# Machine Learning in Bioinformatics

## Feature Selection

Philipp Benner
*philipp.benner@bam.de*

S.3 - eScience
Federal Institute for Materials Research and Testing (BAM)

June 25, 2023

## Feature selection problem

$$\hat{\theta} = \begin{cases} \underset{\theta}{\arg\min} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

- Required are computationally efficient methods to approximate the feature selection problem
- *Offline methods*: Select features before estimating parameters
- *Online methods*: Features are selected during parameter estimation

# Feature selection methods

- Offline methods:
  - Safe and Strong rules
  - Sure independence screening (SIS)
  - Estimation of mutual information
- Online methods:
  - (Orthogonal) matching pursuit
  - Least angle regression (LARS) / Homotopy algorithm
  - Penalty methods

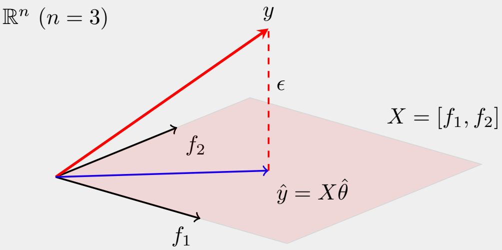
$$y = X\theta + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\begin{aligned} \text{response}: &\quad y \in \mathbb{R}^n \\ \text{covariates}: &\quad X \in \mathbb{R}^{n \times p} \\ \text{coefficients}: &\quad \theta \in \mathbb{R}^p \\ \textit{residuals}: &\quad \epsilon \in \mathbb{R}^n \end{aligned}$$

Geometric interpretation of ordinary least squares [Hastie et al., 2009]:

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \|\epsilon\|_2^2 \\ &= \arg\min_{\theta} \|y - X\theta\|_2^2\end{aligned}$$



# Sure Independence Screening (SIS)

# Sure independence screening (SIS)

- Consider the case of ultrahigh-dimensional data, where the number of features $p$ is much larger than the number of observations $n$
- Specifically, we assume that $p$ is so large that we cannot compute an estimate of $\theta$
- Assuming $\theta$ is sparse, we can first select a *promising* subset of $q$ features $M_q$ (called *feature screening*)
- The coefficients $\theta$ are estimated based on the subset $M_q$

- Consider the solution of rigde regression:

$$\hat{\theta}(\lambda) = (X^\top X + \lambda I)^{-1} X^\top y$$

- For $\lambda \to 0$ we obtain the OLS solution
- For $\lambda \to \infty$ it follows that $\lambda \hat{\theta}(\lambda)$ converges to the componentwise regression estimator

$$\hat{\theta}_k(\lambda) = \tilde{X}^\top y$$

  where $\tilde{X}$ is the data matrix $X$ with normalized columns $f_j$ such that $f_j^\top f_j = 1$
- Traditionally, for very large $p$ we would select $\lambda$ large in order to decrease the variance of $\hat{\theta}$

- $\tilde{X}^\top y = (f_1^\top y, \ldots, f_p^\top y)$ can be interpreted as the correlation of features $f_j$ with $y$
- Sure independence screening (SIS) [Fan and Lv, 2008] selects a subset of features

$$\Omega = \left\{ j \;\middle|\; |f_j^\top y| > t \right\} \tag{1}$$

based on their correlation with $y$, where $t$ is a threshold such that $|\Omega| = q < p$

- The OLS estimate $\hat{\theta}$ is computed using only the selected features $\Omega$
- All remaining components of $\hat{\theta}$ are set to zero

- The same idea can be applied to more complex models [Fan and Song, 2010], such as logistic regression, where

$$\hat{\theta} = \arg\max_{\theta} \mathrm{pr}_{\theta}(y \mid X)$$

- Select a subset of features

$$\Omega = \{\, j \mid \mathrm{score}(f_j, y) > t \} \tag{2}$$

- The score is given by the independent estimate

$$\mathrm{score}(f_j, y) = \arg\max_{\theta_j} \mathrm{pr}_{\theta_j}(y \mid f_j)$$

for all $j = 1, \ldots, p$

# Matching Pursuit for Linear Regression

## Feature selection problem

$$\hat{\theta} = \begin{cases} \underset{\theta}{\arg\min} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

## Feature selection problem

$$\hat{\theta} = \begin{cases} \arg\min\limits_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

## Matching Pursuit

Greedy approximation to feature selection problem.

