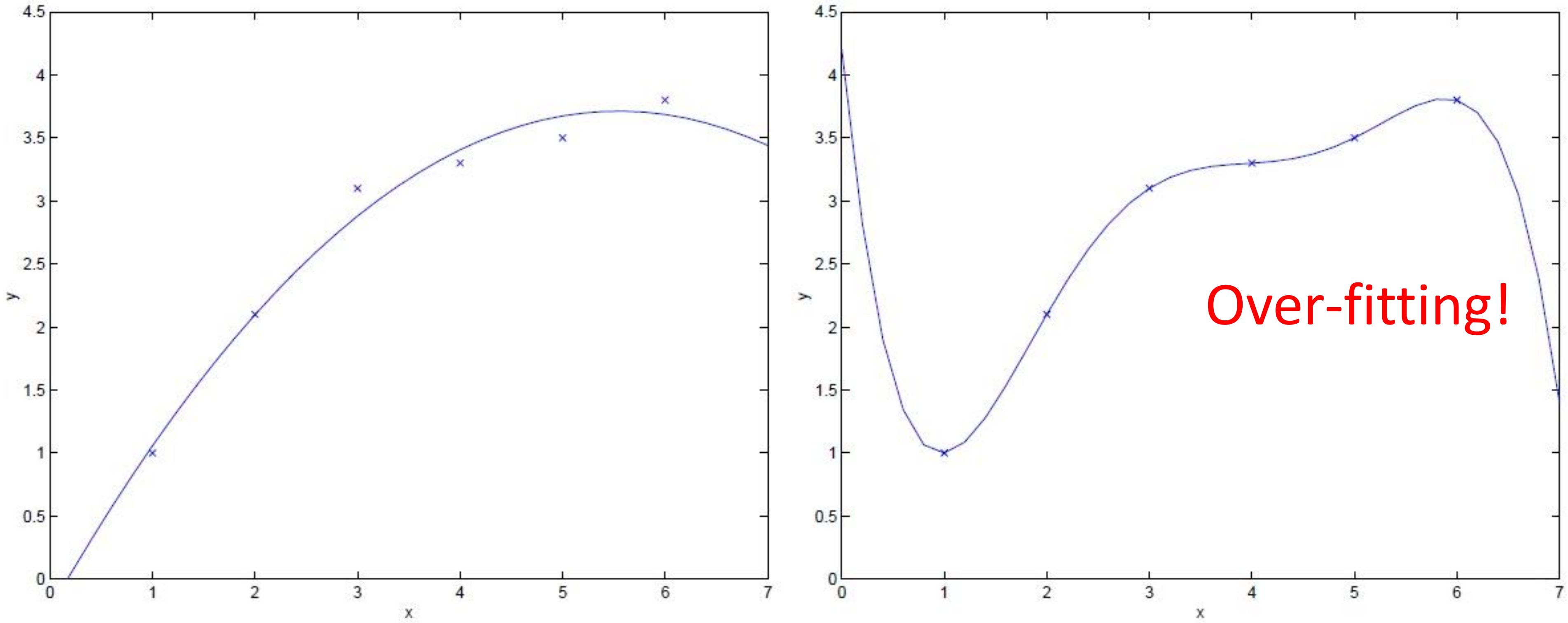
Regularization & Cross Modal

Reporter: Yunfei WANG
Date: March/04/2014
E-mail: yunfeiwang@hust.edu.cn

Ordinary Least Squares



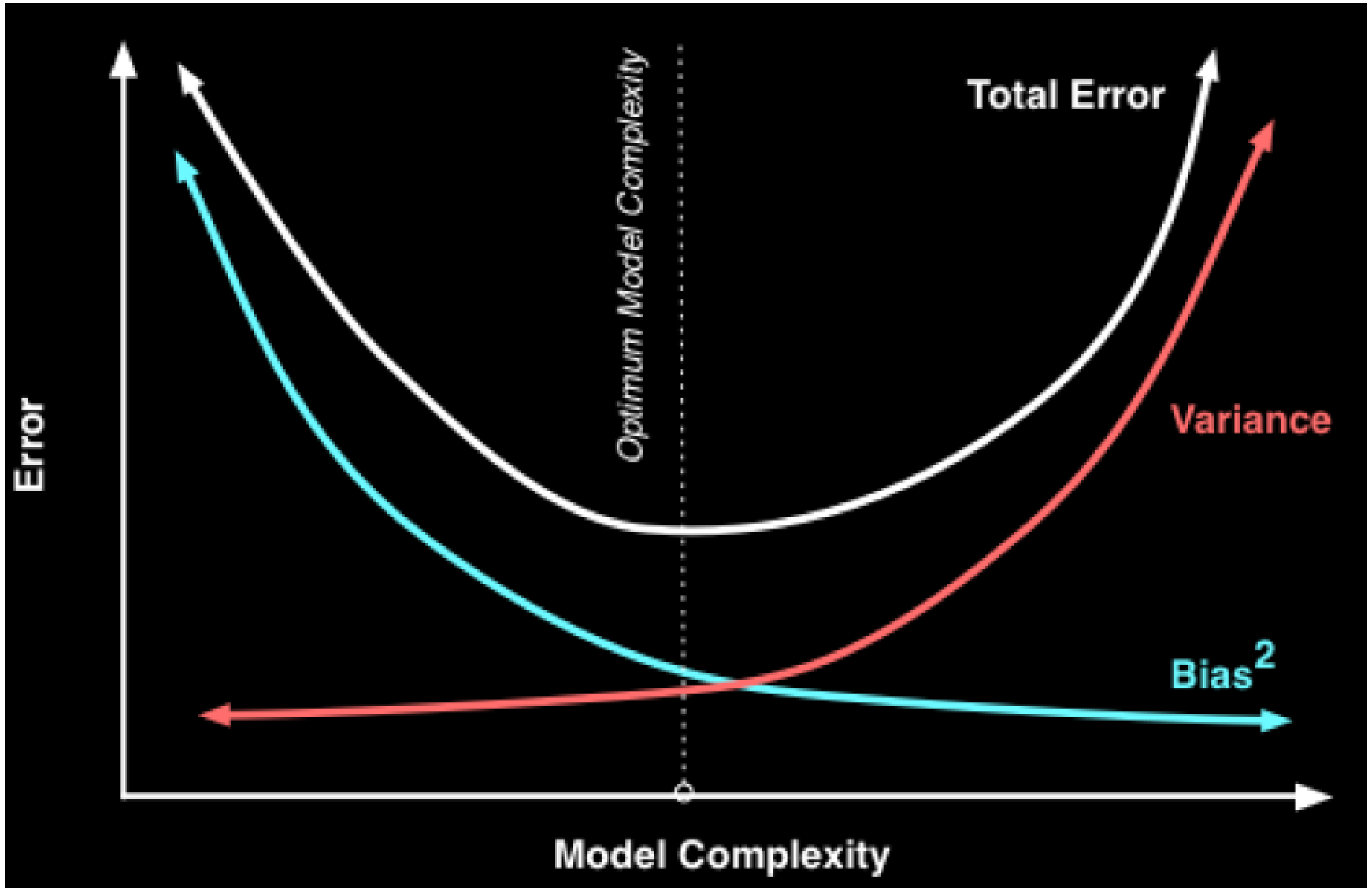
Low Bias

High Variance

$$J(w) = \min_{w} \frac{1}{2} ||Xw - y||_2^2$$

Shrinking or setting to zero some coefficients can improve prediction accuracy and lead to reasonable interpretation with several the most important features.

