ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Structured sparsity in genetics

Niko Beerenwinkel

Outline

- Linear regression
- The Lasso
- Grouped lasso
- Elastic net
- Fused lasso
- Generalized fused lasso

Linear regression

Regression

- Let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ be random variables with joint distribution P(X,Y).
- We refer to X as input, X_j as features, and Y as output.
- Regression means prediction of Y from X.
- The performance of a regression function f is usually evaluated by its squared error loss $(Y - f(X))^2$.
- With this loss function, the expected prediction error

$$\mathrm{EPE}(f) = \mathrm{E}_{(X,Y)}[(Y - f(X))^2]$$

is minimal for

$$f(x) = \mathrm{E}_Y(Y \mid X = x)$$

Linear regression

- In linear regression, the output Y is predicted from input $X = (X_1, \ldots, X_p)^T$ via the linear function f

$$\hat{Y} = f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

  with coefficients β_j.
- The intercept (or bias) β_0 is often included in the vector $\beta^T = (\beta_0, \beta_1, \ldots, \beta_p)$ and $X^T = (1, X_1, X_2, \ldots, X_p)^T$, such that

$$\hat{Y} = f(X) = X^T \beta$$

- Thus, in linear regression, we make the approximation

$$\mathrm{E}_Y(Y \mid X = x) \approx x^T \beta$$

Least squares

- How to estimate the parameters of the linear regression problem **y** = **X**β with *N* observations (**X, y**) $\in \mathbb{R}^{N \times p} \times \mathbb{R}^{N}$?
- We minimize the residual sum of squares

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

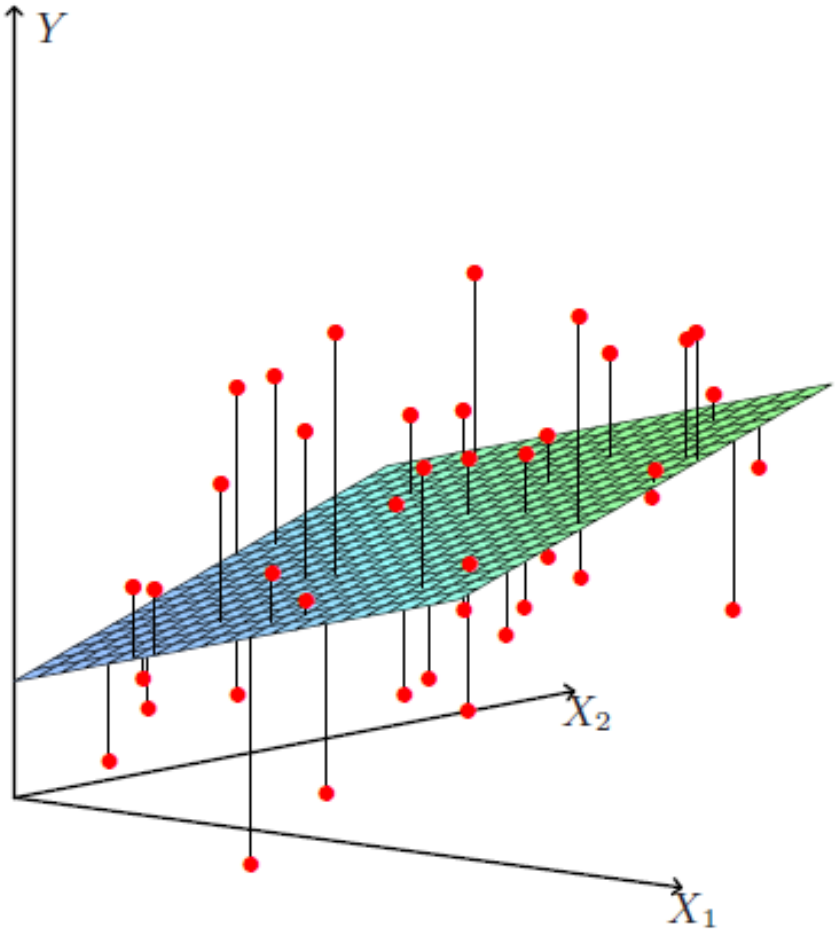
- d RSS(β) / d β = 0 gives the normal equations

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

which can be solved if $\mathbf{X}^T\mathbf{X}$ is non-singular:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Least squares



Bias-variance decomposition

- Assume $Y = f(X) + \varepsilon, \quad \mathrm{E}(\varepsilon) = 0, \ \mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$
- Then the expected prediction error at an input x is

$$\begin{aligned} \mathrm{Err}(x) &= \mathrm{E}[(Y - \hat{f}(x))^2 \mid X = x] \\ &= \sigma_\varepsilon^2 + [\mathrm{E}\,\hat{f}(x) - f(x)]^2 + \mathrm{E}[\hat{f}(x) - \mathrm{E}\,\hat{f}(x)]^2 \\ &= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x)) + \mathrm{Var}(\hat{f}(x)) \end{aligned}$$

irreducible error (variance around the target; cannot be avoided)

Bias: Difference between average prediction and true mean

Variance: Expected squared deviation of the prediction

- Typically, more complex models have lower bias but higher variance. (→ trade-off for model selection)