## Feature selection problem

$$\hat{\theta} = \begin{cases} \arg\min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

## Matching Pursuit

Greedy approximation to feature selection problem.

If we must represent $y$ with only one feature, which one should we take?

## Feature selection problem

$$\hat{\theta} = \begin{cases} \arg\min_{\theta} & \|y - X\theta\|_2^2 \\ \text{subject to} & \|\theta\|_0 = m \end{cases} \quad \text{with } \binom{p}{m} \text{ possible subsets}$$

## Matching Pursuit

Greedy approximation to feature selection problem.

If we must represent $y$ with only one feature, which one should we take?

$$j_1 = \arg\min_{j} \left\|y - f_j\hat{\theta}_j\right\|_2^2, \quad \text{where} \quad \hat{\theta}_j = \arg\min_{\theta_j} \|y - f_j\theta_j\|_2^2$$

# Matching Pursuit for linear regression

$$j_1 = \arg\min_{j} \left\| y - f_j \hat{\theta}_j \right\|_2^2 , \quad \text{where} \quad \hat{\theta}_j = \arg\min_{\theta_j} \left\| y - f_j \theta_j \right\|_2^2$$

$$j_1 = \arg\min_j \left\| y - f_j \hat{\theta}_j \right\|_2^2 , \quad \text{where} \quad \hat{\theta}_j = \arg\min_{\theta_j} \left\| y - f_j \theta_j \right\|_2^2$$
$$= \arg\max_j \frac{(f_j^\top y)^2}{f_j^\top f_j}$$
$$= \arg\max_j \left| f_j^\top y \right|$$

[assuming normalized data, i.e. $f_j^\top f_j = 1$]

$\Rightarrow$ select feature $j$ with maximal scalar projection of $y$ onto $f_j$

$$
\begin{aligned}
\epsilon &= y - X\theta \\
&= \underbrace{\underbrace{\underbrace{y}_{r_0} - f_{j_1}\theta_{j_1}}_{r_1} - f_{j_2}\theta_{j_2}}_{r_2} - \ldots - f_{j_p}\theta_{j_p}
\end{aligned}
$$

$$
\begin{aligned}
j_1 &= \arg\min_j \left\| y - f_j\hat{\theta}_j \right\|_2^2 \qquad = \arg\min_j \left\| r_0 - f_j\hat{\theta}_j \right\|_2^2 \\
&= \arg\max_j \left| f_j^\top r_0 \right|
\end{aligned}
$$

$$\begin{aligned}\epsilon &= y - X\theta \\ &= \underbrace{\underbrace{\underbrace{y}_{r_0} - f_{j_1}\theta_{j_1}}_{r_1} - f_{j_2}\theta_{j_2}}_{r_2} - \ldots - f_{j_p}\theta_{j_p}\end{aligned}$$

$$\begin{aligned}j_1 &= \arg\min_j \left\|y - f_j\hat{\theta}_j\right\|_2^2 && = \arg\min_j \left\|r_0 - f_j\hat{\theta}_j\right\|_2^2 \\ &= \arg\max_j \left|f_j^\top r_0\right| \\ j_2 &= \arg\min_j \left\|y - f_{j_1}\hat{\theta}_{j_1} - f_j\hat{\theta}_j\right\|_2^2 && = \arg\min_j \left\|r_1 - f_j\hat{\theta}_j\right\|_2^2 \\ &= \arg\max_j \left|f_j^\top r_1\right|\end{aligned}$$

## Matching pusuit (MP) [Tropp et al., 2007]

The MP feature selection rule is given by

$$j_k = \arg\max_j \left| f_j^\top r_{k-1} \right| \qquad k = 1, \dots, m$$

where $r_k$ are the residuals at step $k$:

$$\begin{aligned} \epsilon &= y - X\theta \\ &= \underbrace{\underbrace{\underbrace{y}_{r_0} - f_{j_1}\theta_{j_1}}_{r_1} - f_{j_2}\theta_{j_2}}_{r_2} - \dots - f_{j_p}\theta_{j_p} \end{aligned}$$

## Orthogonal Matching Pursuit

Orthogonal Matching Pursuit: Re-estimate parameters after every iteration.

