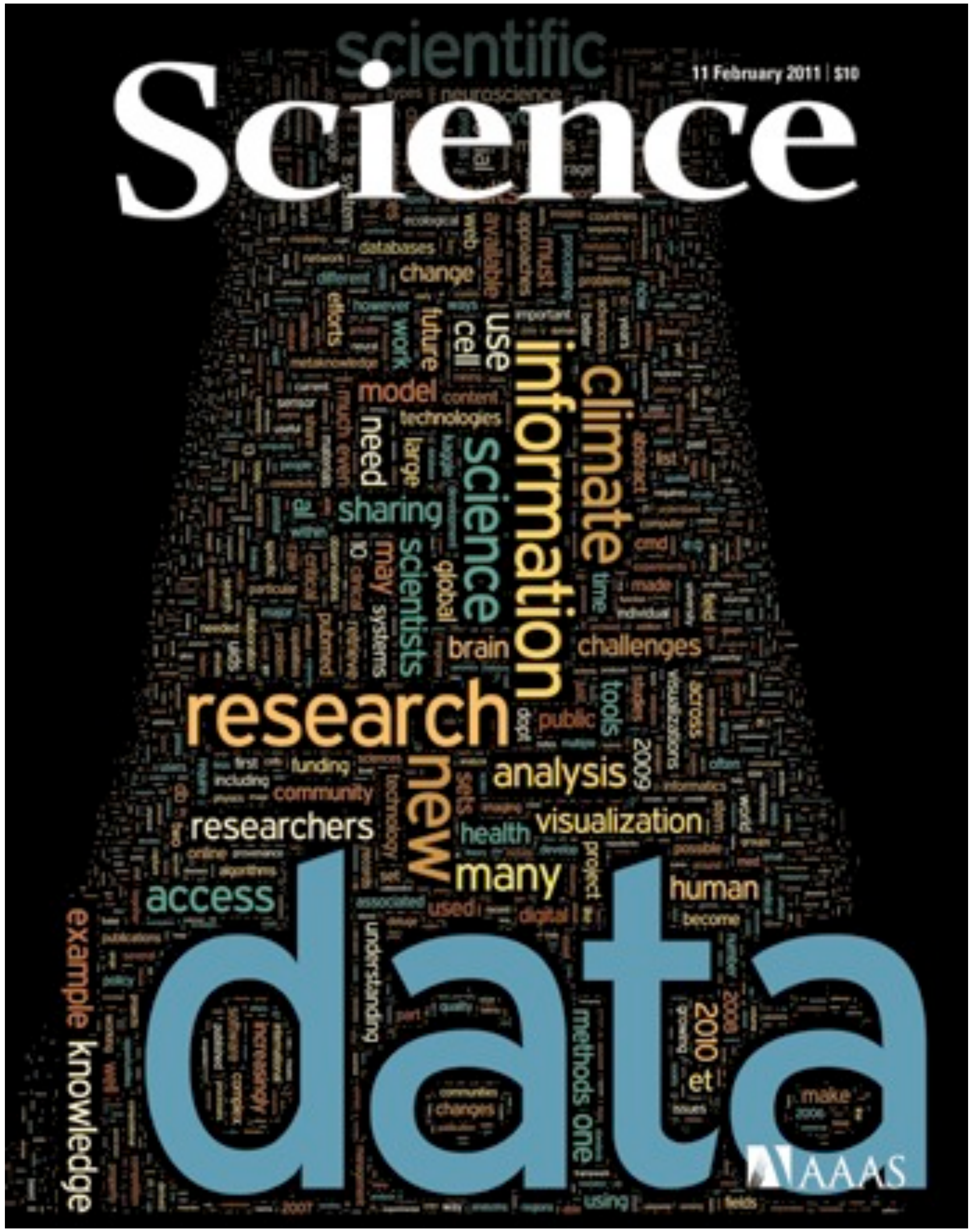
Robustly finding the needles in a haystack of high-dimensional data

Eric Chi

Department of Statistics,
Rice University

July 21, 2011

The Haystack of high-dimensional data



The Haystack of high-dimensional data

PERSPECTIVE

More Is Less: Signal Processing and the Data Deluge

Richard G. Baraniuk

The data deluge is changing the operating environment of many sensing systems from data-poor to data-rich—so data-rich that we are in jeopardy of being overwhelmed. Managing and exploiting the data deluge require a reinvention of sensor system design and signal processing theory. The potential pay-offs are huge, as the resulting sensor systems will enable radically new information technologies and powerful new tools for scientific discovery.

A lot of sensor data...

DARPA Autonomous Real-Time Ground Ubiquitous Surveillance Imaging System

- 1.8 gigapixels
- 160 km^2 (Greater LA)
- 30-cm ground resolution
- Video at 15 frames/sec = 770 gigabits per second

“Data, data everywhere, but not a thought to think”

Q: Are all measurements equally informative?

A: Probably not.

"Data, data everywhere, but not a thought to think"

Q: Are all measurements equally informative?

A: Probably not.

The key notion: Pareto Principle or 80/20 Rule

- 80% of an effect comes from 20% of the possible causes.
  - Garden: 80% of the peas came from 20% of the pea pods
  - Econ: 80% of the land in Italy was owned by 20% of the population
  - Business: 80% of your $$$ come from 20% of your clients

