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)

- **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?
Sparse Statistical Models: Prediction + Interpretability

Data
\[ x - fMRI \text{ voxels}, \]
\[ y - \text{mental state} \]

Small number of Predictive Variables?

Predictive Model
\[ y = f(x) \]

- Sparsity \rightarrow \text{variable selection} \rightarrow \text{model interpretability}
- Sparsity \rightarrow \text{regularization} \rightarrow \text{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?
``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):

$$\min_{x} \|x\|_1 \text{ s.t. } y = Ax$$
And what if there is noise in observations?
Still, can reconstruct the input accurately (in \(l_2\)-sense), for \(A\) satisfying RIP; just solve a noisy version of our \(l_1\)-optimization:

\[
\min_x \|x\|_1 \quad \text{s.t.} \quad \|y - Ax\|_2^2 \leq \epsilon
\]

\[
\uparrow
\]

\[
\min_x \|y - Ax\|_2^2 \quad \text{s.t.} \quad \|x\|_1 \leq t
\]  (Basis Pursuit, aka Lasso)
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 ($l_1/l_q$, 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).
Regularization constrains the model space to avoid overfitting:

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

\[
\uparrow
\]

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

- \( Z = \{Z^1, \ldots, Z^n\} \) - data (e.g., \( Z^i = (X_{(i,:)}, y_i) \))
- \( \beta \) - 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)
- \( \lambda \) - regularization parameter
Bayesian Interpretation: MAP Estimation

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

\[
\arg \max_{\beta} \log P(Z|\beta) P(\beta|\lambda) \\
\iff \\
\arg \min_{\beta} - \log P(Z|\beta) - \log P(\beta|\lambda) \\
\iff \\
\arg \min_{\beta} L(Z, \beta) + R(\beta, \lambda)
\]
Best Subset Selection

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

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

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 $q$

$q = 4$ \hspace{1cm} $q = 2$ \hspace{1cm} $q = 1$ \hspace{1cm} $q = 0.5$ \hspace{1cm} $q = 0.1$

- 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** \( \iff \) **variable selection**
Equivalent Constrained Formulation: A Geometric View

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

- **$p \leq n$**:
  - 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 \( \lambda \)) 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 \( \lambda \), optimizes each parameter at a time; using warm-starts, it can compute the solutions on a grid of \( \lambda \) 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](http://dsp.rice.edu/cs)
At step $k$, LARS estimate $\mu_k$ moves towards the current OLS estimate $\bar{y}_k$ in the direction $\mathbf{u}_k$ equiangular among the current predictors.

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

Three scenarios (Tibshirani, 1996):

<table>
<thead>
<tr>
<th></th>
<th>Best Subset</th>
<th>Ridge</th>
<th>Lasso</th>
</tr>
</thead>
<tbody>
<tr>
<td>a few large $\beta_i$</td>
<td>best</td>
<td>worst</td>
<td>2nd</td>
</tr>
<tr>
<td>medium number of moderate $\beta_i$</td>
<td>worst</td>
<td>2nd</td>
<td>best</td>
</tr>
<tr>
<td>large number of small $\beta_i$</td>
<td>worst</td>
<td>best</td>
<td>2nd</td>
</tr>
</tbody>
</table>
Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Modeling: Beyond Lasso
  - Consistency-improving extensions
  - Beyond $l_1$-regularization ($l_1/l_q$, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues
Let $X_S$ be the columns of the nonzero variables in the true model (support), and let $X_{Sc}$ be the remaining columns (complement).

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

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

states that the least-squares regression coefficients (i.e., correlations) for the non-essential variables ($X_{Sc}$ 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 $\lambda$ parameter.
Outline

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

\[ \text{Loss}(\mathbf{x}) + \lambda \| \mathbf{x} \|_1 \]

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

Other likelihoods (loss functions)

Adding structure beyond sparsity

- Elastic Net
- Fused Lasso
- Block l1-lq 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)

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

Elastic Net penalty:

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

where \( \alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1} \)

- \( l_1 \) keeps singularities at vertices ⇒ sparsity
- \( l_2 \) enforces strictly convex edges ⇒ 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

Predicting ‘Instructions’ (auditory stimulus)

Small grouping effect: $\lambda_2 = 0.1$

Larger grouping effect: $\lambda_2 = 2.0$

Higher $\lambda_2 \rightarrow$ selection of more voxels from correlated clusters $\rightarrow$ larger, more spatially coherent clusters