After every iteration $t$, update all $\theta_{\Omega_t}$ entries, where $\Omega_t = \{j_1, j_2, \ldots, j_t\}$, i.e. compute

$$\theta_{\Omega_t} = \arg\min_{\theta} \|y_{\Omega_t} - X_{\Omega_t}\theta\|_2^2 \ .$$

This update changes the residuals

$$r_t = y - f_{j_1}\theta_{j_1} - f_{j_2}\theta_{j_2} - \cdots - f_{j_t}\theta_{j_t}$$

used in the next iteration of the algorithm.

# Matching Pursuit for Logistic Regression

$$\begin{bmatrix} \mathrm{pr}_\theta(y_1 = 1) \\ \mathrm{pr}_\theta(y_2 = 1) \\ \vdots \\ \mathrm{pr}_\theta(y_n = 1) \end{bmatrix} = \sigma\left( \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \right)$$

$$\begin{aligned} \text{class labels}: &\quad y \in \{0, 1\}^n \\ \text{covariates}: &\quad X \in \mathbb{R}^{n \times p} \\ \text{coefficients}: &\quad \theta \in \mathbb{R}^p \end{aligned}$$

Parameter estimation for logistic regression:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \mathrm{pr}_{\theta}(y) \approx \arg\min_{\theta} \|y - \sigma(X\theta)\|_2^2 \quad \text{[but not convex]} \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log \mathrm{pr}_{\theta}(y_i) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \{y_i \log \sigma(x_i\theta) + (1 - y_i) \log(-x_i\theta)\} \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log \sigma(\tilde{y}_i x_i \theta),\end{aligned}$$

where $\tilde{y}_i = 2y_i - 1 \in \{-1, 1\}$

## Pseudo-residuals

$$
\begin{aligned}
r_k &= y - \sigma(f_{j_1}\theta_{j_1} + f_{j_2}\theta_{j_2} + \cdots + f_{j_k}\theta_{j_k}) \\
X^\top r_p &= \nabla \log \mathrm{pr}_\theta(y)
\end{aligned}
$$

## Pseudo-residuals

$$
\begin{aligned}
r_k &= y - \sigma(f_{j_1}\theta_{j_1} + f_{j_2}\theta_{j_2} + \cdots + f_{j_k}\theta_{j_k}) \\
X^\top r_p &= \nabla \log \mathrm{pr}_\theta(y)
\end{aligned}
$$

$$
\begin{aligned}
j_1 &= \arg\min_j \left\| y - \sigma(f_j \hat{\theta}_j) \right\|_2^2 \\
&\approx \arg\max_j \left| f_j^\top r_0 \right|
\end{aligned}
$$

# Matching Pursuit for Logistic Regression

## Pseudo-residuals

$$r_k = y - \sigma(f_{j_1}\theta_{j_1} + f_{j_2}\theta_{j_2} + \cdots + f_{j_k}\theta_{j_k})$$
$$X^\top r_p = \nabla \log \mathrm{pr}_\theta(y)$$

$$
\begin{aligned}
j_1 &= \arg\min_j \left\| y - \sigma(f_j\hat{\theta}_j) \right\|_2^2 \\
&\approx \arg\max_j \left| f_j^\top r_0 \right| \\
j_2 &= \arg\min_j \left\| y - \sigma(f_{j_1}\hat{\theta}_{j_1} - f_j\hat{\theta}_j) \right\|_2^2 \\
&\approx \arg\max_j \left| f_j^\top r_1 \right|
\end{aligned}
$$

## Matching pursuit feature selection rule [Lozano et al., 2011]

Assuming normalized data, i.e. $f_j^\top f_j = 1$, the OMP rule is given by

$$j_k = \arg\max_j \left| f_j^\top r_{k-1} \right|$$

where $r_k$ are the $k$th pseudo-residuals

$$\begin{aligned} r_k &= y - \sigma(f_{j_1}\theta_{j_1} + f_{j_2}\theta_{j_2} + \cdots + f_{j_k}\theta_{j_k}) \\ X^\top r_p &= \nabla \log \mathrm{pr}_\theta(y) \end{aligned}$$