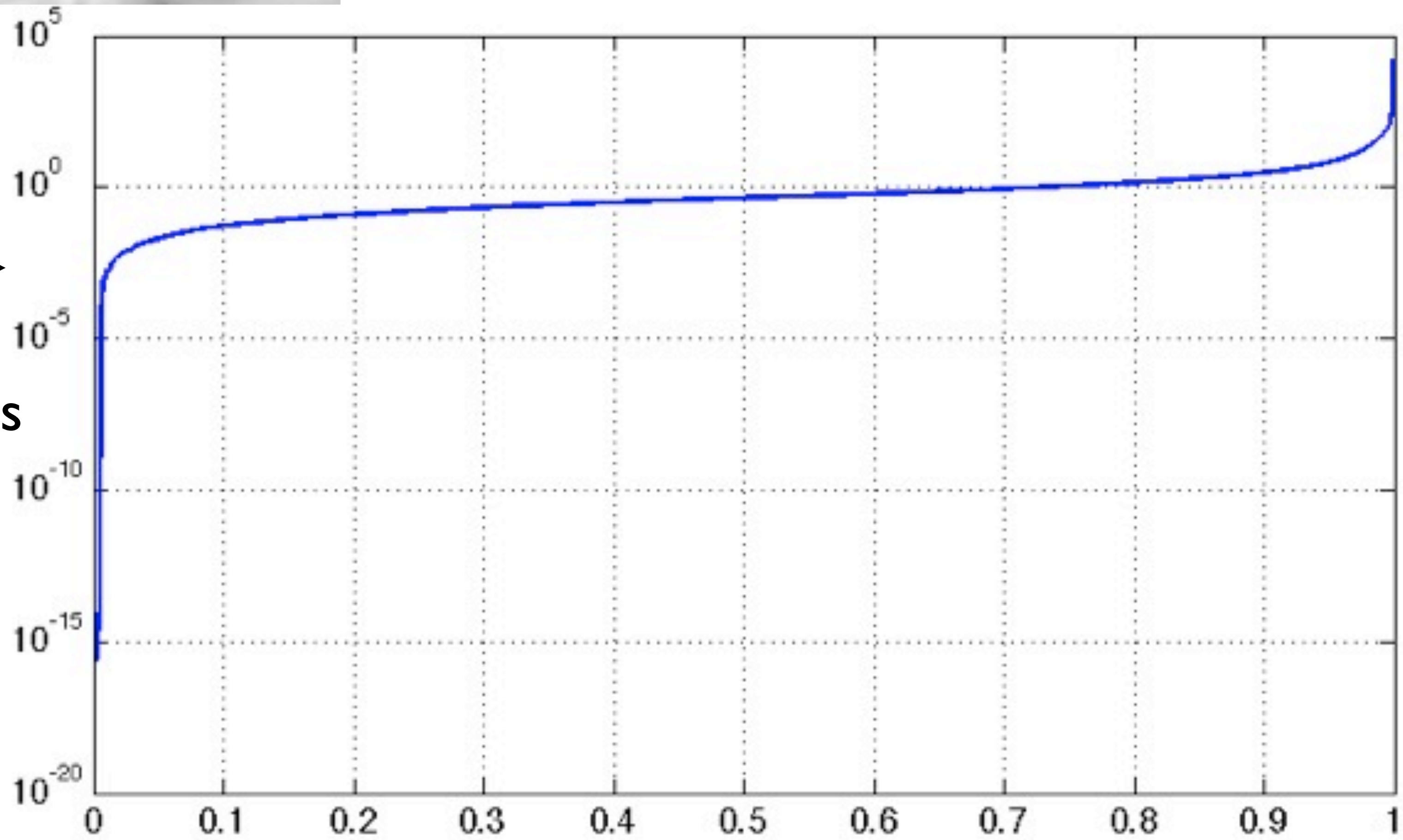
Look at data through the lens of sparsity

Majority of systematic variation in data is due to a minority of possible sources

