Convex Relaxation Formulations for Structured Sparsity

Tong Zhang

Statistics Department
Rutgers University

Outline of the Talk

- Motivation of Structured Sparsity and how to quantify "structure"
  - information theoretical characterization
  - other possibilities
- How much can we benefit from "structures" ?
  - review of group Lasso analysis: from info. theoretical view
- Convex relaxation for composite structures
  - additive composition of regularizers
  - additive composition of covariance
    - theoretical justification

Outline of the Talk

- Motivation of Structured Sparsity and how to quantify "structure"
  - information theoretical characterization
  - other possibilities
- How much can we benefit from "structures" ?
  - review of group Lasso analysis: from info. theoretical view
- Convex relaxation for composite structures
  - additive composition of regularizers
  - additive composition of covariance
    - theoretical justification
- Work in progress :
  - talk focuses on high level ideas
  - no empirical results yet
  - theory suggests additive covariance formulation is a worthy alternative

Sparse Regression Problem

- Model: $Y = X\bar{\beta} + \epsilon$
  - $Y \in R^n$: observation
  - $X \in R^{n \times p}$: design matrix
  - $\bar{\beta} \in R^p$: parameter vector to be estimated
  - $\epsilon \in R^n$: zero mean sub-Gaussian noise with variance σ^2
- Sparsity: $\bar{\beta}$ has few nonzero components
  - $\mathrm{supp}(\bar{\beta}) = \{j : \bar{\beta}_j \neq 0\}$.
  - $\|\bar{\beta}\|_0 = |\mathrm{supp}(\bar{\beta})|$ is small: $\ll n$

Structured sparsity

- Wavelet domain: sparsity pattern not random (structured)

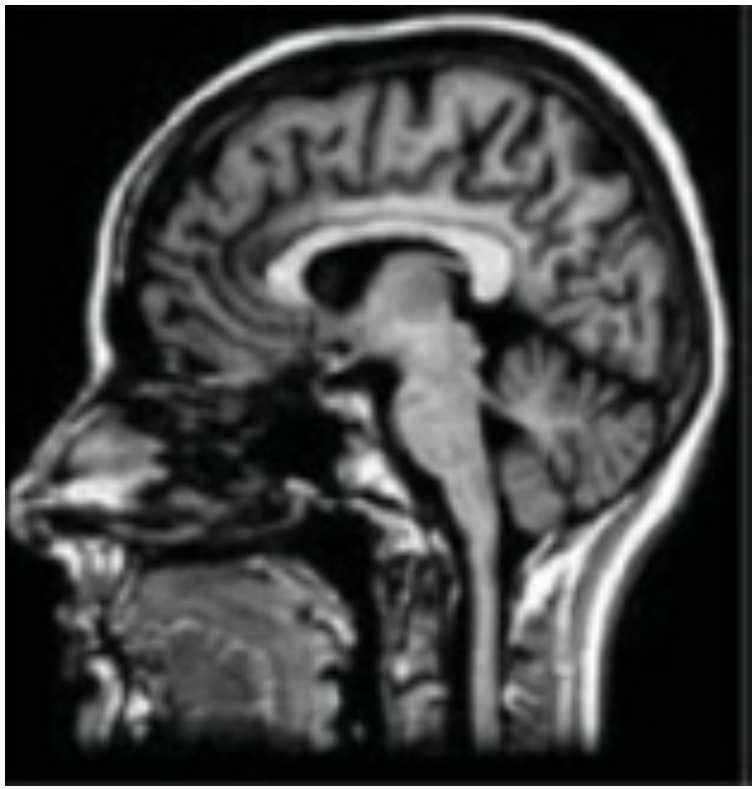
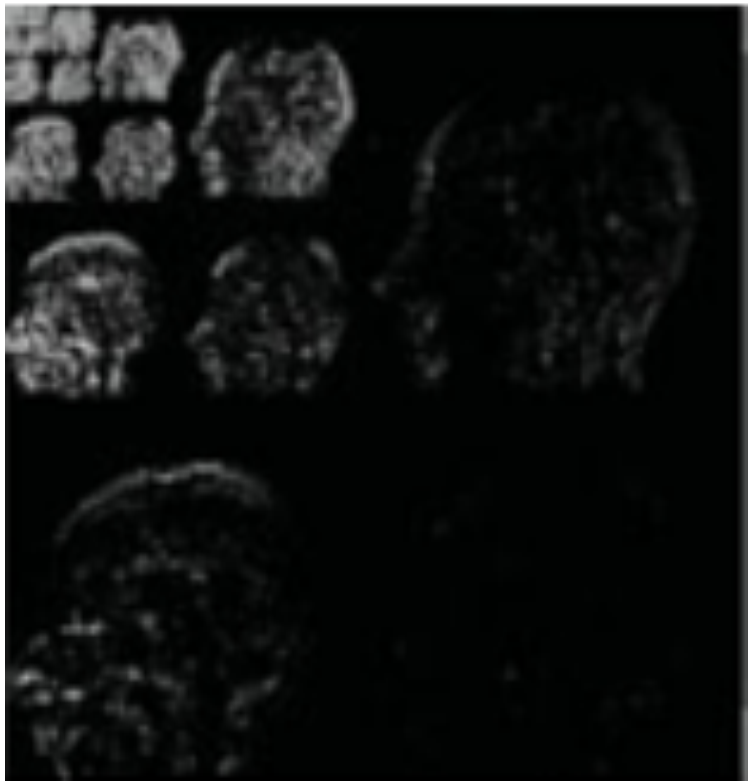


Image domain



Wavelet domain

- can we take advantage of structure?