## Matching pursuit feature selection rule [Lozano et al., 2011]

Assuming normalized data, i.e. $f_j^\top f_j = 1$, the OMP rule is given by

$$j_k = \arg\max_j \left| f_j^\top r_{k-1} \right|$$

where $r_k$ are the $k$th pseudo-residuals

$$\begin{aligned} r_k &= y - \sigma(f_{j_1}\theta_{j_1} + f_{j_2}\theta_{j_2} + \cdots + f_{j_k}\theta_{j_k}) \\ X^\top r_p &= \nabla \log \mathrm{pr}_\theta(y) \end{aligned}$$

## OMP Performance

Greedy strategy causes poor performance of Orthogonal Matching Pursuit in practice

# Least Angle Regression (LARS)

# Least Angle Regression (LARS)

- Consider $\ell_1$-penalized linear regression (LASSO) where

$$\hat{\theta}(\lambda) = \arg\min_{\theta} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$$

- There exists a regularization strength $\lambda = \lambda_{\max}$ for which all estimated coefficients are zero
- Least Angle Regression (LARS) [Efron et al., 2004] is a method to efficiently compute $\hat{\theta}(\lambda)$ for all $0 \leq \lambda \leq \lambda_{\max}$
- LARS computes breakpoints $\lambda_k$ at which individual coefficients $\hat{\theta}_j(\lambda_k) \in \mathbb{R}$ change its value from
  - zero to non-zero, or from
  - non-zero to zero
- Between breakpoints the values of coefficients can be linearly interpolated

- Remember that the OLS solution $\hat{\theta}(0)$ for $\lambda = 0$ requires that

$$\nabla_\theta \|y - X\theta\|_2^2 = 2X^\top(y - X\theta) = 0$$

- For $\lambda > 0$ the solution requires

$$X^\top(y - X\theta) \in \tfrac{\lambda}{2}\partial \|\theta\|_1$$

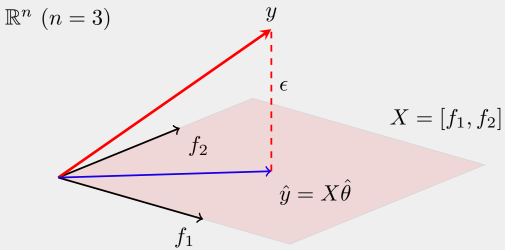
  where $\partial \|\theta\|_1$ is the subgradient with respect to $\theta$

- We define

$$c(\theta) = X^\top(y - X\theta)$$

  which is interpreted as the correlation of features $X = [f_1, f_2, \ldots, f_p]$ with the residuals $\epsilon = y - X\theta$

- The correlation $\hat{c}(\lambda) = c(\hat{\theta}(\lambda))$ varies with $\lambda$ as follows:
  - $\hat{c}(\lambda) = c_{\max}$ for $\lambda = \lambda_{\max}$
  - $\hat{c}(\lambda) = 0$ for $\lambda = 0$



- LARS maintains a set of active features $\Omega \subset \{1, \ldots, p\}$ all equally correlated with the residuals $y - X\hat{\theta}(\lambda)$ for the current estimate $\hat{\theta}(\lambda)$
- Let $X_\Omega = (f_j)_{j \in \Omega}$ denote the covariate matrix and $\theta_\Omega = (\theta_j)_{j \in \Omega}$ the coefficients restricted to the features in the active set $\Omega$
- In each iteration, the coefficients $\theta$ are updated

$$\theta \leftarrow \theta + \gamma^* v \,,$$

where $\gamma^*$ is the amount by which the correlation $c_\Omega(\theta)$ is reduced and $v \in \mathbb{R}^p$ defines the direction and relative size of the update

- The vector $v$ is selected so that for features in $\Omega$ the difference in correlation $c_\Omega(\theta) - c_\Omega(\theta + \gamma v)$ shrinks uniformly towards zero with rate $\gamma$, i.e.

$$\begin{aligned} c_\Omega\,(\theta) - c_\Omega\,(\theta + \gamma v) &= \gamma \operatorname{sign} c_\Omega(\theta)\,, \quad \text{while} \\ c_{\Omega^c}(\theta) - c_{\Omega^c}(\theta + \gamma v) &= 0\,. \end{aligned}$$

