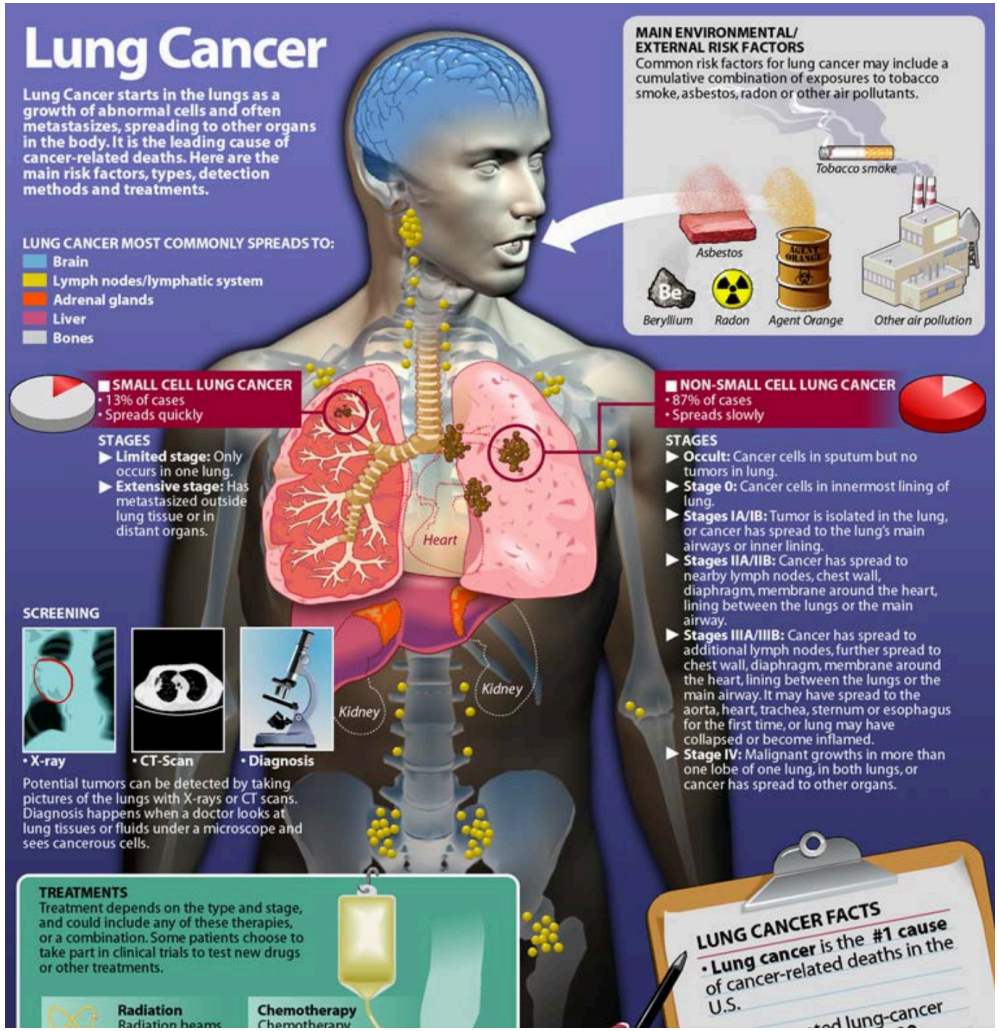
Feature Selection (FS)

- **FS is one of 3 main tasks for Machine Learning**
  - A patient may have cancer or not (disease class label)
  - An image have keywords descriptors (class labels)
  - A data instance (image, patients) have class labels due to many factors/reasons.
- **FS finds the most important factors for classification**
  - Most relevant genes for a diseases
  - e.g. select smoking for lung cancer
- **Many FS methods/algorithms**
  - T-test, F-test, Chi-statistic
  - Mutual information, ReliefF, mRMR
  - Sparse coding(e.g. L_1-norm, L_{21}-norm, $L_{1\infty}$-norm, L_{12}-norm)

