping effect: $\lambda_2 = 0.1$



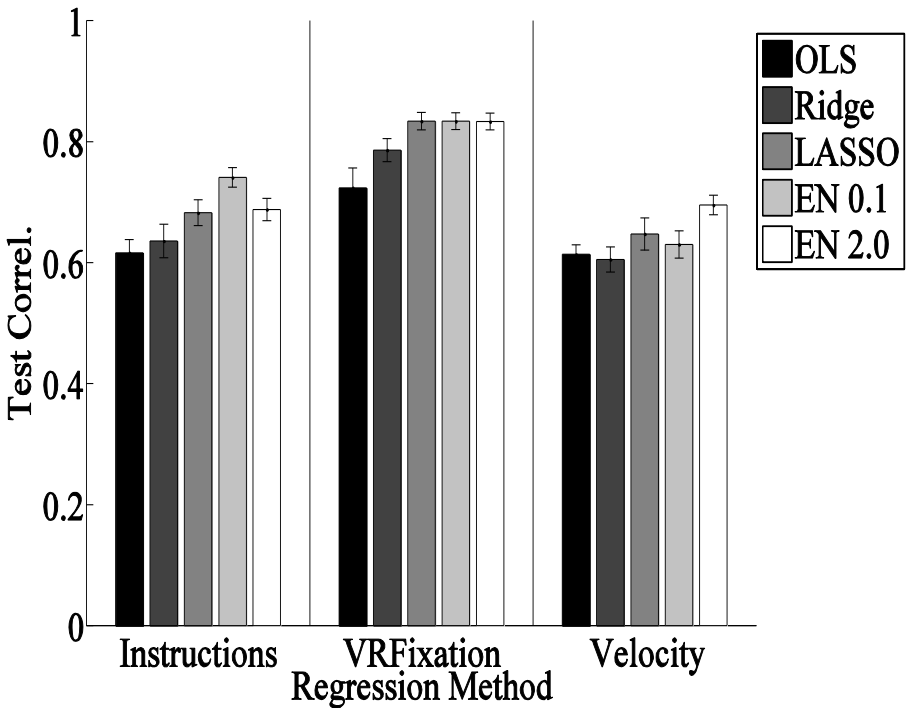
Larger grouping effect: $\lambda_2 = 2.0$

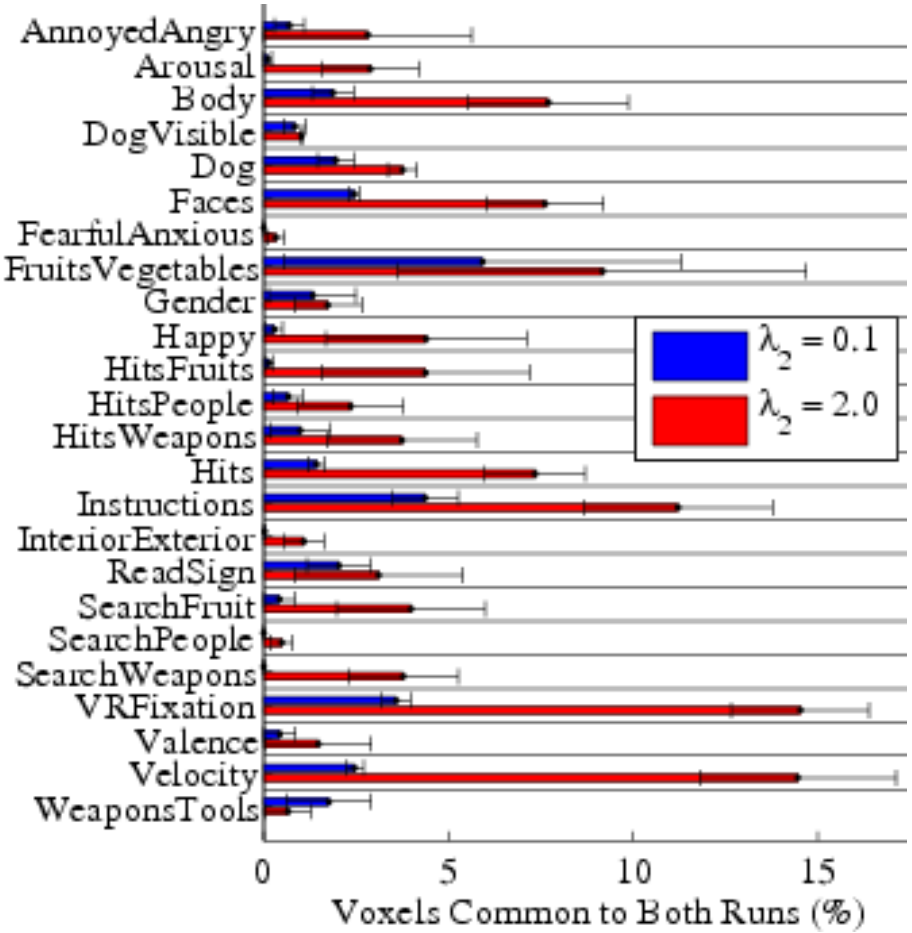
Higher λ_2→ selection of more voxels from correlated clusters →
larger, more spatially coherent clusters

Grouping Tends to Improve Model Stability

(Carroll, Cecchi, Rish, Garg, Rao 2009)

Stability is measured here by average % overlap between models for 2 runs by same subject



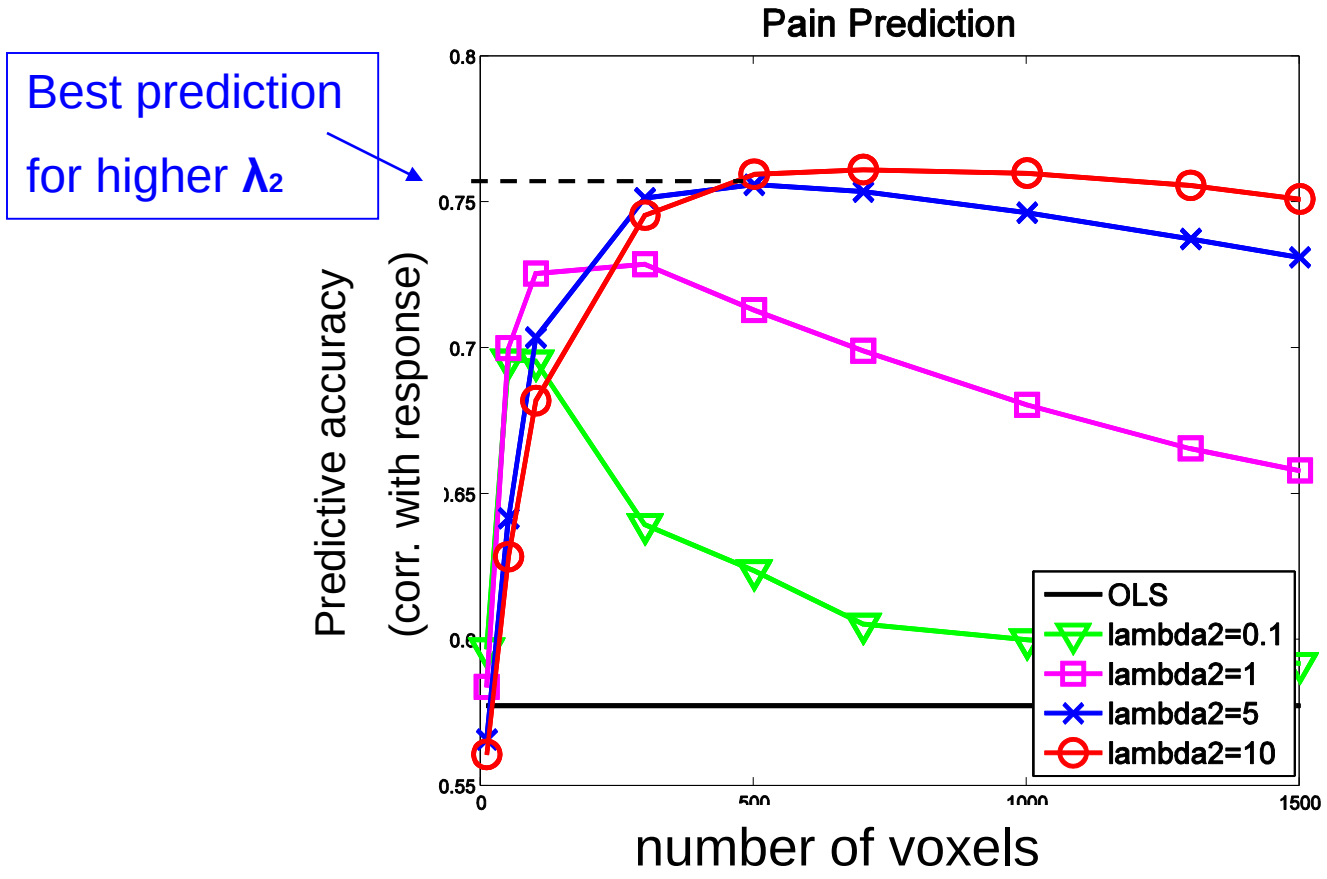


Among almost equally predictive models,

increasing λ_2 can significantly improve model stability

Another Application: Sparse Models of Pain Perception from fMRI

Predicting pain ratings from fMRI in presence of thermal pain stimulus
(Rish, Cecchi, Baliki, Apkarian, BI-2010)



Including more correlated voxels (increasing λ_2) often improves the prediction accuracy as well

Fused Lasso (Tibshirani et al., 2005)

- EN smoothes coefficients uniformly
- But what if there is a natural ordering of the predictors?
- Fused Lasso encourages smoothness along such ordering (besides sparsity):

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 \sum^{p-1} |\beta_{i+1} - \beta_i|$$

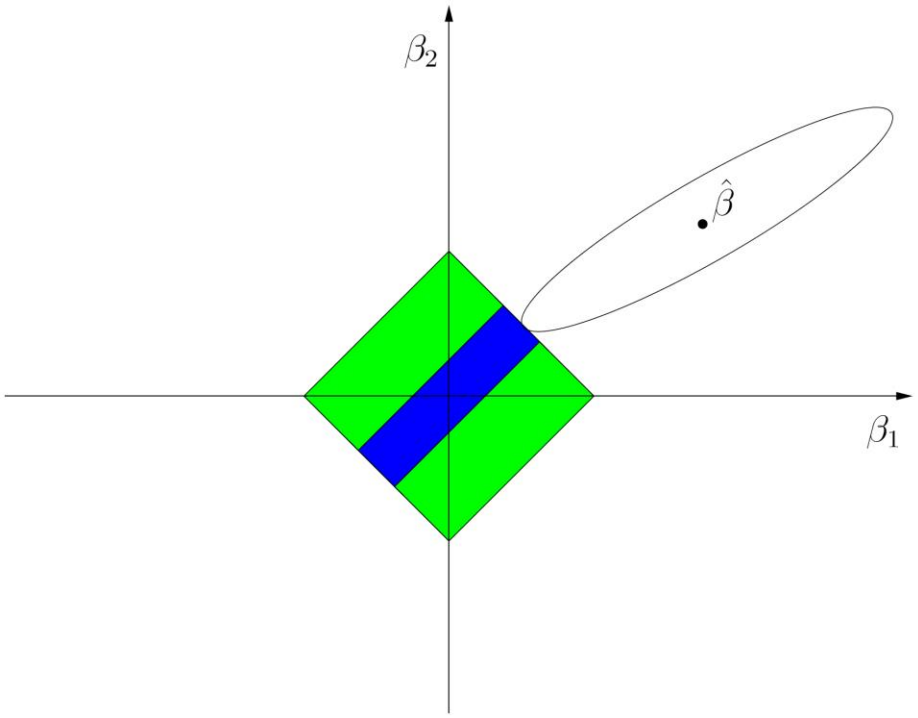


Image courtesy of [Tibshirani et al, 2005]