Structured Sparsity Characterization

- Observation:
  - sparsity pattern: $\mathrm{supp}(\bar{\beta})$
  - not all sparse patterns are equally likely
- Approaches to formalize the intuition
  - approach 1: information theoretical characterization on sparsity pattern
  - approach 2: also look into the correlation of features (variables)
  - approach 3: also look into the magnitude of coefficients
- We focus on approach 1 in this talk.
  - consider the situation where features are weakly correlated (e.g. RIP)

Structured Sparsity: Information Theoretical Characterization

- $F = \mathrm{supp}(\beta)$: sparsity pattern
- Prior knowledge on the sparsity pattern F:
  - e.g., neighboring variables are likely to become nonzeros simultaneously
- Each sparsity pattern F is associated with cost $c(F)$
  - $c(F)$: proportional to negative log-likelihood of (prior of) F
- Nonconvex formulation: complexity penalization

$$\min_{\beta} \|X\beta - Y\|_2^2 \qquad \text{subject to } \|\beta\|_0 + c(\mathrm{supp}(\beta)) \leq s.$$

  can obtain recovery results using empirical processes
- We are interested in convex relaxation

Example: Group Structure

- Variables are divided into pre-defined groups $G_1, \ldots, G_{p/m}$
  - m variables per group
- Assumption:
  - coefficients in each group are simultaneously zeros or nonzeros
- Group sparsity pattern cost: $\|\beta\|_0 + m^{-1}\|\beta\|_0 \ln p$.
- Standard sparsity pattern cost (for Lasso): $\|\beta\|_0 \ln p$
- Convex relaxation for group sparsity: group Lasso
  - how does group Lasso improve over Lasso?

Theory of Convex Relaxation for group sparsity

- Recovery result for group Lasso: convex relaxation for group sparsity

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \|X\beta - Y\|_2^2 + \lambda \sum_{\ell} \|\beta_{G_\ell}\|_2.$$

  If coefficients $\bar{\beta}$ in each group are simultaneously zeros or nonzeros (condition can be relaxed). With appropriate λ:

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O(\sigma^2(\|\bar{\beta}\|_0 + m^{-1}\|\bar{\beta}\|_0 \ln p)/n).$$

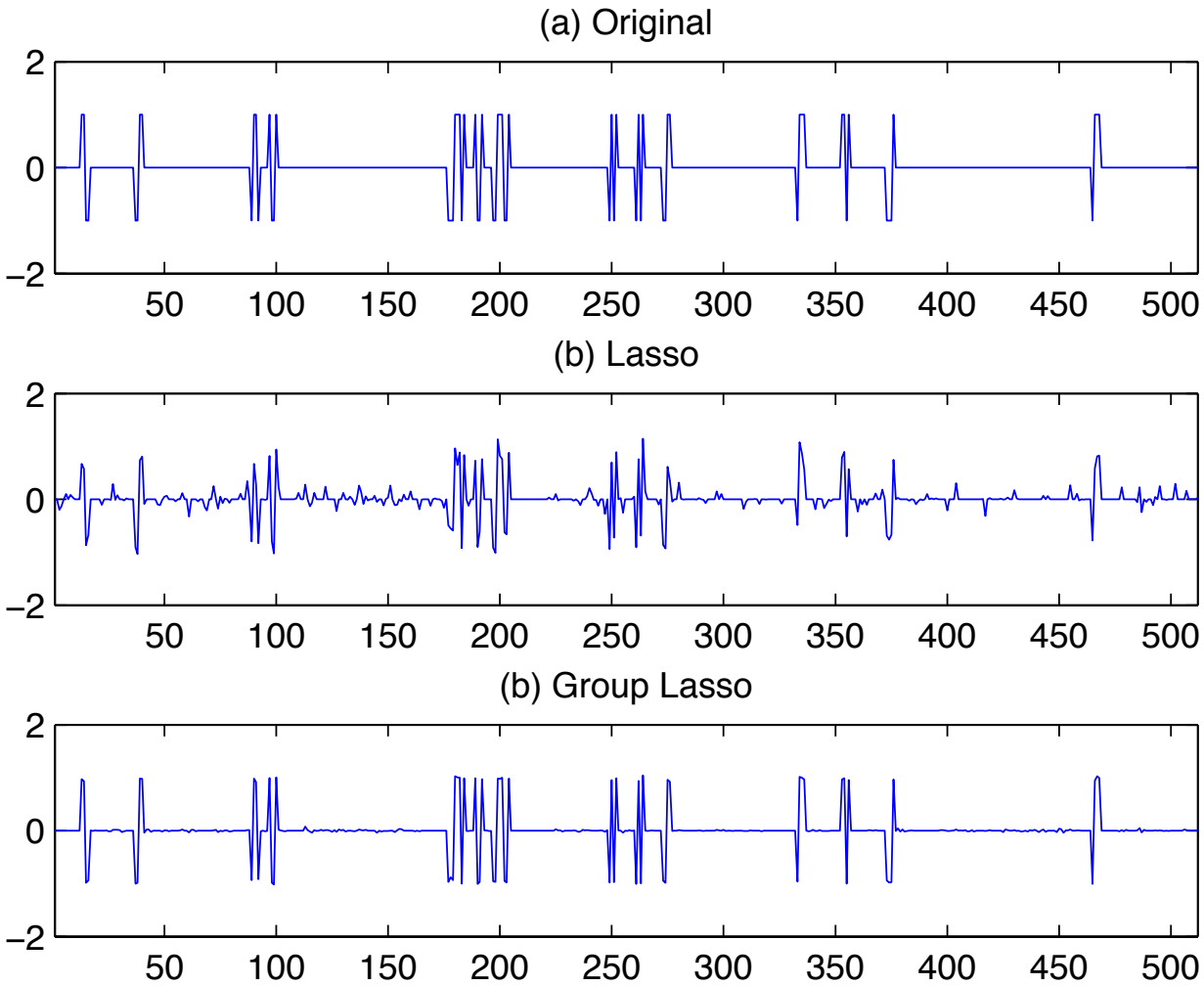
  Moreover, the bound matches information theoretical lower-bound.