Model Complexity & Regularization



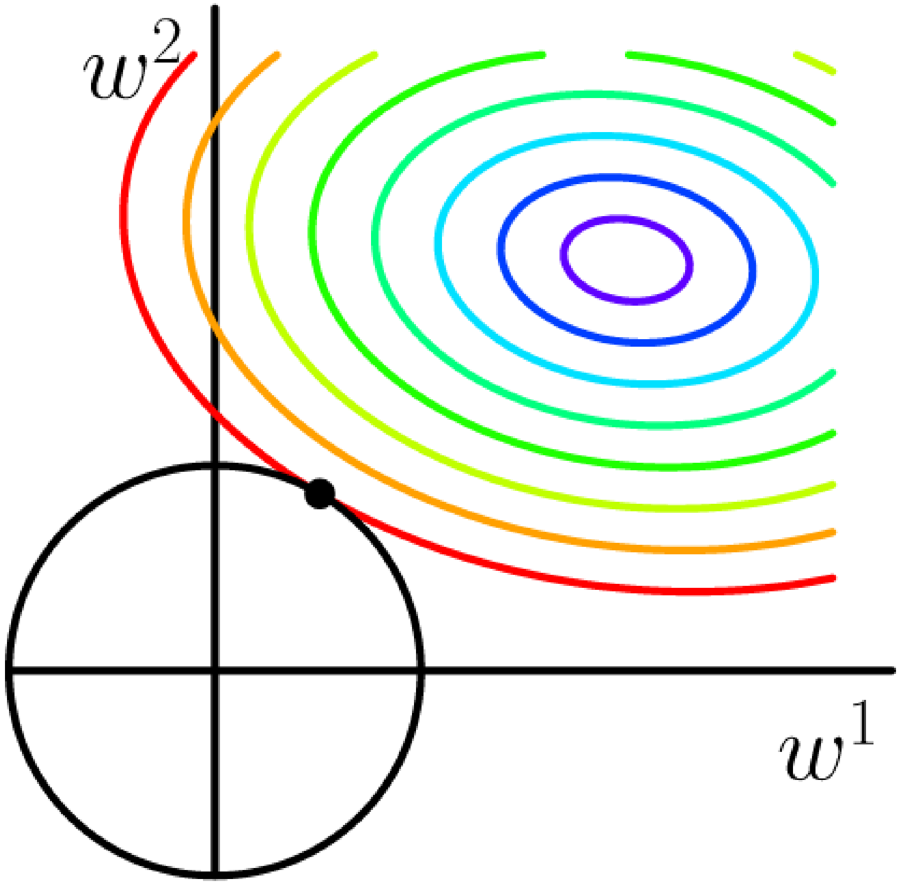
$$J(w) = loss(X, y, w) + \lambda\Omega(w)$$

Regularization helps to control the complexity of models by limiting the scope of parameters

1)reduce the risk of over-fitting

2)improve the ability of generalization

Ridge Regression



ℓ_2-ball meets quadratic function. ℓ_2- ball has no corner. It is very unlikely that the meet-point is on any of axes."

$$J(w, \lambda) = \min_{w} \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

$$J(w) = \min_{w} \frac{1}{2}\|Xw - y\|_2^2,$$
$$s.t. \quad \|w\|_2^2 \leq C$$

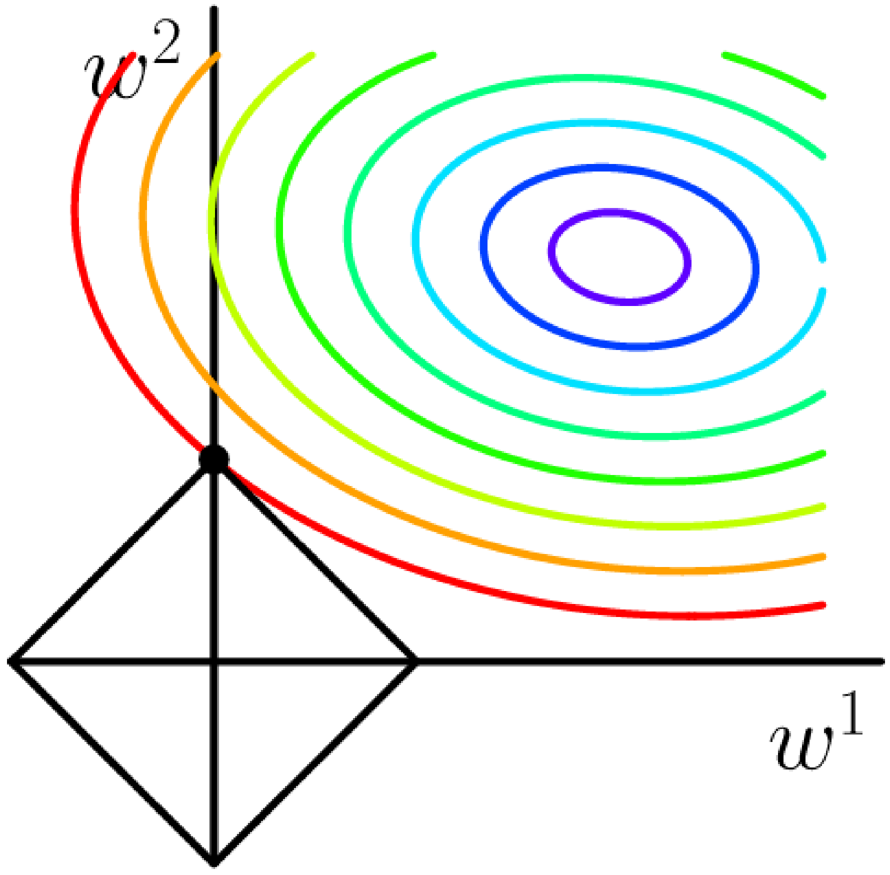
$$w = (X^T X + \lambda I)^{-1} X^T y$$

No feature selection

Perform badly in sparse high-dimensional space

http://freemind.pluskid.org/machine-learning/sparsity-and-some-basics-of-l1-regularization/

LASSO



ℓ_1-ball meets quadratic function. ℓ_1-ball has corners. It's very likely that the meet-point is at one of the corners.

$$J(w,\lambda) = \min_w \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_1$$

$$J(w) = \min_w \frac{1}{2}\|Xw - y\|_2^2,$$
$$s.t. \quad \|w\|_1 \leq C$$

Limitations:

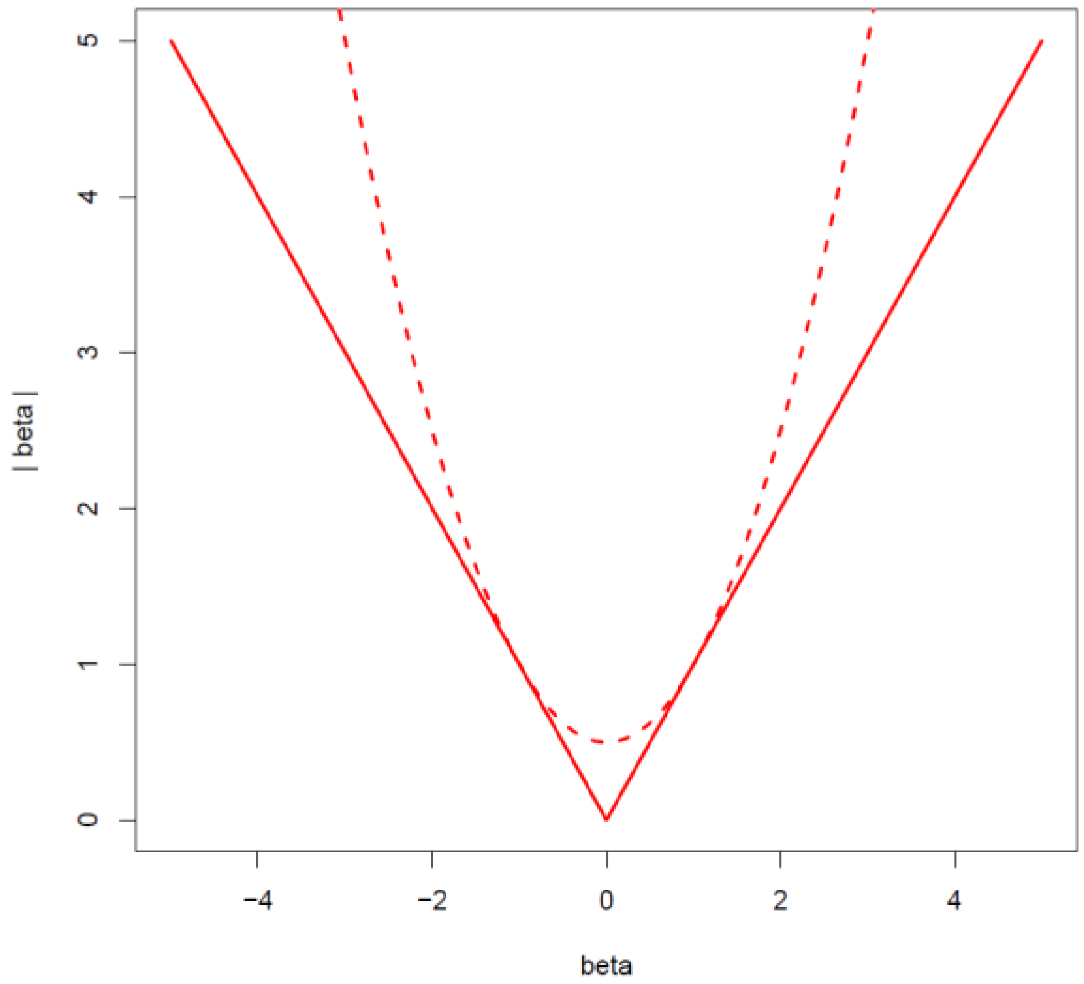
① In the p>n case, owing to the nature of convex optimization problem it can select at most *n* out of the *p* variables[2].