Sparsity and your digital camera

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$$

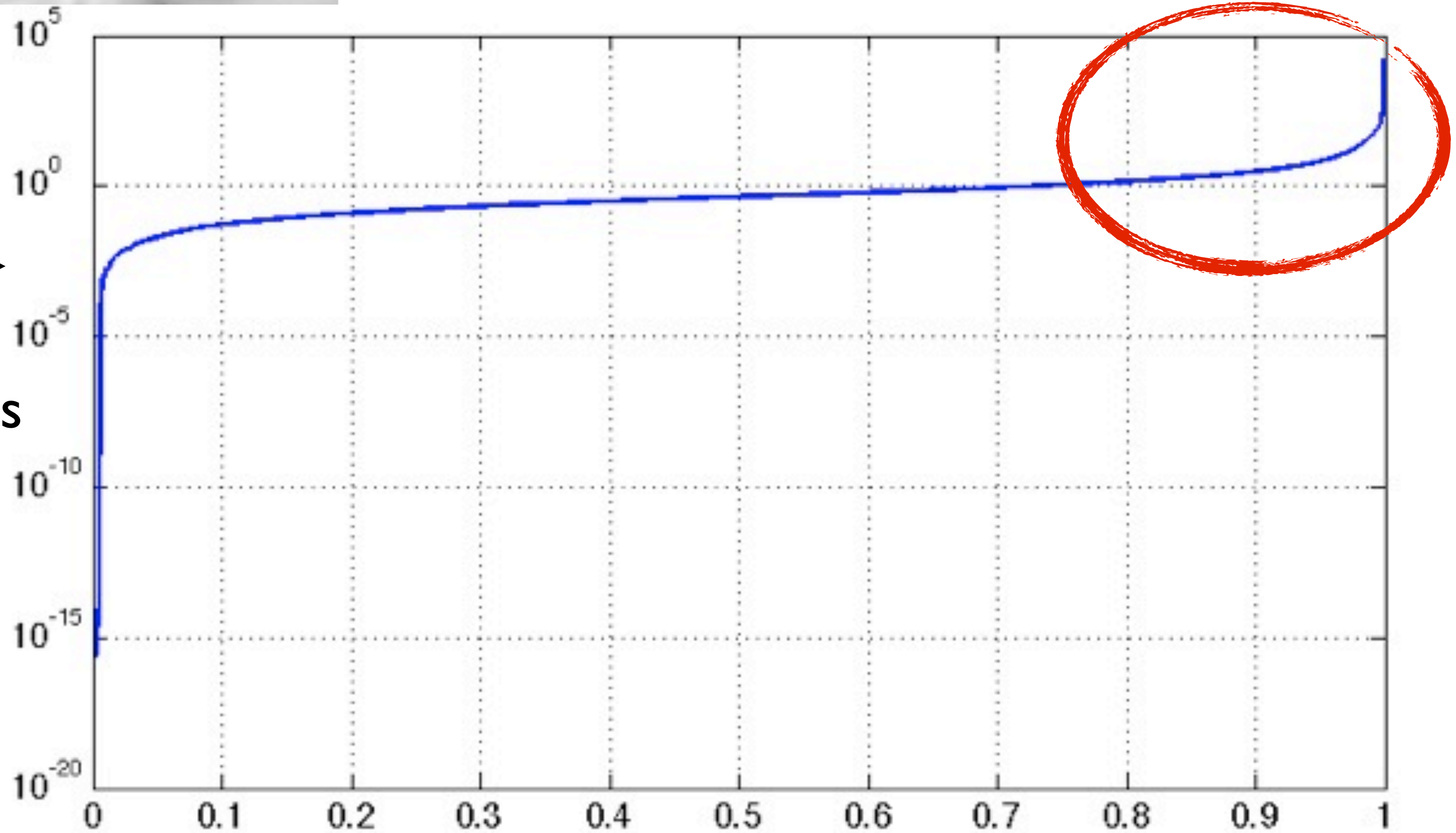
Sorted wavelet coefficient magnitudes



Sparsity and your digital camera

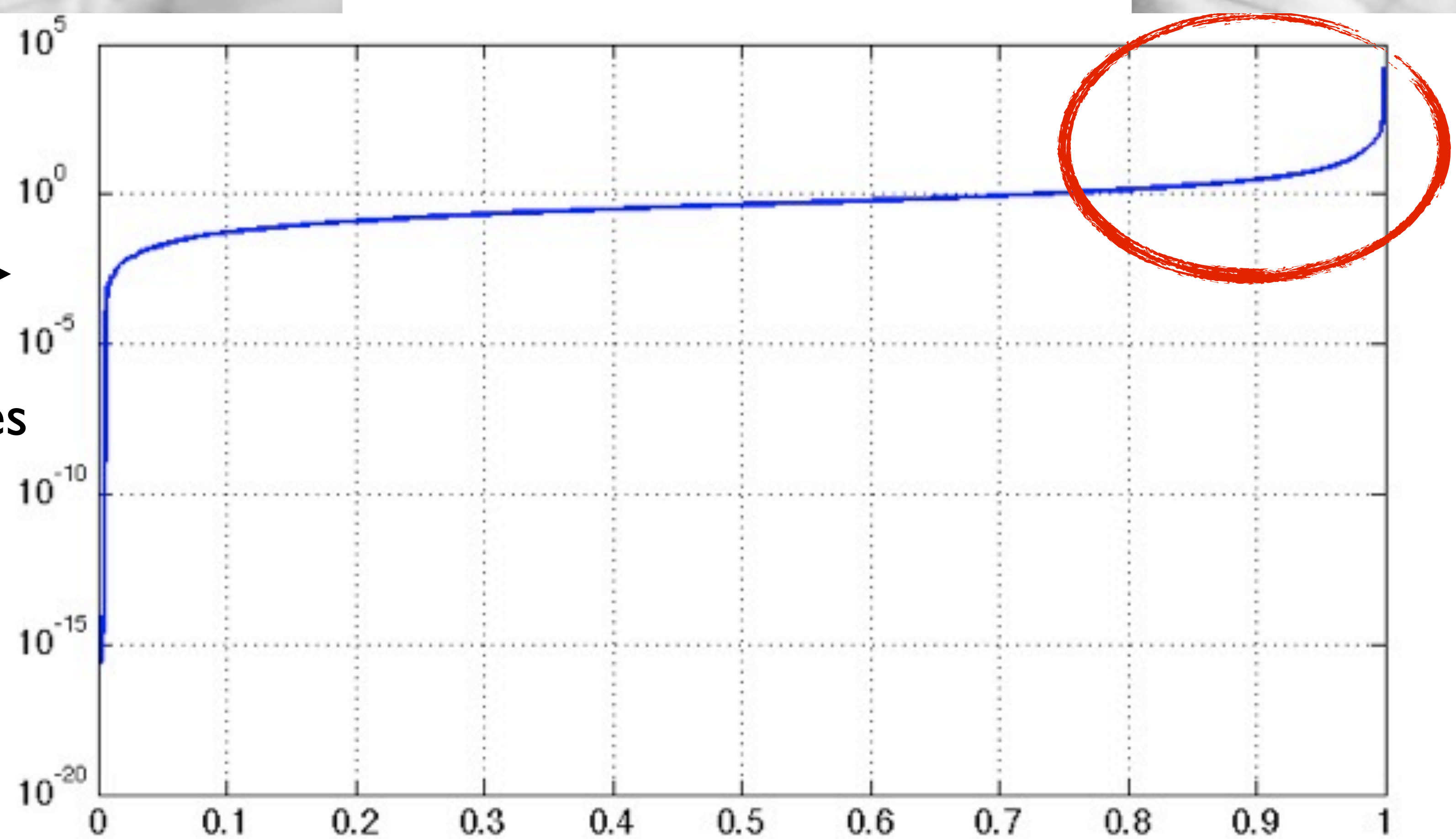
$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$$

Sorted wavelet coefficient magnitudes



Sparsity and your digital camera

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$$



Sorted wavelet coefficient magnitudes

Keep largest 20%
and
invert transform

Sparsity and genomic data

A question about infectious diseases

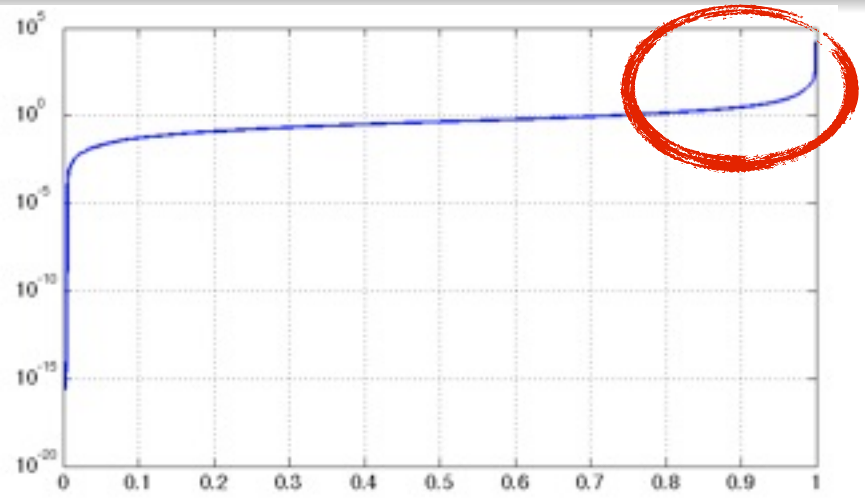
Why do most people have innate immunity to leprosy? NEJM Dec 31, 2009

Which genes explain most of the systematic variation?

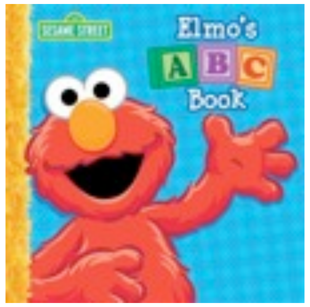
Predict or explain $\mathbf{y} \in \{0, 1\}^n$ using $\mathbf{X} \in \mathbb{R}^{n \times p}$; $n << p$.

- SNP: $n = 1000$s, $p = 100,000$s

Summary



- **Ss** is for **S**parsity.
  - Haystack = all possible sources of variation.
  - Needle = minority of sources (**s**parse set of variables) that explain majority of systematic variation.