- Compare to Lasso bound:

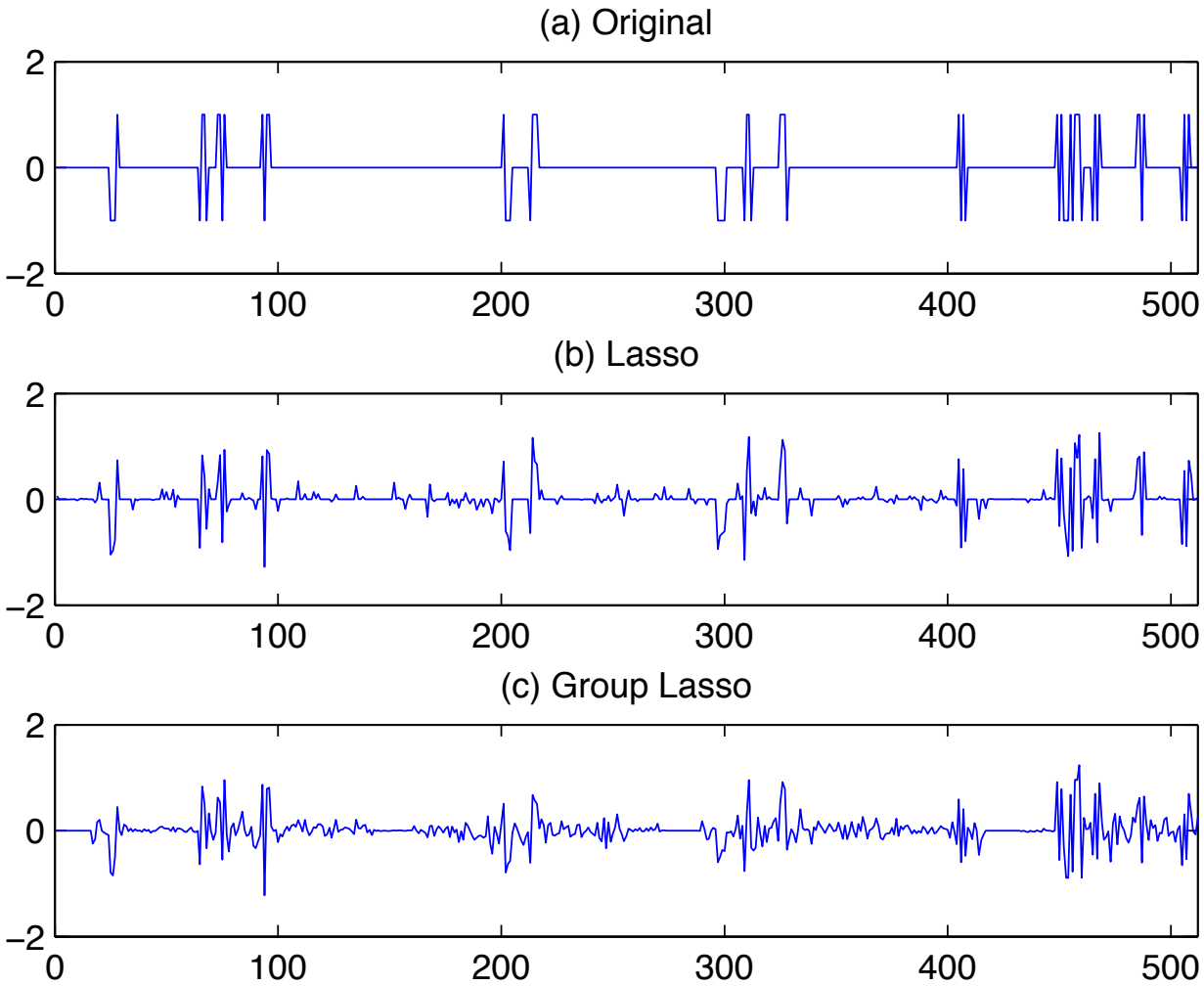
$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O(\sigma^2\|\bar{\beta}\|_0 \ln p/n).$$