② For a group of highly correlated variables, LASSO tend to select only arbitrary one of them[2,3].

③ In the case of n>p, LASSO is dominated by ridge regression if there are high correlations between variables[4].

http://freemind.pluskid.org/machine-learning/sparsity-and-some-basics-of-l1-regularization/

Parameters Estimation of LASSO



Approximation of the LASSO penalty[5]: $||\beta|| \approx ||\beta_0|| + \frac{1}{2||\beta_0||}(\beta^2 - \beta_0^2)$

Parameters Estimation of LASSO

Loss function of LASSO regression:

$$J(w,\lambda) = \min_w \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_1$$

Plug the approximation into the LASSO loss function:

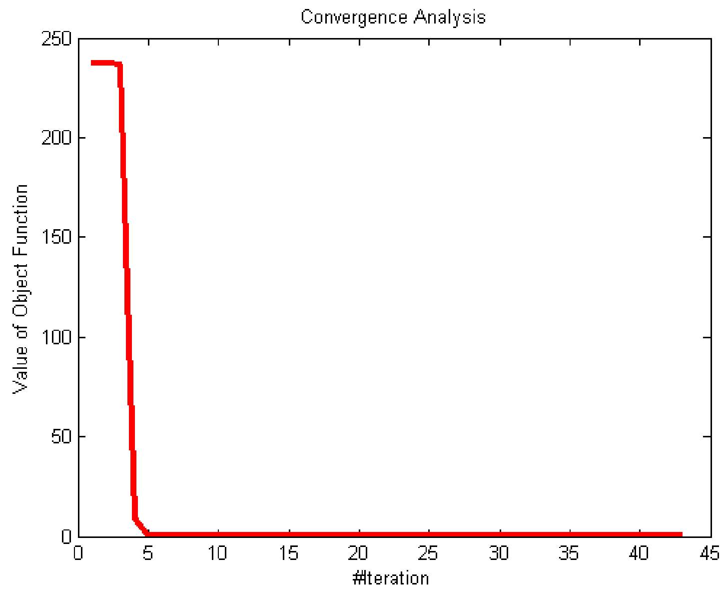
$$\mathcal{L}(w^{(t+1)})$$

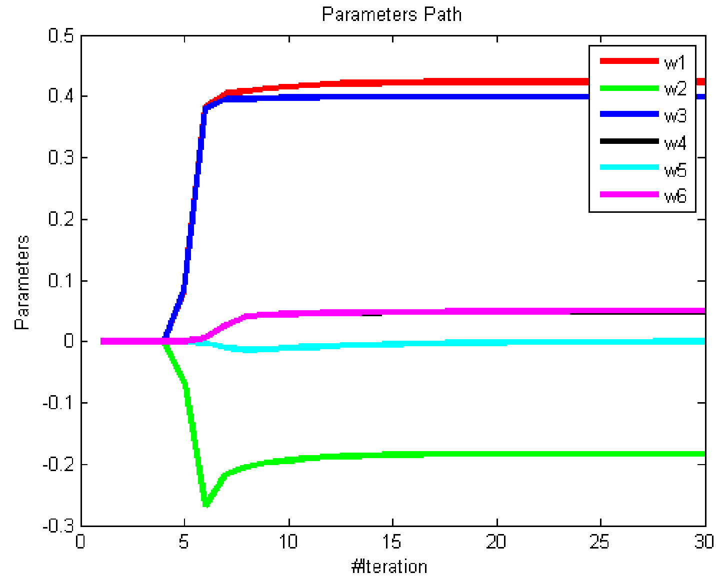
$$= \frac{1}{2}\|Xw^{(t+1)} - y\|_2^2 + \lambda\|w^{(t+1)}\|_1$$

$$= \frac{1}{2}\|Xw^{(t+1)} - y\|_2^2 + \lambda\|w^{(t)}\|_1 + \frac{\lambda}{2}\sum_{i=1}^{p}\frac{[w_i^{(t+1)}]^2 - [w_i^{(t)}]^2}{|w_i^{(t)}|}$$

$$\propto \boxed{\frac{1}{2}\|Xw^{(t+1)} - y\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{p}\frac{[w_i^{(t+1)}]^2}{|w_i^{(t)}|}}$$

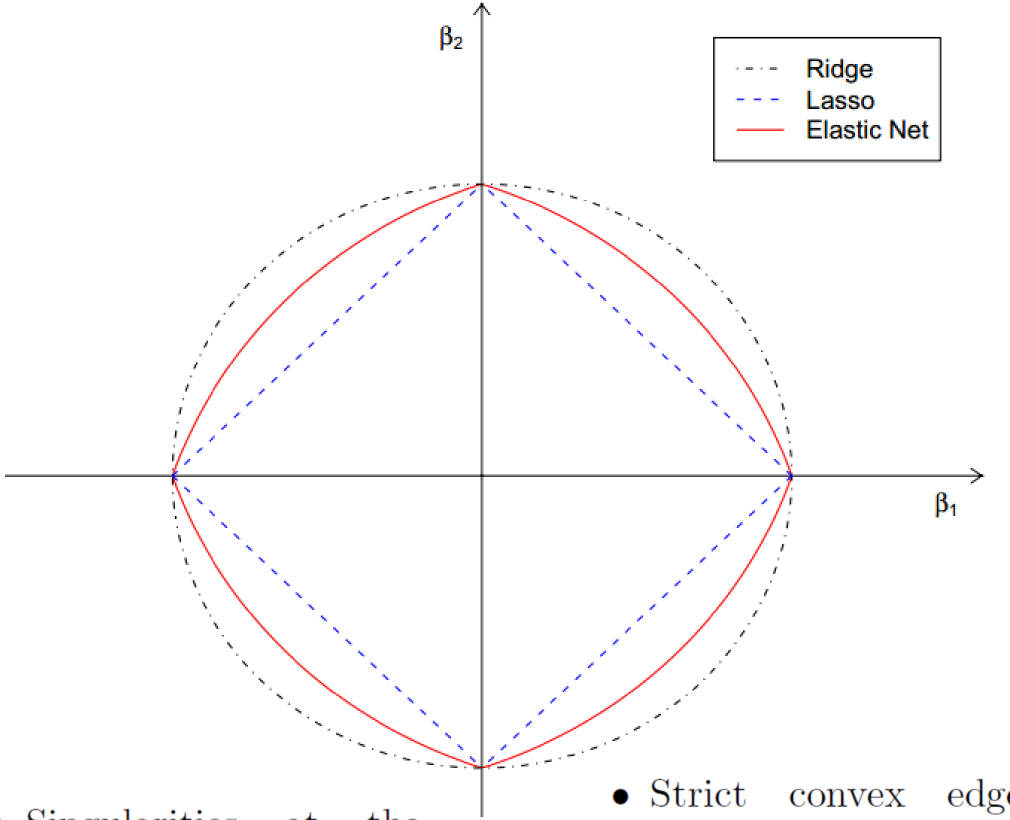
Weighted rigid regression





Elastic Net

2-dimensional illustration $\alpha = 0.5$



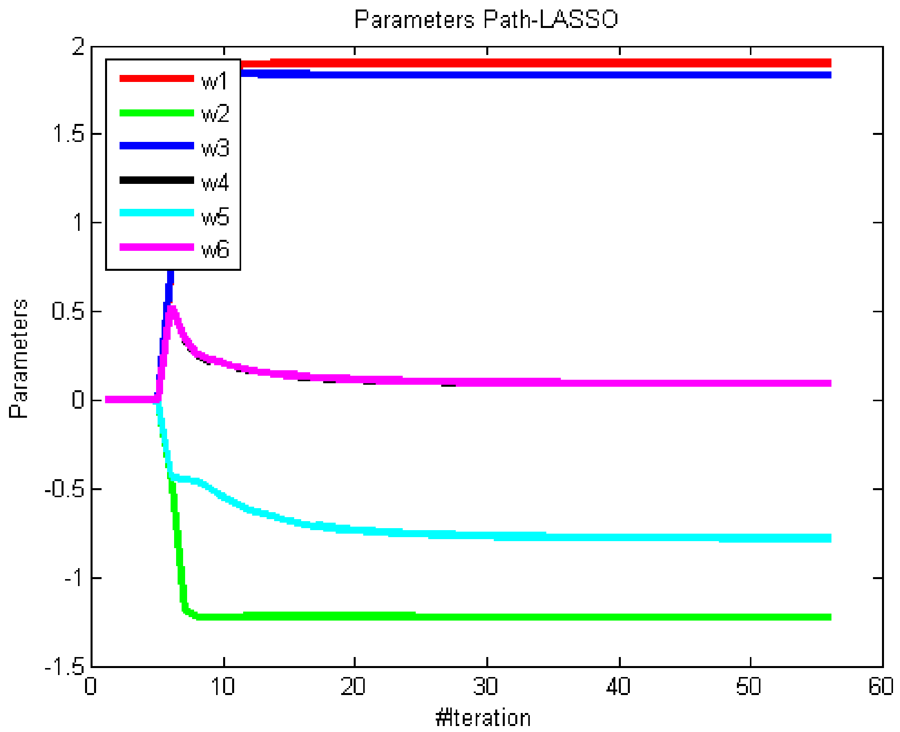
- Singularities at the vertexes (necessary for sparsity)
- Strict convex edges. The strength of convexity varies with α (grouping)