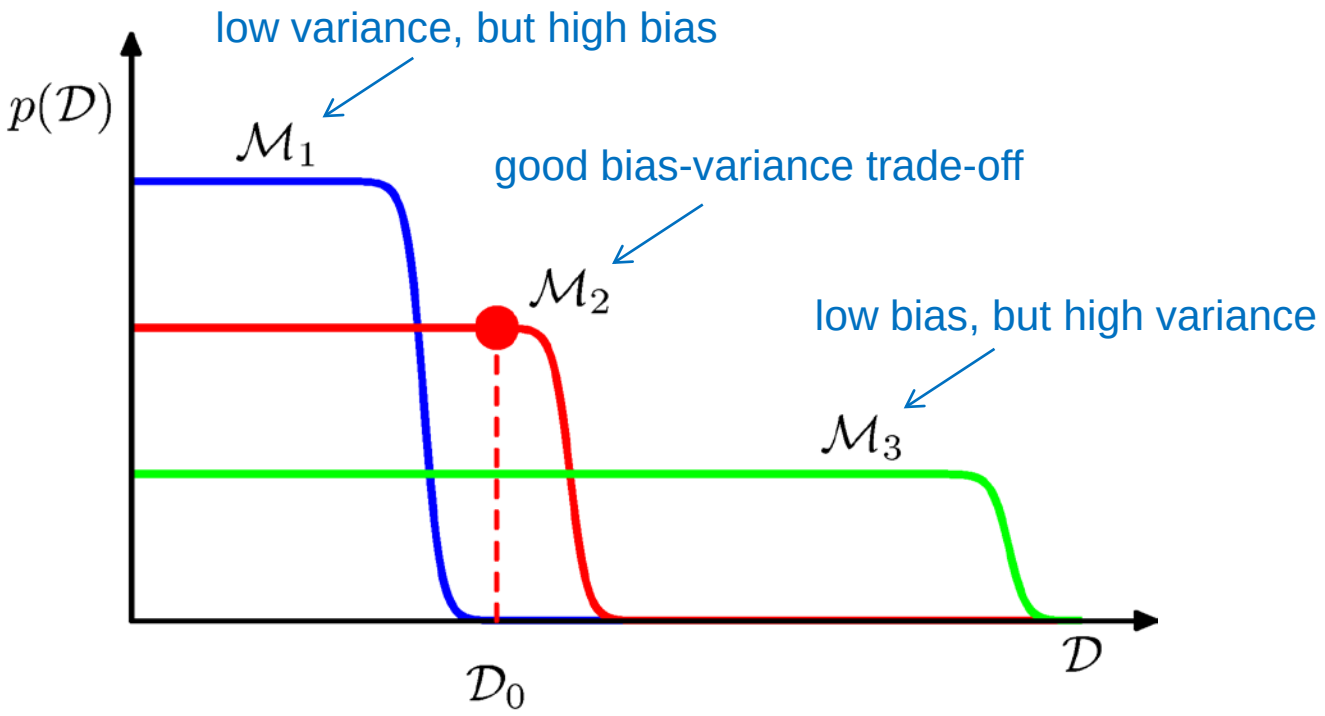
Bias-variance trade-off

- Assume $Y = f(X) + \varepsilon, \quad \mathrm{E}(\varepsilon) = 0, \ \mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$
- If the true mean of Y is $f(\mathrm{X}) = \mathrm{X}^\mathsf{T}\beta$, then the least squares estimator has no bias, and it is the unbiased estimator with minimal variance (Gauss-Markov theorem).
- However, other *biased* estimators may have smaller mean squared error:

$$\mathrm{Err}(x) \ = \ \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x)) + \mathrm{Var}(\hat{f}(x))$$

- For example, if we exclude some features ($\beta_j = 0$), the estimator will be biased, but often of lower variance.
- We seek a good trade-off between bias and variance.

Model complexity



Data sets

L_1 regularization
(least absolute shrinkage and selection operator, lasso)

The Lasso

- The lasso is a shrinkage method, defined by the quadratic program

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2$$

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t$$

or, equivalently, by its Lagrangian dual

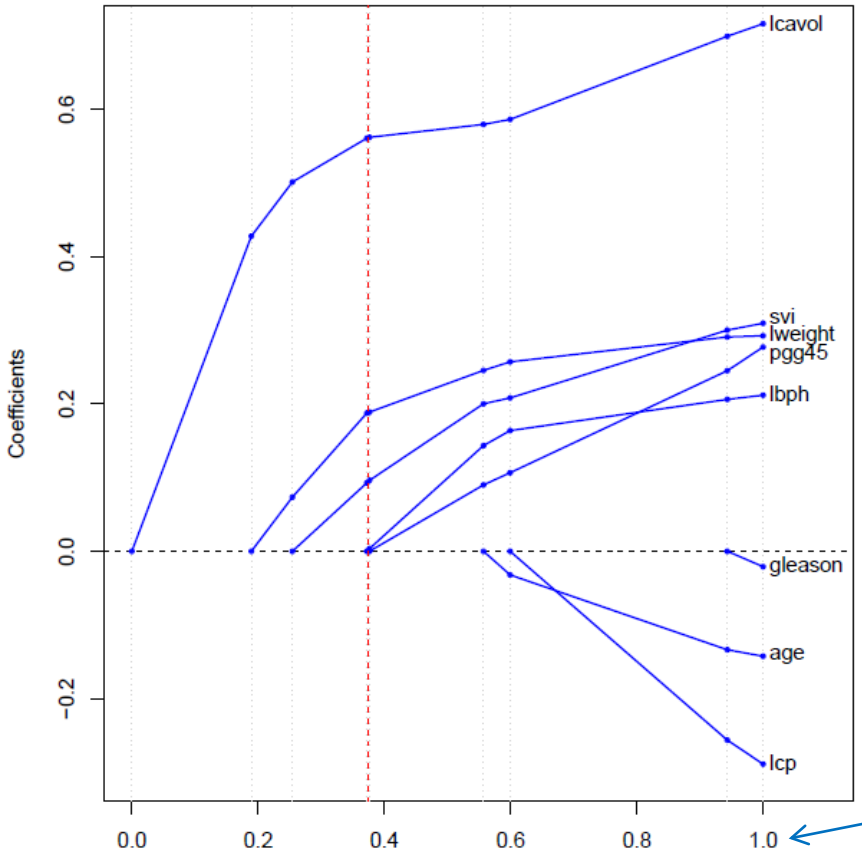
$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Sparsity of lasso solutions