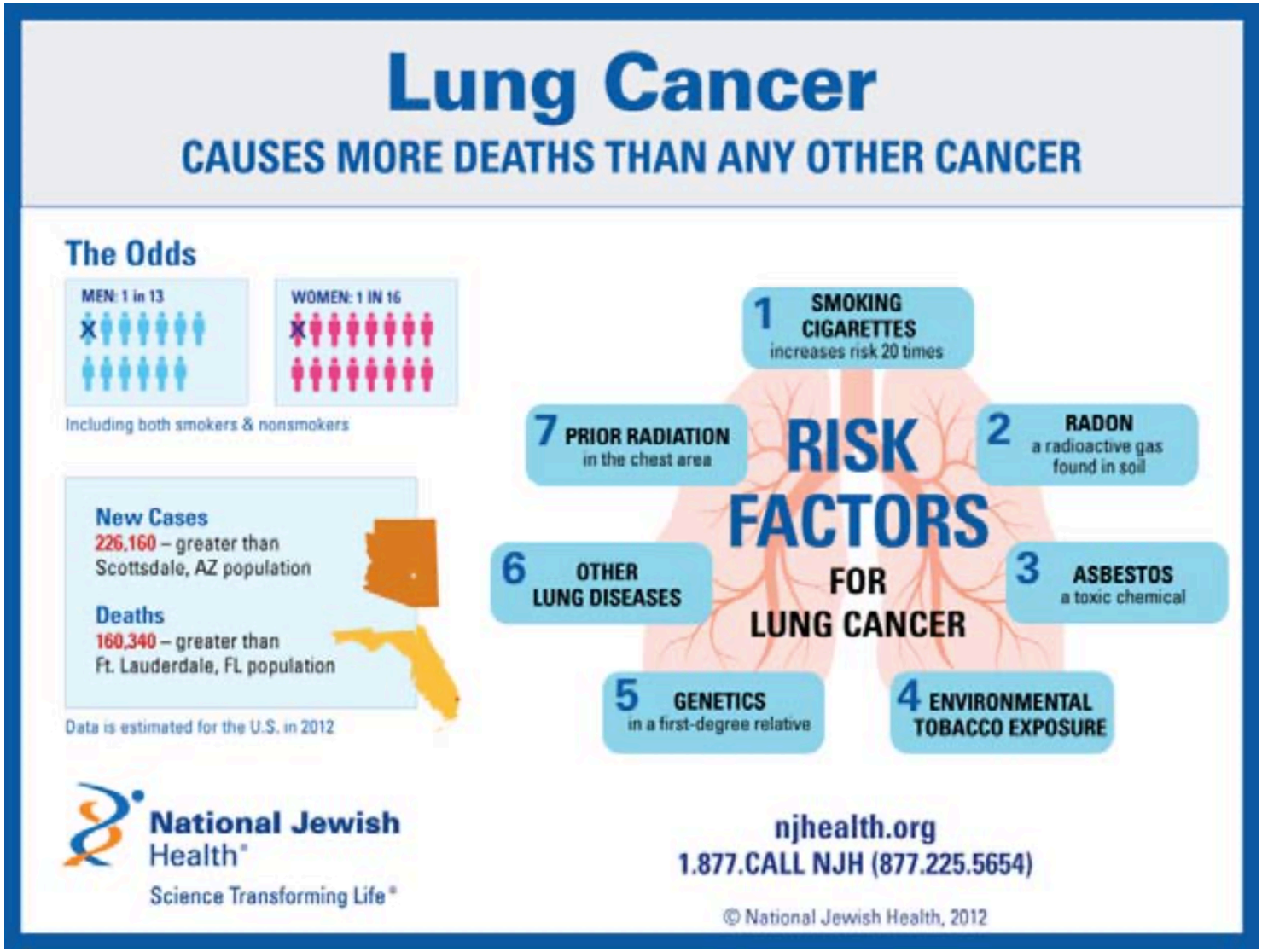
FS example: lung cancer

No.1 cause of cancer-related death in US!

Feature selection → finding related factors



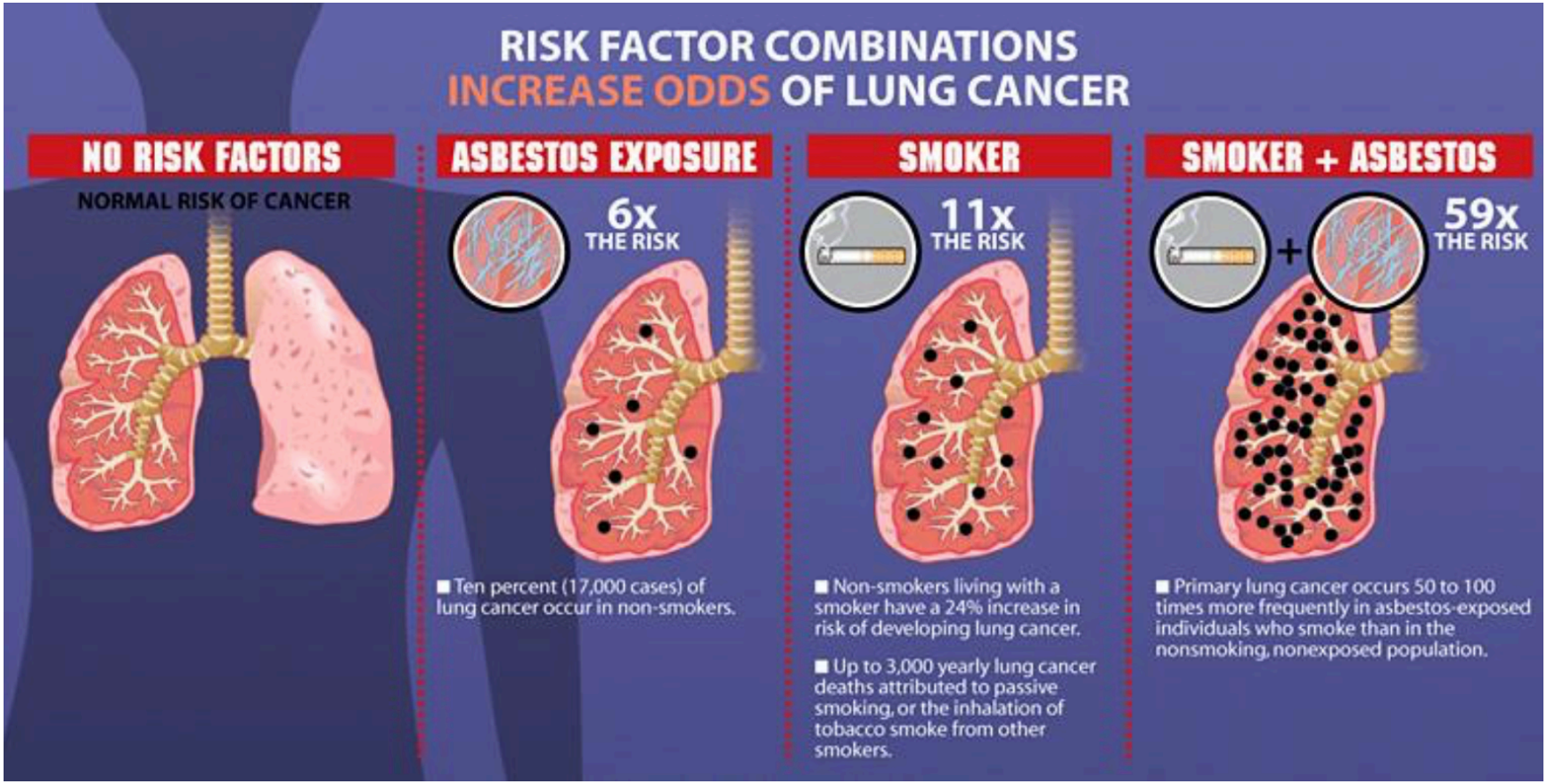
FS example: lung cancer



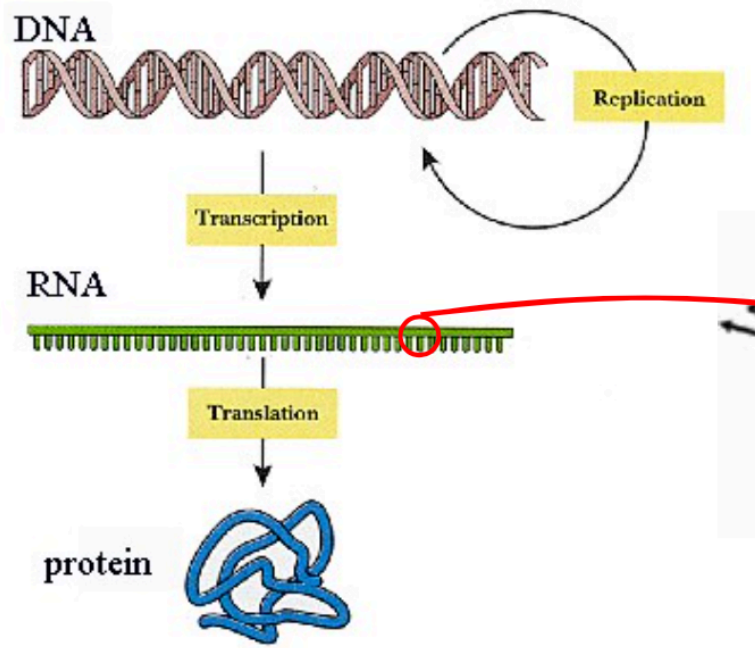
FS example: lung cancer

Single factors: smoking, expose to asbestos, etc.