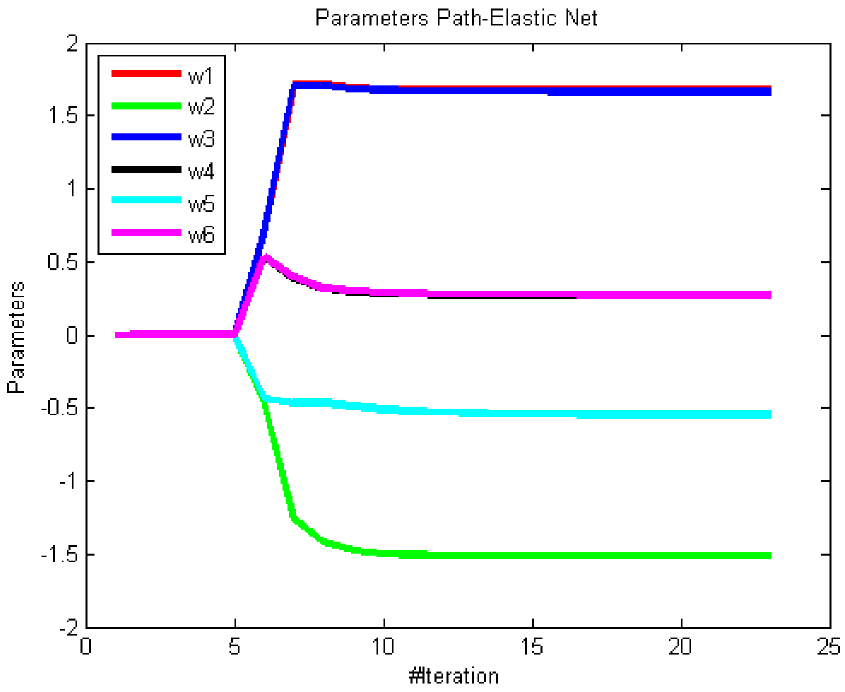
$$J(w, \lambda) = \min_w \frac{1}{2}\|Xw - y\|_2^2 + \boxed{\lambda_1\|w\|_2^2 + \lambda_2\|w\|_1}$$

$$J(w) = \min_w \frac{1}{2}\|Xw - y\|_2^2,$$
$$s.t. \quad \alpha\|w\|_1 + (1-\alpha)\|w\|_2^2 \leq C$$

- The ℓ_1 part of the penalty generates a sparse model.
- The quadratic part of the penalty
  - Removes the limitation on the number of selected variables;
  - Encourages *grouping effect*;
  - Stabilizes the ℓ_1 regularization path.

A simple illustration: elastic net vs. LASSO





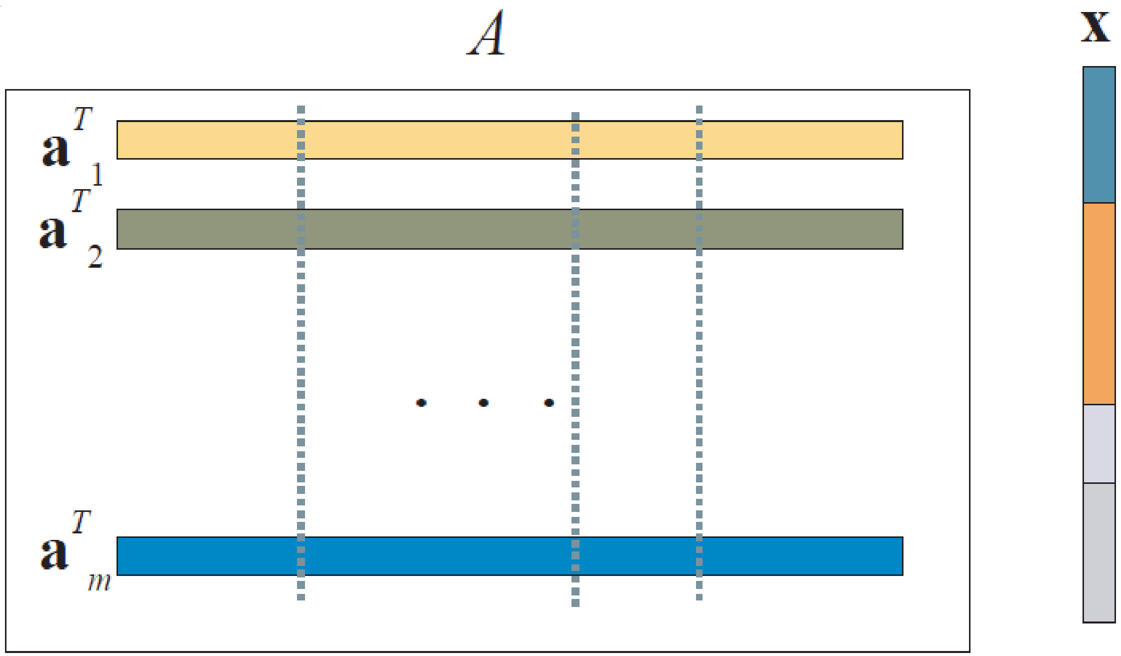
$$\mathbf{z}_1 \sim U(0,20), \quad \mathbf{z}_2 \sim U(0,20)$$

$$\mathbf{y} = \mathbf{z}_1 + 0.1 \cdot \mathbf{z}_2 + N(0,1)$$

$$\mathbf{x}_1 = \mathbf{z}_1 + \epsilon_1, \quad \mathbf{x}_4 = \mathbf{z}_2 + \epsilon_4,$$
$$\mathbf{x}_2 = -\mathbf{z}_1 + \epsilon_2, \quad \mathbf{x}_5 = -\mathbf{z}_2 + \epsilon_5,$$
$$\mathbf{x}_3 = \mathbf{z}_1 + \epsilon_3 \quad \mathbf{x}_6 = \mathbf{z}_2 + \epsilon_6$$

- An "oracle" would identify $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ (the $\mathbf{z}_1$ group) as the most important variables.

Group LASSO for Group Variable Selection



$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2 + \lambda \sum_{i=1}^{k} w_i^g \|\mathbf{x}_{G_i}\|_q$$

$\mathbf{x} \in \mathbb{R}^{n\times 1}$ is divided into k non-overlapping groups $\mathbf{x}_{G_1}, \mathbf{x}_{G_2}, \ldots, \mathbf{x}_{G_k}$

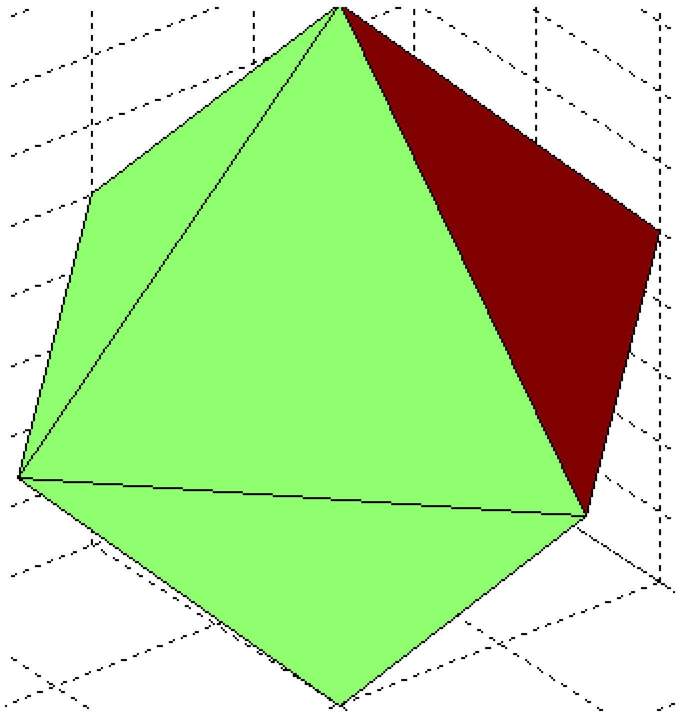
The parameter w_i^g denotes the weight for the i-th group