- Both conditions can be combined into

$$c(\theta) - c(\theta + \gamma v) = \gamma \operatorname{sign} c(\theta)\,,$$

since $\operatorname{sign} c_{\Omega^c}(\theta) = 0$

- It follows that

$$v_\Omega = [X_\Omega^\top X_\Omega]^{-1} \operatorname{sign} c_\Omega(\theta)$$

and $v_{\Omega^c} = 0$

# Least Angle Regression (LARS)

- LARS stop shrinking the correlations whenever:
  - Case 1: A non-active feature becomes equally correlated with the residuals
  - Case 2: A coefficient of an active feature becomes zero[1]
- Case 1: More formally, $\gamma$ is increased until some feature $j' \in \Omega^c$ outside the active group satisfies

$$\begin{aligned} |c_{j'}(\theta + \gamma v)| &= |c_j(\theta + \gamma v)| \\ &= \lambda - \gamma \,, \end{aligned}$$

  where $j \in \Omega$, and $\lambda = |c_j(\theta)|$ is the absolute correlation of the active features

[1]This case was not part of the initial LARS algorithm but was later on added in order to ensure equivalence with the LASSO (see also Homotopy algorithm [Osborne et al., 2000])

- The solution is given by

$$\gamma^+ = \min_{j\in\Omega^c}{}^{+} \left\{ \frac{\lambda - c_j(\theta)}{1 - f_j^\top X v}, \frac{\lambda + c_j(\theta)}{1 + f_j^\top X v} \right\} ,$$

  where $\min^+$ is the minimum over positive elements and note that $f_j^\top X v = f_j^\top X_\Omega v_\Omega$

- Case 2: The algorithm also removes a feature $j$ from the active set when for some $\gamma$

$$\theta_j + \gamma v_j = 0$$

  so that $\gamma^- = \min_{j\in\Omega}\{-\theta_j / v_j\}$

- The subsequent breakpoint is given by $\gamma^* = \min\{\gamma^+, \gamma^-\}$

# SAFE and Strong Rules

## Penalized regression

$$\omega(\theta) = -\log \mathrm{pr}_\theta(y) \qquad \text{(logistic regression), or}$$
$$\omega(\theta) = \|y - X\theta\|_2^2 \qquad \text{(linear regression)}$$

$$\hat{\theta} = \begin{cases} \arg\min\limits_{\theta} & \omega(\theta) \\ \text{subject to} & \|\theta\|_1 = \Lambda \end{cases}$$

Basic idea: Select $\Lambda$ such that $\|\theta\|_0 = m$

## Numerical solution of penalized regression

Identify saddle points of Lagrangian

$$\mathcal{L}(\theta, \lambda) = \omega(\theta) + \lambda(\|\theta\|_1 - \Lambda)$$

## Numerical solution of penalized regression

Identify saddle points of Lagrangian

$$\mathcal{L}(\theta, \lambda) = \omega(\theta) + \lambda(\|\theta\|_1 - \Lambda)$$

In practice the constraint $\|\theta\|_1 = \Lambda$ is ignored, but $\lambda$ is chosen such that classification performance is optimal:

## Penalized regression in practice

$$\hat{\theta}(\lambda) = \arg\min_{\theta} \omega(\theta) + \lambda \|\theta\|_1$$

# SAFE Rule for linear regression

SAFE rule: What features can we neglect for a fixed $\lambda$?

SAFE rule: What features can we neglect for a fixed $\lambda$?

**SAFE rule [Ghaoui et al., 2010, Kim et al., 2007] for $\ell_1$-penalized linear regression**

$j$th component of $\hat{\theta}$ must be zero if

$$|f_j^\top y| < \lambda - \|f_j\|_2 \, \|y\|_2 \, \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$

$$\lambda_{\max} = \max_j |f_j^\top y|$$

SAFE rule: What features can we neglect for a fixed $\lambda$?