- The L_1 penalty shrinks (scaled) coefficients such that some will become zero and hence are removed from the model.
- Lasso coefficients are biased towards zero and generally not statistically consistent.
- Thus, the lasso is a 'continuous' model selection procedure.
- The regularization parameter t (or λ) should be chosen to minimize an estimate of the prediction error (for example, obtained from cross-validation).

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso shrinkage paths



least squares solution

Shrinkage factor $t / \sum_{j=1}^{p} |\beta_j|$

Generalizations

- For any $q \geq 0$, we can consider

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

- $q = 0$: subset selection
  - combinatorial model selection
  - penalty counts the number of non-zero parameters
- $q = 1$: lasso
- $q = 2$: ridge regression
  - shrinkage to small, non-zero coefficients

q norm

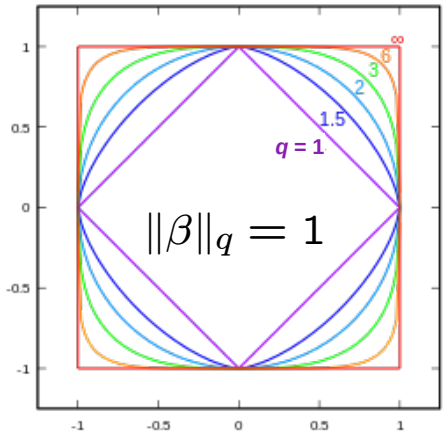
- For any $q \geq 0$,

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

$$= \arg\min_{\beta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_q^q$$

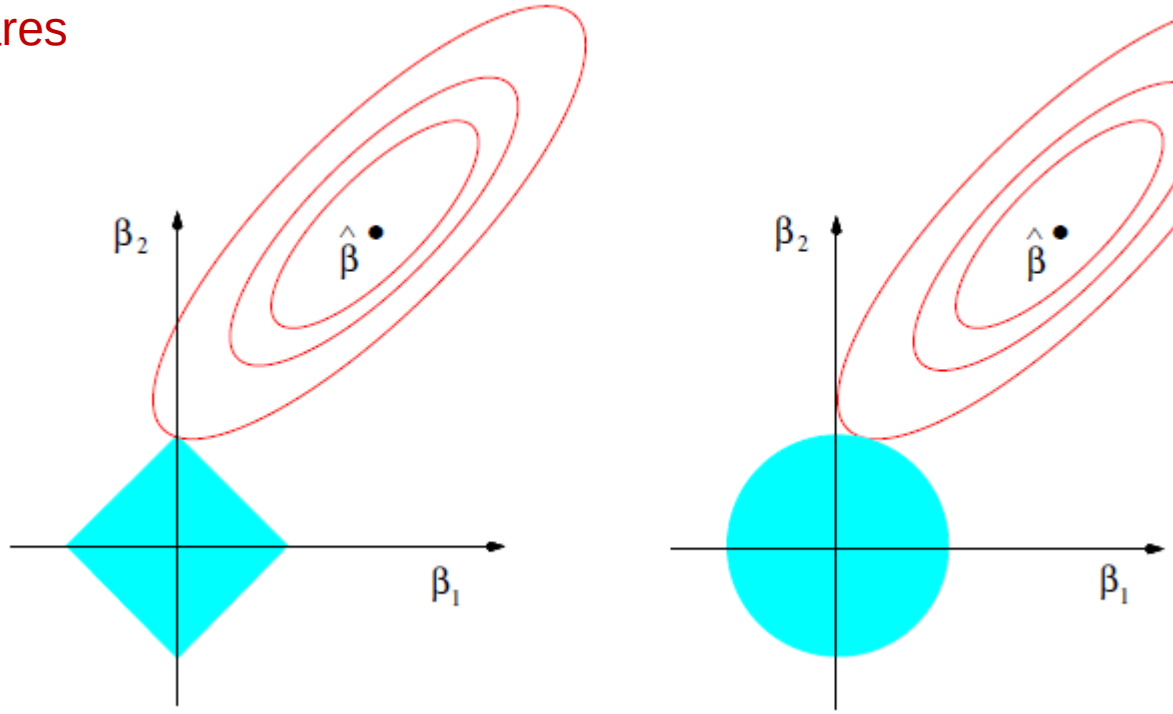
where the *q* norm is

$$\|\beta\|_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{1/q}$$



Lasso versus ridge (L_1 versus L_2)