Features are grouped into four non-overlapping groups

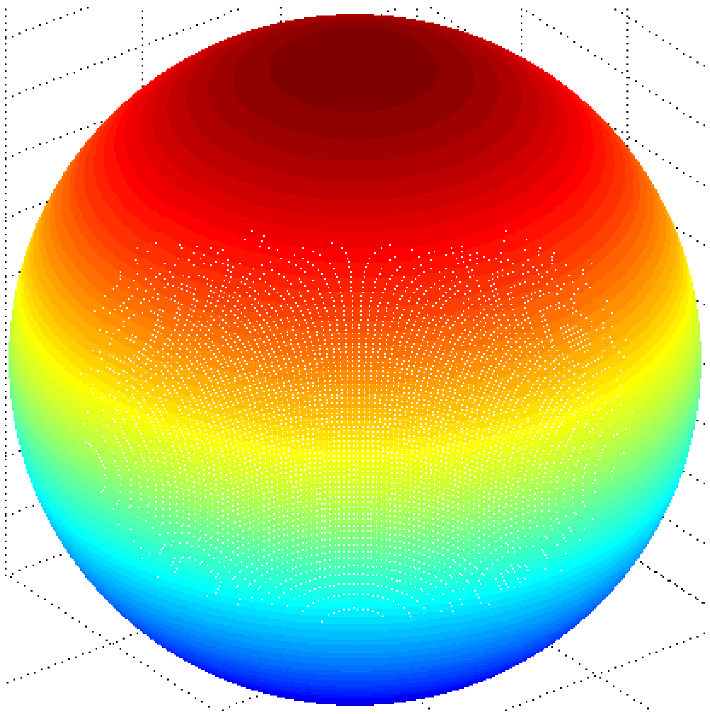
General strategy for grouped variable selection is to use block L_1 norm regularization. For variables with each block(group) L_q norm is applied and different blocks are combined by L_1 norm.[7]

SLEP Package: http://www.public.asu.edu/~jye02/Software/SLEP/

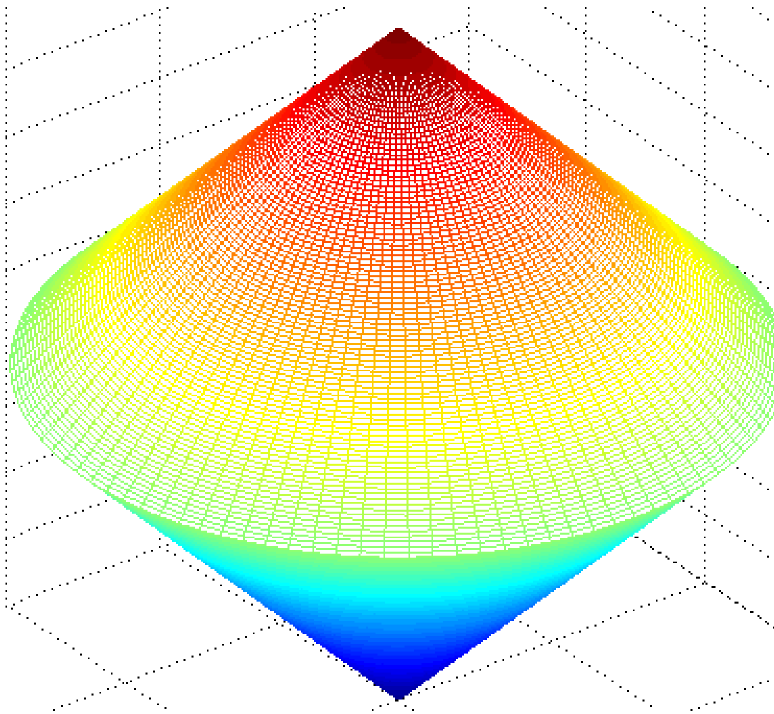
Comparison in 3D



L1-norm



L2-norm



Group LASSO

Bridge Regression

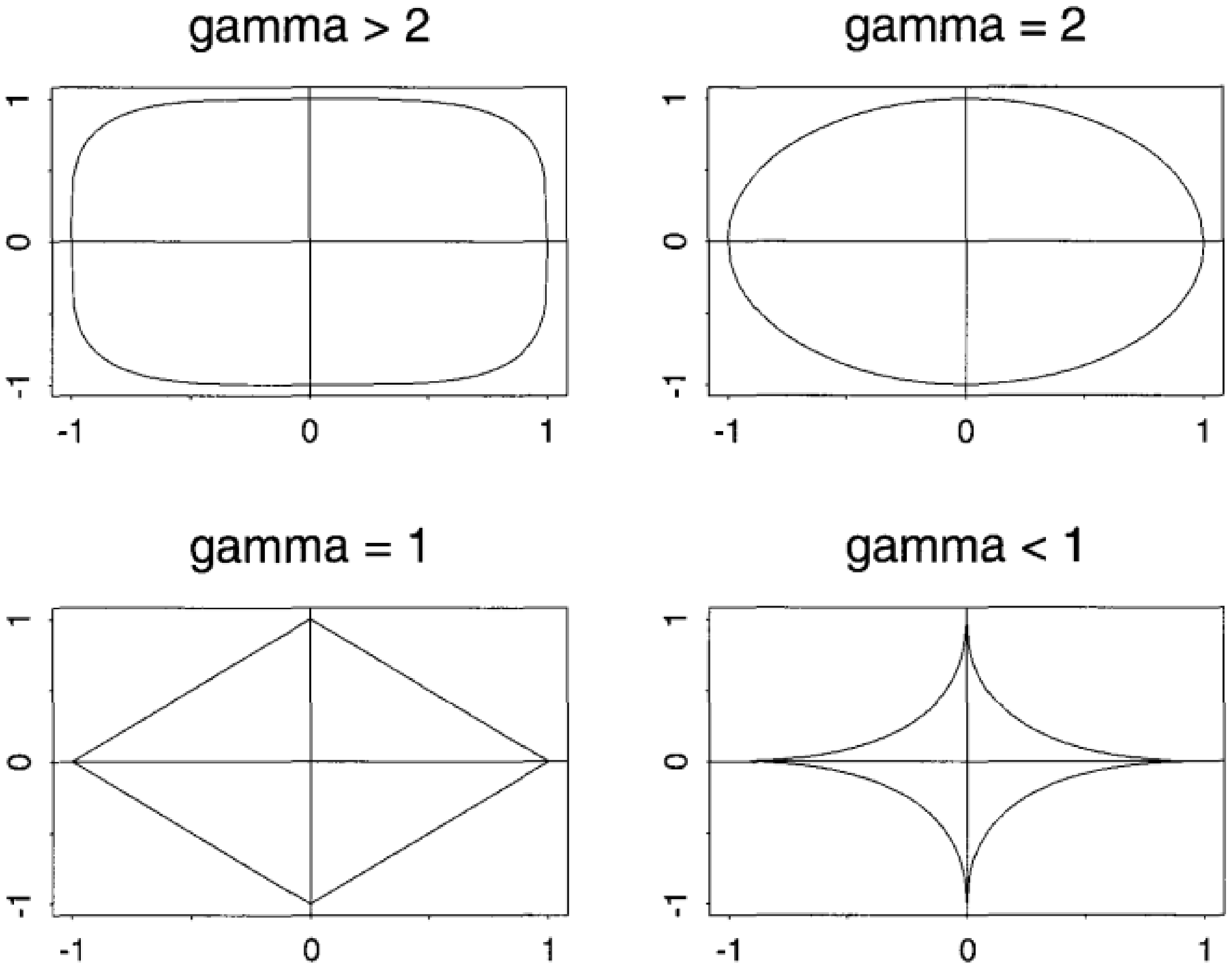
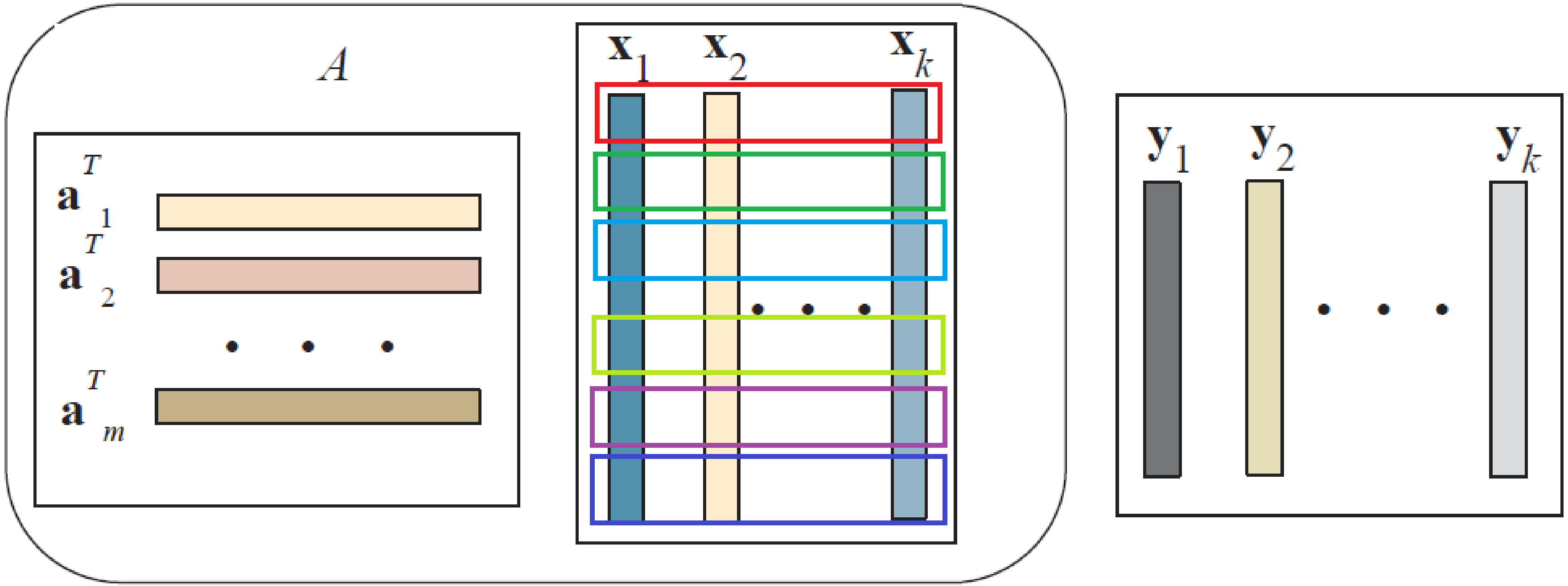