- Group Lasso can be inferior when the group structure is incorrect
  - question: can we combine Lasso and group Lasso?

Group sparsity: correct group structure



Group sparsity: incorrect group structure



Simplified High Level Analysis

- Sparsity regularization
  - shrink the coefficients toward zero.
- Two effects for shrinkage:
  - positive effect: shrink noise toward zero — leads to sparse solution
  - negative effect: shrink nonzero coefficients toward zero — cause bias
- Balancing:
  - regularization has to be strong enough to dominate the noise
  - regularization should not be too strong to cause excessive bias
- Simplified analysis (assume weak correlation; such as RIP, etc):
  - assume regularization is strong enough to dominate the noise
  - then recovery performance is the bias of the regularization

Simplified noise domination condition

- Consider a regularizer $R(\beta)$
- Consider projection of noise to the variables $\xi = n^{-1} X^\top \epsilon$
  - X is weakly correlated
- $R(\beta)$ dominates noise if (roughly) $\beta = 0$ is the unique solution of

$$\min_\beta [-2\beta^\top \xi + R(\beta)].$$

- Sub-Gaussian noise property: if ϵ is sub-Gaussian, then
  - $|\xi_j| = O(\sigma\sqrt{\ln p/n})$ for all variable j
  - $\|\xi_G\|_2 = O(\sigma\sqrt{(m + \ln p)/n})$ for all p/m groups G

Lasso Analysis

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \|X\beta - Y\|_2^2 + \lambda \|\beta\|_1.$$

- Regularization needed to dominate sub-Gaussian noise:

$$\lambda = \Omega(\sigma \sqrt{\ln p / n})$$

- Regularizer bias
  - Each coefficient: $O(\lambda^2)$
  - Multiply by sparsity $\|\bar{\beta}\|_0$: $\lambda^2 \|\bar{\beta}\|_0$
- Recovery performance — overall bias with optimal λ

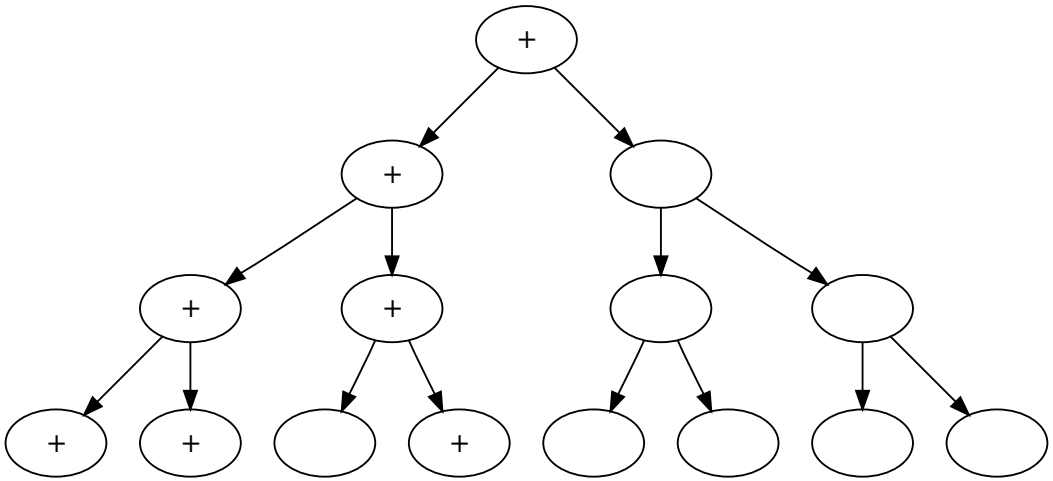
$$\|\hat{\beta} - \bar{\beta}\|_2^2 = \Omega(\sigma^2 \|\bar{\beta}\|_0 \ln p / n).$$

Group Lasso Analysis

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \|X\beta - Y\|_2^2 + \lambda \sum_{\ell} \|\beta_{G_\ell}\|_2.$$