least squares contours

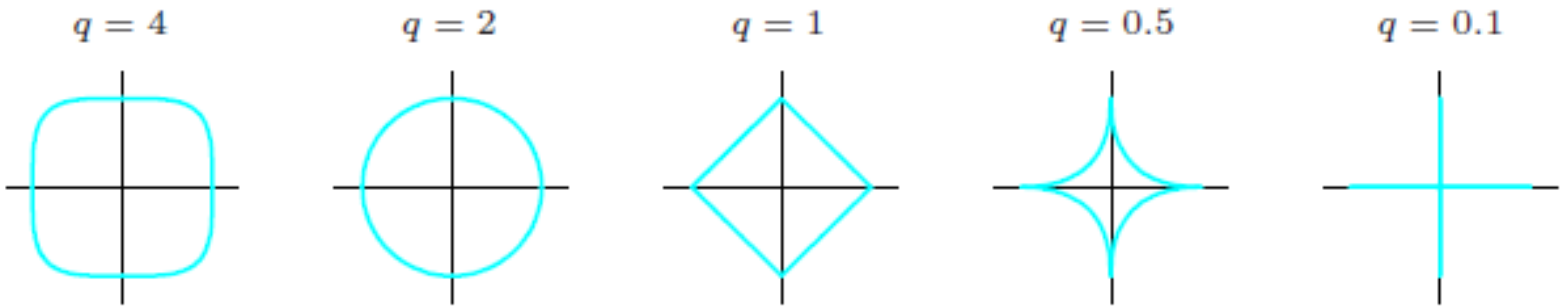


$$|\beta_1| + |\beta_2| \leq t \qquad \beta_1^2 + \beta_2^2 \leq t^2$$

Properties

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \sum_{i=1}^{N} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

- For q = 0, 1, 2, the optimization problem can be regarded as finding the mode of the posterior for different priors.
- q = 1 (lasso) is the smallest q for which the constraint region is convex (allowing efficient convex optimization)



A generic genomics problem

Many genomic predictors
(genes, SNPs, CNAs, …)

Few observations
(patients, conditions, …)

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \cdots & x_{Np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Genotype, X
(DNA sequencing, SNP arrays, microarrays, epigenomics, transcriptomics, proteomics, … etc.)

Phenotype (trait), Y
(disease status, disease subtype, survival probability, height, eye color, cognitive phenotypes, molecular phenotypes, … etc.)

$$p \gg N$$

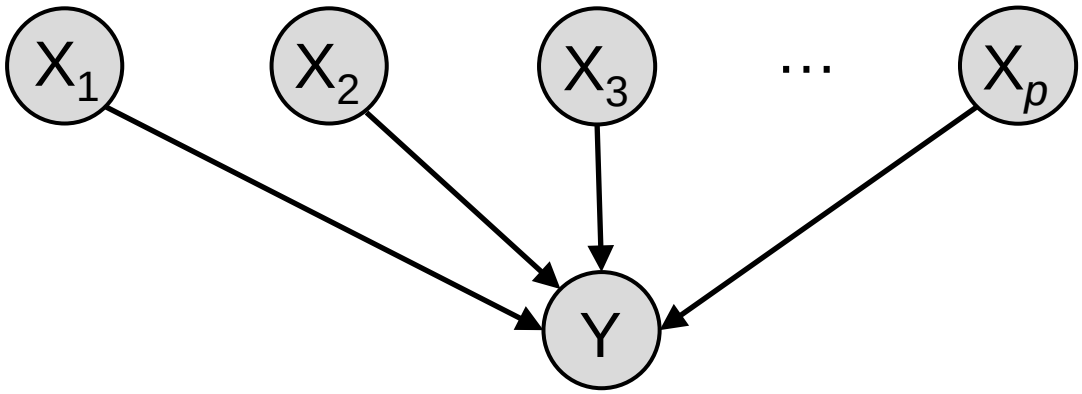
If $p \gg N$, less fitting is better

- In genomics, we often have many more features than observations, for example
  - the expression values of p = 20,000 genes in N = 100 patients
  - the $p = 10^6$ SNPs of N = 1000 patients
- In this setting, high variance and overfitting are a major concern, and strong regularization is required.
- Lasso is a severe regularizer, because for any λ, the number of non-zero coefficients is at most N.

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

Sparsity is a common assumption.

- Sparsity makes statistical sense:
  - Learning becomes feasible in high dimensions with small sample size.
- Sparsity makes biological sense:
  - Each phenotype is likely to be associated with a small number of genomic features (e.g., SNPs), rather than all.