Figure 1. Constrained Areas of Bridge Regressions with t = 1.

$$J(w, \lambda) = \min_{w} \tfrac{1}{2} \|Xw - y\|_2^2 + \lambda \textstyle\sum_{i=1}^{p} |w_i|^\gamma$$

L_1/L_q -norm for Multi-task Feature Selection



$$\min_X \frac{1}{2}\|AX - Y\|^2 + \|X\|_{\ell_1/\ell_q} \qquad \|X\|_{\ell_1/\ell_q} = \sum_{i=1}^{n} \|X^{(i)}\|_{\ell_q} = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} X_{ij}^q \right)^{\frac{1}{q}}$$

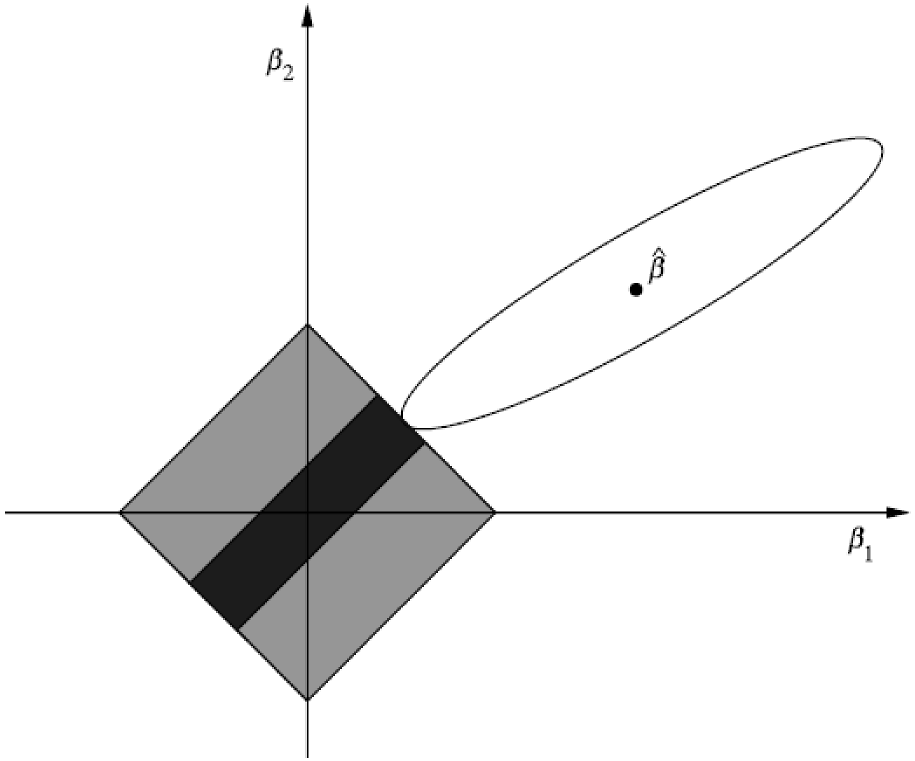
Based on the assumption that all the tasks share the common set of features which is not realistic in many real-word applications. [13]

A subset of highly related outputs may share a common set of relevant inputs, whereas weakly related outputs are less likely to be affected by the same inputs.[13]

Fused Lasso

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{x}\|_1 + \lambda_2\sum_{i=1}^{p-1}\boxed{|x_i - x_{i+1}|}$$

Successive difference

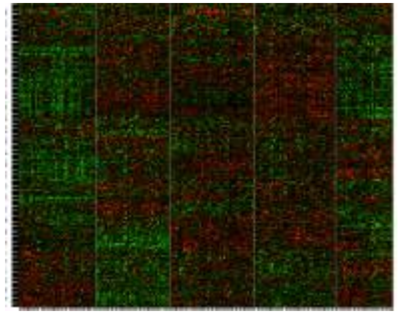


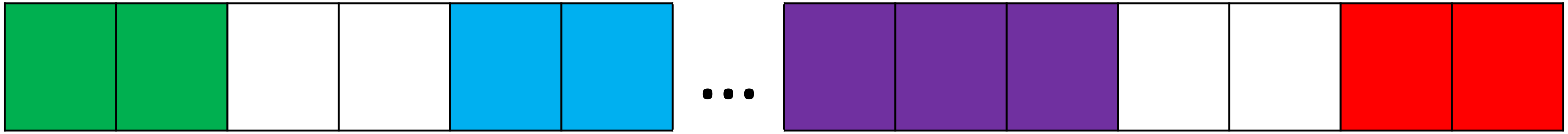
LASSO ignores ordering of the features[8].

Fused Lasso produces sparse and blocky solution[9], which is useful when features are ordered in some meaningful way or the number of features is much greater than the sample size[8].

Unordered features?[8]

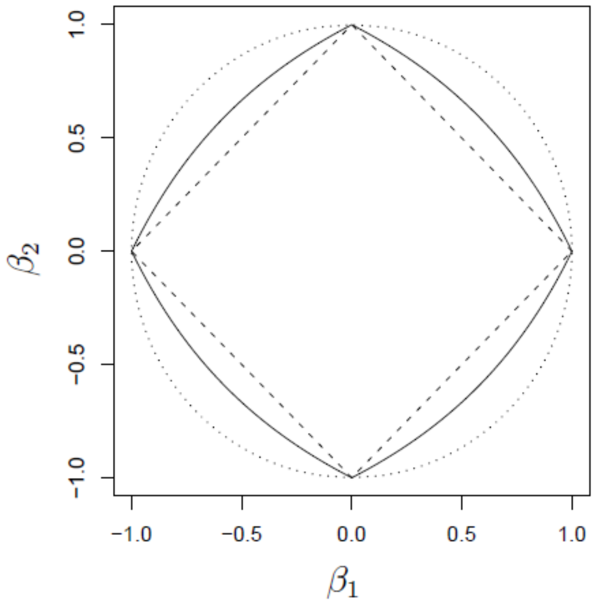
- Multidimentional scaling(MDS) or Hierarchical clustering
- Heat map displays