Review: Penalized Regression

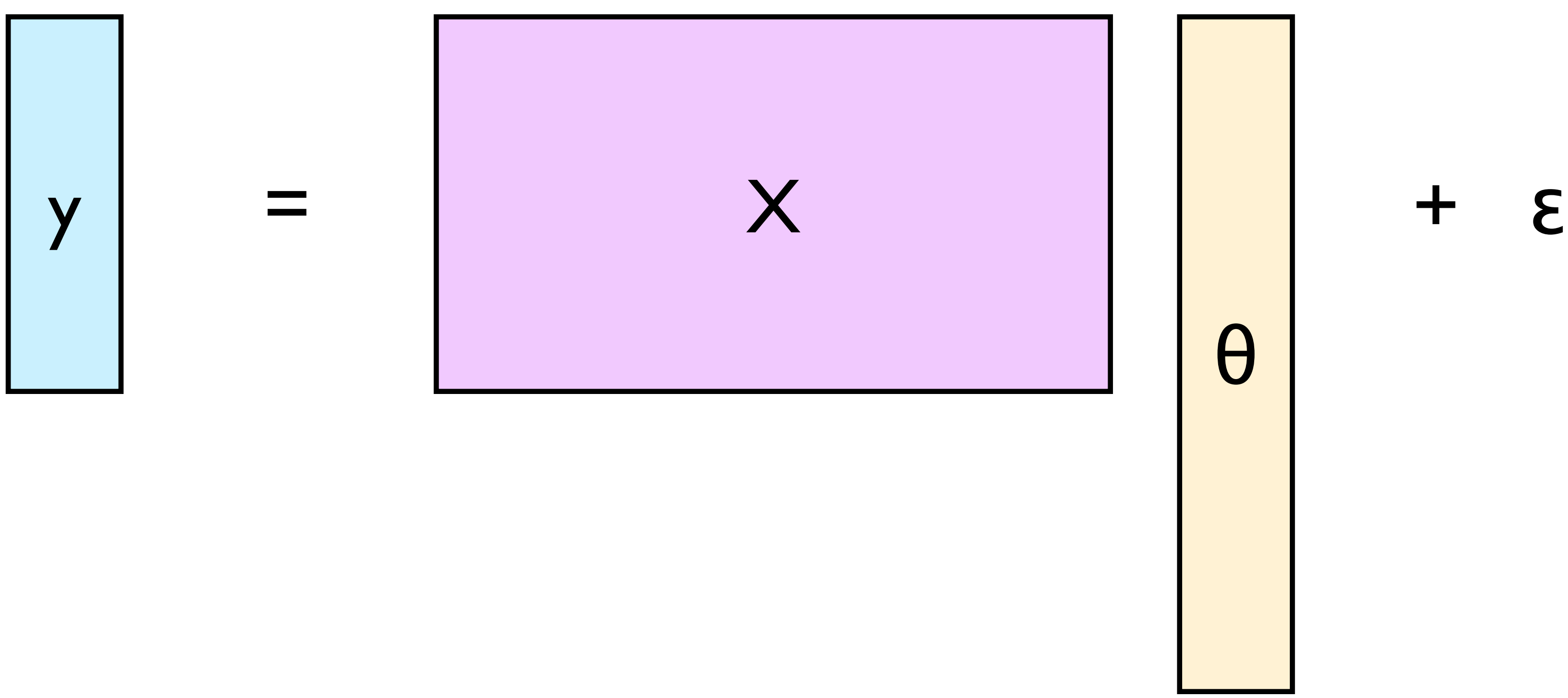
$$\hat{\theta} = \underset{\boldsymbol{\theta}}{\arg\min}\ \underbrace{L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta})}_{\text{Lack of fit}} + \underbrace{\lambda J(\boldsymbol{\theta})}_{\text{Complexity}}$$

Review: Penalized Regression

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \underbrace{L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta})}_{\text{Lack of fit}} + \underbrace{\lambda J(\boldsymbol{\theta})}_{\text{Complexity}}$$

Least Squares Regression

$$L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

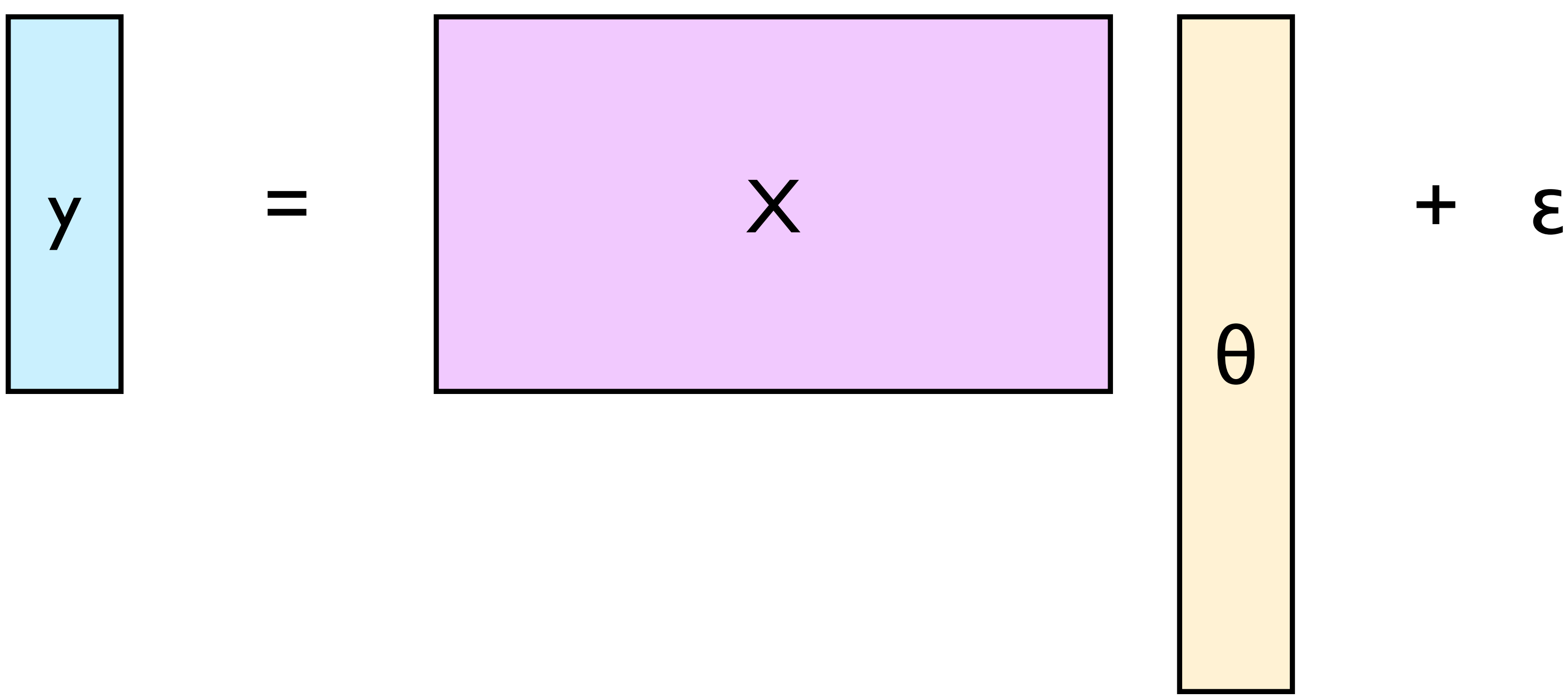


Review: Penalized Regression

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \underbrace{L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta})}_{\text{Lack of fit}} + \underbrace{\lambda J(\boldsymbol{\theta})}_{\text{Complexity}}$$