Group Lasso (Yuan and Lin, 2006)

- What if there is a natural group structure among the variables?
  - functional clusters of genes, or brain voxels
  - categorical variables encoded by groups of indicator variables
  - multi-task learning: parameters for same feature across all tasks

- Block l_1-l_2 penalty selects groups of variables from $G = \bigcup_{i=1}^{K} G_i$, a partition of $\{1, ..., p\}$:

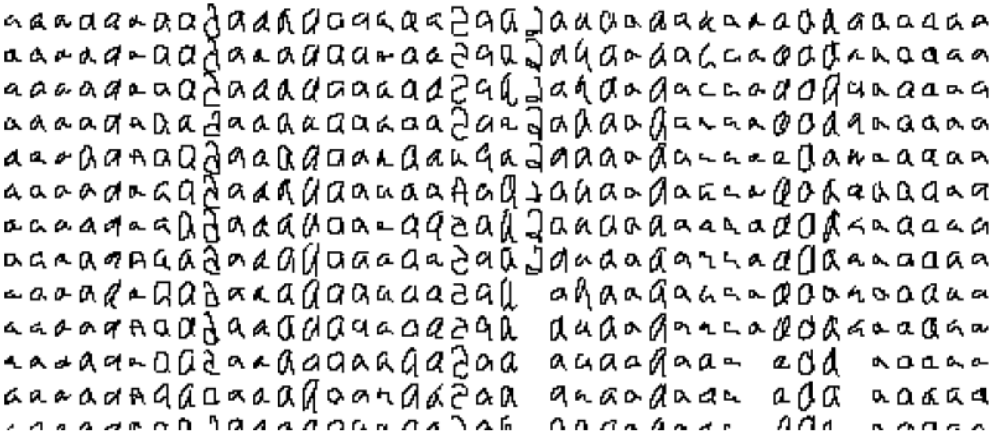
  l_1 promotes sparsity between the groups,
  l_2 discourages sparsity within the groups:

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda \sum_{i=1}^{K} ||\beta_{G_i}||_2$$

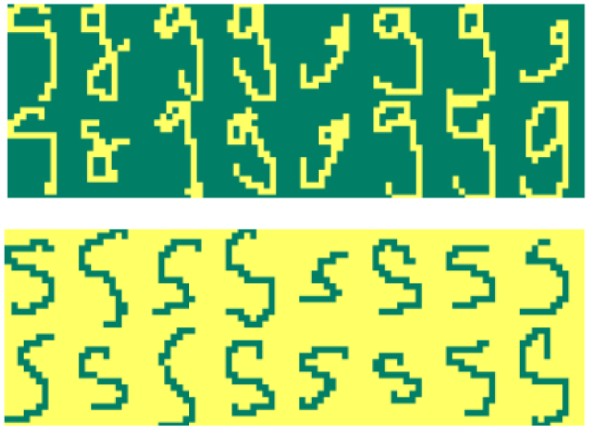
Multi-Task (Simultaneous) Variable Selection

- Select a common subset of variables for k problems
- Example: joint feature selection for character-recognition problems for multiple writers (Obozinski et al., 2010); variables: pixels or strokes

The letter '*a*' *written by 40 different people*



Samples of the letters *s and g for one writer*



- Group-Lasso approach: groups $\Leftrightarrow$ same-variable coefficients across tasks (Obozinski et al., 2010, 2009; Liu et al., 2009b)

Beyond Lasso: General Log-likelihood Losses

$$Loss(\mathbf{x}) + \lambda ||\mathbf{x}||_1$$

$$\downarrow$$

$$-\log P(y|\mathbf{x}) + \lambda ||\mathbf{x}||_1$$

$$\downarrow$$

1. Gaussian $\Leftrightarrow$ Lasso
2. Bernoulli $\Leftrightarrow$ logistic regression
3. Exponential-family $\Leftrightarrow$ Generalized Linear Models (includes 1 and 2)
4. Multivariate Gaussian $\Leftrightarrow$ Gaussian MRFs

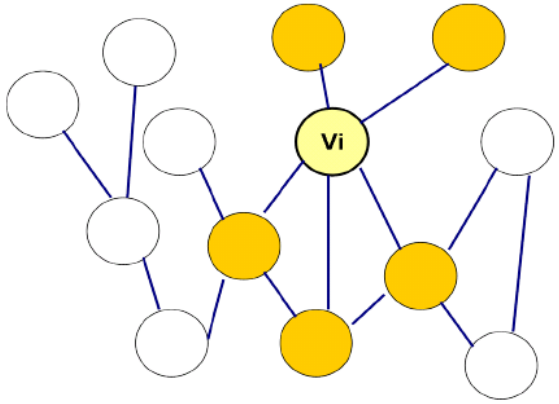
l_1-regularized M-estimators

Markov Networks (Markov Random Fields)

$X = \{X_1, ..., X_p\}, \quad G = (V, E)$

$$P(\boldsymbol{X}) = \frac{1}{\boldsymbol{Z}} \prod_{\mathcal{C} \in Cliques} \Phi_{\mathcal{C}}(\boldsymbol{X}_{\mathcal{C}})$$

Lack of edge $(i, j) \rightarrow$
conditional independence $X_i \perp X_j | rest$

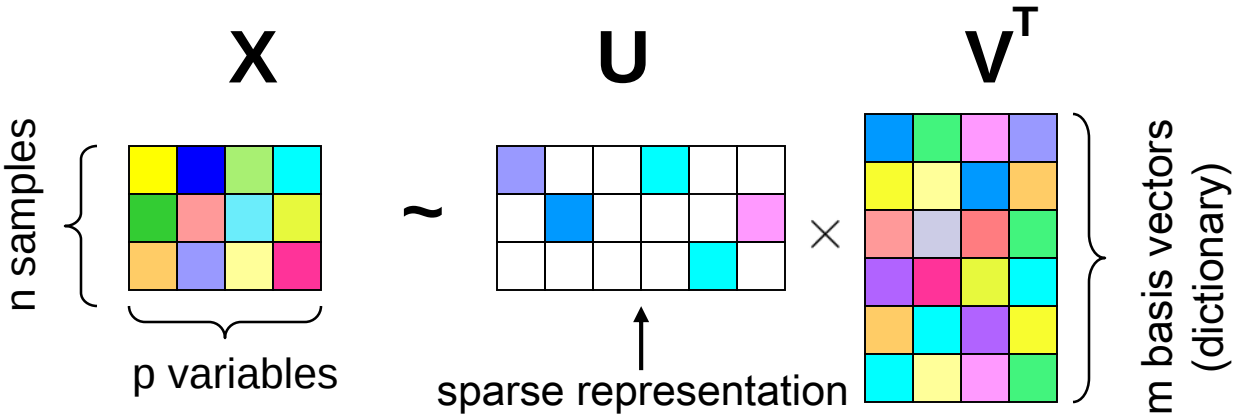


Gaussian Markov Networks (GMRFs):

- $P(\mathbf{x}) = (2\pi)^{-\frac{P}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$
- Σ - covariance matrix, Σ^{-1} - precision (concentration) matrix
- Zeros in Σ: marginal independence
- Zeros in $\Sigma^{-1} \Leftrightarrow$ conditional independence $\Leftrightarrow$ lack of edge (Lauritzen, 1996)
- Sparse $\Sigma^{-1} \Leftrightarrow$ sparse Markov network

Sparse Matrix Factorization

- Dictionary learning
  (Elad and Aharon, 2006; Raina et al., 2007; Mairal et al., 2009):



$$\min_{U,V} ||X - UV^T||^2 + \lambda \sum_{i=1}^{n} ||U(i,:)||_1$$

$$\forall j = 1, ..., m, \ ||V(:,j)||_2 \leq 1$$

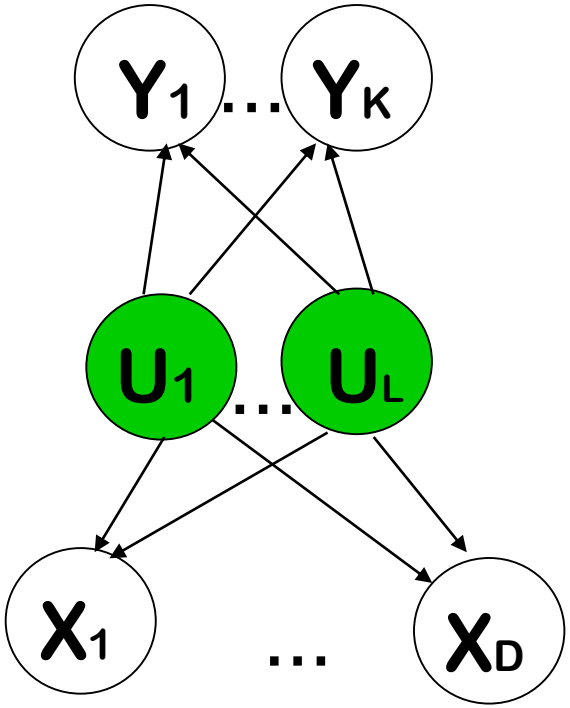
sparse $U(i,:) \Leftrightarrow$ sparse representation in dictionary V

- Sparse PCA (Zou et al., 2006; d'Aspremont et al., 2007):
  sparse $V(:,j)$ (loadings/coordinates of components) $\rightarrow$ interpretability

- other sparse matrix factorization methods:
  sparse CCA (Sriperumbudur et al., 2009; Hardoon and Shawe-Taylor, 2008), sparse NMF (Hoyer, 2004), with applications to blind-source separation and diagnosis (Chandalia and Rish, 2007)