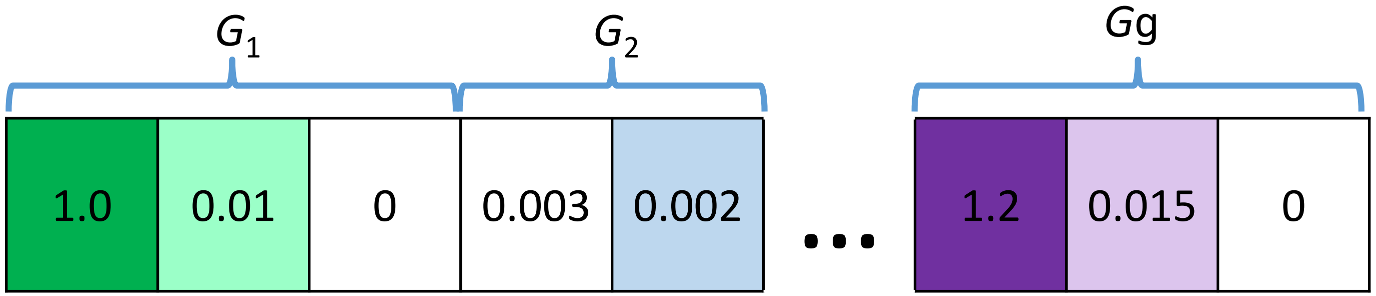




Sparse Group Lasso

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{x}\|_1 + \lambda_2\sum_{i=1}^{g} w_i^g\|\mathbf{x}_{G_i}\|_2$$





① Have the property of Elastic Net: highly correlated features can be selected simultaneously.

② The group information here is more flexible than that in LASSO(correlation).

Figure 1: *Contour lines for the penalty for the group lasso (dotted), lasso (dashed) and sparse group lasso penalty (solid), for a single group with two predictors.*

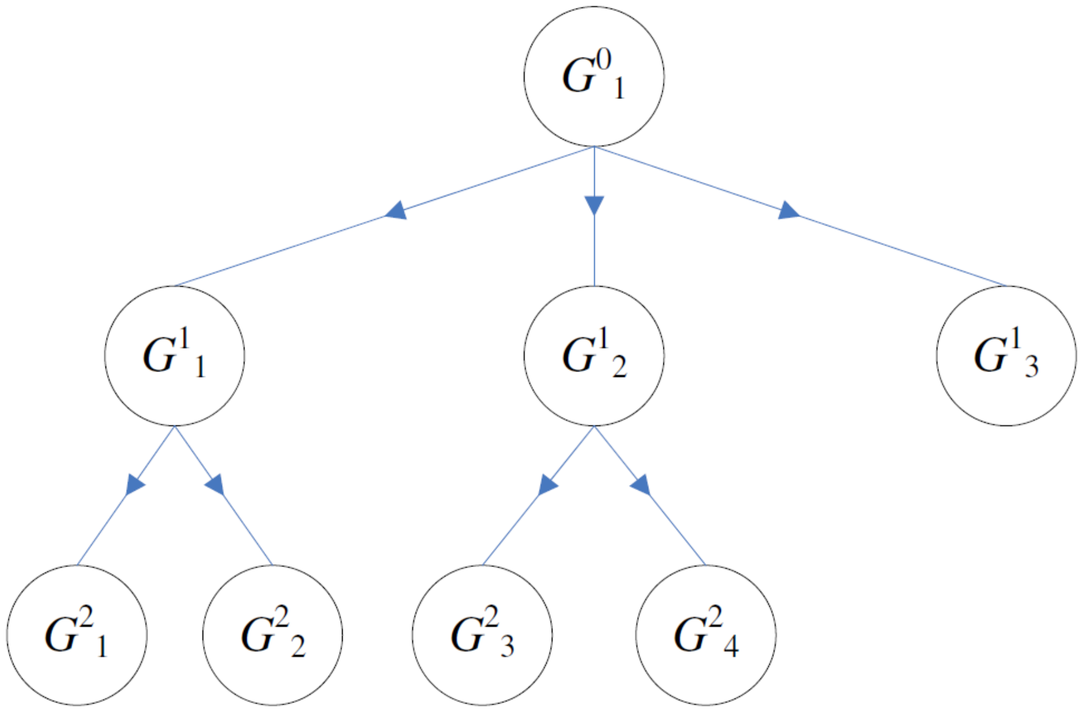
Sparse Group Lasso yields sparsity at both the group and individual feature levels, in order to select groups and features within a group.[10]

Tree Structured Group Lasso

$$\phi_\lambda(\mathbf{v}) = \min_{\mathbf{x}} \left\{ f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{v}\|^2 + \lambda \sum_{i=0}^{d} \sum_{j=1}^{n_i} w_j^i \|\mathbf{x}_{G_j^i}\| \right\}$$

Based on group lasso

Structure over features be express as a tree with leaf nodes as features and internal nodes as clusters of the features[11].

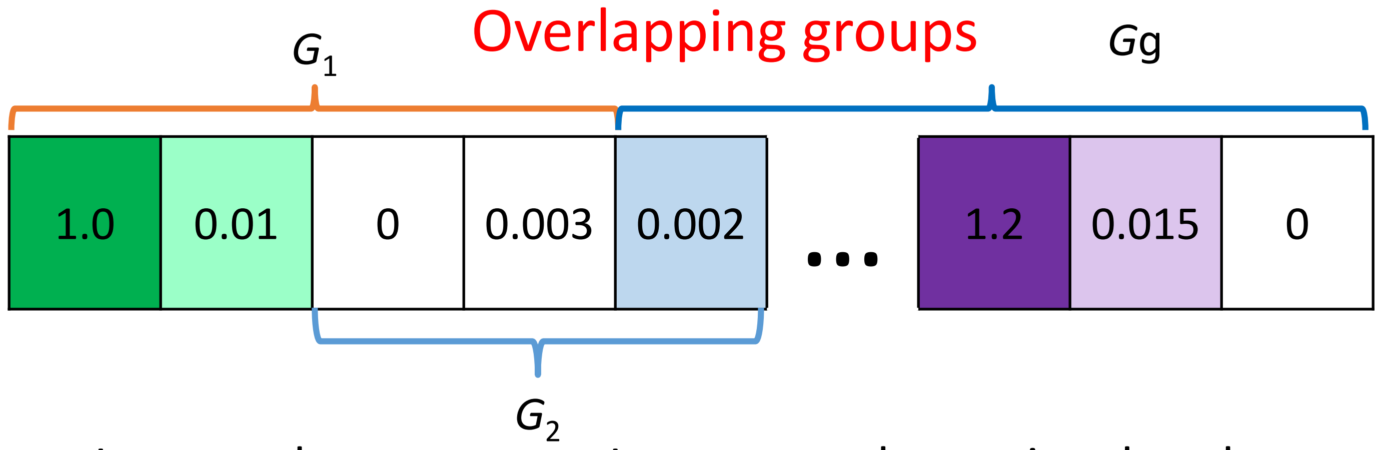


This tree structure may be available as prior knowledge, or can be learned from data using methods such as a hierarchical agglomerative clustering algorithm.[13]

We can see the true model parameters within the tree hierarchy in Figure 7 in [12].

Overlapped Group Lasso

$$\min_{\mathbf{x}} \frac{1}{2}\|A\mathbf{x}-\mathbf{y}\|_2^2 + \lambda_1\|\mathbf{x}\|_1 + \lambda_2 \sum_{i=1}^{g} w_i^g \|\mathbf{x}_{G_i}\|_2$$