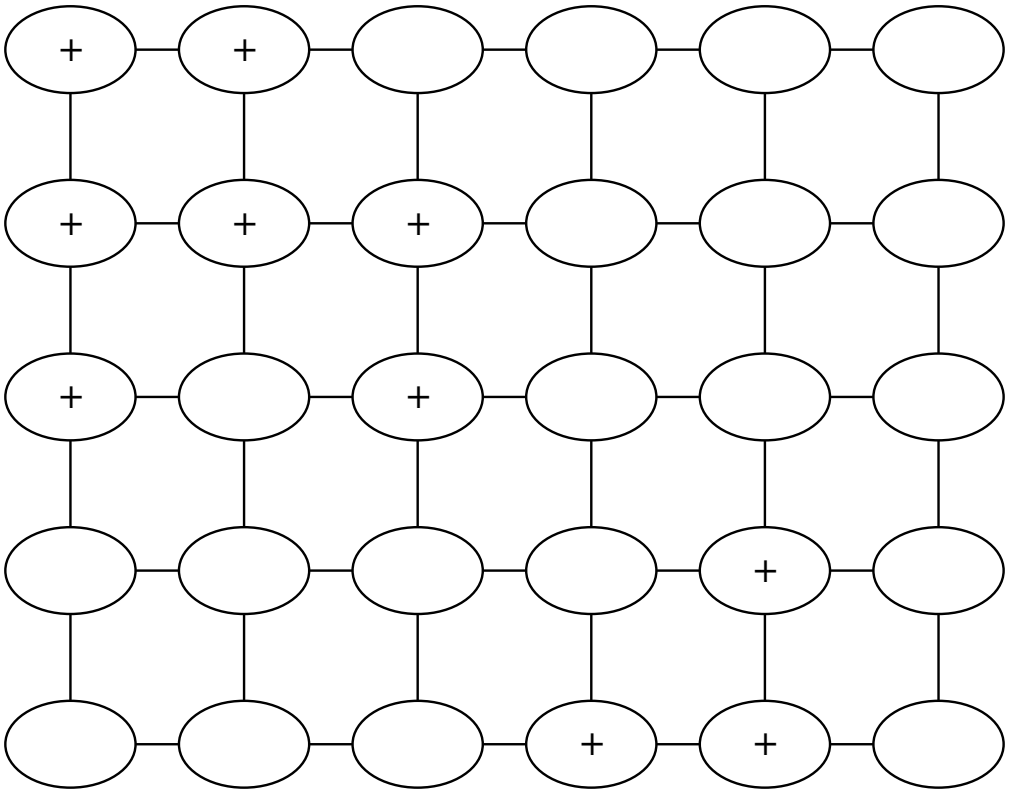
- Regularization needed to dominate sub-Gaussian noise in each group of size m:
$$\lambda = \Omega(\sigma\sqrt{(m + \ln p)/n})$$
- Regularization bias:
  - Each group: $O(\lambda^2)$
  - Multiply by nonzero-groups $\|\bar{\beta}\|_0/m$: $\lambda^2\|\bar{\beta}\|_0/m$
- Recovery performance – overall bias with optimal λ:
$$\|\hat{\beta} - \bar{\beta}\|_2^2 = \Omega(\sigma^2\|\bar{\beta}\|_0(1 + m^{-1}\ln p)/n).$$
- Remark:
  - group Lasso benefits: can dominate noise with weaker regularization

More Complex Structure: hierarchical



- Parent is nonzero if a child is nonzero
- Complexity: $O(\|\bar{\beta}\|_0)$
  - assuming if a child is a feature, then its parent is also a feature

More Complex Structure: connected region



- a nonzero pixel implies adjacent pixels more likely to be nonzeros
- Complexity: $O(\|\bar{\beta}\|_0 + q \ln p)$ (q: number of connected components)

Complex Structure by combining simple structures

- Form complex structure by combining simpler structures
  - focus on overlapping groups
- Semantics (under information theoretical framework)
  - coefficients within each group likely to be simultaneously nonzeros

Complex Structure by combining simple structures

- Form complex structure by combining simpler structures
  - focus on overlapping groups
- Semantics (under information theoretical framework)
  - coefficients within each group likely to be simultaneously nonzeros
- Desired recovery performance (info. Th. view)
  - each group has a sufficiently large regularization to dominate noise
  - recovery performance: cover $\mathrm{supp}(\beta)$ by groups with smallest overall bias
  - assume weak correlation

Complex Structure by combining simple structures

- Form complex structure by combining simpler structures
  - focus on overlapping groups
- Semantics (under information theoretical framework)
  - coefficients within each group likely to be simultaneously nonzeros
- Desired recovery performance (info. Th. view)
  - each group has a sufficiently large regularization to dominate noise
  - recovery performance: cover $\mathrm{supp}(\beta)$ by groups with smallest overall bias
  - assume weak correlation
- Two different convex relaxation approaches
  - additive composition of simpler regularizers (popular)
  - additive composition of covariance (newer or new)
- Question: which approach gives desired recovery performance?