From Variable Selection to Variable Construction

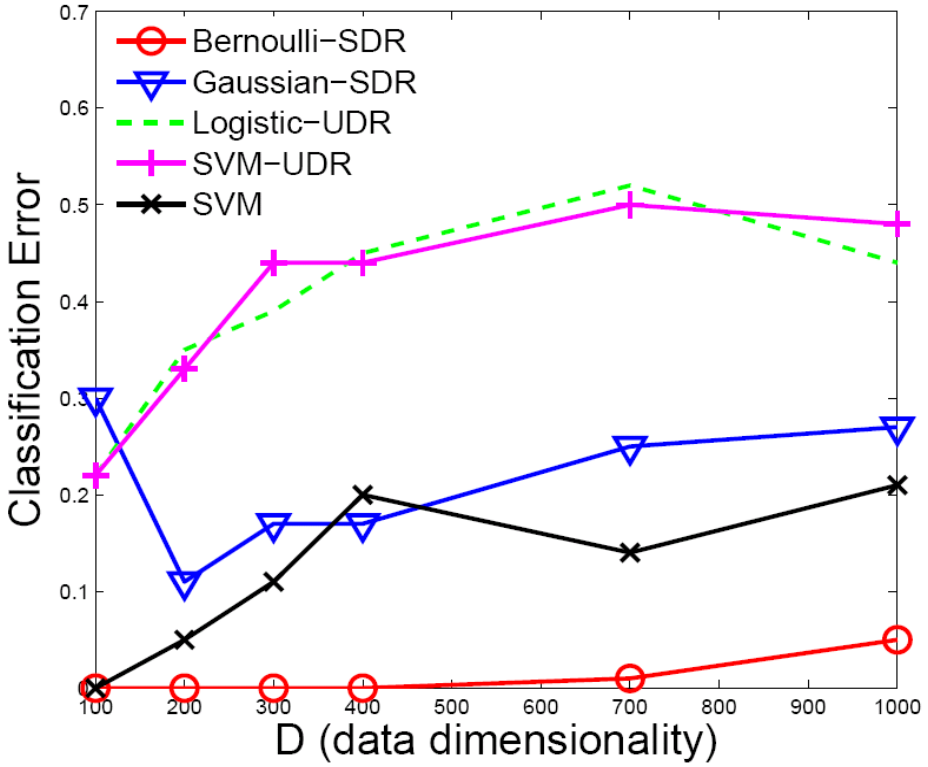
Supervised Dimensionality Reduction (SDR):



- Assume there is an inherent low-dimensional structure in the data that is predictive about the target Y
- Learn a predictor (mapping from U to Y) simultaneously with dimensionality reduction
- Idea: dimensionality reduction (DR) guided by the class label may result into better predictive features than the unsupervised DR

Supervised DR Outperforms Unsupervised DR on Simulated Data

- Generate a separable 2-D dataset U
- Blow-up in D dimensional data X by adding exponential-family noise (e.g., Bernoulli)
- Compare SDR w/ different noise models (Gaussian, Bernoulli) vs. unsupervised DR (UDR) followed by SVM or logistic regression



- SDR outperforms unsupervised DR by 20-45%
- Using proper data model (e.g., Bernoulli-SDR for binary data) matters
- SDR ``gets'' the structure (0% error), SVM does not (20% error)

…and on Real-Life Data from fMRI Experiments

Real-valued data, Classification Task

Predict the type of word (tools or buildings) the subject is seeing

84 samples (words presented to a subject), 14043 dimensions (voxels)

Latent dimensionality L = 5, 10, 15, 20, 25

method\ L	5	10	15	20	25
Gaussian-SDR	0.21	0.26	0.23	**0.20**	0.23
Logistic-UDR	0.44	0.42	**0.29**	0.30	0.26
SVM-UDR	**0.49**	0.52	0.56	0.57	0.55
SVDM	0.32	0.25	**0.21**	0.23	0.23
SVM	0.21				

- Gaussian-SDR achieves overall best performance
- SDR matches SVM's performance using only 5 dimensions, while SVDM needs 15
- **SDR greatly outperforms unsupervised DR followed by learning a classifier**

Summary and Open Issues

- Common problem: small-sample, high-dimensional inference
- Feasible if the input is structured – e.g. sparse in some basis
- Efficient recovery of sparse input via l_1- relaxation
- Sparse modeling with l_1-regularization: interpretability + prediction
- Beyond l_1-regularization: adding more structure
- Beyond Lasso: M-estimators, dictionary learning, variable construction
- Open issues, still:
  - choice of regularization parameter?
  - choice of proper dictionary?
  - Is interpretability $\Leftrightarrow$ sparsity? (NO!)

Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Signal Recovery and Lasso: Some Theory
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

Beyond LASSO

$$Loss(\mathbf{x}) + \lambda ||\mathbf{x}||_1$$

Other likelihoods (loss functions)

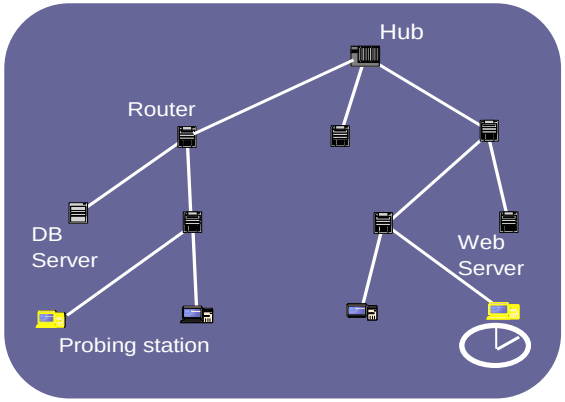
- Generalized Linear Models (exponential family noise)
- Multivariate Gaussians (Gaussian MRFs)

Adding structure beyond sparsity

- Elastic Net
- Fused Lasso
- Block l_1-l_q norms:
  - group Lasso
  - simultaneous Lasso

Why Exponential Family Loss?

- Network Management – Problem Diagnosis:
  - binary failures - Bernoulli
  - non-negative delays – exponential



- Collaborative prediction:
  - discrete rankings - multinomial

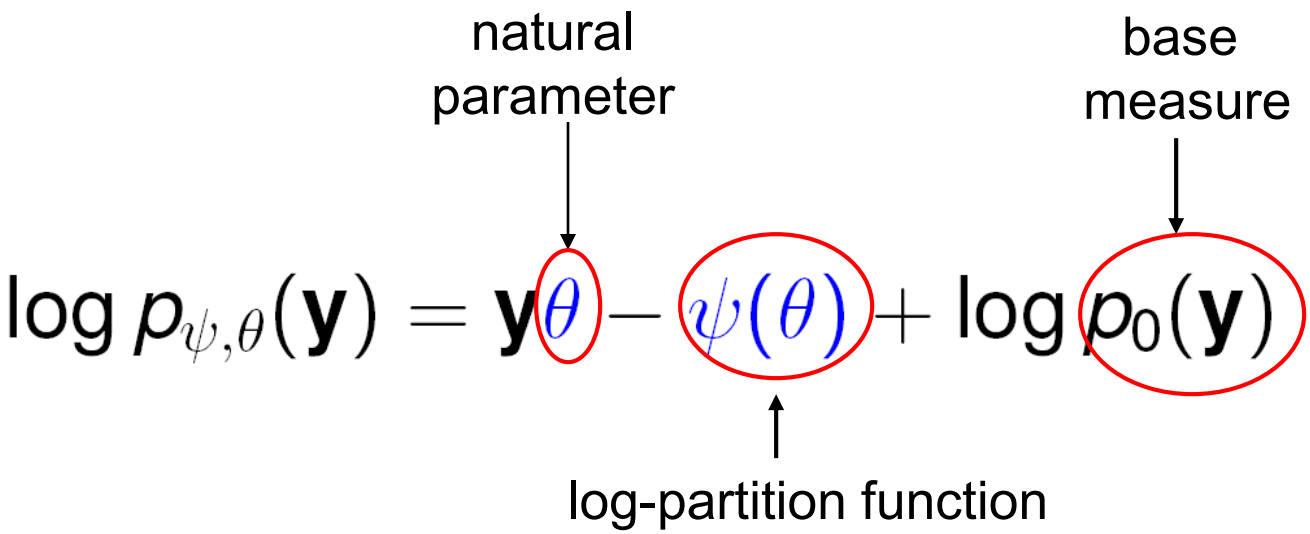
- DNA microarray data analysis:
  - Real-valued expression level – Gaussian

- fMRI data analysis
  - Real-valued voxel intensities, binary, nominal and continuous responses

Variety of data types: real-valued, binary, nominal, non-negative, etc.

Noise model: exponential-family

Exponential Family Distributions



Examples: Gaussian, exponential, Bernoulli, multinomial, gamma, chi-square, beta, Weibull, Dirichlet, Poisson, etc.

Generalized Linear Models (GLMs)

$$E_{p_{\psi,\theta}}(\mathbf{y}) = f^{-1}(\mathbf{Ax})$$

$f(\theta)$ - *link function*, where $f^{-1}(\theta) = \nabla\psi(\theta)$

1. Gaussian noise - *identity* function $f(\mu) = \mu$ (linear regression):

$$E(\mathbf{y}) = \mathbf{Ax}$$

2. Bernoulli noise - *logit* function $f(\mu) = \log \frac{\mu}{1-\mu}$ (logistic regression)

$$E(\mathbf{y}) = \frac{1}{1 + e^{-\mathbf{Ax}}}$$

Exponential Family, GLMs, and Bregman Divergences

Bijection Theorem (Banerjee et al, 2005):

$$p_{\psi,\theta}(\mathbf{y}) = e^{-d_\phi(\mathbf{y},\mu(\theta))} f_\phi(\mathbf{y})$$

Bregman divergence

Domain	Distribution	Divergence
$\mathbb{R}$	1*D* Gaussian	square loss
$\{0,1\}$	Bernoulli	logistic loss
R_{++}	Exponential	Itakura-Saito distance
n-simplex	*nD* Multinomial	KL-divergence
$\mathbb{R}^n$	*nD* Sph. Gaussian	squared Euclidean distance
$\mathbb{R}^n$	*nD* Gaussian	Mahalanobis distance

Fitting GLM $\Leftrightarrow$ maximizing exp-family likelihood $\Leftrightarrow$

$\Leftrightarrow$ minimizing Bregman divergence

Sparse Signal Recovery from Noisy Observations

Euclidean distance (Candes, Romberg and Tao, 2006):

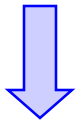
If

- small observation noise: $||y - Ax^0||_{l_2} \leq \epsilon$
- A satisfies the restricted isometry property (RIP)