(Carroll, Cecchi, Rish, Garg, Rao 2009)
Among almost equally predictive models, increasing $\lambda_2$ 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 $\lambda_2$) 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):

\[
\min_{\beta} \| y - X \beta \|_2^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|
\]
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 $\{1, \ldots, p\}$:

  $l_1$ 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 \( \iff \) same-variable coefficients across tasks (Obozinski et al., 2010, 2009; Liu et al., 2009b)
Beyond Lasso: General Log-likelihood Losses

\[ \text{Loss}(\mathbf{x}) + \lambda \| \mathbf{x} \|_1 \]

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

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)

\[ X = \{X_1, \ldots, X_p\}, \quad G = (V, E) \]

\[ P(X) = \frac{1}{Z} \prod_{C \in \text{Clique}} \Phi_C(X_C) \]

Lack of edge \((i, j) \rightarrow\) conditional independence \(X_i \perp X_j \mid \text{rest}\)

Gaussian Markov Networks (GMRFs):

- \[ P(x) = (2\pi)^{-\frac{p}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \]

- \( \Sigma \) - covariance matrix, \( \Sigma^{-1} \) - precision (concentration) matrix

- Zeros in \( \Sigma \): 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):

  \[
  X \approx UV^T
  \]

  \[
  \min_{U,V} \|X - UV^T\|^2 + \lambda \sum_{i=1}^n \|U(i,:)\|_1 \quad \forall j = 1, \ldots, m, \|V(:,j)\|_2 \leq 1
  \]

  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)
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)
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 $L = 5, 10, 15, 20, 25$

<table>
<thead>
<tr>
<th>method</th>
<th>$L$</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian-SDR</td>
<td>0.21</td>
<td>0.26</td>
<td>0.23</td>
<td><em><strong>0.20</strong></em></td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>Logistic-UDR</td>
<td>0.44</td>
<td>0.42</td>
<td><em><strong>0.29</strong></em></td>
<td>0.30</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>SVM-UDR</td>
<td><em><strong>0.49</strong></em></td>
<td>0.52</td>
<td>0.56</td>
<td>0.57</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>SVDM</td>
<td>0.32</td>
<td>0.25</td>
<td><em><strong>0.21</strong></em></td>
<td>0.23</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.21</td>
</tr>
</tbody>
</table>

- 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 ($l_1/l_q$, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
  - Beyond variable-selection: variable construction
- Summary and Open Issues
Beyond LASSO

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

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

Other likelihoods (loss functions)

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

Adding structure beyond sparsity
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

\[
\log p_{\psi, \theta}(y) = y \theta - \psi(\theta) + \log p_0(y)
\]

Examples: Gaussian, exponential, Bernoulli, multinomial, gamma, chi-square, beta, Weibull, Dirichlet, Poisson, etc.
Generalized Linear Models (GLMs)

\[ E_{\rho_{\psi, \theta}}(y) = f^{-1}(Ax) \]

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

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

\[ E(y) = Ax \]

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

\[ E(y) = \frac{1}{1 + e^{-Ax}} \]
Bijection Theorem (Banerjee et al, 2005):

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

Fitting GLM $\iff$ maximizing exp-family likelihood $\iff$ minimizing Bregman divergence
Sparse Signal Recovery from Noisy Observations

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

If

- small observation noise: \( ||y - Ax^0||_2 \leq \epsilon \)
- \( A \) satisfies the restricted isometry property (RIP)

Then the solution to the sparse linear regression problem

\[
x^* = \arg\min_x ||x||_1 \quad s.t. \quad ||y - Ax||_2 \leq \epsilon
\]

is a good approximation of \( x^0 \), i.e. \( ||x^* - x^0||_2 \leq C_S \cdot \epsilon \).

Generalized Linear Models: (Rish and Grabarnik, 2009)

replace Euclidean distances \( ||y - Ax^0||_2 \) and \( ||y - Ax||_2 \) 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)

1. Noise is small:  
   \[ d_{\phi_i}(y_i, \mu(A_i; x^0)) \leq \epsilon \]

2. Restricted Isometry Property (RIP)  
   \[ ||x^0||_1 \leq s \]

3. 

4. Bounded \( \phi_i''(y) \)

Then the solution \( x^* \) to the sparse GLM regression problem

\[
\min ||x||_1 \quad \text{subject to} \quad \sum_i d(y_i, \mu(A_i x)) \leq \epsilon
\]

is a good approximation of \( x^0 \), i.e.  
\[ ||x^* - x^0||_2 \leq C_S \cdot \delta(\epsilon) \]
\( \delta(\epsilon) \) - continuous monotone increasing function, and \( \delta(0) = 0 \) (i.e. \( \delta(\epsilon) \) is small when \( \epsilon \) 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 $\phi$.