**SAFE rule [Ghaoui et al., 2010, Kim et al., 2007] for $\ell_1$-penalized linear regression**

$j$th component of $\hat{\theta}$ must be zero if

$$|f_j^\top y| < \lambda - \|f_j\|_2 \, \|y\|_2 \, \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$

$$\lambda_{\max} = \max_j |f_j^\top y|$$

$$|f_j^\top (y - \underbrace{X\theta}_{\theta=0})| < \lambda - \|f_j\|_2 \, \|y\|_2 \, \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$

SAFE rule for linear regression: $j$th component of $\hat{\theta}$ must be zero if

$$|f_j^\top y| < \lambda - \|f_j\|_2 \|y\|_2 \frac{\lambda_{\max} - \lambda}{\lambda_{\max}}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

**Strong rule for $\ell_1$-penalized linear regression [Tibshirani et al., 2012]**

Discard $j$th component if

$$|f_j^\top y| < \lambda - (\lambda_{max} - \lambda) = 2\lambda - \lambda_{max}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

## Strong rule for $\ell_1$-penalized linear regression [Tibshirani et al., 2012]

Discard $j$th component if

$$|f_j^\top y| < \lambda - (\lambda_{max} - \lambda) = 2\lambda - \lambda_{max}$$

$$\lambda_{\max} = \max_j |f_j^\top y|$$

## Remark

Strong rule may drop features that should not be discarded $\Rightarrow$ KKT conditions must be checked, i.e.

$$X^\top (y - X\hat{\theta}) \in \lambda \partial_{\theta=\hat{\theta}} \|\theta\|_1$$

## Strong rule for $\ell_1$-penalized linear regression [Tibshirani et al., 2012]

Discard $j$th component if

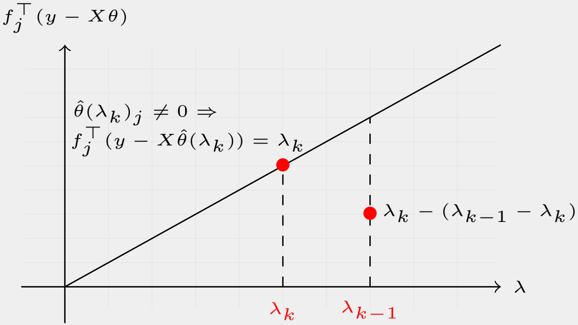
$$|f_j^\top y| < \lambda - (\lambda_{max} - \lambda) = 2\lambda - \lambda_{max}$$
$$\lambda_{\max} = \max_j |f_j^\top y|$$

## Strong sequential rule for $\ell_1$-penalized linear regression[Tibshirani et al., 2012]

Discard $j$th feature if

$$|f_j^\top \{y - \sigma(X\hat{\theta}(\lambda_{k-1}))\}| < 2\lambda_k - \lambda_{k-1}$$

Compute $\hat{\theta}(\lambda_k)$ for all $\lambda_1 > \cdots > \lambda_k > \cdots > \lambda_K$




$$\text{Assumption} : |f_j^\top(y - X\hat{\theta}(\lambda_{k-1} - \epsilon)) - f_j^\top(y - X\hat{\theta}(\lambda_{k-1}))| \leq \epsilon$$

$$\Rightarrow |f_j^\top(y - X\hat{\theta}(\lambda_{k-1}))| < 2\lambda_k - \lambda_{k-1}$$

# References I


Boyd, S. and Vandenberghe, L. (2004).
**Convex optimization.**
Cambridge university press.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014).
**Saga: A fast incremental gradient method with support for non-strongly convex composite objectives.**
*Advances in neural information processing systems*, 27.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004).
**Least angle regression.**
*The Annals of statistics*, 32(2):407–499.

Fan, J. and Lv, J. (2008).
**Sure independence screening for ultrahigh dimensional feature space.**
*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

# References II


Fan, J. and Song, R. (2010).
**Sure independence screening in generalized linear models with NP-dimensionality.**
*The Annals of Statistics*, 38(6):3567–3604.

Ghaoui, L. E., Viallon, V., and Rabbani, T. (2010).
**Safe feature elimination for the lasso and sparse supervised learning problems.**
*arXiv preprint arXiv:1009.4219*.

Hastie, T., Tibshirani, R., and Friedman, J. (2009).
***The elements of statistical learning: data mining, inference, and prediction.***
Springer Science & Business Media.