Desirable Recovery Bound (Info. Theoretical View)

- Consider a set of groups $\mathcal{G} = \{G_\ell\}$.
  - each group j associated with complexity $c(G_\ell) = O(|G_\ell| + \ln|\mathcal{G}|)$,
  - penalty of $\sigma^2 n^{-1} c(G_\ell)$ is strong enough to dominate noise.
- F: set of nonzero coefficients; define its minimum covering cost

$$c_{\mathcal{G}}(F) = \min_{G' \subset \mathcal{G}} \sum_{G_\ell \in G'} c(G_\ell) \quad \text{subject to } F \subset \cup_{G_\ell \in G'} G_\ell$$

  - $\sigma^2 n^{-1} c_{\mathcal{G}}(F)$: bias of overlapping groups.
  - the problem itself is set covering number which is NP hard.
- Desirable recovery performance: bias under noise domination

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O(\sigma^2 n^{-1} c_{\mathcal{G}}(F)).$$

  - can be achieved with nonconvex regularization:
  - question: can this be achieved through convex relaxation?

Simple example: combine standard & group sparsity

- Combining sparsity and group sparsity to benefit from both
  - non-overlapping m-element groups $\{G_\ell\}_{\ell=1,\ldots,p/m}$
  - standard sparsity structure: single-element groups $G'_j = \{j\}_{j=1,\ldots,p}$.
- The combination has overlapping group structure $\{G_\ell\} \cup \{G'_j\}$
  - intuition 1: encourage group sparsity as well as within group sparsity
  - intuition 2: $\mathrm{supp}(\bar{\beta})$ is partially group sparse and partially standard sparse
- We are more interested in intuition 2.
  - e.g. multi-task learning with shared features plus individual features
- Minimum bias cover: $\mathrm{supp}(\bar{\beta})$ can be covered by
  - K m-element groups in $\{G_\ell\}$;
  - plus L single-element groups in $\{G'_j\}$
- Desired recovery performance

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O(\sigma^2(K(m + \ln p) + L \ln p)/n).$$

Additive regularizer composition: properties

- Additive composition of regularizers:

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \|X\beta - Y\|_2^2 + \lambda(\|\beta\|_1 + \alpha \sum_{\ell} \|\beta_{G_\ell}\|_2).$$

- Question: can we achieve the bound?

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O\left(\sigma^2(K(m + \ln p) + L \ln p)/n\right).$$

Additive regularizer composition: properties

- Additive composition of regularizers:

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \|X\beta - Y\|_2^2 + \lambda(\|\beta\|_1 + \alpha \sum_{\ell} \|\beta_{G_\ell}\|_2).$$

- Question: can we achieve the bound?

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O\left(\sigma^2(K(m + \ln p) + L \ln p)/n\right).$$

answer is no.

Additive regularizer composition: properties

- Additive composition of regularizers:

$$\hat{\beta} = \arg\min_{\beta} n^{-1} \|X\beta - Y\|_2^2 + \lambda(\|\beta\|_1 + \alpha \sum_{\ell} \|\beta_{G_\ell}\|_2).$$

- Question: can we achieve the bound?

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O\left(\sigma^2(K(m + \ln p) + L \ln p)/n\right).$$

  answer is no.
- Problems of additive composition under *weak correlation*:
  - over-counting: each coefficient is covered by multiple groups
    - extra bias with multiple penalization
  - If regularizers dominate noise, then L_1 is always stronger than group-L_1.

Additive regularizer composition: high level analysis

- Formulation:

$$\hat{\beta} = \arg\min_{\beta} n^{-1}\|X\beta - Y\|_2^2 + \lambda(\|\beta\|_1 + \alpha\sum_{\ell}\|\beta_{G_\ell}\|_2).$$

- Require L_1 & group-L_1 to dominate noise (assume weak correlation)
  - $\lambda = \Omega(\sigma\sqrt{\ln p/n})$
  - $\alpha\lambda = \Omega(\sigma\sqrt{(m+\ln p)/n})$
  - The L_1 term always cause bias of order $\sigma\|\bar{\beta}\|_0\sqrt{\ln p/n}$.
- Conclusion:
  - additive regularization isn't too help for some models (weak correlation)
    - maybe constant
  - can prove something under intuition 1, but not entirely satisfactory either
    - intuition 1: encourage group sparsity and within group sparsity
  - can be more helpful when features are "highly correlated" within group.
    - situation not considered in this talk

Alternative formulation of sparse regularization

- Main idea:
  - function β^2/μ is jointly convex in β and $\mu \geq 0$.
  - μ: covariance term which can be jointly optimized with β
- Recently explored in [Micchelli,Morales,Pontil] NIPS 2010.