Grouped lasso, Elastic net

Grouped lasso

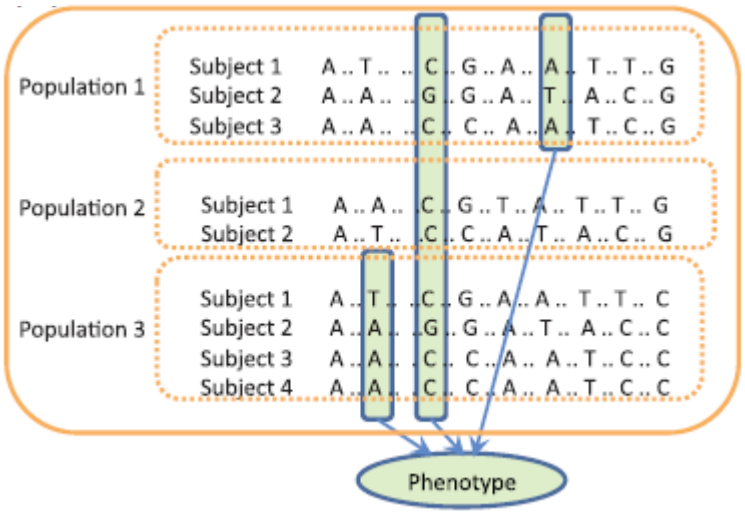
- Suppose the *p* predictors are divided into *L* groups of sizes p_ℓ with corresponding predictors $\mathbf{X}_\ell$ and coefficients β_ℓ.
- Grouped lasso solves the convex optimization problem

$$\min_{\beta} \|\mathbf{y} - \beta_0\mathbf{1} - \sum_{\ell=1}^{L} \mathbf{X}_\ell\beta_\ell\|_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell}\|\beta_\ell\|_2$$

- $\|\beta_\ell\|_2 = 0$ iff $\beta_\ell = (0,\ldots,0)$, thus entire groups are shrunken to 0.
- Examples:
  - Genes grouped into functional pathways
  - Indicator variables for the levels of a categorical variable
- Variations: grouping of output variables, samples

Example: Multi-population group lasso

- Genome-wide association study (GWAS)
- Inputs: single-nucleotide polymorphisms (SNPs)
- Problem: Structure in human population, due to its evolutionary history, can confound associations.



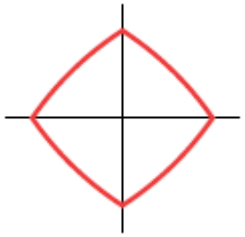
- Consider one feature vector for each subpopulation, $\mathbf{B} = [\beta^1, \ldots, \beta^C]$ with rows β_j, and the group lasso problem

$$\min_{\mathbf{B}} \sum_{c=1}^{C} \|\mathbf{y^c} - \mathbf{X^c}\beta^c\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_2$$

Elastic net

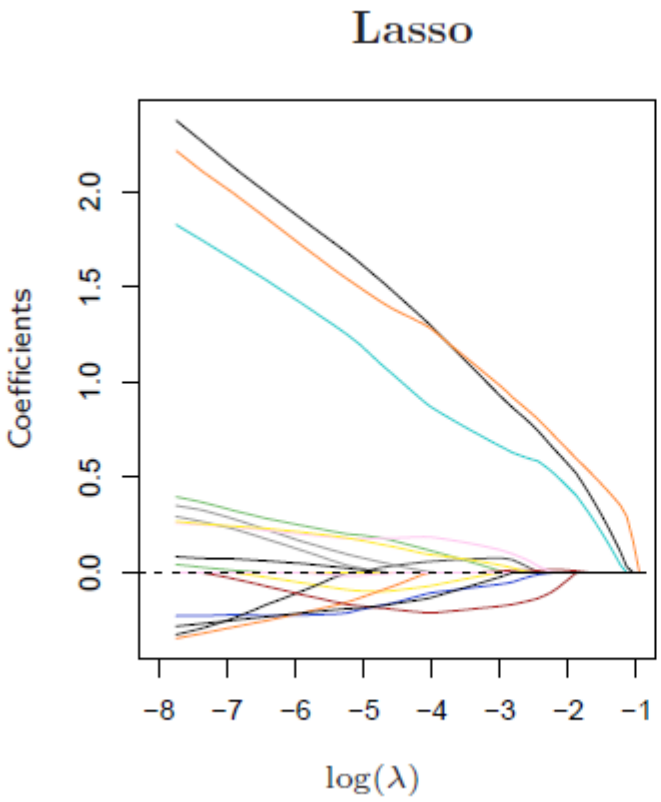
- Genomic features often have strong correlations (gene-gene interactions, e.g., in functional pathways).
- Lasso will select one of them arbitrarily. But ridge regression will shrink correlated features to each other.
- The elastic net combines the L_1 and L_2 penalties:

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \left[\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right]$$

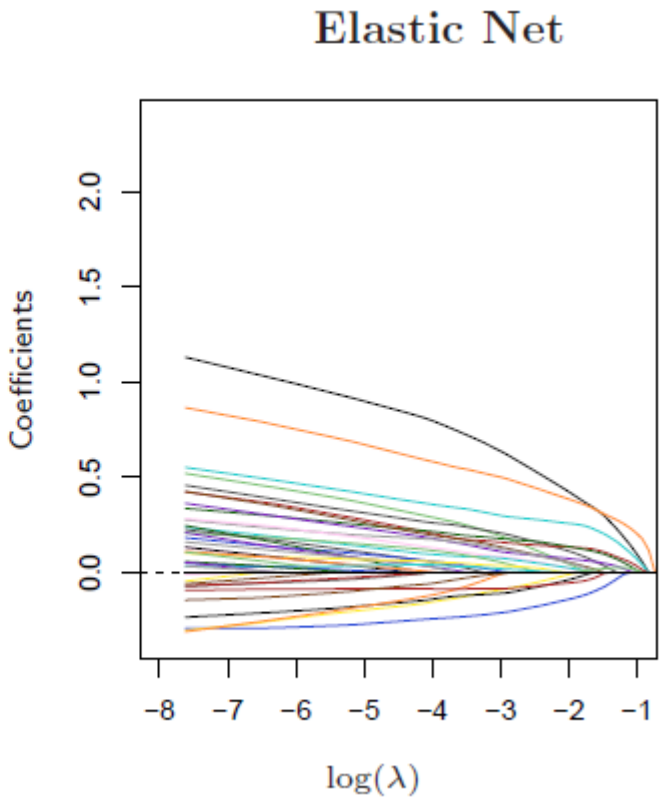


$\alpha = 0.8$

Lasso versus elastic net



19 non-zero coefficients



39 non-zero coefficients

Fused lasso

Fused lasso

- If the features have a natural order (e.g., in time, or in space), we may also care about the smoothness of β.
- The fused lasso solves