Least Squares Regression + Ridge/Tikhonov Penalization

$$L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$



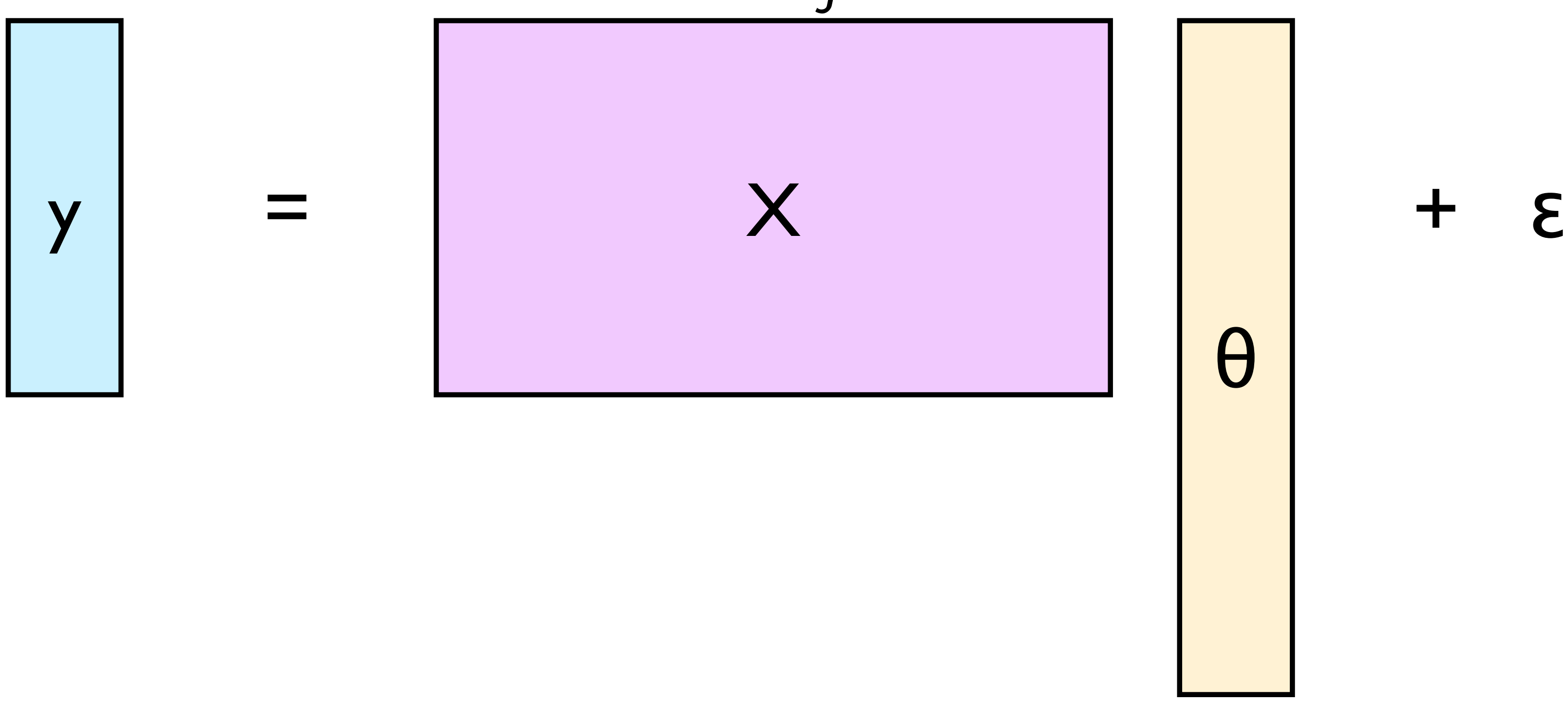
Review: Penalized Regression

$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \underbrace{L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta})}_{\text{Lack of fit}} + \underbrace{\lambda J(\boldsymbol{\theta})}_{\text{Complexity}}$$

Least Squares Regression + ℓ_1-**Penalization**[1].

$$L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$$

$$\|\boldsymbol{\theta}\|_1 = \sum_j |\theta_j|$$



[1] Tibshirani 1996, Chen, Donoho 1995

Global optimality

Objective	Solution
$\frac{1}{2}\lVert \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \rVert_2^2$	$\theta_j^* = \mathbf{x}_j^\mathsf{T}\mathbf{r}^{(j)}$
$\frac{1}{2}\lVert \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \rVert_2^2 + \frac{\lambda}{2}\lVert \boldsymbol{\theta} \rVert_2^2$	$\theta_j^* = \mathbf{x}_j^\mathsf{T}\mathbf{r}^{(j)}\,(1+\lambda)^{-1}$
$\frac{1}{2}\lVert \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \rVert_2^2 + \lambda\lVert \boldsymbol{\theta} \rVert_1$	$\theta_j^* = S(\mathbf{x}_j^\mathsf{T}\mathbf{r}^{(j)}, \lambda)$

*j*th partial residual

$$r_i^{(j)} = y_i - \sum_{l \neq j} x_{il}\theta_l^*.$$

Residual variation in $\mathbf{y}$ unexplained after adjusting for the effect of all other predictors, $l \neq j$.