Alternative formulation of sparse regularization

- Main idea:
  - function β^2/μ is jointly convex in β and $\mu \geq 0$.
  - μ: covariance term which can be jointly optimized with β
- Recently explored in [Micchelli,Morales,Pontil] NIPS 2010.
- Equivalent formulation of Lasso:

$$[\hat{\beta}, \hat{\mu}] = \arg\min_{\beta,\mu} n^{-1} \|X\beta - Y\|_2^2 + 0.5\lambda(\sum_j \beta_j^2/\mu_j + \mu_j) \quad \mu_j \geq 0.$$

- Equivalent formulation of group Lasso:

$$[\hat{\beta}, \hat{\mu}] = \arg\min_{\beta,\mu} n^{-1} \|X\beta - Y\|_2^2 + 0.5\lambda(\sum_\ell \|\beta_{G_\ell}\|_2^2/\mu_\ell + \alpha\mu_\ell) \quad \mu_\ell \geq 0.$$

Additive Covariance Composition

- Dealing with overlapping groups by combining covariance μ.
- Consider groups $\{G_\ell\}$ that might overlap.
  - μ_ℓ: covariance parameter associated with G_ℓ
- Overlapping group formulation: additively combine μ_ℓ

Additive Covariance Composition

- Dealing with overlapping groups by combining covariance μ.
- Consider groups $\{G_\ell\}$ that might overlap.
  - μ_ℓ: covariance parameter associated with G_ℓ
- Overlapping group formulation: additively combine μ_ℓ

$$[\hat{\beta}, \hat{\mu}] = \arg\min_{\beta,\mu} n^{-1}\|X\beta - Y\|_2^2 + 0.5\lambda \left(\sum_j \frac{\beta_j^2}{\underbrace{\sum_\ell \mu_\ell I(j \in G_\ell)}_{\text{groups covering } j}} + \sum_\ell \alpha_\ell \mu_\ell \right)$$

$$\mu_\ell \geq 0.$$

Also equivalent to [Jacob, Obozinski, Vert], ICML 2009.

- α_ℓ: sufficiently large to dominate noise within each group G_ℓ.

Simple Example: combine group & standard sparsity

- Combine group and standard sparsity
  - non-overlapping *m*-element groups $\{G_\ell\}$
  - standard sparsity structure, with single-element groups $G'_j = \{j\}_{j=1,\ldots,p}$.
- Additive composition of covariance formulation:

$$[\hat{\beta}, \hat{\mu}, \hat{\mu}'] = \arg \min_{\beta,\mu,\mu'} n^{-1} \|X\beta - Y\|_2^2 + 0.5\lambda \sum_{j=1}^{p} \left[\frac{\beta_j^2}{\mu'_j + \mu_{\ell(j)}} + \mu'_j + \alpha\mu_{\ell(j)} \right]$$

$$\mu_\ell \geq 0.$$

where $\ell(j)$ is the group ℓ that contains j: $j \in G_\ell$.

Simple Example: combine group & standard sparsity

- Combine group and standard sparsity
  - non-overlapping *m*-element groups $\{G_\ell\}$
  - standard sparsity structure, with single-element groups $G'_j = \{j\}_{j=1,\ldots,p}$.
- Additive composition of covariance formulation:

$$[\hat{\beta}, \hat{\mu}, \hat{\mu}'] = \arg \min_{\beta,\mu,\mu'} n^{-1} \|X\beta - Y\|_2^2 + 0.5\lambda \sum_{j=1}^{p} \left[\frac{\beta_j^2}{\mu'_j + \mu_{\ell(j)}} + \mu'_j + \alpha \mu_{\ell(j)} \right]$$

$$\mu_\ell \geq 0.$$

  where $\ell(j)$ is the group ℓ that contains j: $j \in G_\ell$.
- Does it achieve desired info theoretical recovery performance under weak correlation?
  - yes — we will go through a simplified high level analysis

Additive Covariance: bias

- Combining group sparsity and standard sparsity.
- The regularizer satisfies

$$\frac{\beta_j^2}{\mu_j' + \mu_{\ell(j)}} + \mu_j' + \alpha\mu_{\ell(j)} \leq \min\left[\frac{\beta_j^2}{\mu_j'} + \mu_j', \frac{\beta_j^2}{\mu_{\ell(j)}} + \alpha\mu_{\ell(j)}\right].$$

  - Informally, bias is the minimum of Lasso/group Lasso bias
- Minimum cover of $\mathrm{supp}(\bar{\beta})$:
  - L single-element groups
  - K m-element groups.
- The bias for the cover is of order

$$\lambda^2(L + \alpha mK),$$

which matches the form of desired bound: need to estimate λ, α

Additive Covariance: noise domination

- Combining group sparsity and standard sparsity
- Sub-Gaussian noise: $\xi = n^{-1} X^\top \epsilon$ — projection of noise to variables.
  - $\xi_j = O(\sigma\sqrt{\ln p/n})$ for each individual variable $j = 1, \ldots, p$
  - $\xi_{G_\ell} = O(\sigma\sqrt{(m + \ln p)/n})$ for each group G_ℓ
- Noise domination: the following problem has unique solution $\beta = 0$

$$\min_{\beta,\mu,\mu'} -2\beta^\top \xi + 0.5\lambda \sum_j \left[\frac{\beta_j^2}{\mu'_j + \mu_{\ell(j)}} + \mu'_j + \alpha\mu_{\ell(j)} \right]$$

  - main question: how large λ and α need to be ?
  - only need to look at single group situation