Then the solution to the sparse linear regression problem

$$x^* = \arg\min_x ||x||_{l_1} \quad s.t. \; ||y - Ax||_{l_2} \leq \epsilon$$

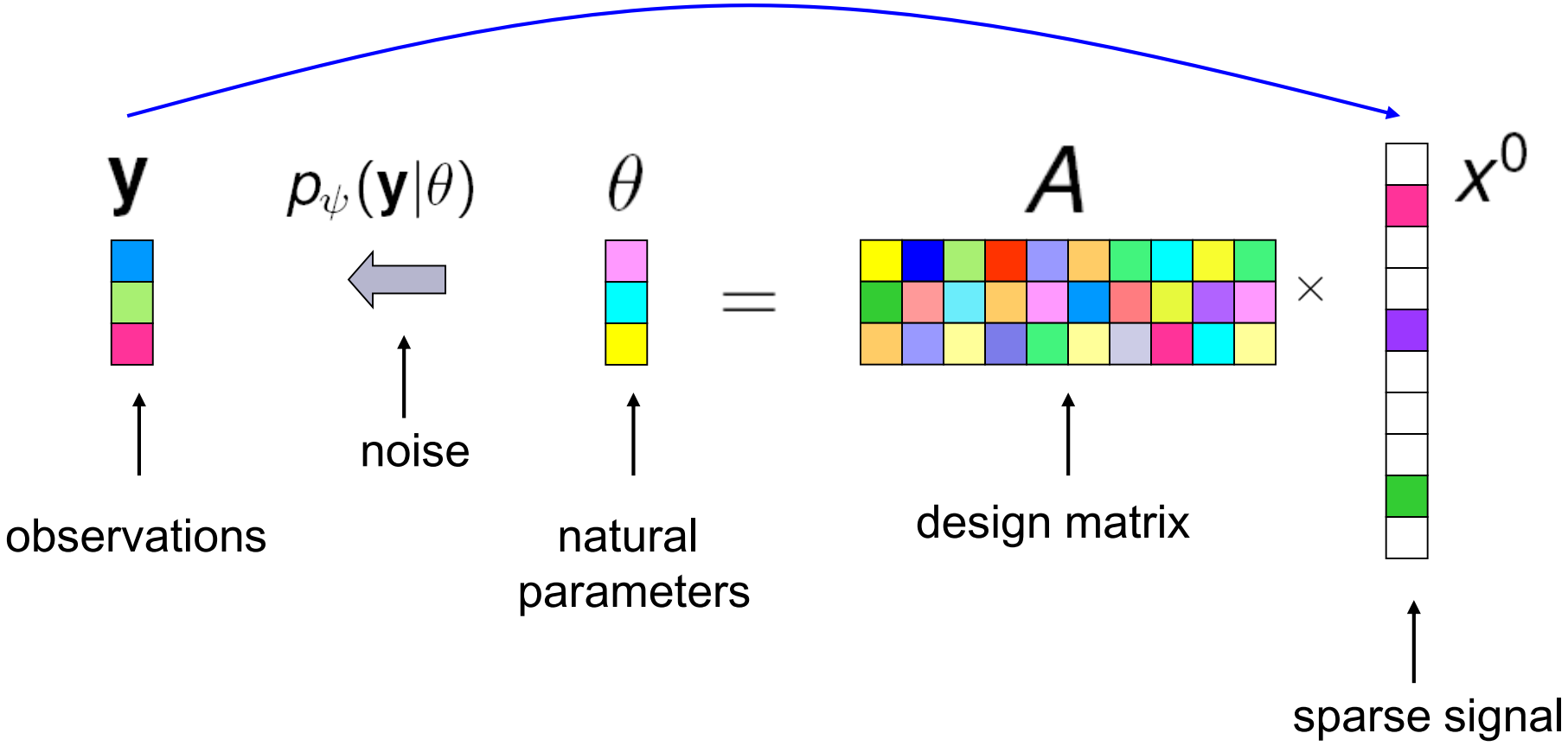
is a good approximation of x^0, i.e. $||x^* - x^0||_{l_2} \leq C_S \cdot \epsilon$.



Generalized Linear Models: (Rish and Grabarnik, 2009)

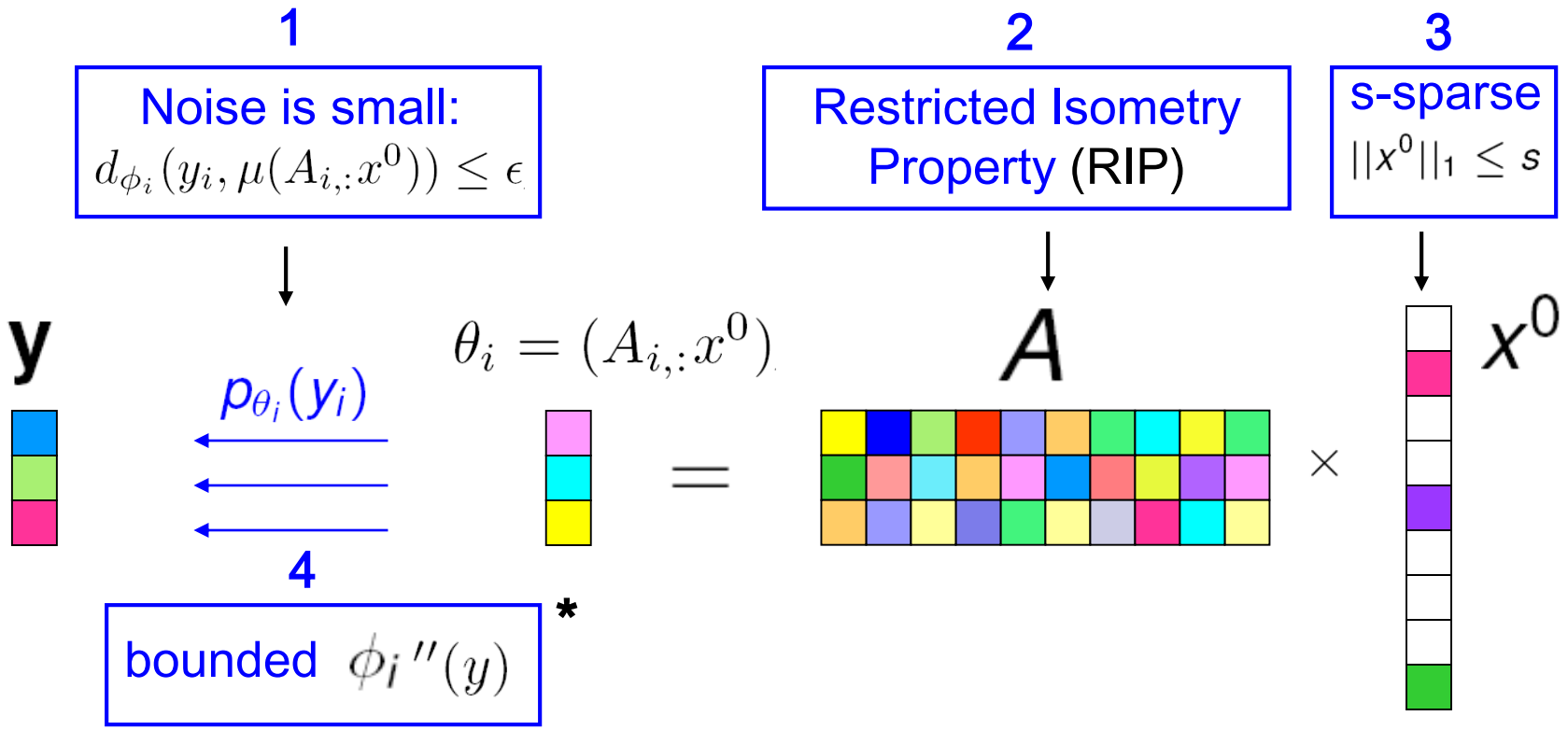
replace Euclidean distances $||y - Ax^0||_{l_2}$ and $||y - Ax||_{l_2}$ by the corresponding Bregman divergences $d(y, \mu(Ax^0))$ and $d(y, \mu(Ax))$.

Sparse Signal Recovery with Exponential-Family Noise



Can we recover a sparse signal
from a small number of noisy observations?

Sufficient Conditions (Rish and Grabarnik, 2009)



Then the solution x^* to the sparse GLM regression problem

$$min||x||_1 \quad subject\ to \quad \sum_i d(y_i, \mu(A_i x)) \leq \epsilon$$

is a good approximation of x^0, i.e. $||x^* - x^0||_{l_2} \leq C_S \cdot \delta(\epsilon)$

$\delta(\epsilon)$ - continuous monotone increasing function, and $\delta(0) = 0$ (i.e. $\delta(\epsilon)$ is small when ϵ is small).

*otherwise, different proofs for some specific cases (e.g., Bernoulli, exponential, etc.)

Summary

- sparse signal recovery (Candes, Romberg & Tao, 2006) can be extended from linear to generalized linear models (*exponential-family* observation noise)
- signal recovery requires solving an l_1-regularized *Generalized Linear Model (GLM)* regression problem
- recovery conditions include, besides standard RIP for design matrix:
  (1) small noise (Bregman divergence) $d_\phi(y_i, \mu(A_{i,:}x^0)) \leq \epsilon$
  (2) certain conditions on ϕ
- results also hold for compressible (rather than sparse) signals

Beyond LASSO

$$Loss(\mathbf{x}) + \lambda ||\mathbf{x}||_1$$

Other likelihoods (loss functions)

- Generalized Linear Models (exponential family noise)
- Multivariate Gaussians (Gaussian MRFs)

Adding structure beyond sparsity

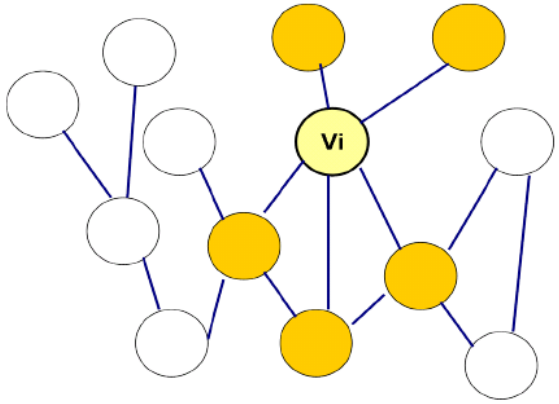
- Elastic Net
- Fused Lasso
- Block l_1-l_q norms:
  - group Lasso
  - simultaneous Lasso

Markov Networks (Markov Random Fields)

$X = \{X_1, ..., X_p\}, \quad G = (V, E)$

$$P(\boldsymbol{X}) = \frac{1}{\boldsymbol{Z}} \prod_{\mathcal{C} \in Cliques} \Phi_{\mathcal{C}}(\boldsymbol{X}_{\mathcal{C}})$$

Lack of edge $(i, j) \rightarrow$
conditional independence $X_i \perp X_j | rest$

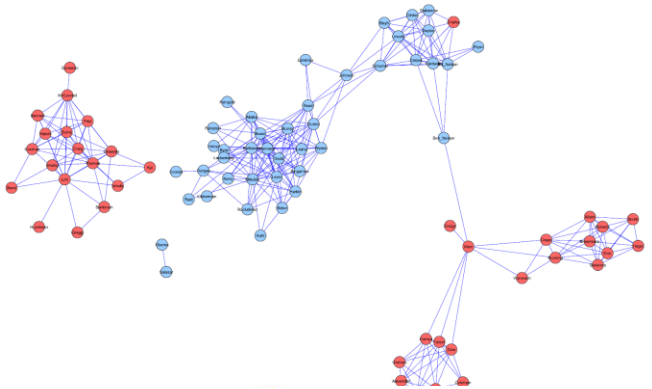


Gaussian Markov Networks (GMRFs):