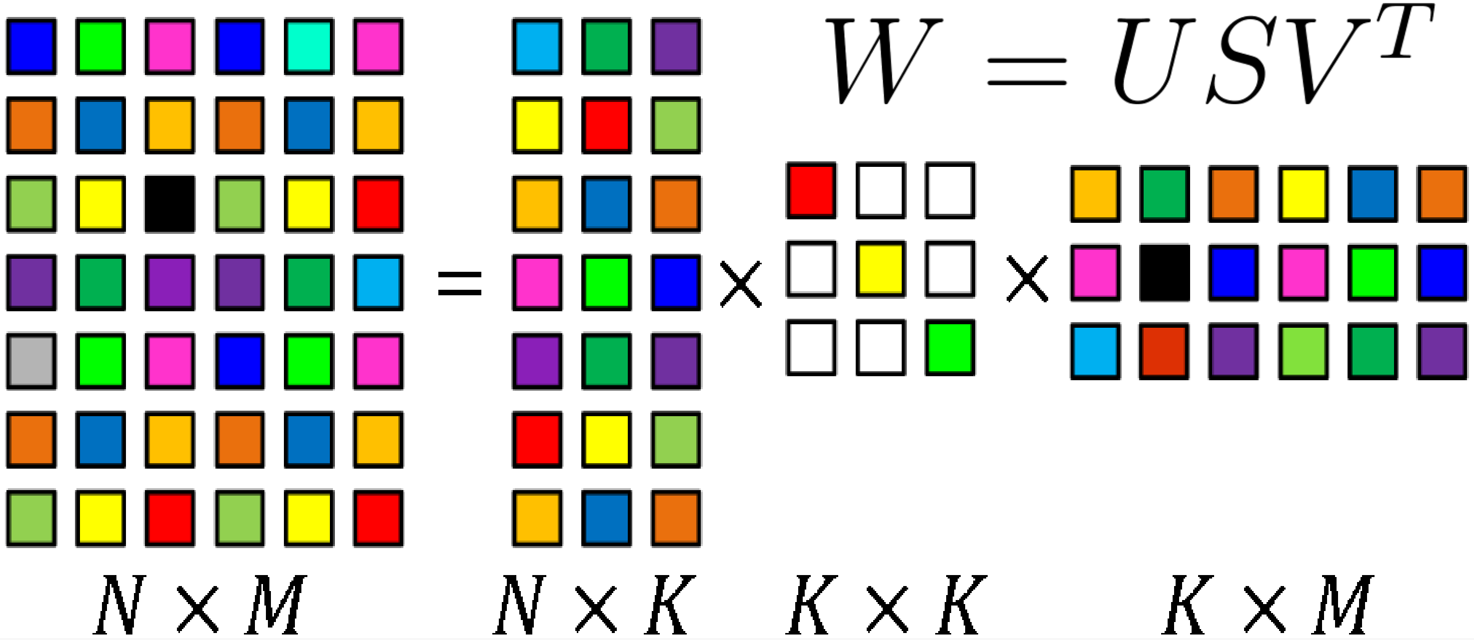


Group Lasso and sparse group Lasso are only restricted to the non-overlapping groups of features;
Tree structured group Lasso is restricted to the tree structured groups with no overlapping in the same level.

In some applications, a more flexible overlapping group structure is desired.

Trace Norm

$$W = USV^T$$



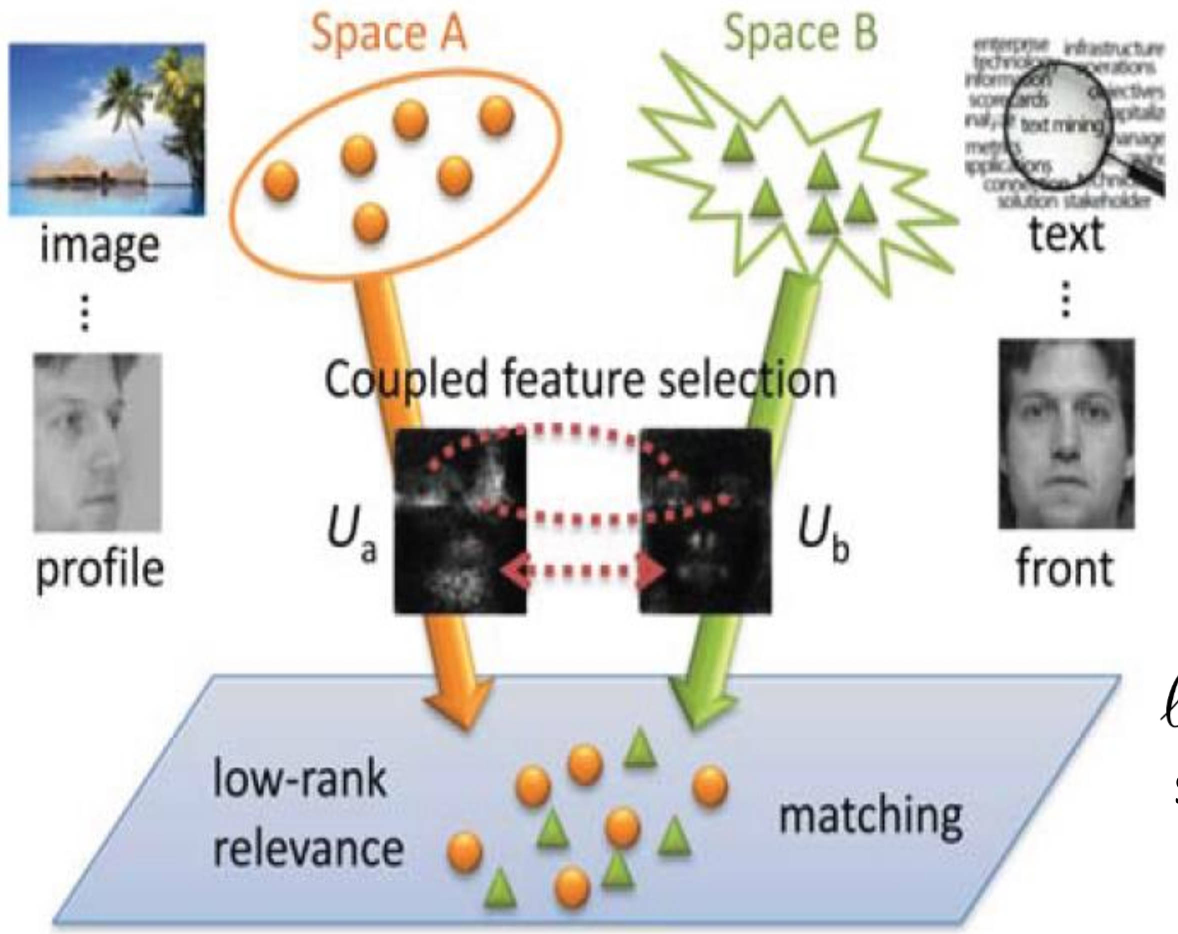
Rank of X gives the number of basis functions and measures the correlation between them, so the selection of a *joint subspace* of low dimension is equivalent to choosing a low-rank parameter matrix.

$$\|W\|_* = Tr(W) = \sum_{i=1}^{K} \sigma_i = Tr\left((WW^T)^{\frac{1}{2}}\right)$$

$$\min_W \frac{1}{2}||AW - Y||_F^2 + \lambda||W||_*$$

Rank constraint is non-convex, so we replace it with trace norm just as replacing L0-norm with L1-norm.

Overview of the Cross Modal Matching[14]



Coupled linear regression, learning projection matrices for mapping different modal data into a common space.

$$\min_{\mathbf{U}_a,\mathbf{U}_b} \boxed{\frac{1}{2}(\|\mathbf{X}_a^T\mathbf{U}_a - \mathbf{Y}\|_F^2 + \|\mathbf{X}_b^T\mathbf{U}_b - \mathbf{Y}\|_F^2)}$$
$$+\boxed{\lambda_1(\|\mathbf{U}_a\|_{21} + \|\mathbf{U}_b\|_{21})} + \boxed{\lambda_2\|[\mathbf{X}_a^T\mathbf{U}_a \quad \mathbf{X}_b^T\mathbf{U}_b]\|_*}$$

ℓ_{21}-norms play a role of feature selection on two feature spaces

Trace norm enforce the relevance of projected data with connections.

References

- [1]Efron, Bradley, et al. "Least angle regression." *The Annals of statistics* 32.2 (2004): 407-499.
- [2] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005): 301-320.
- [3] Flexeder, Claudia. *Generalized Lasso Regularization for Regression Models*. Diss. Institut für Statistik, 2010.
- [4] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.
- [5] Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association* 96.456 (2001): 1348-1360.
- [6] Fu, Wenjiang J. "Penalized regressions: the bridge versus the lasso." *Journal of computational and graphical statistics* 7.3 (1998): 397-416.
- [7] Liu, Han, and Jian Zhang. *On the l1− lq regularized regression*. Technical Report, 2008.
- [8] Tibshirani, Robert, et al. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1 (2005): 91-108.
- [9] Rinaldo, Alessandro. "Properties and refinements of the fused lasso." *The Annals of Statistics* 37.5B (2009): 2922-2952.
- [10] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "A note on the group lasso and a sparse group lasso." *arXiv preprint arXiv:1001.0736* (2010).
- [11] Liu, Jun, and Jieping Ye. "Moreau-Yosida Regularization for Grouped Tree Structure Learning." *NIPS*. Vol. 23. 2010.
- [12] Zhao, Peng, Guilherme Rocha, and Bin Yu. "The composite absolute penalties family for grouped and hierarchical variable selection." *The Annals of Statistics*(2009): 3468-3497.
- [13] Kim, Seyoung, and Eric P. Xing. "Tree-guided group lasso for multi-task regression with structured sparsity." *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010.
- [14] Wang, Kaiye, et al. "Learning Coupled Feature Spaces for Cross-modal Matching."