Additive Covariance: single group noise domination

- Noise domination condition with $\beta \in R^m$ as single group:

$$0 = \arg\min_{\beta,\mu,\mu'} -2\beta^\top \xi + 0.5\lambda \sum_{j=1}^{m} \left[\frac{\beta_j^2}{\mu'_j + \mu} + \mu'_j + \alpha\mu \right]$$

- Without loss of generality, assume at solution:

$$|\beta_1| \geq |\beta_2| \geq |\beta_L| > \mu \geq |\beta_{L+1}| \geq \cdots \geq |\beta_m|$$

- Solution property of μ and μ' given β:

$$\mu'_j = \max[|\beta_j| - \mu, 0], \quad \mu = \sqrt{\alpha m - L} \sqrt{\sum_{j=L+1}^{m} \beta_j^2}.$$

- Eliminating μ, μ', we get lower bound of the regularizer:

$$\sum_{j=1}^{m} \left[\frac{\beta_j^2}{\mu'_j + \mu} + \mu'_j + \alpha\mu \right] \geq \sum_{j=1}^{L} |\beta_j| + 2\sqrt{\alpha m} \sqrt{\sum_{j=L+1}^{m} \beta_j^2}.$$

Additive Covariance: single group noise domination

- Simplified noise domination condition: $\beta = 0$ is the unique solution of

$$\min_{\beta} -2\beta^\top \xi + 0.5\lambda \left[\sum_{j=1}^{L} |\beta_j| + 2\sqrt{\alpha m} \sqrt{\sum_{j=L+1}^{m} \beta_j^2} \right]$$

- Reduces separately to sparse and group regularizations:

$$\min_{\beta_{1:L}} -2\beta_{1:L}^\top \xi_{1:L} + 0.5\lambda \sum_{j=1}^{L} |\beta_j|$$

$$\min_{\beta_{L+1:m}} -2\beta_{L+1:m}^\top \xi_{L+1:m} + \lambda\sqrt{\alpha m} \sqrt{\sum_{j=1}^{L} \beta_j^2}$$

  now apply sub-Gaussian noise property

- Only need $\lambda = \Omega(\sigma\sqrt{\ln p/n})$ and $\lambda\sqrt{\alpha m} = \Omega(\sigma\sqrt{(m + \ln p)/n})$.

Additive Covariance Composition: recovery result

- Formulation to combine group sparsity and standard sparsity

$$[\hat{\beta}, \hat{\mu}, \hat{\mu}'] = \arg\min_{\beta,\mu,\mu'} n^{-1}\|X\beta - Y\|_2^2 + 0.5\lambda \sum_j \left[\frac{\beta_j^2}{\mu_j' + \mu_{\ell(j)}} + \mu_j' + \alpha\mu_{\ell(j)}\right]$$

$$\mu_\ell \geq 0.$$

- Pick $\lambda = \Omega(\sigma\sqrt{\ln p/n})$ and $\lambda\sqrt{\alpha m} = \Omega(\sigma\sqrt{(m + \ln p)/n})$.
  - noise domination holds (under weak correlation assumption)
- Recovery performance is the bias (with optimal λ and α)

$$\|\hat{\beta} - \bar{\beta}\|_2^2 = O\left(\sigma^2(L\ln p + K(m + \ln p))/n\right).$$

  - $\mathrm{supp}(\beta)$ covered by L single-element groups and K m-element groups
  - match desired information theoretical result.

Extension

- Similar argument applies to tree structures with cascaded groups:
  - partial ordering of groups such that two groups are either non-overlapping or one is the subset of the other.
- Similar results can be derived by cascading the argument.
- Further question:
  - need to study more complex overlapping structures
  - harder to decouple

Comparison of two approaches

- Problem: how to combine two sparse regularizers
- Approach 1: additive regularizer composition
  - Combined regularizer is roughly equivalent to the stronger of the two
  - Useful when neither is strong enough: complementary regularizers (e.g. non-overlapping groups)
- Approach 2: additive covariance composition
  - Combined regularizer is roughly equivalent to the weaker of the two
  - Reduce bias: useful when both regularizers are strong enough (pick one)
- Both are useful, but should be employed in the right context.
- We focus on the weak correlation case; what about correlation?

Some Remarks about Correlation

- Assume within group features are correlated
- Consider within group strongly convex regularization (e.g. L_2)
  - reduces within group model complexity
  - this effect is much weaker when features are uncorrelated.
- Additive composition of regularizers can be beneficial
  - within group complexity reduction and sparse regularization are complementary
- Question: analyze it more formally.

Summary

- Structured sparsity problem
  - focus on composite regularization from simpler sparse regularizers
- Two Convex Relaxation Formulations
  - Additive regularizer composition versus additive covariance composition
- Additive regularizer combination
  - hard to prove improvement under some assumptions (weak correlation)
  - futher study: what if variables are correlated within groups.
- Additive covariance combination
  - theoretically can be beneficial under the weak correlation case
  - a valid approach worthy of serious consideration
- Work in progress: empirical study?