Kim, S.-j., Koh, K., Lustig, M., Boyd, S., and Gorinevsky, D. (2007).
**An interior-point method for large-scale $\ell$1-regularized logistic regression.**
In *Journal of Machine learning research*. Citeseer.

- Lozano, A., Swirszcz, G., and Abe, N. (2011).
  **Group orthogonal matching pursuit for logistic regression.**
  In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 452–460. JMLR Workshop and Conference Proceedings.
- Osborne, M. R., Presnell, B., and Turlach, B. A. (2000).
  **A new approach to variable selection in least squares problems.**
  *IMA journal of numerical analysis*, 20(3):389–403.
- Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012).
  **Strong rules for discarding predictors in lasso-type problems.**
  *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266.

# References IV


Tropp, J., Gilbert, A. C., et al. (2007).
**Signal recovery from partial information via orthogonal matching pursuit.**
*IEEE Trans. Inform. Theory*, 53(12):4655–4666.

# Derivation of the SAFE Rule for linear regression

$$\hat{\theta} = \arg\min_{\theta} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$$

Define

$$\beta = y - X\theta$$

Equivalent optimization problem

$$\hat{\theta} = \begin{cases} \arg\min_{\theta} & \beta^\top \beta + \lambda \|\theta\|_1 \\ \text{subject to} & \beta = y - X\theta \end{cases}$$

# Derivation of the SAFE Rule for linear regression

Lagrangian

$$\mathcal{L}(\theta, \beta, \nu) = \beta^\top \beta + \lambda \left\|\theta\right\|_1 + \nu^\top (y - X\theta - \beta)$$

Dual function

$$\inf_{\theta,\beta} \mathcal{L}(\theta, \beta, \nu) = \begin{cases} G(\nu) & \text{if } |f_j^\top \nu| \leq \lambda \,,\ j = 1, \dots, p \\ -\infty & \text{otherwise} \end{cases}$$

where $G(\nu) = -\frac{1}{4}\nu^\top \nu + \nu^\top y$. Lagrange dual

$$\hat{\theta}^* = \begin{cases} \arg\max\limits_{\nu} & G(\nu) \\ \text{subject to} & |f_j^\top \nu| \leq \lambda \,,\ j = 1, \dots, p \end{cases}$$

# Derivation of the SAFE Rule for linear regression

Side note: Since the primal problem satisfies Slater's condition, we know that the duality gap $\gamma = \hat{\theta} - \hat{\theta}^*$ is zero, i.e.

$$\hat{\theta} = \hat{\theta}^*$$

For a dual feasible point $\nu_0$, we solve for each $j = 1, \ldots, p$

$$\begin{aligned}\xi_j(\nu_0) &= \begin{cases} \arg\max_{\nu} & |f_j^\top \nu| \\ \text{subject to} & G(\nu) \geq G(\nu_0) \end{cases} \\ &= |f_j^\top y| + \sqrt{(y^\top y - 2G(\nu_0))f_j^\top f_j}\end{aligned}$$

If $\xi_j(\nu_0) < \lambda$ we know that $\hat{\theta}_j = 0$. A simple dual feasible point is $\nu_0 = y\lambda/\lambda_{max}$. The SAFE rule is obtained from

$$\xi_j(y\lambda/\lambda_{max}) < \lambda$$

# Logistic Regression Classifier

SAGA algorithm [Defazio et al., 2014]: select $j \in \{1, \ldots, n\}$ at random

$$
\begin{aligned}
\vartheta_{j,t+1} &= \theta_t \\
\vartheta_{i,t+1} &= \vartheta_{i,t+1} \text{ for all } i \neq j \\
\theta^*_{t+1} &= \theta_t - \gamma \left[ \nabla \ell_j(\vartheta_{j,t+1}) - \nabla \ell_j(\vartheta_{j,t}) + \frac{1}{n} \sum_{i=1}^{n} \nabla \ell_i(\vartheta_{i,t}) \right] \\
\theta_{t+1} &= \arg\min_{\theta} \left\{ \lambda \|\theta\|_1 + \frac{1}{2\gamma} \|\theta - \theta^*_{t+1}\|_2^2 \right\}
\end{aligned}
$$