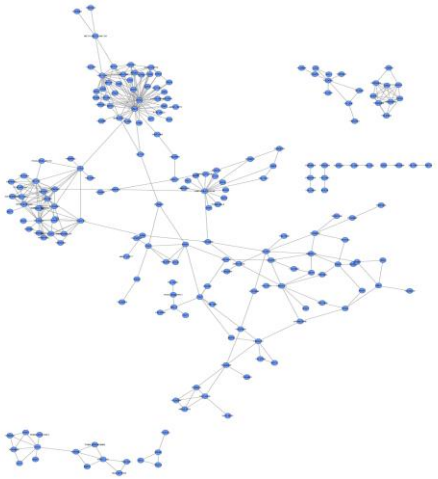
- $P(\mathbf{x}) = (2\pi)^{-\frac{P}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$
- Σ - covariance matrix, Σ^{-1} - precision (concentration) matrix
- Zeros in Σ: marginal independence
- Zeros in $\Sigma^{-1} \Leftrightarrow$ conditional independence $\Leftrightarrow$ lack of edge (Lauritzen, 1996)
- Sparse $\Sigma^{-1} \Leftrightarrow$ sparse Markov network

Sparse Markov Networks in Practical Applications

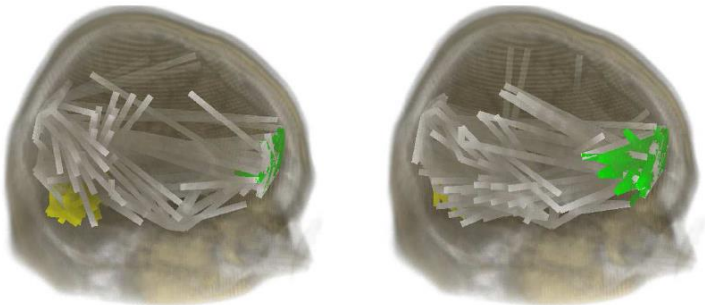
- Social Networks
  - US senate voting data (Banerjee et al, 2008): democrats (blue) and republicans (red)



- Genetic Networks
  - Rosetta Inpharmatics Compendium of gene expression profiles (Banerjee et al, 2008)



- Brain Networks from fMRI
  - Monetary reward task (Honorio et al., 2009)
  - Drug addicts more connections in cerebellum (yellow) vs control subjects (more connections in prefrontal cortex – green)



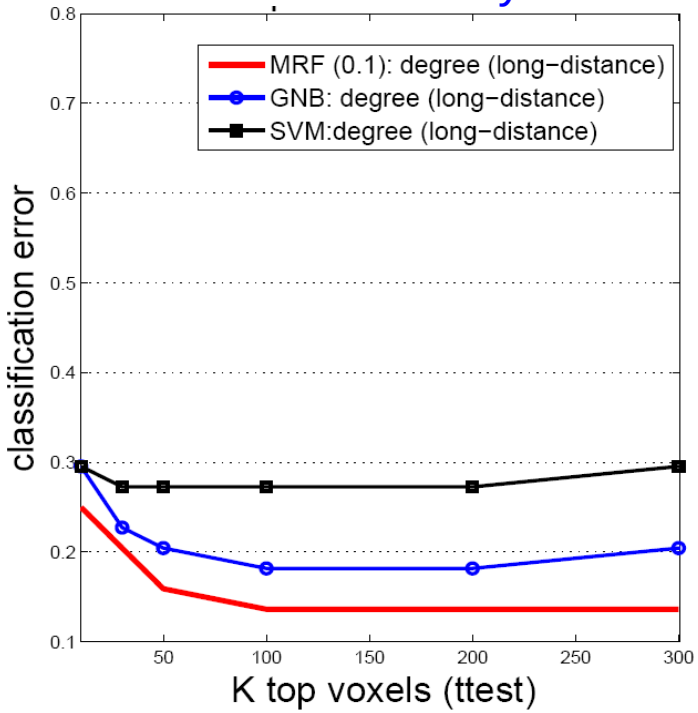
(a) Drug addicts (b) controls

Sparse MRFs Can Predict Well

Classifying Schizophrenia

(Cecchi et al., 2009)

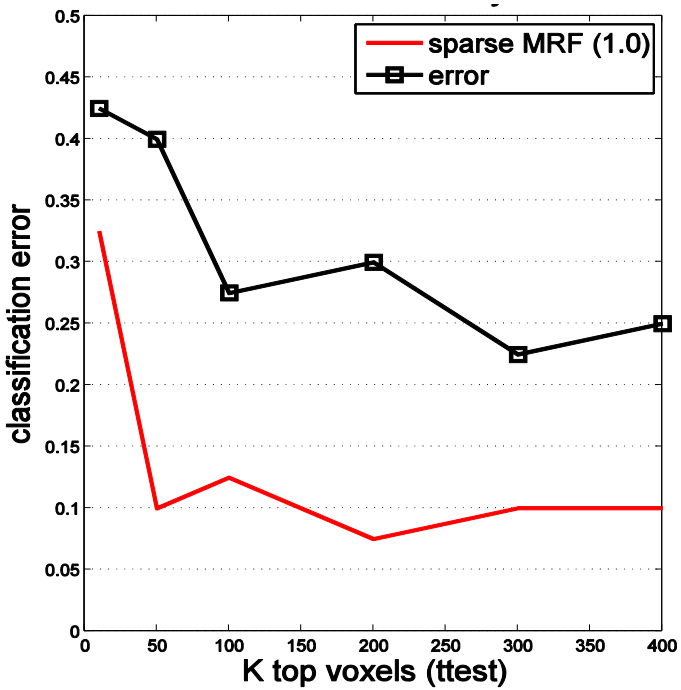
86% accuracy



Mental state prediction (sentence vs picture)*:

(Scheinberg and Rish, submitted)

90% accuracy



MRF classifiers can often exploit informative interactions among variables and often outperform state-of-art linear classifiers (e.g., SVM)

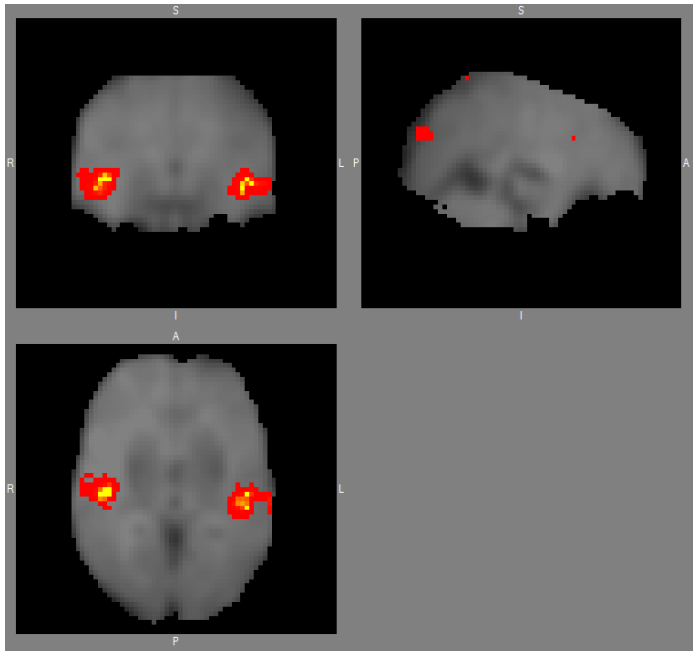
*Data @ www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/ from T. Mitchell et al., *Learning to Decode Cognitive States from Brain Images,* Machine Learning, 2004.

Network Properties as BioMarkers (Predictive Features)

Discriminative Network Models of Schizophrenia (Cecchi et al., 2009)

- Voxel degrees in *functional networks* (thresholded *covariance* matrices) are statistically significantly different in schizophrenic patients that appear to **lack "hubs" in auditory/language areas**

FDR-corrected Degree Maps



2-sample t-test performed for each voxel in degree maps, followed by FDR correction

Red/yellow: Normal subjects have *higher* values than Schizophrenics

Also, abnormal MRF connectivity observed in Alzheimer's patients (Huang 2009), in drug addicts (Honorio 2009), etc.

Maximum Likelihood Estimation

Assume the data $\mathbf{X}$ are centered to have zero mean. Then:

$$\hat{\Sigma}^{-1} = \arg\max_{C \succ 0} \log p(C|\mathbf{X}) = \arg\max_{C \succ 0} \log p(\mathbf{X}, C) =$$

$$\boxed{= \arg\max_{C \succ 0} \log\det(C) - \mathrm{tr}(SC)}$$

where $S = \frac{1}{N}\sum_{i=1}^{N} x_i^T x_i$ is the empirical covariance matrix (MLE of Σ)

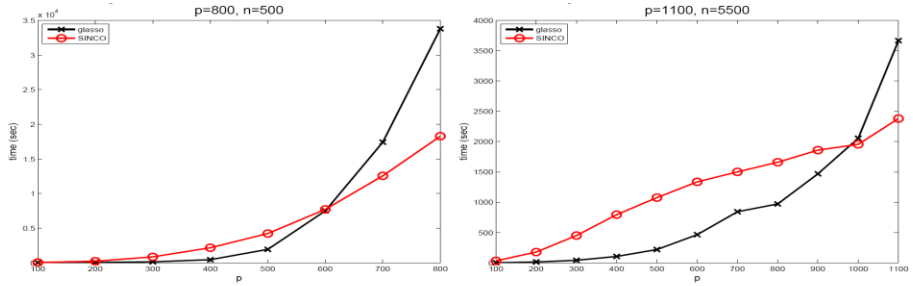
Why not just use $\hat{\Sigma}^{-1} = S^{-1}$?

- in small-sample case ($n < p$), S may not be even invertible
- even if it is, S^{-1} almost never contains exact zeros
- l_1-regularization takes care of both issues!

Solving Primal Problem Directly

1. Greedy coordinate ascent approach: SINCO (Scheinberg et al., 2009)

   - updates one diagonal or two (symmetric) off-diagonal elements of C at each step
   - evaluating each C_{ij} takes constant time (solving quadratic equation), thus each step takes $O(p^2)$ time and can be easily parallelized
   - naturally preserves the sparsity of a solution; can reduce false-positive error by not including "weak" edges not contributing much to the objective
   - Speedwise, comparable to glasso; outperforms glasso on large-scale problems



(a) Random (b) Scale-free

CPU time comparison: SINCO vs *glasso* on (a) random networks ($N = 500$, fixed range of λ) and (b) scale-free networks (density 21%, N and λ scaled by the same factor with p, $N = 500$ for $p = 100$).