Multiple factors: increase risk much more!



FS example: gene expressions



Microarray Analysis

Expression Microarray

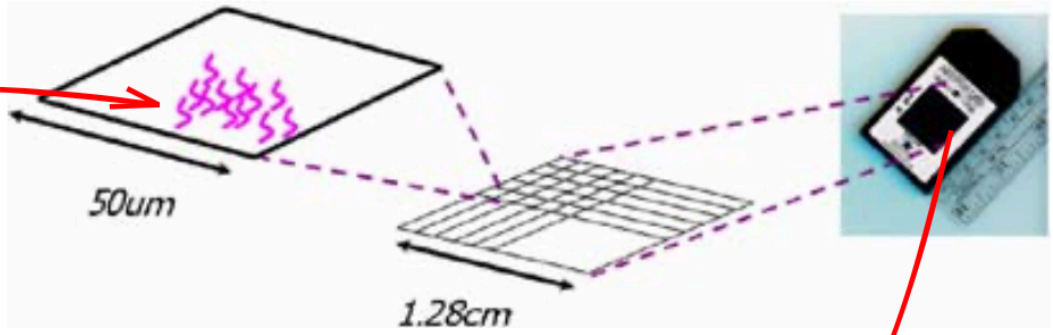


Image Courtesy of Affymetrix

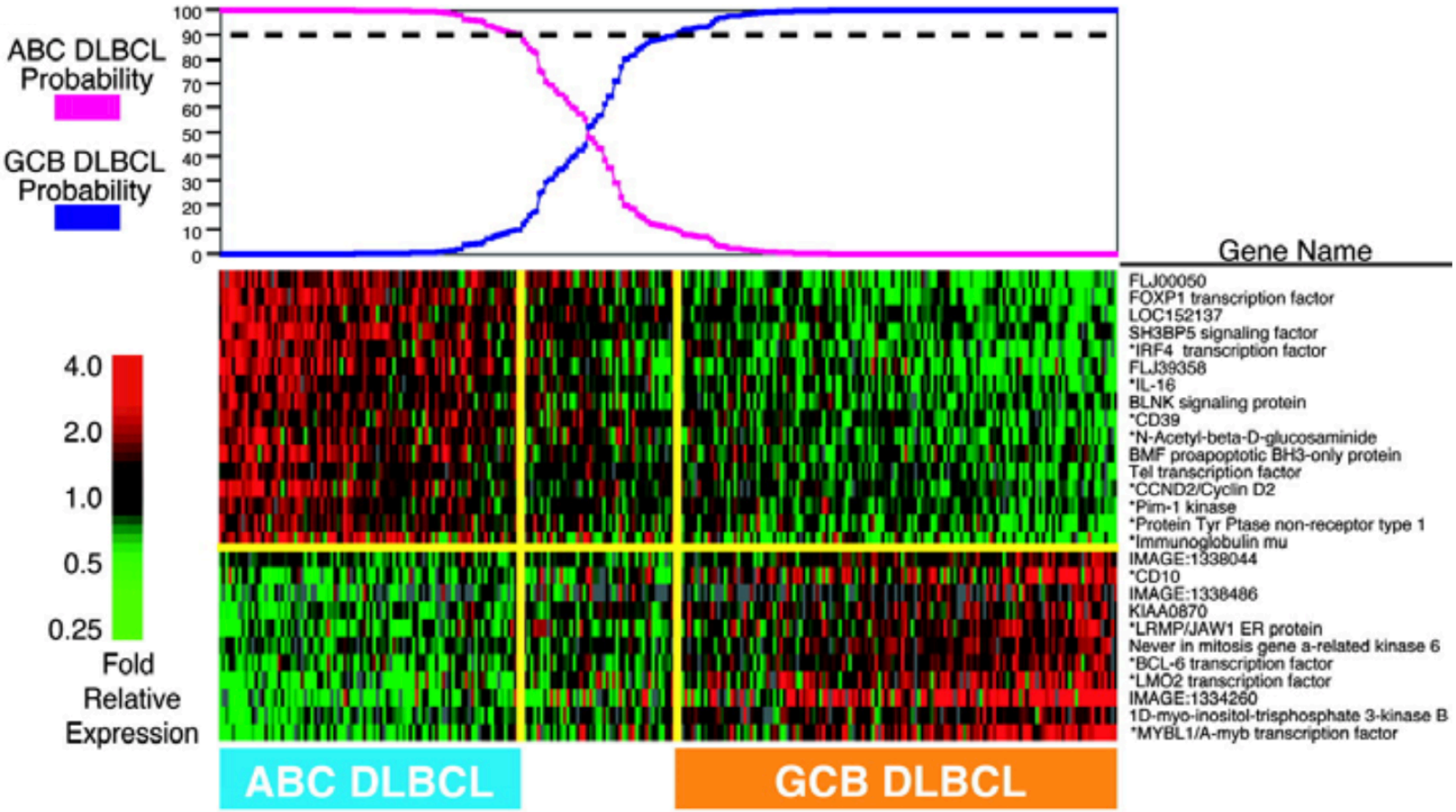
- **Analyze gene expressions of** disease types (disease diagnosis)
- **Challenge:** thousands of genes, few patients samples
- **Solution:** select several most relevant genes

Sample \ Gene	M23197_at	U66497_at	M92287_at	...	Class
Sample 1	261	88	4778	...	ALL
Sample 2	101	74	2700	...	ALL
Sample 3	1450	34	498	...	AML
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.