- Results also hold for compressible (rather than sparse) signals.
Beyond LASSO

\[ \text{Loss}(x) + \lambda \| x \|_1 \]

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

Other likelihoods (loss functions)

Adding structure beyond sparsity

- Elastic Net
- Fused Lasso
- Block $l_1$-$l_q$ norms:
  - group Lasso
  - simultaneous Lasso
Markov Networks (Markov Random Fields)

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

\[ P(X) = \frac{1}{Z} \prod_{C \in \text{Clique}} \Phi_C(X_C) \]

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

Gaussian Markov Networks (GMRFs):

- \(P(x) = (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)\)
- \(\Sigma\) - covariance matrix, \(\Sigma^{-1}\) - precision (concentration) matrix
- Zeros in \(\Sigma\): 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)
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)

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 $X$ are centered to have zero mean. Then:

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

$$= \arg \max_{\Sigma > 0} \log \det(C) - \text{tr}(SC)$$

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

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

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 (Vandenberghhe 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 $\lambda$ (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 posteriori 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 \( \lambda \)

ROC Curve

Bayesian Lambda

True Positive

Cross Validation Lambda

Banerjee's Lambda

False Positive
Outline

- Introduction
- Sparse Linear Regression: Lasso
- Sparse Signal Recovery and Lasso: Some Theory
- **Sparse Modeling: Beyond Lasso**
  - Consistency-improving extensions
  - Beyond $l_1$-regularization ($l_1/l_q$, 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):

  \[ X \sim UV^T \]

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

  \[
  \min_{U,V} \|X - UV^T\|^2 + \lambda \sum_{i=1}^{n} ||U(i,:)||_1 \\
  \forall j = 1, \ldots, m, ||V(:,j)||_2 \leq 1
  \]

- **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 ($l_1/l_q$, Elastic Net, fused Lasso)
  - Beyond linear model (GLMs, MRFs)
  - Sparse Matrix Factorizations
    - Beyond variable-selection: variable construction
- Summary and Open Issues
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)

![Diagram](attachment:image.png)

Generalized Linear Models (GLMs)

\[
E(X_d) = f_d^{-1}(UV_d)
\]

\[
E(Y_k) = f_k^{-1}(UW_k)
\]

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

\[
X \sim U V \quad \text{and} \quad Y \sim U V
\]
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)
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

<table>
<thead>
<tr>
<th>method</th>
<th>L = 5</th>
<th>L = 10</th>
<th>L = 15</th>
<th>L = 20</th>
<th>L = 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian-SDR</td>
<td>0.21</td>
<td>0.26</td>
<td>0.23</td>
<td><strong>0.20</strong></td>
<td>0.23</td>
</tr>
<tr>
<td>Logistic-UDR</td>
<td>0.44</td>
<td>0.42</td>
<td><strong>0.29</strong></td>
<td>0.30</td>
<td>0.26</td>
</tr>
<tr>
<td>SVM-UDR</td>
<td><strong>0.49</strong></td>
<td>0.52</td>
<td>0.56</td>
<td>0.57</td>
<td>0.55</td>
</tr>
<tr>
<td>SVDM</td>
<td>0.32</td>
<td>0.25</td>
<td><strong>0.21</strong></td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.21</td>
</tr>
</tbody>
</table>

- 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

Predictive Model
\[ y = f(x) \]

Interpretable Predictive Patterns

+ happy
- sad


References


Donoho, D., July 2006a. For most large underdetermined systems of linear equations, the minimal $\ell_1$ norm near-solution approximates the sparsest near-solution. Communications on Pure and Applied Mathematics 59 (7), 907–934.
Donoho, D., June 2006b. For most large underdetermined systems of linear equations, the minimal 
\( \ell_1 \) norm solution is also the sparsest solution. Communications on Pure and Applied 
Mathematics 59 (6), 797–829.


Donoho, D. L., 2006d. For most large underdetermined systems of linear equations the minimal 
\( \ell_1 \)-norm solution is also the sparsest solution. Comm. Pure Appl. Math. 59, no. 6, 797–829.


49, 906–931.

In: Proc. of UAI-08.

407–499.

Elad, M., Aharon, M., 2006. Image denoising via sparse and redundant representations over 

Fan, J., Li, R., 2005. Variable selection via nonconcave penalized likelihood and its oracle 

Technometrics 35 (2), 109–148.


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


References