2. (Honorio et al., 2009) also solve the primal problem:

   - Optimize over each column (node) at a time
   - Exploit "local constancy" structure adding a regularizer similar to fused Lasso

Additional Related Work

- (Yuan and Lin, 2007) solve the primal problem (1) using interior-point method for the maxdet problem (Vandenberghe et al., 1998)
- (Lee et al., 2007) learn MRFs using clique selection heuristic and approximate inference
- (Wainwright et al., 2007) extend the approach of (Meinshausen and Buhlmann, 2006) to binary MRFs Ising models, applying sparse logistic regression at each node, and derive asymptotic consistency results
- (Schmidt et al., 2007) apply l_1-regularization to structure learning in Bayesian networks
- (Huang et al., 2009) prove the monotone property of (1) under decreasing λ (i.e., connected nodes stay connected with decreasing sparsity levels)
- (Lin et al., 2009) propose an alternative approach based on ensemble-of-trees that is shown to sometimes outperform l_1-regularization approaches of (Banerjee et al., 2008) and (Wainwright et al., 2007)
- (Schmidt and Murphy, 2010) learn log-linear models with higher-order (beyond pairwise) potentials; use group-l_1 regularization with overlapping groups to enforce hierarchical structure over potentials

Selecting the Proper Regularization Parameter

"…the general issue of selecting a proper amount of regularization for getting a right-sized structure or model has largely remained a problem with unsatisfactory solutions" (Meinshausen and Buehlmann , 2008)

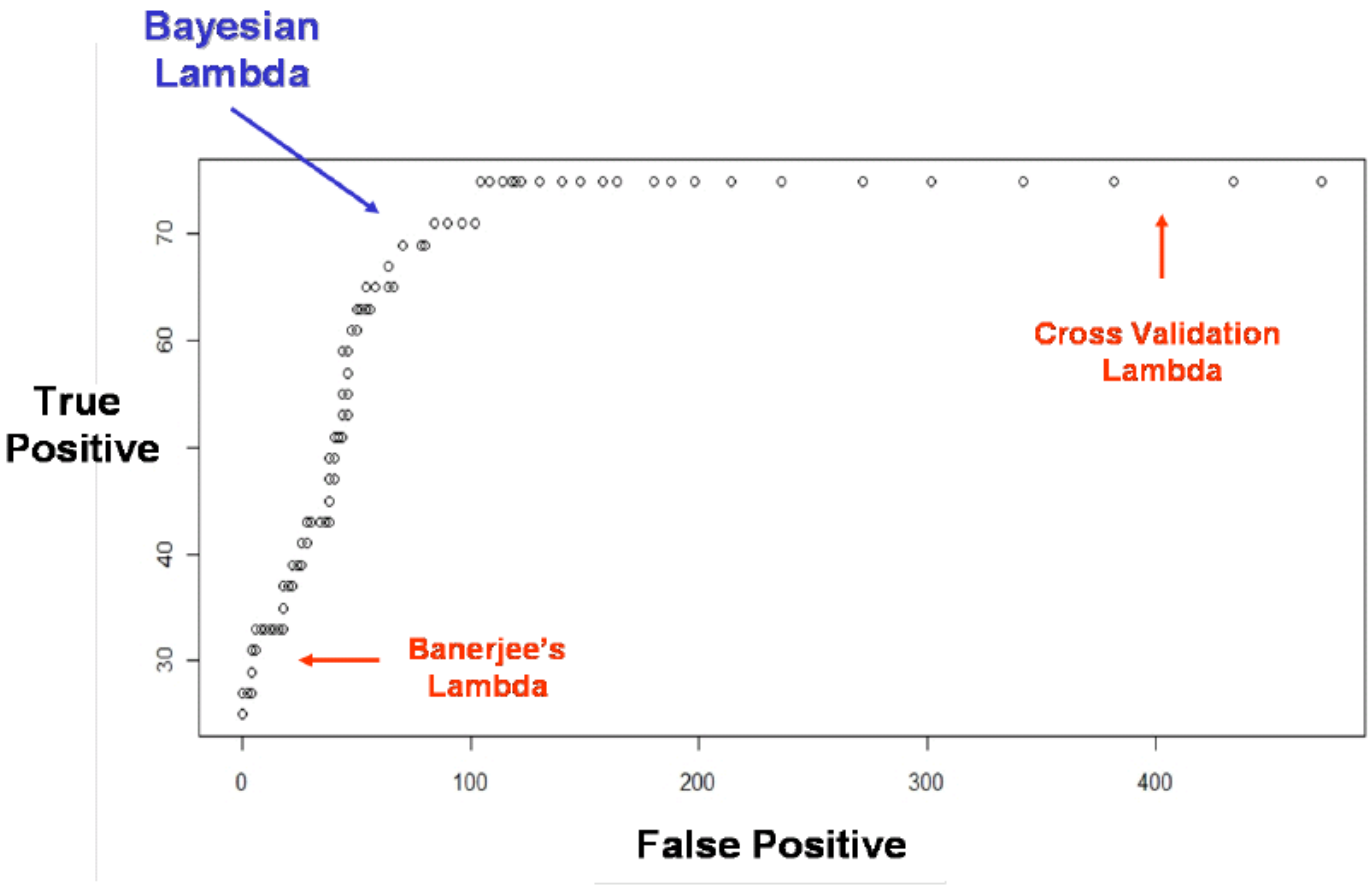
"asymptotic considerations give little advice on how to choose a specific penalty parameter for a given problem" (Meinshausen and Buehlmann , 2006)

- Bayesian Approach (N.Bani Asadi, K. Scheinberg and I. Rish, 2009)
  - Assume a Bayesian prior on the regularization parameter
  - Find maximum a posteriory probability (MAP) solution

- Result:
  - more ``balanced'' solution (False Positive vs False Negative error) than
    - *cross-validation* - too dense, and
    - *theoretical* (Meinshausen & Buehlmann 2006, Banerjee et al 2008) - too sparse
  - Does not require solving multiple optimization problems over data subsets as compared to the *stability selection* approach (Meinshausen and Buehlmann 2008)

The Bayesian λ

ROC Curve

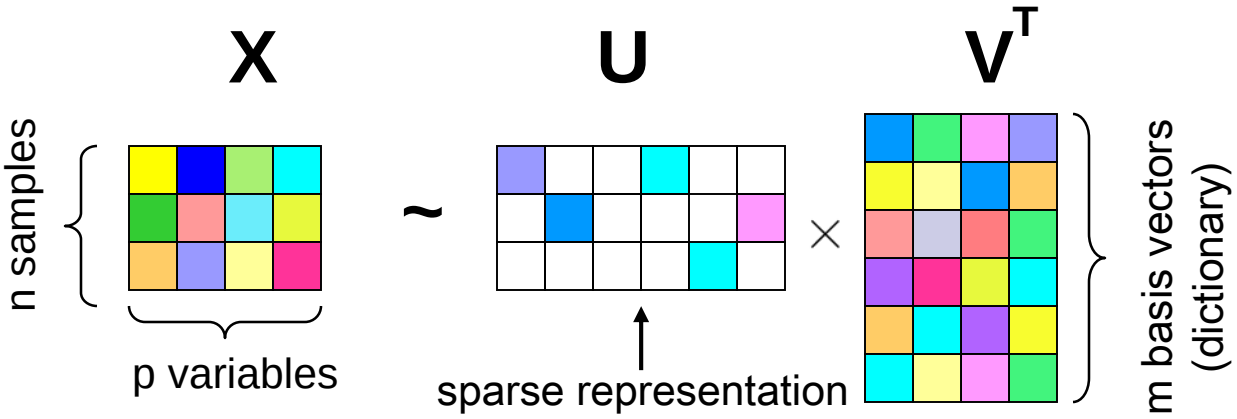


Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Signal Recovery and Lasso: Some Theory
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

Sparse Matrix Factorization

- Dictionary learning
  (Elad and Aharon, 2006; Raina et al., 2007; Mairal et al., 2009):



$$\min_{U,V} ||X - UV^T||^2 + \lambda \sum_{i=1}^{n} ||U(i,:)||_1$$

$$\forall j = 1, ..., m, \ ||V(:,j)||_2 \leq 1$$

sparse $U(i,:) \Leftrightarrow$ sparse representation in dictionary V

- Sparse PCA (Zou et al., 2006; d'Aspremont et al., 2007):
  sparse $V(:,j)$ (loadings/coordinates of components) $\rightarrow$ interpretability

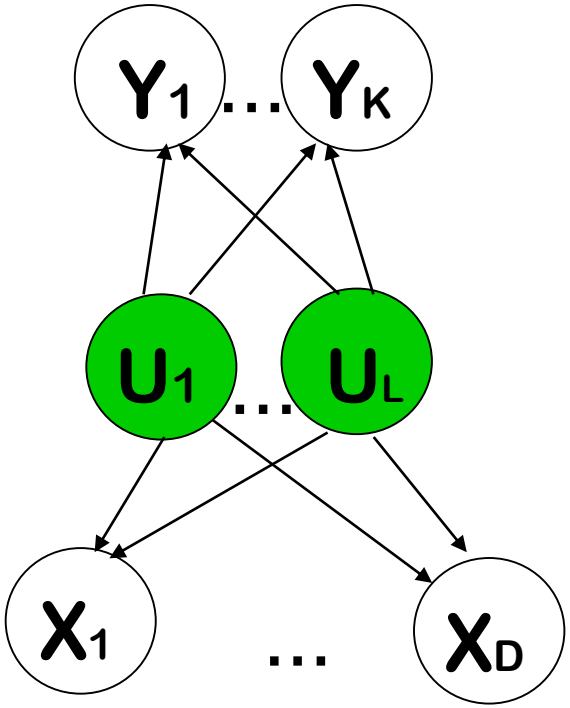
- other sparse matrix factorization methods:
  sparse CCA (Sriperumbudur et al., 2009; Hardoon and Shawe-Taylor, 2008), sparse NMF (Hoyer, 2004), with applications to blind-source separation and diagnosis (Chandalia and Rish, 2007)

Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Signal Recovery and Lasso: Some Theory
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond l_1-regularization (l1/lq, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues

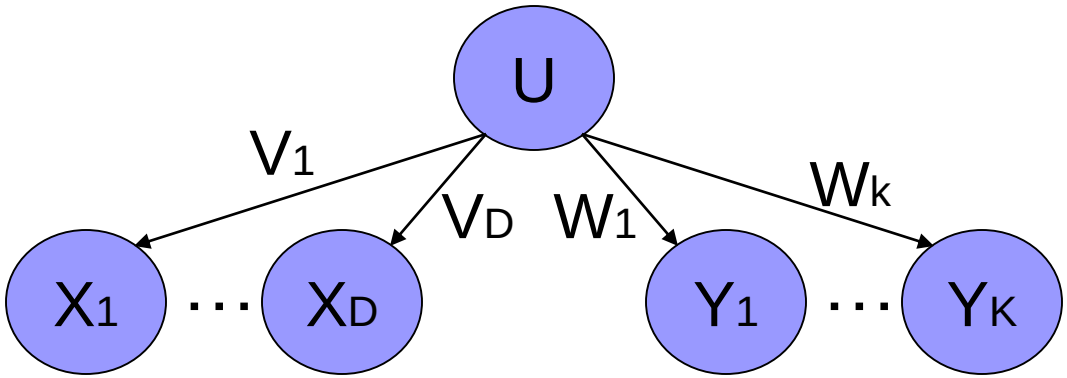
From Variable Selection to Variable Construction