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p-1} |\beta_{j+1} - \beta_j|$$

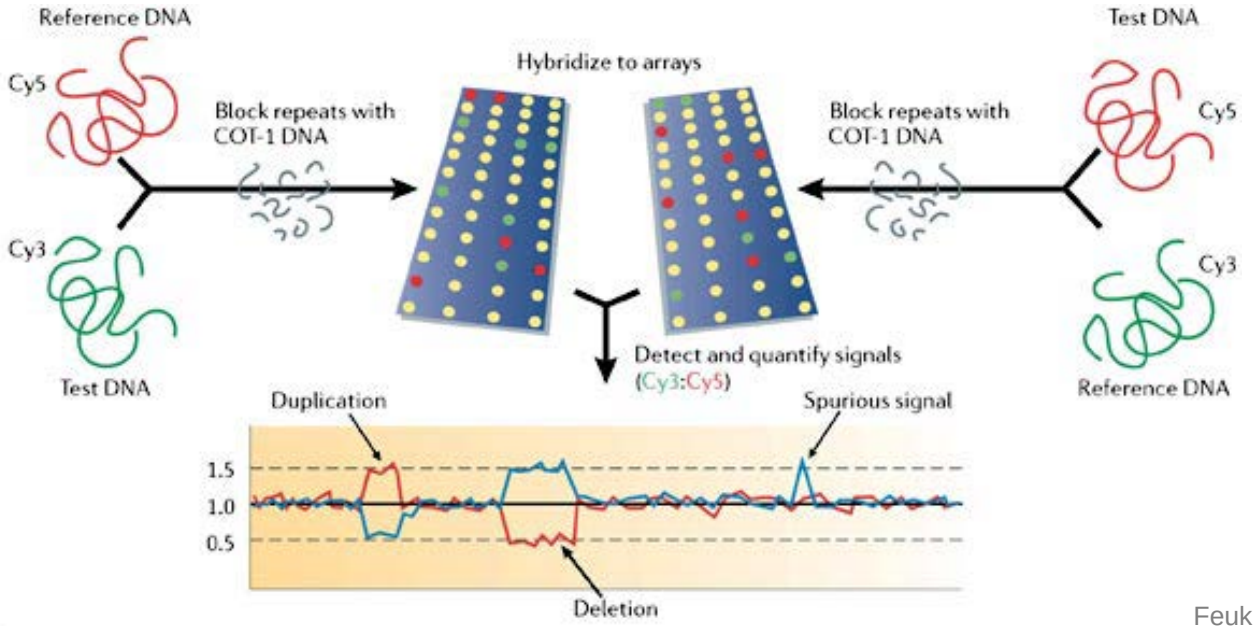
sparsity smoothness

- For $\mathbf{X} = \mathbf{I}_{N\times N}$, we obtain the *fused lasso signal approximator*

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_i)^2 + \lambda_1 \sum_{i=1}^{N} |\beta_i| + \lambda_2 \sum_{i=1}^{N-1} |\beta_{i+1} - \beta_i|$$