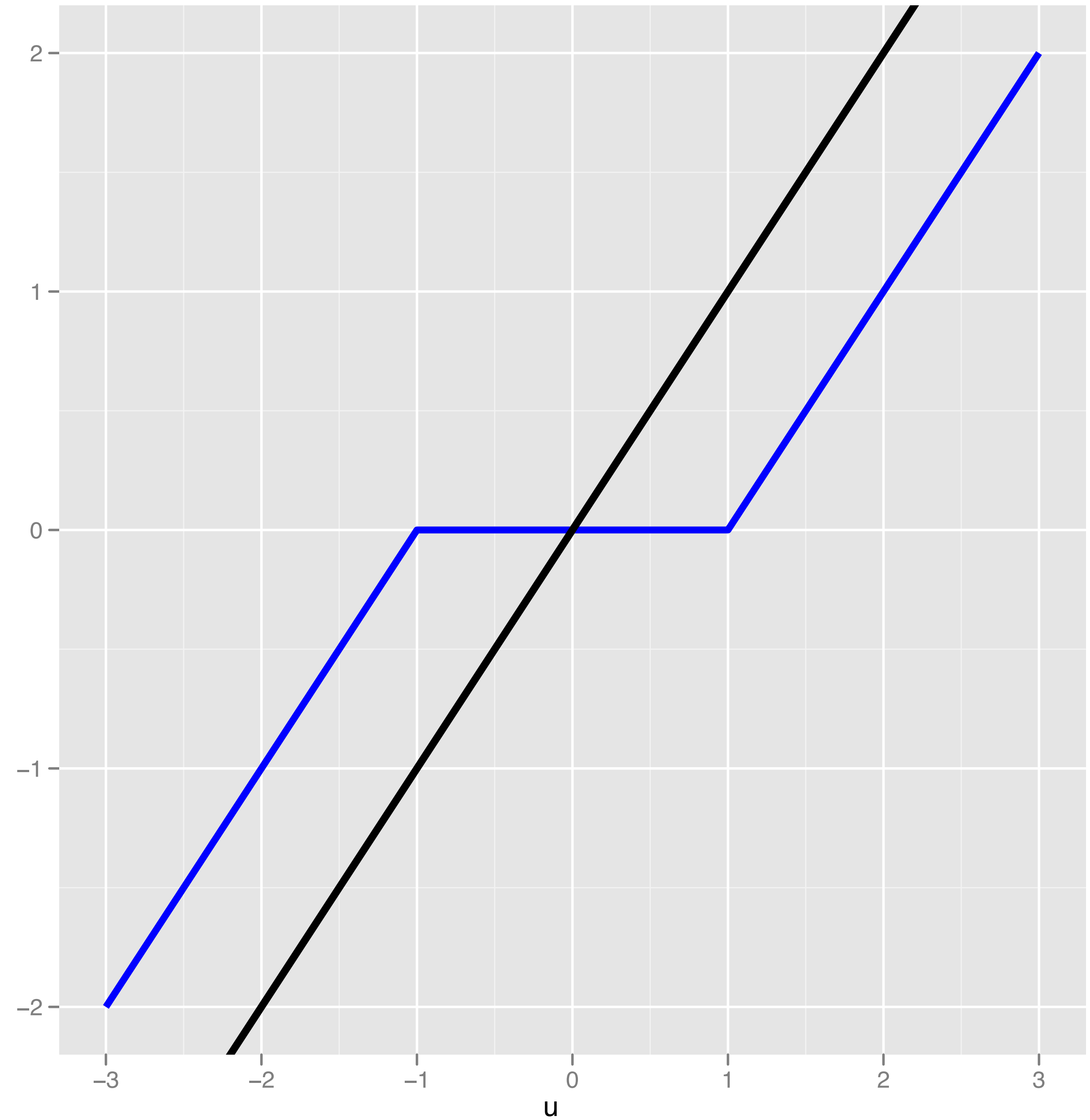
The inner product

$\mathbf{x}_j^\mathsf{T}\mathbf{r}^{(j)}$ = correlation between the *j*th predictor and the *j*th partial residual.

ℓ_1-Penalization and Soft Thresholding

Soft Thresholding

$$S(u,\lambda) = \begin{cases} u-\lambda & u>\lambda \\ u+\lambda & u<-\lambda \\ 0 & |u|\le\lambda \end{cases}.$$



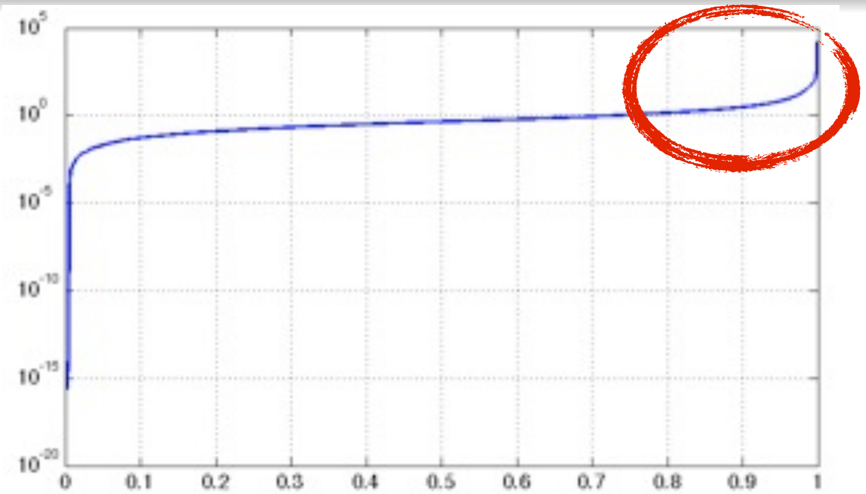
Recall the optimization

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

$$\theta_j^* = S(\mathbf{x}_j^{\mathsf{T}}\mathbf{r}^{(j)},\lambda)$$

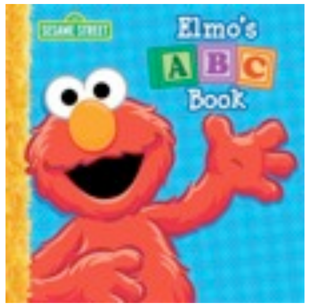
N.B. Solutions are biased towards zero!