Expression Microarray Data Set

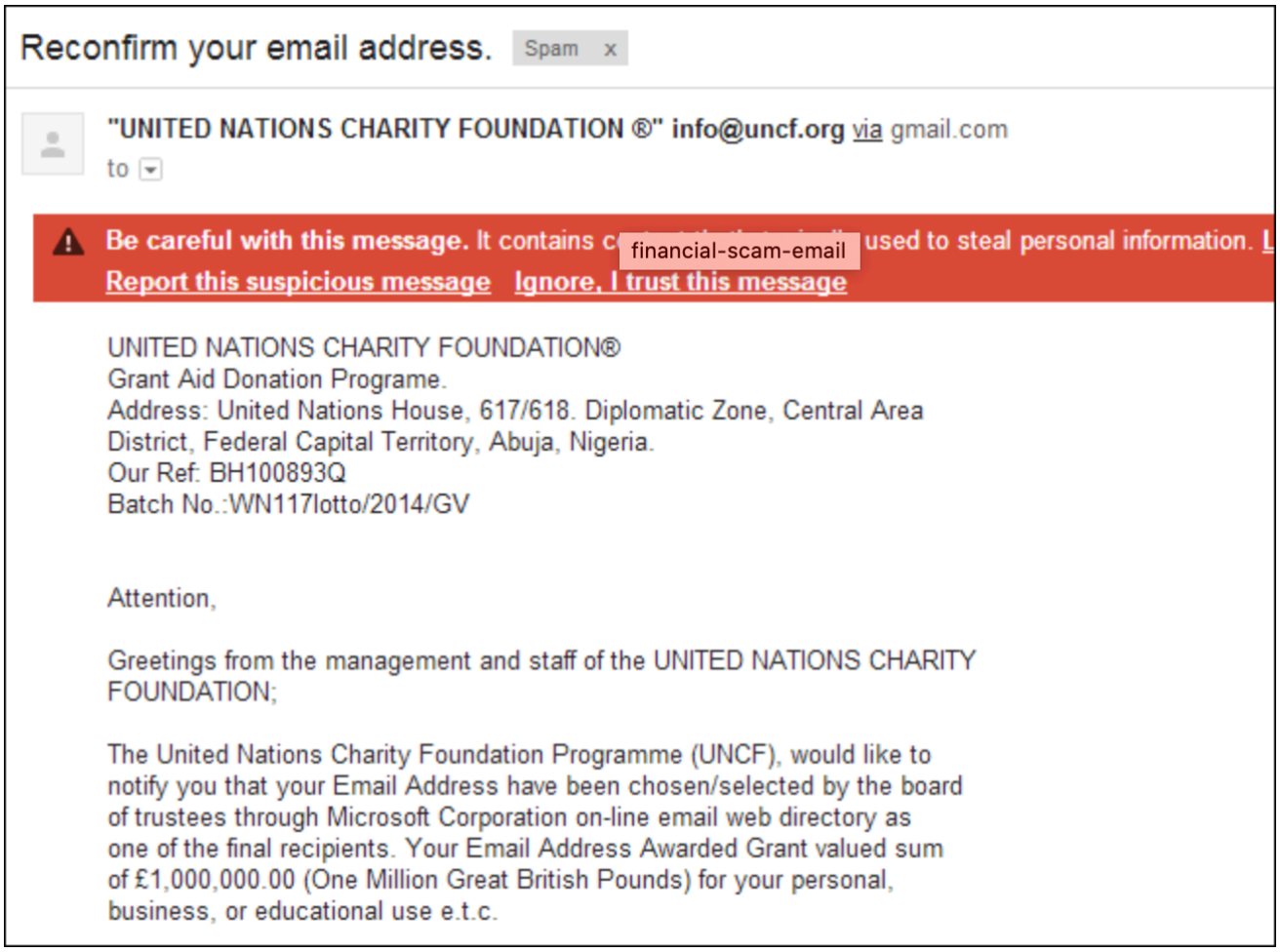
FS example: gene expressions

Pick most relevant genes for each groups: ABC vs GCB



FS example: email filtering

Pick the most relevant words: Spam vs Non-Spam



FS example: object recognition

Pick the most distinctive features: Men vs Women

St Marco Square, Venice, Italy

Why Feature Selection?

- Some algorithms scale (computationally) poorly with increased dimension.
- Irrelevant features can confuse some algorithms.
- Redundant features adversely affect regularization.
- Removal of features can increase (relative) margin (and generalization).
- Reduces data set and resulting model size.

Feature Selection Methods

- **Wrapper Methods**
  - Repeated runs of learning algorithm with different set of features
  - Can be computationally expensive,
  - e.g. SVM-RFE
- **Filter Methods**
  - Uses heuristics but is much faster than wrapper methods
  - Rank features in order of their correlation with the labels.
  - e.g. Mutual information, F-statistic, ReliefF, mRMR.
- **Sparse Coding based Methods (a.k.a. Embedded Method)**
  - Find a sparse representation of weights.
  - Add a penalty on weight in the learning model.
  - Achieve a balance between loss and penalty.
  - e.g. LASSO (L_1-norm), Multi-task FS (L_{21}-norm, L_{12}-norm)

F-statistics

- For continuous data variables (or attributes), we can choose the **F-statistic** between the **genes** and the **classification variable** as the **score of maximum relevance**.
- The F-test value of gene variable g_i in K classes denoted by h has the following form:

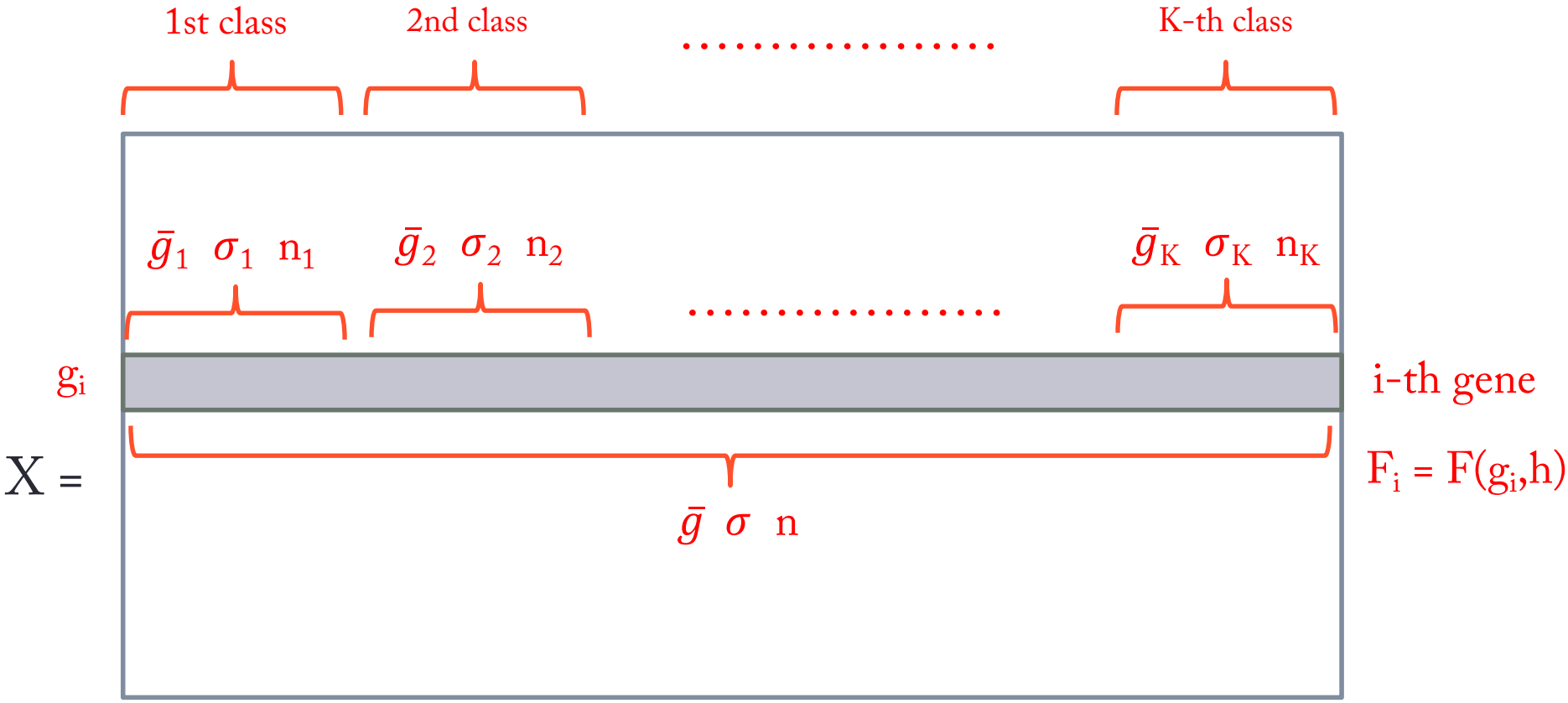
$$F(g_i, h) = \left[\sum_k n_k (\bar{g}_k - \bar{g})^2 / (K-1) \right] \Big/ \sigma^2$$

where $\bar{g}$ is the mean value of g_i in all tissue samples, $\bar{g}_k$ is the mean value of g_i within the k-th class, σ^2 is the pooled variance defined as

$$\sigma^2 = \left[\sum_k (n_k - 1)\sigma_k^2 \right] \Big/ (n - K)$$

where n_k and σ_k are the size and the variance of the k-th class.

F-statistics



- Sort F_1, F_2, …… , F_{1000} in a decreasing order
- Select the top 100 features

Sparse Coding

- Given data instances $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where $\mathbf{x}_i \in \mathbb{R}^d$, and class labels $\mathbf{Y} = [\mathbf{y}^1; \cdots; \mathbf{y}^n] \in \mathbb{R}^{n \times k}$, where $\mathbf{y}^i \in \mathbb{R}^k$, sparse coding based method can be mathematically formulated as:

$$\min_{\mathbf{W}} \ f(\mathbf{W}) + \lambda \Omega(\mathbf{W})$$

  - $\mathbf{W} \in \mathbb{R}^{d \times k}$ is weight to be learned.
  - $f(\mathbf{W})$ is loss function, e.g. $\|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2$ (Linear Regression).
  - $\Omega(\mathbf{W})$ is sparse-inducing penalty, e.g. $||\mathbf{W}||_1$ (L_1-norm).
  - λ is hyperparameter, controlling the level of sparsity in $\mathbf{W}$.

Recap: Ridge Regression

- Ridge Regression:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{||\mathbf{X}^T\mathbf{w}-\mathbf{y}||_2^2}_{\text{loss}} + \underbrace{\lambda||\mathbf{w}||_2^2}_{\text{penalty}}$$

  - Have better prediction error than linear regression in a variety of scenarios, depending on the choice of lambda.
  - But it will never sets coefficients to zero exactly, and therefore cannot perform variable selection in the linear model.
  - While this didn't seem to hurt its prediction ability, it is not desirable for the purposes of interpretation, especially if the number of variables p is large.

Sparse Coding based Feature Selection: LASSO

- LASSO:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{||\mathbf{X}^T\mathbf{w}-\mathbf{y}||_2^2}_{\text{loss}} + \underbrace{\lambda||\mathbf{w}||_1}_{\text{penalty}}$$

- The only difference between the lasso problem and ridge regression is the penalty term.
- But even though these problems look similar, their solutions behave very differently!
- The nature of L_1 penalty causes some coefficients to be shrunken to zero exactly.
- This is what makes the lasso substantially different from ridge regression: it is able to perform variable selection in the linear model.

Sparse Coding based Feature Selection: LASSO

- LASSO vs Ridge Regression

Unconstrained forms

$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{X}^T\mathbf{w}-\mathbf{y}||_2^2+\lambda||\mathbf{w}||_2^2$$

$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{X}^T\mathbf{w}-\mathbf{y}||_2^2+\lambda||\mathbf{w}||_1$$