Supervised Dimensionality Reduction (SDR):



- Assume there is an inherent low-dimensional structure in the data that is predictive about the target Y
- Learn a predictor (mapping from U to Y) simultaneously with dimensionality reduction
- Idea: dimensionality reduction (DR) guided by the class label may result into better predictive features than the unsupervised DR

Example: SDR with Generalized Linear Models (Rish et al., 2008)



Generalized Linear Models (GLMs)

$$E(\mathbf{X}_d) = f_d^{-1}(\mathbf{UV_d})$$

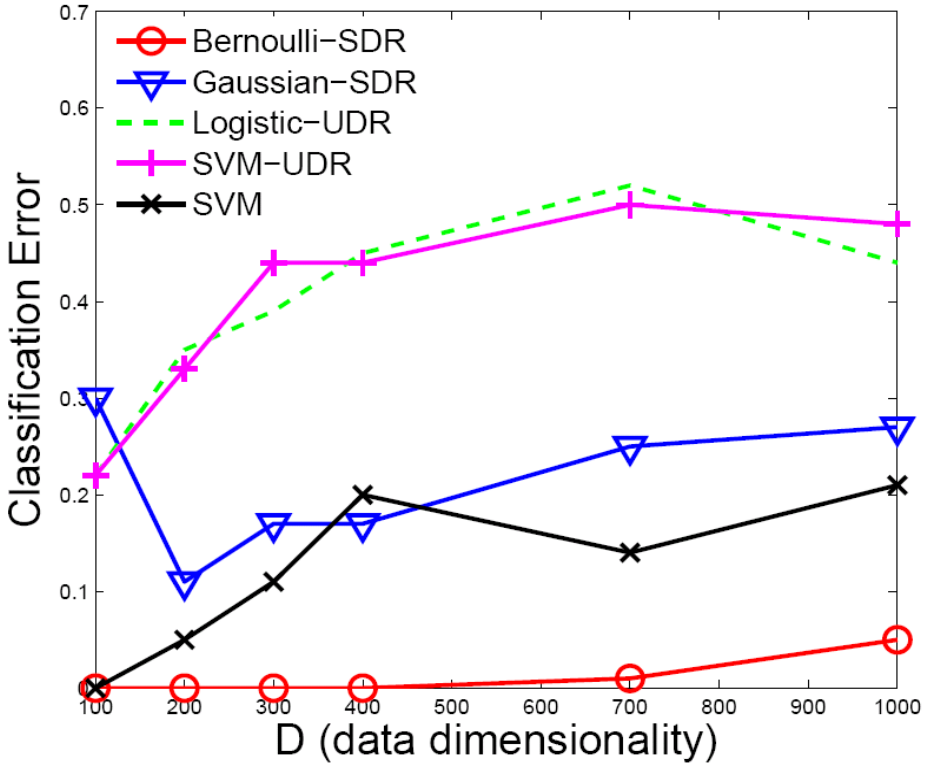
$$E(\mathbf{Y}_k) = f_k^{-1}(\mathbf{UW_k})$$

E.g., in linear case, we have:

X ~ U V and Y ~ U V

Supervised DR Outperforms Unsupervised DR on Simulated Data

- Generate a separable 2-D dataset U
- Blow-up in D dimensional data X by adding exponential-family noise (e.g., Bernoulli)
- Compare SDR w/ different noise models (Gaussian, Bernoulli) vs. unsupervised DR (UDR) followed by SVM or logistic regression



- SDR outperforms unsupervised DR by 20-45%
- Using proper data model (e.g., Bernoulli-SDR for binary data) matters
- SDR ``gets'' the structure (0% error), SVM does not (20% error)

…and on Real-Life Data from fMRI Experiments

Real-valued data, Classification Task

Predict the type of word (tools or buildings) the subject is seeing

84 samples (words presented to a subject), 14043 dimensions (voxels)

Latent dimensionality L = 5, 10, 15, 20, 25

method\ L	5	10	15	20	25
Gaussian-SDR	0.21	0.26	0.23	**0.20**	0.23
Logistic-UDR	0.44	0.42	**0.29**	0.30	0.26
SVM-UDR	**0.49**	0.52	0.56	0.57	0.55
SVDM	0.32	0.25	**0.21**	0.23	0.23
SVM	0.21				

- Gaussian-SDR achieves overall best performance
- SDR matches SVM's performance using only 5 dimensions, while SVDM needs 15
- **SDR greatly outperforms unsupervised DR followed by learning a classifier**

Summary and Open Issues

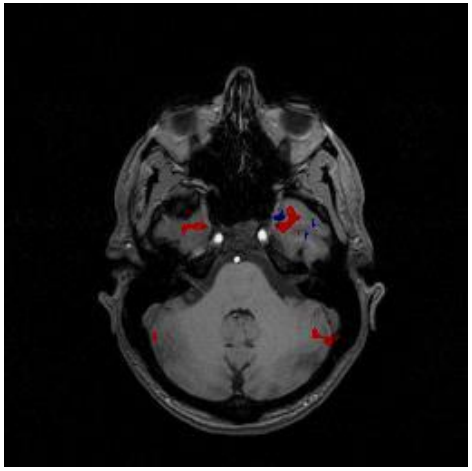
- Common problem: small-sample, high-dimensional inference
- Feasible if the input is structured – e.g. sparse in some basis
- Efficient recovery of sparse input via l_1- relaxation
- Sparse modeling with l_1-regularization: interpretability + prediction
- Beyond l_1-regularization: adding more structure
- Beyond Lasso: M-estimators, dictionary learning, variable construction
- Open issues, still:
  - choice of regularization parameter?
  - choice of proper dictionary?
  - Is interpretability $\Leftrightarrow$ sparsity? (NO!)

Interpretability: Much More than Sparsity?

Data

x - fMRI voxels,

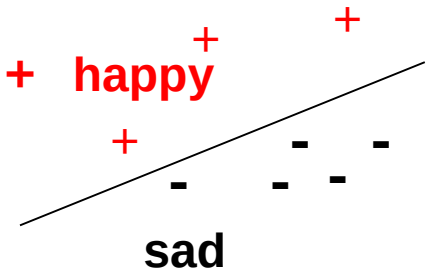
y - mental state



Interpretable Predictive Patterns

Predictive Model

y = f(**x**)



References

Bach, F., 2008a. Bolasso: model consistent lasso estimation through the bootstrap. In: ICML '08: Proceedings of the 25th international conference on Machine learning. ACM, New York, NY, USA, pp. 33–40.

Bach, F., 2008b. Consistency of the group Lasso and multiple kernel learning. Journal of Machine Learning Research 9, 1179–1225.

Bach, F. R., Lanckriet, G. R. G., Jordan, M. I., 2004. Multiple kernel learning, conic duality, and the smo algorithm. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning. ACM, New York, NY, USA, p. 6.

Bakin, S., 1999. Adaptive Regression and Model Selection in Data Mining Problems. Ph.D. thesis, Australian National University, Canberra, Australia.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., March 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research 9, 485–516.

Banerjee, O., Ghaoui, L. E., d'Aspremont, A., Natsoulis, G., 2006. Convex optimization techniques for fitting sparse Gaussian graphical models. In: ICML. pp. 89–96.

Baraniuk, R., Davenport, M., DeVore, R., Wakin, M., 2008. A simple proof of the restricted isometry property for random matrices. Constructive Approximation 28 (3), 253–263.

Beygelzimer, A., Kephart, J., Rish, I., 2007. Evaluation of optimization methods for network bottleneck diagnosis. In: In Proc. of International Conference on Autonomic Computing (ICAC-07).

Bickel, P., Ritov, Y., Tsybakov, A., 2009. Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37 (4), 1705–1732.

Boyd, S., Vandenberghe, L., 2004. Convex optimization. Cambridge Univ Pr.

Bunea, F., Tsybakov, A., , Wegkamp, M., 2007. Sparsity oracle inequalities for the lasso. Electron. J. Statist. 1, 169–194.

Candès, E., 2006. Compressive sampling. In: Proceedings of the Int. Congress of Mathematics. pp. 1433–1452.

Candès, E., Romberg, J., 2007. Sparsity and incoherence in compressive sampling. Inverse Problems 23(3), 969–985.

References

Candès, E., Romberg, J., Tao, T., 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory 52(2), 489–509.

Candès, E., Tao, T., 2006a. The dantzig selector: Statistical estimation when *p* is much larger than *n*. Annals of Statistics To appear.

Candès, E., Tao, T., 2006b. Decoding by linear programming,. IEEE Trans. Inform. Theory 51, 4203–4215, j.

Candès, E., Tao, T., 2006c. Near optimal signal recovery from random projections: Universal encoding strategies? IEEE Trans. Inform. Theory 52(12), 5406–5425.

Carroll, M., G.A.Cecchi, Rish, I., Garg, R., Rao, A., 2009. Prediction and Interpretation of Distributed Neural Activity with Sparse Models. Neuroimage (44(1)), 112–22.

Cecchi, G., Rish, I., Thyreau, B., Thirion, B., Plaze, M., Paillere-Martinot, M., C. Martelli, J.L. Martinot, J. P., 2009. Discriminiative network models of schizophrenia. In: Proc. of NIPS-09.

Chandalia, G., Rish, I., 2007. Blind Source Separation Approach to Performance Diagnosis and Dependency Discovery. In: In Proceedings of IMC-2007.

Chen, S. S., Donoho, D. L., Saunders, M. A., 1999. Atomic Decomposition by Basis Pursuit. SIAM Journal on Scientific Computing (20), 33–61.
URL `http://ww~-stat.stanford.ed~~donoho~s/l995/3040l`

d'Aspremont, A., Ghaoui, L., 2008. Testing the Nullspace Property using Semidefinite Programming. Arxiv preprint arXiv:0807.3520.

d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., Lanckriet, G. R. G., 2007. A direct formulation for sparse pca using semidefinite programming. SIAM Review 49 (3), 434Ű448.

Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics 57, 1413–1457.

Dempster, A. P., March 1972. Covariance selection. Biometrics 28 (1), 157–175.

Donoho, D., July 2006a. For most large underdetermined systems of linear equations, the minimal ℓ_1 norm near-solution approximates the sparsest near-solution. Communications on Pure and Applied Mathematics 59 (7), 907–934.