Summary

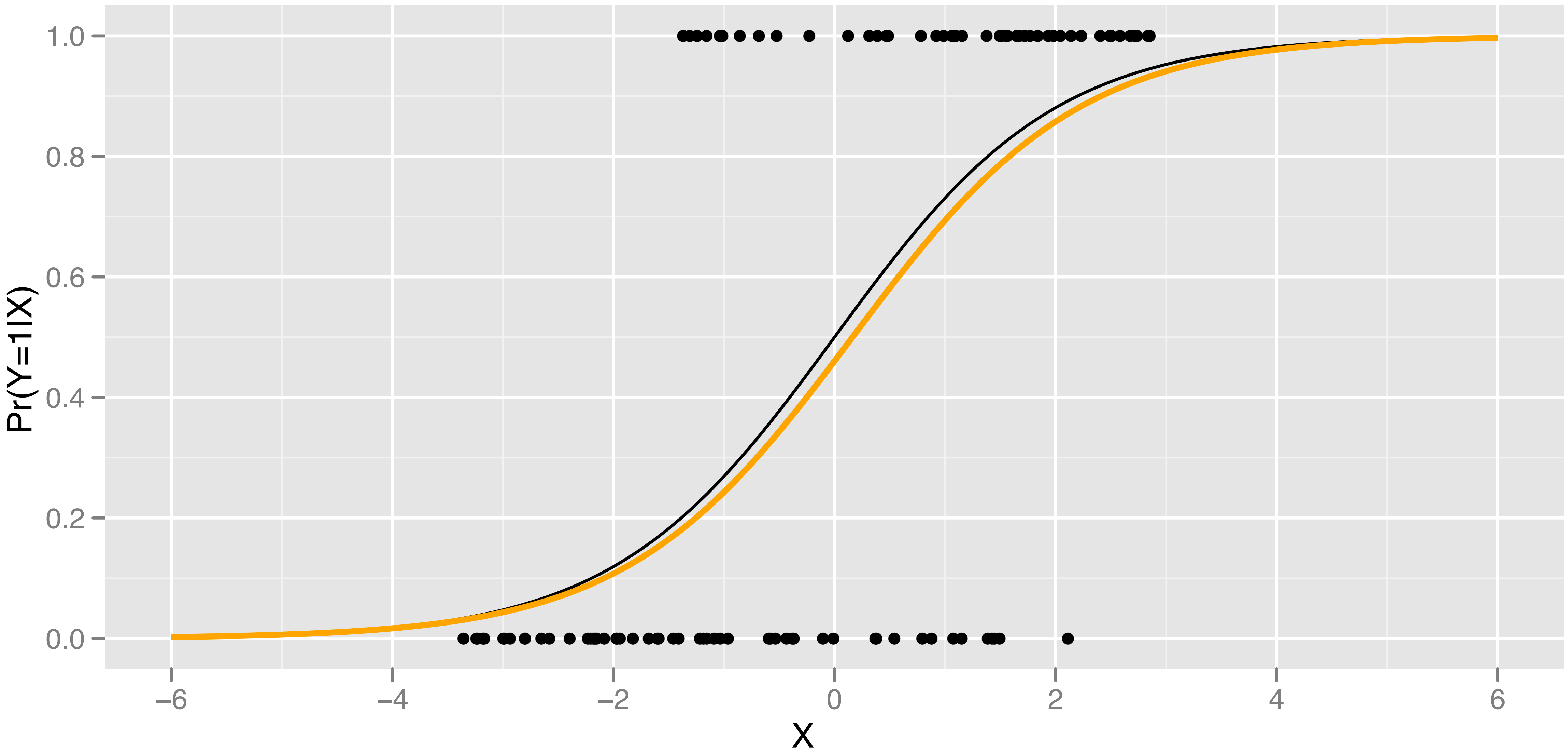


- **Ss** is for **S**parsity.
  - Haystack = all possible sources of variation.
  - Needle = minority of sources (**s**parse set of variables) that explain majority of systematic variation.
- **Ee** is for "**e**ll-one"-penalized regression.

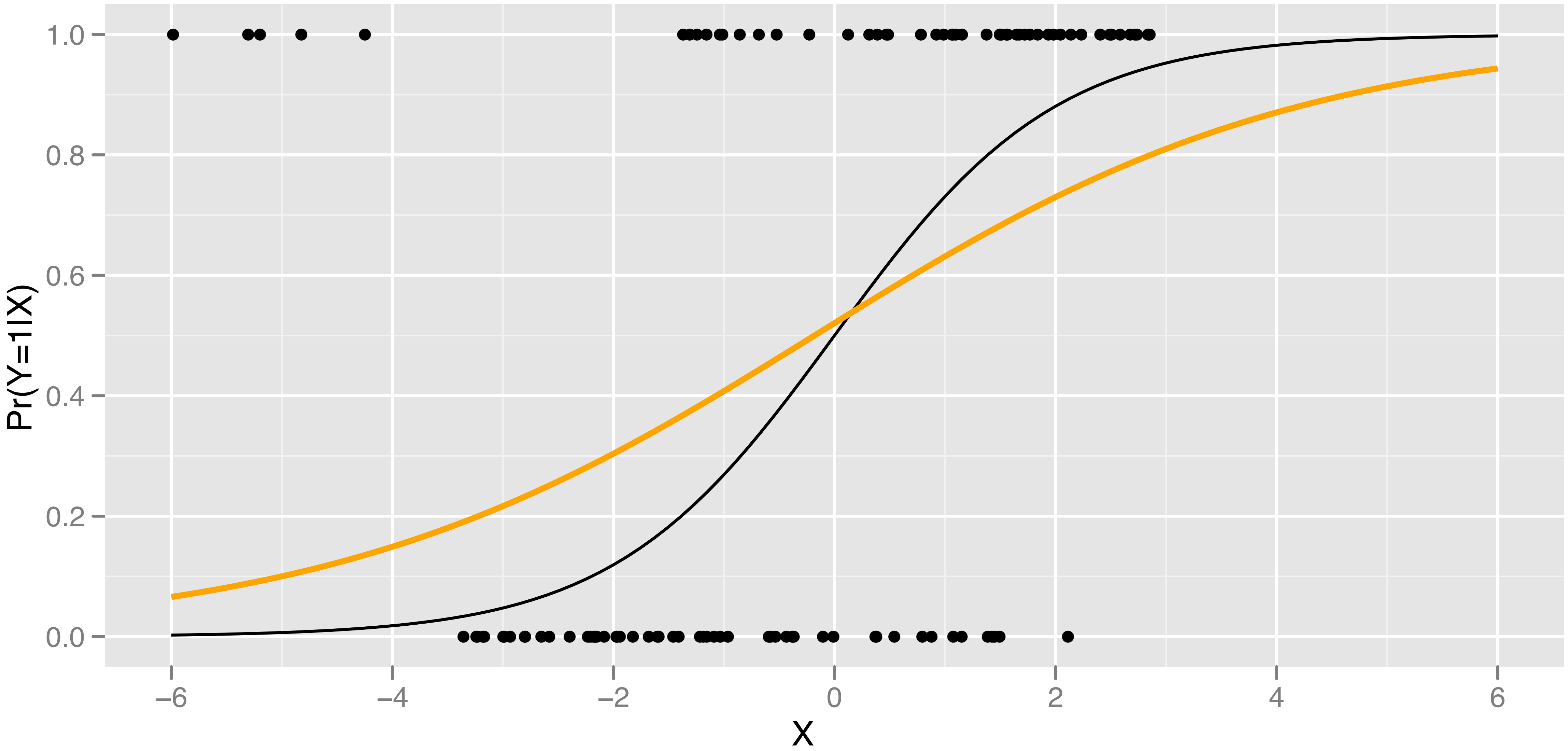
$$\min_{\boldsymbol{\theta}} L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1$$