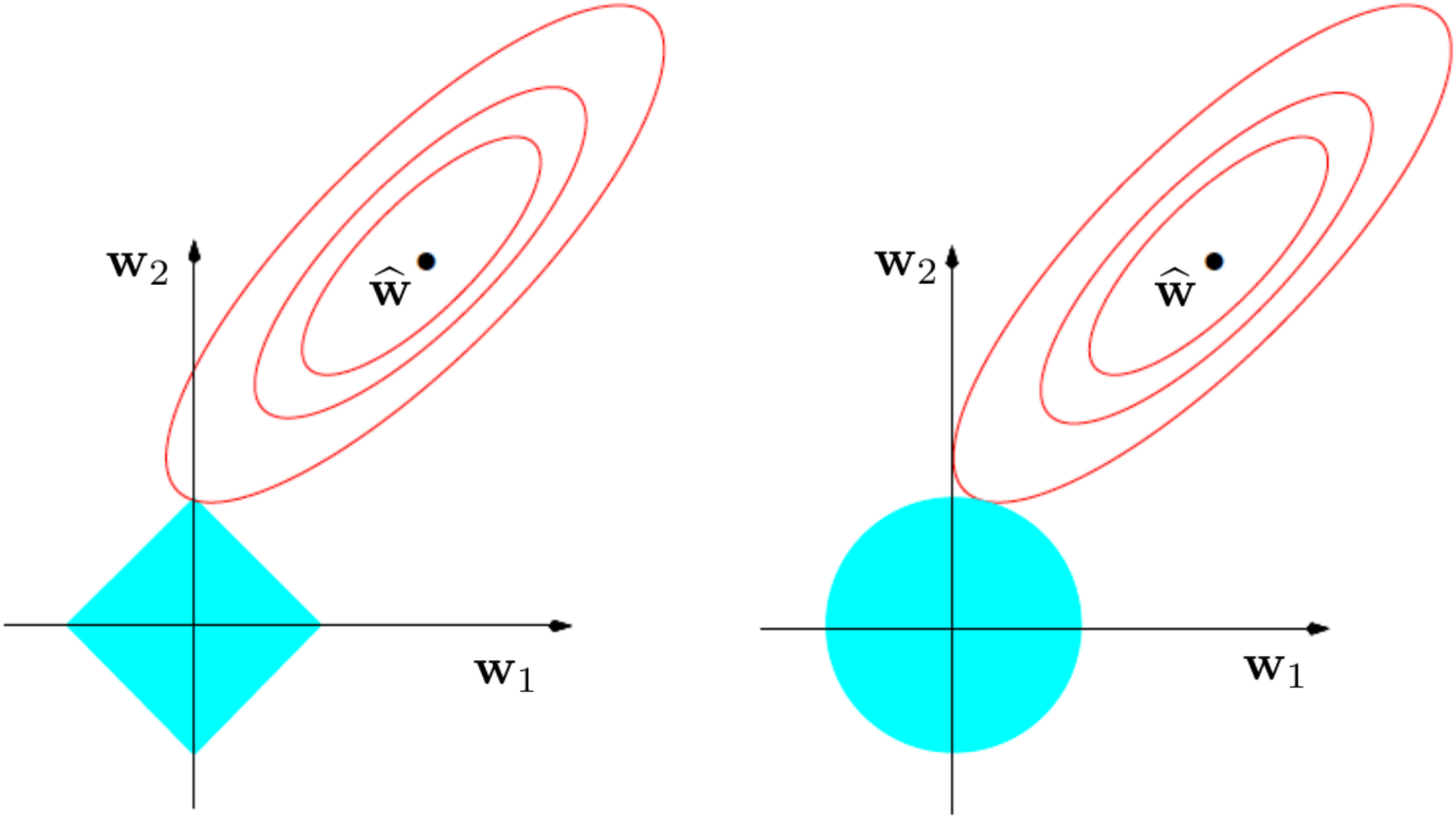
Constrained forms

$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{X}^T\mathbf{w}-\mathbf{y}||_2^2 \quad \text{s.t. } ||\mathbf{w}||_2^2 \leq t$$

$$\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{X}^T\mathbf{w}-\mathbf{y}||_2^2 \quad \text{s.t. } ||\mathbf{w}||_1 \leq t$$

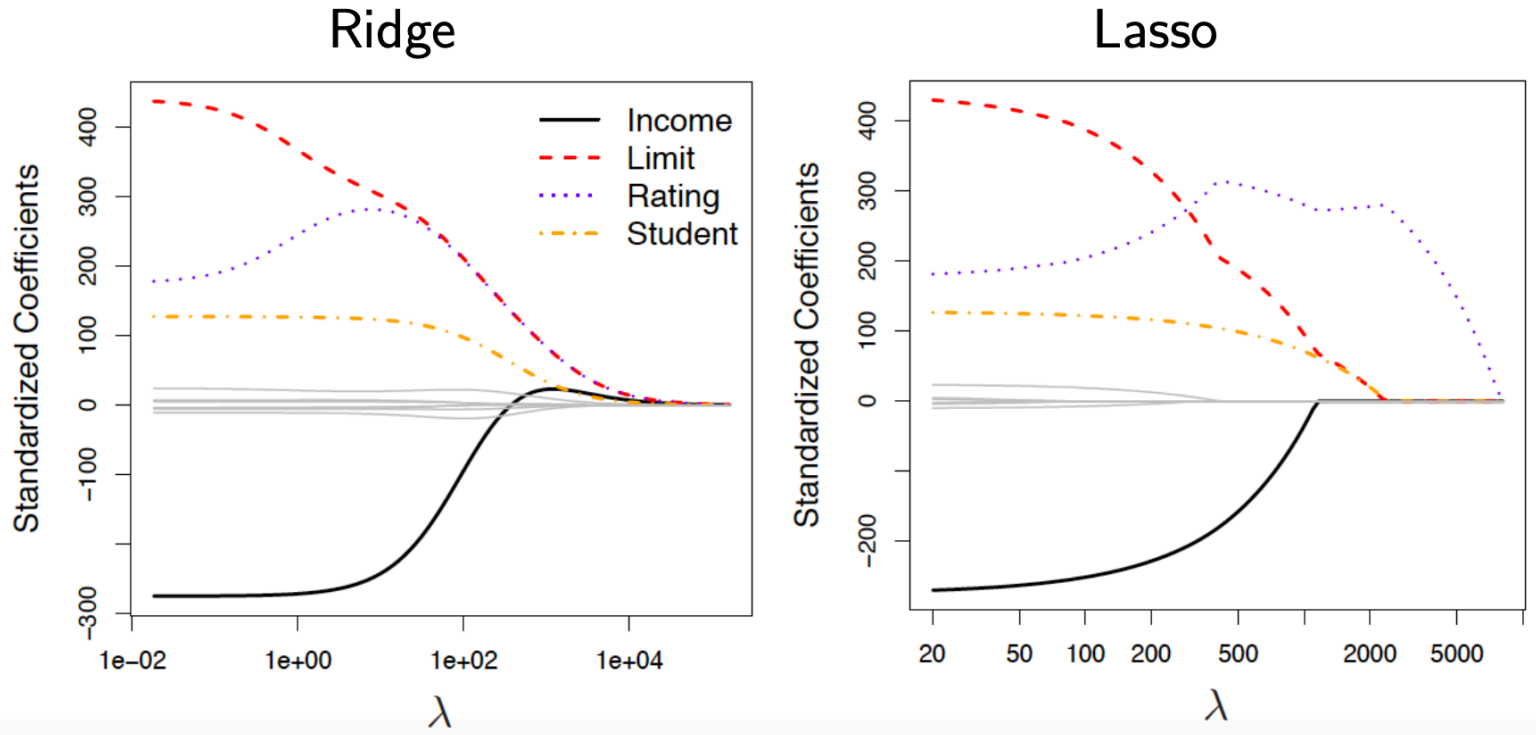
Sparse Coding based Feature Selection: LASSO