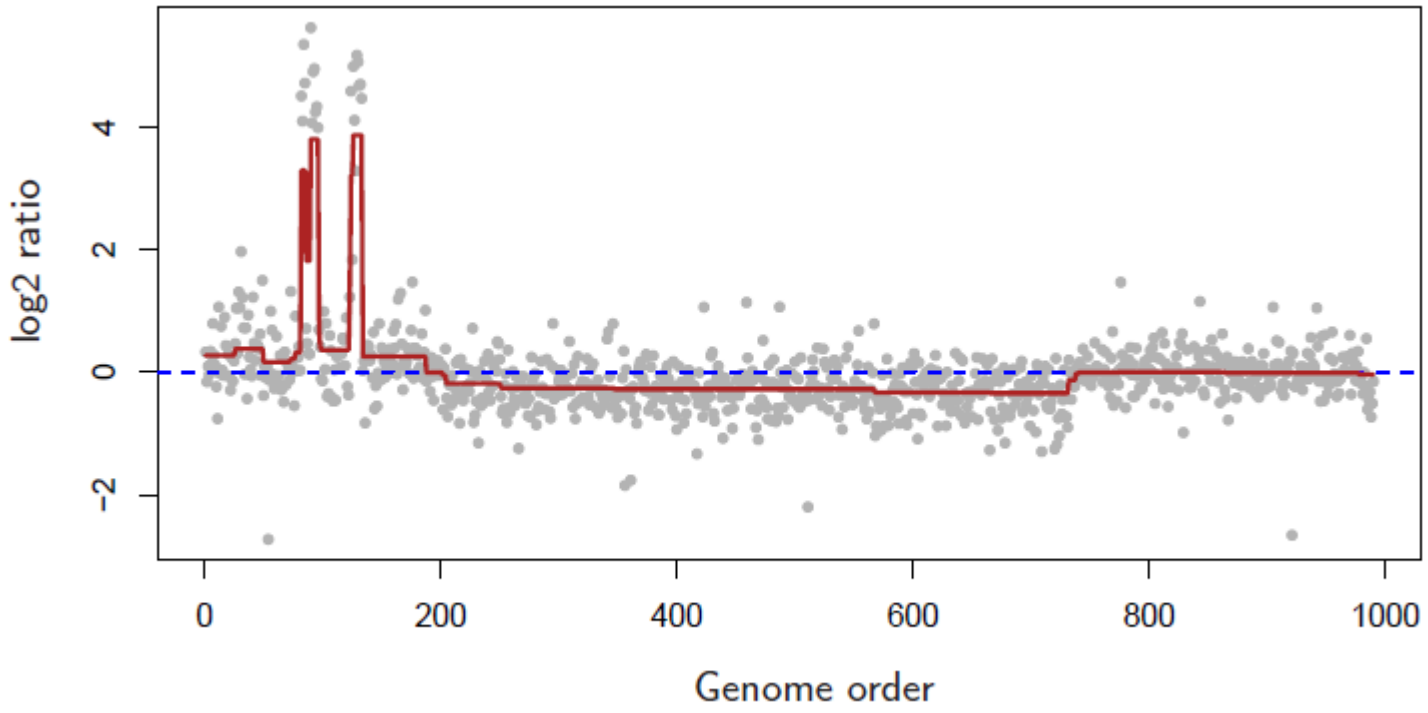
Example: Copy number alterations (CNAs)

- Genomic rearrangements (including insertions and deletions) are common alterations in cancer genomes.
- The resulting CNAs can be detected by comparative genome hybridization (CGH):



Feuk et al., Nat Rev Genet 2006

Fused lasso applied to CGH data



Structural constraints: the generalized lasso

- Let $\mathbf{D} \in \mathbb{R}^{m \times p}$ be a matrix, such that $\mathbf{D}\beta$ corresponds to some desired structural behavior of β, and consider

$$\min_{\beta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\mathbf{D}\beta\|_1$$

- For example, if $\mathbf{X} = \mathbf{I}_{N \times N}$, and

$$\mathbf{D}_{1d} = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 & 0 \\ & & & \ddots & \ddots & & \\ 0 & 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix}$$

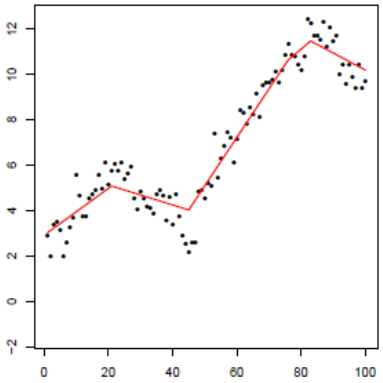
we obtain again the (1d) fused lasso.

Linear trend filtering

- Consider the generalized lasso with $\mathbf{X} = \mathbf{I}_{N \times N}$, and

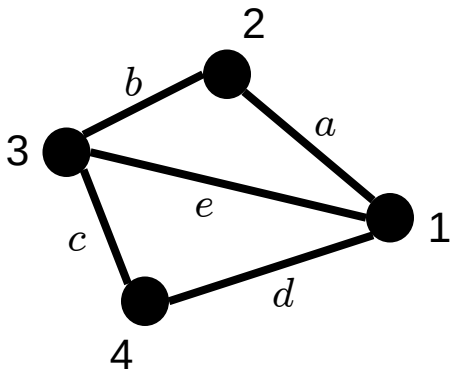
$$\mathbf{D}_{\mathrm{tf},1} = \begin{pmatrix} -1 & 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ & & & \ddots & \ddots & & & \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 & -1 \end{pmatrix}$$

  - The penalty is on the second order discrete derivatives and hence gives a piecewise linear fit.
  - This model can detect linear trends with unknown changepoints.
  - It can be generalized to piecewise polynomial fits.



Arbitrary adjacency matrix

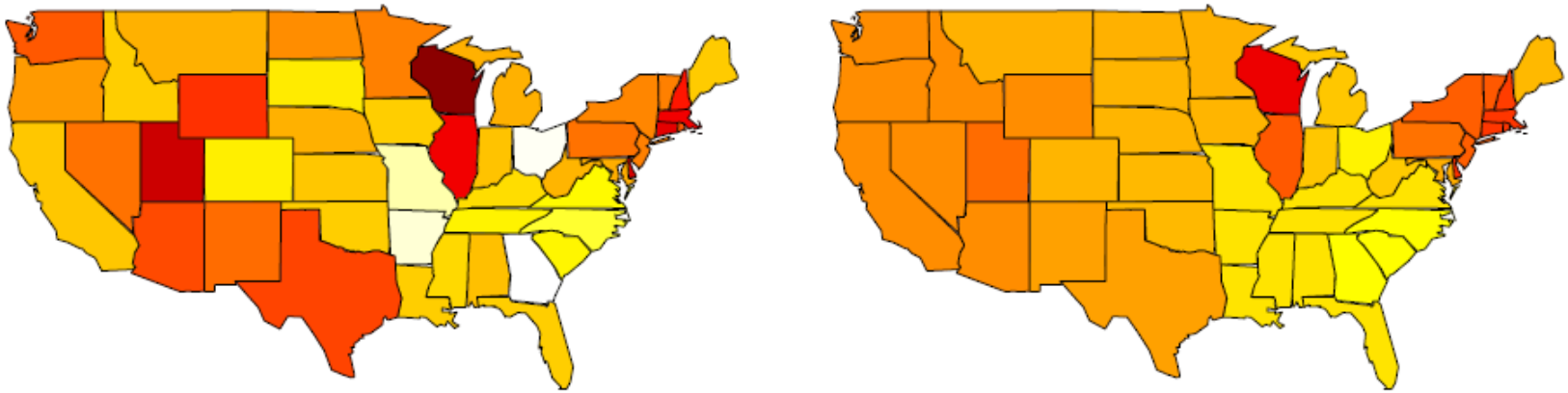
- For any undirected graph with *p* vertices and *m* edges, we can define **D** by placing 1's and −1's in the matrix.