References

Donoho, D., June 2006b. For most large underdetermined systems of linear equations, the minimal ℓ_1 norm solution is also the sparsest solution. Communications on Pure and Applied Mathematics 59 (6), 797–829.

Donoho, D. L., 2006c. Compressed sensing. IEEE Trans. Inform. Theory. 52, n. 4, 1289–1306.

Donoho, D. L., 2006d. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. Comm. Pure Appl. Math. 59, no. 6, 797–829, .

Donoho, D. L., Elad, M., Temlyakov, V., 2006. Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. Theory 52, no. 1, 6–18.

Donoho, D. L., Johnstone, I. M., 1994. Ideal denoising in an orthonormal basis chosen from a library of bases. C. R. Acad. Sci. Paris Sèr. I Math. 319, 1317–1322.

Donoho, D. L., Stark, P. B., 1989. Uncertainty principles and signal recovery. SIAM J. Appl. Math. 49, 906–931.

Duchi, J., Gould, S., Koller, D., 2008. Projected subgradient methods for learning sparse gaussians. In: Proc. of UAI-08.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (1), 407–499.

Elad, M., Aharon, M., 2006. Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing 15 (12), 3736Ű3745.

Fan, J., Li, R., 2005. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

Frank, I., Friedman, J., 1993. A statistical view of some chemometrics regression tools. Technometrics 35 (2), 109–148.

Friedman, J., Hastie, T., Hoefling, H., Tibshirani, R., 2007a. Pathwise coordinate optimization. Annals of Applied Statistics 2 (1), 302–332.

Friedman, J., Hastie, T., Tibshirani, R., 2007b. Sparse inverse covariance estimation with the graphical lasso. Biostatistics.

Fu, W., 1998. Penalized regressions: the bridge vs. the lasso. Journal of Computational and Graphical Statistics 7 (3).

References

Fuchs, J., 2005. Sparsity and uniqueness for some specific underdetermined systems. In IEEE International Conference on Acoustics, Speech and Signal Processing.

Garg, R., Khandekar, R., 2009. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In: ICML. p. 43.

Greenshtein, E., Ritov, Y., 2004. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli 10 (6), 971–988.

Hardoon, D. R., Shawe-Taylor, J., 2008. Sparse canonical correlation analysis. Sparsity and Inverse Problems in Statistical Theory and Econometrics.

Hoerl, A., Kennard, R., 1988. Ridge regression. Encyclopedia of Statistical Sciences 8 (2), 129–136.

Honorio, J., Ortiz, L., Samaras, D., Paragios, N., Goldstein, R., 2009. Sparse and locally constant gaussian graphical models. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. pp. 745–753.

Hoyer, P. O., 2004. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research 5, 1457–1469.

Huang, S., Li, J., Sun, L., Liu, J., Wu, T., Chen, K., Fleisher, A., Reiman, E., Ye, J., 2009. Learning brain connectivity of alzheimer's disease from neuroimaging data. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. pp. 808–816.

Jacob, L., Obozinski, G., Vert, J.-P., 2009. Group lasso with overlap and graph lasso. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, New York, NY, USA, pp. 433–440.

Jogdeo, K., Samuels, S., 1968. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. The Annals of Mathematical Statistics 39 (4), 1191–1195.

Juditsky, A., Karzan, F., Nemirovski, A., 2009. Verifiable conditions of l-recovery of sparse signals with sign restrictions. ArXiv.

Juditsky, A., Nemirovski, A., 2008. On verifiable sufficient conditions for sparse signal recovery via l1 minimization. ArXiv 809.

References

Knight, K., Fu, W., 2000a. Asymptotics for lasso-type estimators. Ann. Statist. 28 (5), 1356–1378.

Knight, K., Fu, W., 2000b. Asymptotics for lasso-type estimators. Annals of Statistics 28 (5), 1356–1378.

Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E., Jordan, M. I., 2004. Learning the Kernel Matrix with Semidefinite Programming. J. Mach. Learn. Res. 5, 27–72.

Lauritzen, S., 1996. Graphical Models. Oxford University Press.

Lee, S., Ganapathi, V., Koller, D., 2007. Efficient structure learning of Markov networks using ℓ_1-regularization. In: NIPS 19.

Lin, Y., Lee, D., Kim, Y., Taskar, B., 2009. Learning Markov Network Structure via Sparse Ensemble-of-Trees Models. In: AISTATS-09.

Liu, H., Palatucci, M., Zhang, J., June 2009a. Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery. In: International Conference on Machine Learning (ICML09).

Liu, J., Ji, S., Ye, J., 2009b. Multi-task feature learning via efficient $\ell_{2,1}$-norm minimization. In: Uncertainty in Artificial Intelligence.

Lozano, A., Abe, N., Liu, Y., Rosset, S., 2009a. Grouped graphical granger modeling methods for temporal causal modeling. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, pp. 577–586.

Lozano, A., Swirszcz, G., Abe, N., 2009b. Grouped orthogonal matching pursuit for variable selection and prediction. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., Culotta, A. (Eds.), Advances in Neural Information Processing Systems 22. pp. 1150–1158.

Lv, J., Fan, Y., 2009. A unified approach to model selection and sparse recovery using regularized least squares. The Annals of Statistics 37 (6A), 3498–3528.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., 2009. Online dictionary learning for sparse coding. In: ICML-09.

Mallat, S., Zhang, Z., 1993. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing 41, 3397–3415.

Meier, L., van de Geer, S., Bühlmann, P., 2008. The group lasso for logistic regression. J. Royal Statistical Society: Series B 70 (1), 53–71.

References

Meinshausen, N., 2007. Lasso with relaxation. Computational Statistics and Data Analysis 52 (1), 374–293.

Meinshausen, N., Buehlmann, P., 2008. Stability Selection. URL http://www.citebase.org/abstract?id=oai:arXiv.org:0809%.2932

Meinshausen, N., Buhlmann, P., 2006. High dimensional graphs and variable selection with the Lasso. Annals of Statistics 34(3), 1436–1462.

Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics 37 (1), 246–270.

Mendelson, S., Pajor, A., Tomczak-Jaegermann, N., 2008. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. Constructive Approximation 28 (3), 277–289.

Moghaddam, B., Weiss, Y., Avidan, S., 2007. Spectral bounds for sparse pca: Exact and greedy algorithms. In: Advances in Neural Information Processing Systems 19.

Nesterov, Y., 2004. Introductory lectures on convex optimization: A basic course. Springer Netherlands.

Obozinski, G., Taskar, B., Jordan, M. I., 2010. Joint covariate selection and joint subspace selection for multiple classification problems. Statistics and Computing 20 (2), 231–252.

Obozinski, G., Wainwright, M., Jordan, M., 2009. High-dimensional support union recovery in multivariate regression. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. pp. 1217–1224.

Osborne, M., Presnell, B., Turlach, B., 2000a. On the LASSO and its dual. Journal of Computational and Graphical Statistics 9 (2), 319–337.

Osborne, M., Presnell, B., Turlach, B., 2000b. A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis 20 (3), 389–403.

Quattoni, A., Carreras, X., Collins, M., Darrell, T., 2009. An efficient projection for $\ell_{1,\infty}$ regularization. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, New York, NY, USA, pp. 857–864.

Raina, R., Battle, A., Lee, H., Packer, B., Ng, A., 2007. Self-taught learning: Transfer learning from unlabeled data. In: In Proc. ICML-07.

Rish, I., Cecchi, G. A., Baliki, M. N., Apkarian, A. V., August 2010. Sparse Regression Models of Pain Perception. In: Proc. of Brain Informatics (BI-2010).

References

Rish, I., Grabarnik, G., September 2009. Sparse signal recovery with exponential-family noise. In: Proc. of Allerton-09.

Rish, I., Grabarnik, G., Cecchi, G., Pereira, F., Gordon, G., July 2008. Closed-form Supervised Dimensionality Reduction with Generalized Linear Models. In: Proc. of ICML-08.

Rockafellar, R., 1996. Convex analysis. Princeton Univ Pr.

Roth, V., Fischer, B., 2008. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In: ICML '08: Proceedings of the 25th international conference on Machine learning. ACM, New York, NY, USA, pp. 848–855.

Rudelson, M., Vershynin, R., 2006. Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In: Information Sciences and Systems, 2006 40th Annual Conference on. pp. 207–212.

Scheinberg, K., Asadi, N. B., Rish, I., 2009. Sparse MRF Learning with Priors on Regularization Parameters. Tech. Rep. RC24812, IBM T.J. Watson Research Center.

Schmidt, M., Murphy, K., 2010. Convex Structure Learning in Log-Linear Models: Beyond Pairwise Potentials. In: Proc. of AISTATS-10.

Schmidt, M., Niculescu-Mizil, A., Murphy, K., 2007. Learning graphical model structure using ℓ_1-regularization paths. In: AAAI-2007.

Sriperumbudur, B. K., Torres, D. A., Lanckriet, G. R. G., 2009. A d.c. programming approach to the sparse generalized eigenvalue problem. Tech. Rep. 0901.1504v2, ArXiv.

Tao, T., 2005. An uncertainty principle for cyclic groups of prime order. Math. Res. Letters 12, 121–127.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58 (1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society Series B, 91–108.

Tropp, J., 2006. Algorithms for simultaneous sparse approximation, Part II: convex relaxation. Signal Proc. 86 (3), 589–602.

Turlach, B., Venables, W., Wright, S., 2005. Simultaneous variable selection. Technometrics 47 (3), 349–363.

References

Vandenberghe, L., Boyd, S., Wu, S., 1998. Determinant maximization with linear matrix inequality constraints. SIAM J. Matrix Anal. Appl. (19), 499–533.

Wainwright, M., 2009a. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). IEEE Trans. Information Theory 55, 2183–2202.

Wainwright, M., May 2009b. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory 55, 2183–2202.

Wainwright, M., Ravikumar, P., Lafferty, J., 2007. High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression. In: NIPS 19. pp. 1465–1472.

Wu, T., Lange, K., 2008. Coordinate descent procedures for lasso penalized regression. Annals of Applied Statistics 2 (1), 224–244.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B 68, 49–67.

Yuan, M., Lin, Y., 2007a. Model Selection and Estimation in the Gaussian Graphical Model. Biometrika 94(1), 19–35.

Yuan, M., Lin, Y., 2007b. On the non-negative garrotte estimator. Journal of The Royal Statistical Society Series B 69 (2), 143–161.

Zhao, P., Rocha, G., Yu, B., 2009. Grouped and hierarchical model selection through composite absolute penalties. Annals of Statistics 37 (6A), 3468–3497.

Zhao, P., Yu, B., November 2006a. On model selection consistency of lasso. J. Machine Learning Research (7), 2541–2567.

Zhao, P., Yu, B., 2006b. On model selection consistency of Lasso. The Journal of Machine Learning Research 7, 2563.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society, Series B 67 (2), 301–320.

Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. Journal of Computational and Graphical Statistics (15), 265–286.