- LASSO vs Ridge Regression (Constrained forms in 2-dim space)



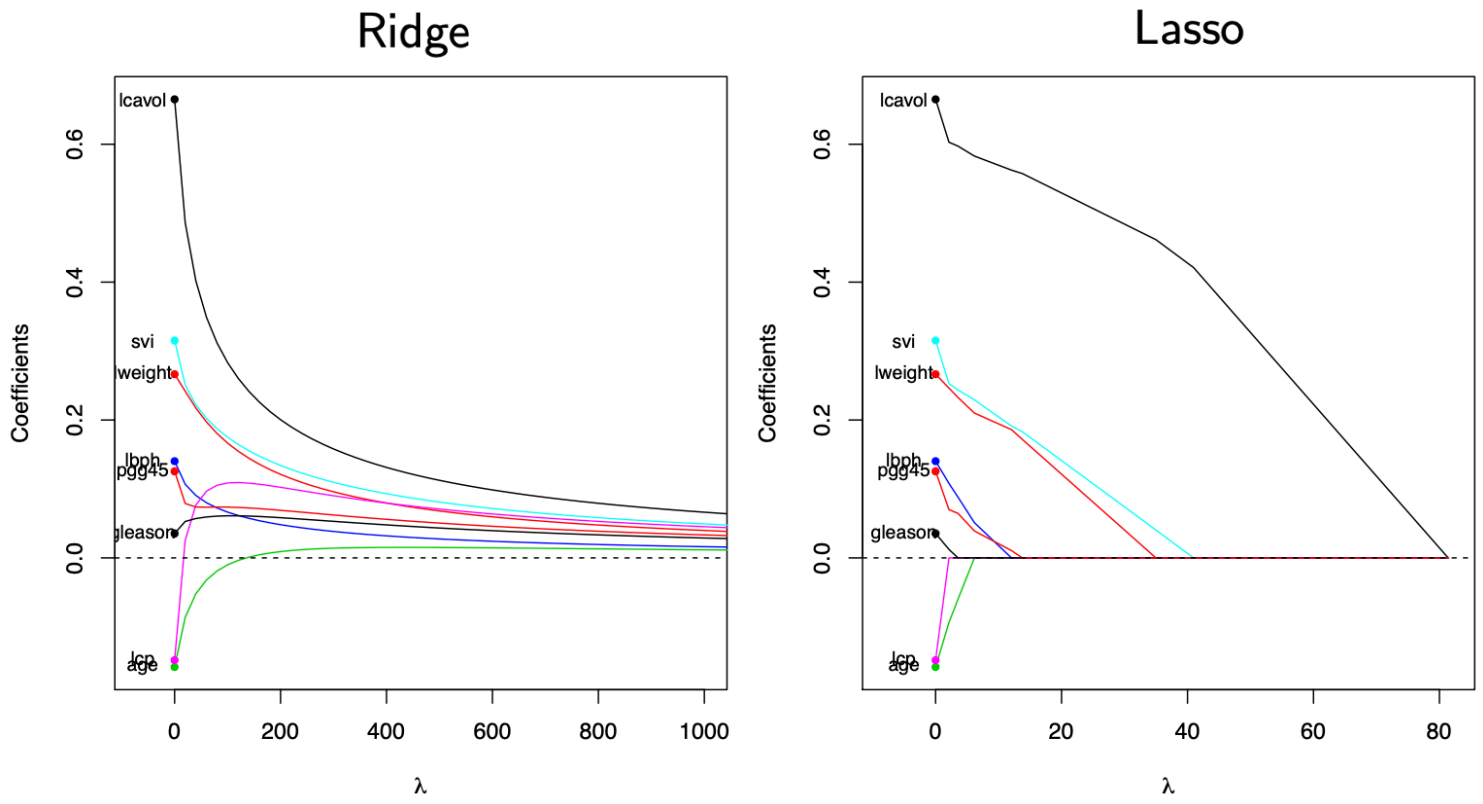
Sparse Coding based Feature Selection: LASSO

- Example: credit data
  - **Response** is average credit debt.
  - **Predictors** are income, limit (credit limit), rating (credit rating), student (indicator), and others.



Sparse Coding based Feature Selection: LASSO

- Example: prostate cancer data
  - **Response** is the level of prostate-specific antigen (PSA).
  - **Predictor** are 8 clinical predictors.



Multi-Task Feature Selection

- **Multi-Task Learning (MTL):**
  - A subfield of feature selection in machine learning.
  - Multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks.
  - This can result in improved learning efficiency and prediction accuracy for the task-specific models, as compared to training the models separately.
  - e.g. Multi-class classification, Multi-label classification

Multi-Task Feature Selection

- **Matrix norm:**

given d-by-k matrix $\mathbf{W} = (W_{ij}) \in \mathbb{R}^{d \times k}$

$L_{p,q}$-norm $\|\mathbf{W}\|_{p,q} = \left(\sum_{i=1}^{d} \left(\sum_{j=1}^{k} |W_{ij}|^p \right)^{q/p} \right)^{1/q}$

Frobenius norm $\|\mathbf{W}\|_F = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{k} W_{ij}^2}$

$L_{2,1}$-norm $\|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d} \left(\sum_{j=1}^{k} W_{ij}^2 \right)^{1/2}$

$L_{1,2}$-norm $\|\mathbf{W}\|_{1,2}^2 = \sum_{i=1}^{d} \left(\sum_{j=1}^{k} |W_{ij}| \right)^2$

Multi-Task Feature Selection

- **Matrix norm (L_{21}-norm):**

$$\mathbf{W} = \begin{bmatrix} W_{11} & \cdots & W_{1k} \\ \vdots & \ddots & \vdots \\ W_{d1} & \cdots & W_{dk} \end{bmatrix} = \begin{bmatrix} \mathbf{w}^1 \\ \vdots \\ \mathbf{w}^d \end{bmatrix}$$