$$\begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Example: H1N1 flu cases in the US in 2009

- Vertices: US (mainland) states
- Edges connect neighboring states
- Data (**y**): log proportion of H1N1 infections (color coded)

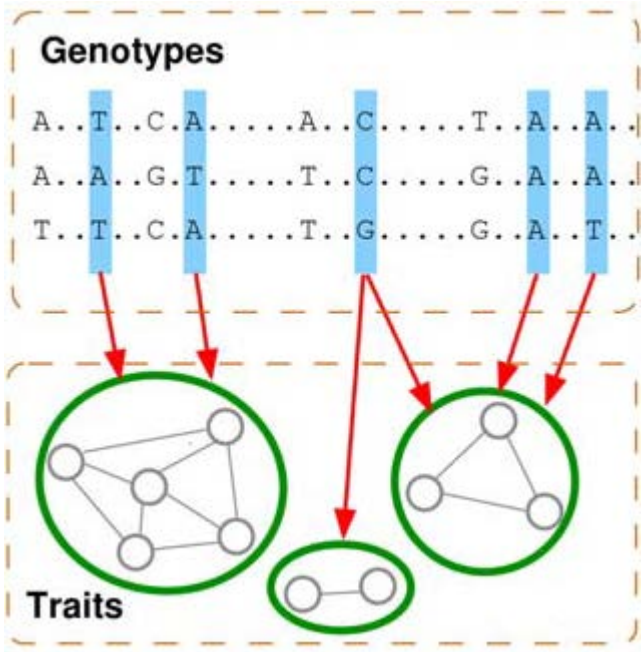


(a) Data

(b) Fused Lasso Solution

Expression quantitative trait loci (eQTL)

- GWAS with outputs gene expression profiles $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_K]$.
- $\mathbf{B} = [\beta_{jk}]$, association of SNP *j* to expression of gene *k*.



- Step 1: Estimate gene expression structure: thresholded correlation network ({1, …, K}, E).
- Step 2: Solve

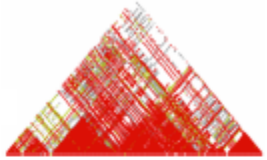
$$\min_{\mathbf{B}} \sum_{k=1}^{K} \|\mathbf{y_k} - \mathbf{X}\beta_k\|_2^2 + \lambda_1 \sum_{j=1}^{p} \sum_{k=1}^{K} |\beta_{jk}| + \lambda_2 \sum_{(m,l)\in E} \sum_{j=1}^{p} |\beta_{jm} - \text{sign}(r_{ml})\beta_{jl}|$$

fusion penalty in output space

Many more variations exist…

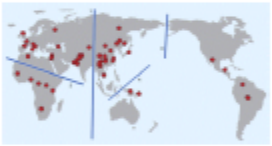
Genome Structure

Linkage Disequilibrium



Stochastic block regression
(Kim & Xing, UAI, 2008)

Population Structure



Multi-population group lasso
(Puniyani, Kim, Xing, ISMB, 2010)

Epistasis

ACGTTTTACTGTACAATT

Group lasso with networks
(Lee, Kim, Xing, Submitted)

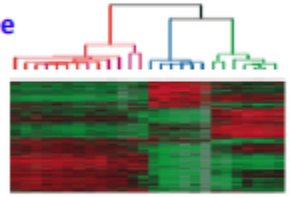
Structured Association

Phenome Structure

Graph

Graph-guided fused lasso
(Kim & Xing, PLoS Genetics, 2009)

Tree



Tree-guided fused lasso
(Kim & Xing, ICML 2010)

Dynamic Trait

Temporally smoothed lasso
(Kim, Howrylak, Xing, Submitted)

Summary

- The lasso is a powerful regularization approach for finding sparse solutions of (linear) regression problems.
- Most genomics problems have $p \gg N$ and hence require strong regularization.
- The grouped lasso allows for joint shrinkage of groups of coefficients.
- The (generalized) fused lasso allows for encouraging specific desired structures among inputs, outputs, features.
- These and other penalties can be combined to formulate complex models with structured sparsity.

References

- Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer, 2013.
- Xing EP, Kim S. ISMB 2011 tutorial.
- Tibshirani RJ, Taylor J (2011). The solution path of the generalized lasso. Annals of Statistics, 39(3), 1335-1371.