- $L_{2,1}$-norm enforces L_2-norm on each row:

$$||\mathbf{w}^i||_2 = (W_{i1}^2 + \cdots + W_{ik}^2)^{1/2}$$

- The summation of all the rows is:

$$\begin{aligned} ||\mathbf{W}||_{2,1} &= ||\mathbf{w}^1||_2 + \cdots + ||\mathbf{w}^d||_2 \\ &= (W_{11}^2 + \cdots + W_{1k}^2)^{1/2} + \cdots + (W_{d1}^2 + \cdots + W_{dk}^2)^{1/2} \\ &= \sum_{i=1}^{d} \left(\sum_{j=1}^{k} W_{ij}^2 \right)^{1/2} \end{aligned}$$

Multi-Task Feature Selection

- **Matrix norm (L_{12}-norm):**

$$\mathbf{W} = \begin{bmatrix} W_{11} & \cdots & W_{1k} \\ \vdots & \ddots & \vdots \\ W_{d1} & \cdots & W_{dk} \end{bmatrix} = \begin{bmatrix} \mathbf{w}^1 \\ \vdots \\ \mathbf{w}^d \end{bmatrix}$$

- $L_{1,2}$-norm enforces squared L_1-norm on each row:

$$||\mathbf{w}^i||_1^2 = (|W_{i1}| + \cdots + |W_{ik}|)^2$$

- The summation of all the rows is:

$$\begin{aligned} ||\mathbf{W}||_{1,2}^2 &= ||\mathbf{w}^1||_1^2 + \cdots + ||\mathbf{w}^d||_1^2 \\ &= (|W_{11}| + \cdots + |W_{1k}|)^2 + \cdots + (|W_{d1}| + \cdots + |W_{dk}|)^2) \\ &= \sum_{i=1}^{d} \left(\sum_{j=1}^{k} |W_{ij}| \right)^2 \end{aligned}$$

Multi-Task Feature Selection

- **L21-norm:**

$$\min_{\mathbf{W}} \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{W}\|_{2,1}$$

$$\text{where } \|\mathbf{W}\|_{2,1} = \sum_{i=1}^{d} \|\mathbf{w}^i\|_2 = \sum_{i=1}^{d}\left(\sum_{j=1}^{k} W_{ij}^2\right)^{1/2}$$

  - **Joint feature selection from multiple tasks.**
  - **Encourages multiple predictors to share similar sparsity patterns.**
  - **Selecting class-shared features.**
    - Shrink all the elements in a row to zeros.
    - One nonzero row (a feature) is selected for all the k classes
    - Eliminate irrelevant features, i.e. a row becomes to zero vector.

Multi-Task Feature Selection

- **Exclusive Lasso (L12-norm):**

$$\min_{\mathbf{W}} \|\mathbf{X}^T\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda\|\mathbf{W}\|_{1,2}^2$$

$$\text{where } \|\mathbf{W}\|_{1,2}^2 = \sum_{i=1}^{d} \|\mathbf{w}^i\|_1^2 = \sum_{i=1}^{d} \left(\sum_{j=1}^{k} |W_{ij}| \right)^2$$

  - **Exclusive feature selection from multiple tasks.**
  - **Multiple predictors DON'T share similar sparsity pattern.**
  - **Selecting discriminative features.**
    - Shrink most of the elements in a row to zeros.
    - But at least one element in each row will be nonzero.
    - Nonzero elements in a row: a feature is selected for certain classes.

Multi-Task Feature Selection

- **An illustration: L21-norm vs Exclusive Lasso**

Synthetic data and labels:

$$\mathbf{X}^T = \begin{bmatrix} 0.463 & 0.319 & -0.100 & 0.526 & 0.535 & 0.329 & 0.475 \\ 0.296 & 0.192 & 0.058 & -0.076 & 0.152 & 0.313 & -0.114 \\ 0.196 & 0.189 & 0.167 & -0.280 & 0.267 & -0.246 & 0.164 \\ 0.330 & 0.357 & 0.027 & -0.001 & 0.118 & 0.058 & 0.191 \\ 0.332 & 0.035 & -0.002 & 0.280 & 0.111 & -0.043 & 0.104 \\ -0.022 & -0.026 & 0.770 & 0.189 & 0.196 & -0.146 & -0.121 \\ -0.217 & 0.028 & 0.404 & 0.359 & 0.335 & -0.282 & -0.235 \\ 0.396 & 0.297 & 0.260 & 0.241 & 0.193 & 0.038 & 0.101 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Multi-Task Feature Selection

- An illustration: L21-norm vs Exclusive Lasso

L21-norm:

$$\mathbf{W}_{21} = \begin{bmatrix} 0.764 & 0.587 & 0.378 \\ 0.097 & 0.033 & 0.082 \\ 0.054 & 0.531 & 1.003 \\ \mathbf{-0.000} & \mathbf{0.000} & \mathbf{0.000} \\ 0.151 & 0.030 & 0.126 \\ \mathbf{0.000} & \mathbf{0.000} & \mathbf{-0.000} \\ \mathbf{0.000} & \mathbf{-0.000} & \mathbf{0.000} \end{bmatrix}$$

selecting all the elements in each non-zero row

Multi-Task Feature Selection

- **An illustration: L21-norm vs Exclusive Lasso**

L21-norm:

	class-1	class-2	class-3
feature-1	✓	✓	✓
feature-2	✓	✓	✓
feature-3	✓	✓	✓
feature-4	✗	✗	✗
feature-5	✓	✓	✓
feature-6	✗	✗	✗
feature-7	✗	✗	✗

Multi-Task Feature Selection

- An illustration: L21-norm vs Exclusive Lasso

Exclusive LASSO:

$$\mathbf{W}_{12} = \begin{bmatrix} 0.336 & 0.352 & \mathbf{0.000} \\ 0.287 & \mathbf{0.000} & 0.358 \\ \mathbf{0.000} & 0.070 & 0.758 \\ -0.009 & 0.173 & \mathbf{0.000} \\ 0.326 & \mathbf{0.000} & 0.298 \\ \mathbf{0.000} & \mathbf{0.000} & -0.344 \\ 0.333 & \mathbf{-0.000} & \mathbf{0.000} \end{bmatrix}$$

selecting non-zero elements in each row

Multi-Task Feature Selection

- **An illustration: L21-norm vs Exclusive Lasso**

Exclusive LASSO:

	class-1	class-2	class-3
feature-1	✓	✓	✗
feature-2	✓	✗	✓
feature-3	✗	✓	✓
feature-4	✓	✓	✗
feature-5	✓	✗	✓
feature-6	✗	✗	✓
feature-7	✗	✓	✓