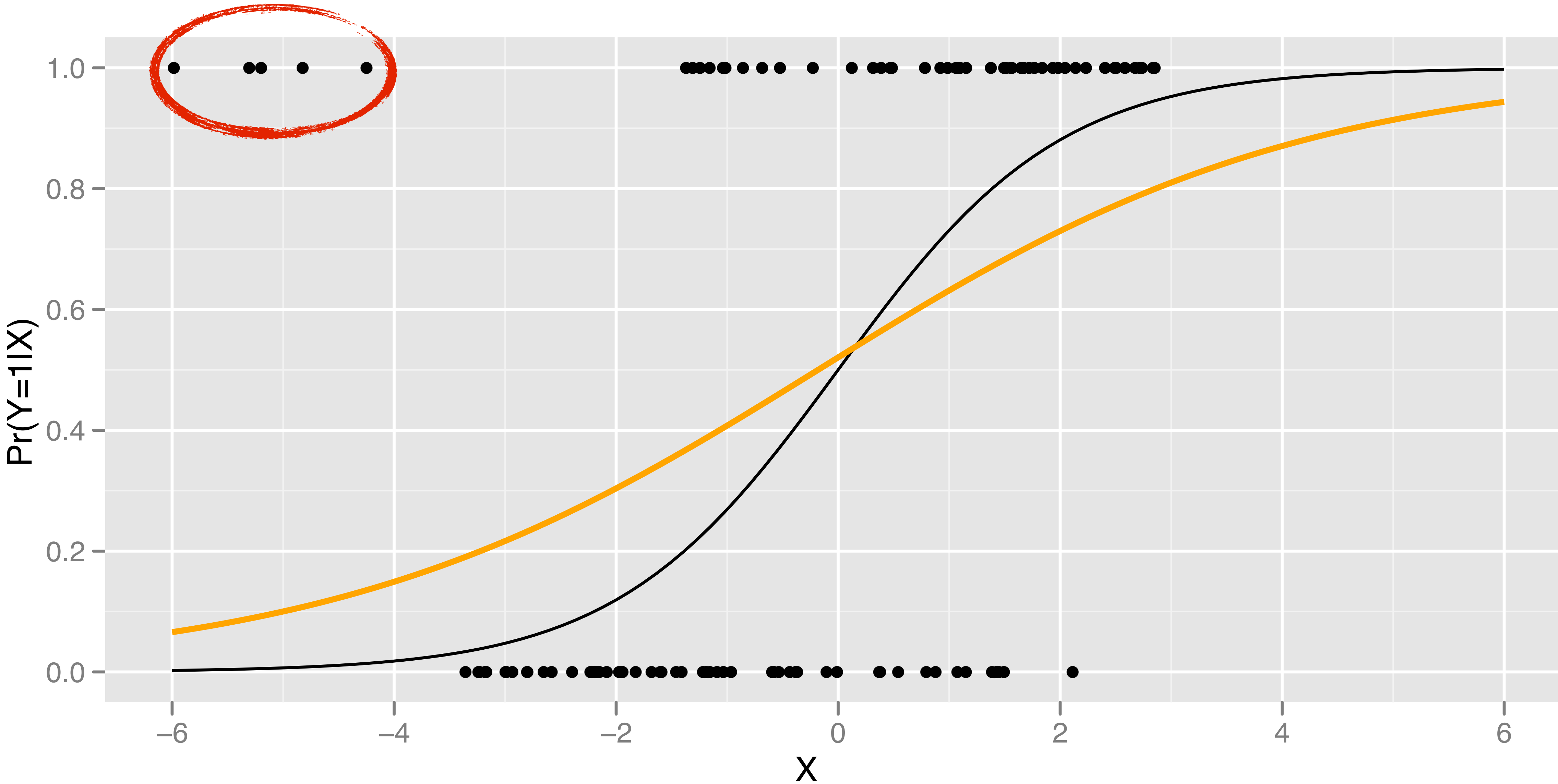
A simple case of logistic regression



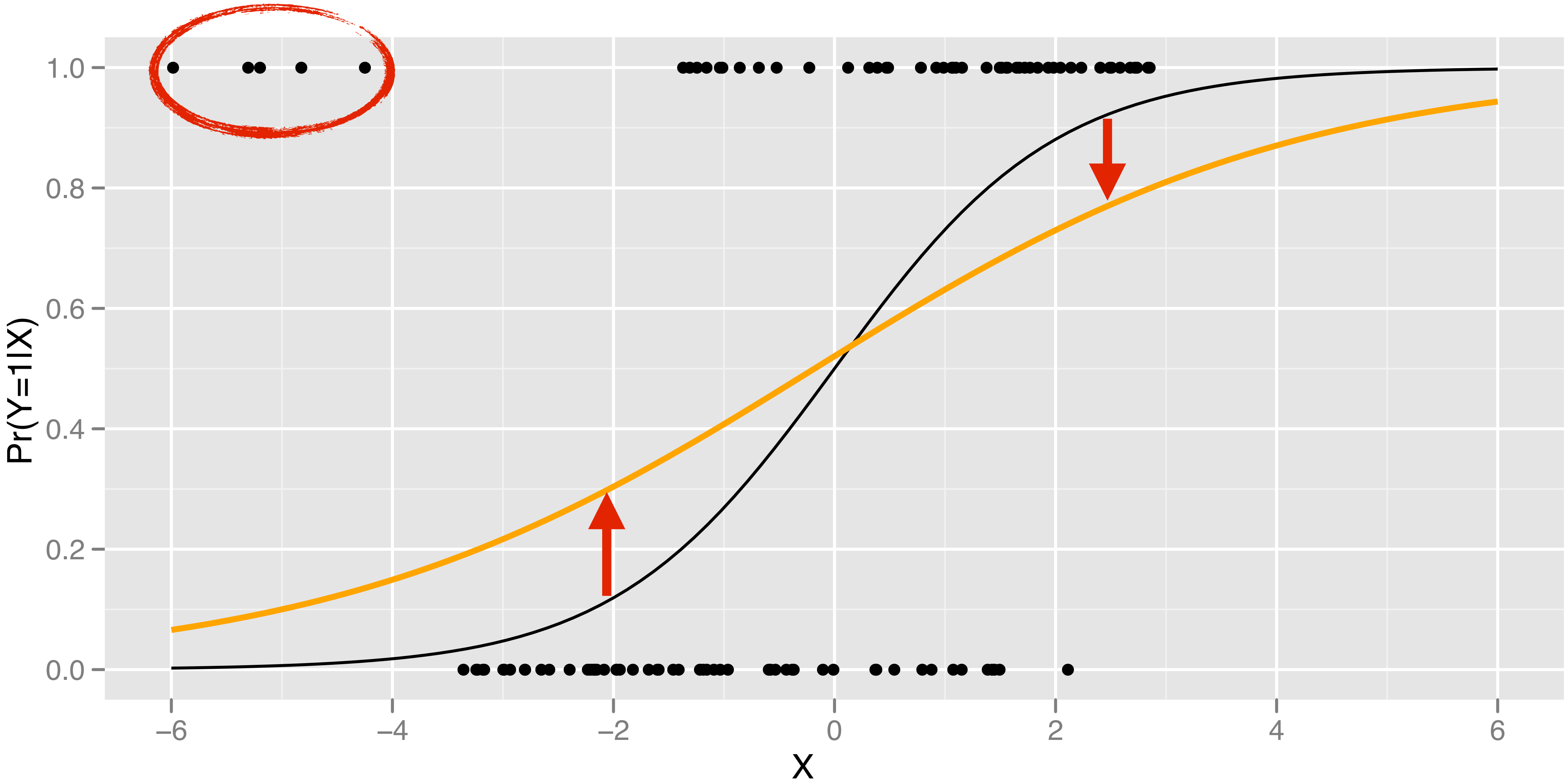
One of these things is not like the others...



One of these things is not like the others...

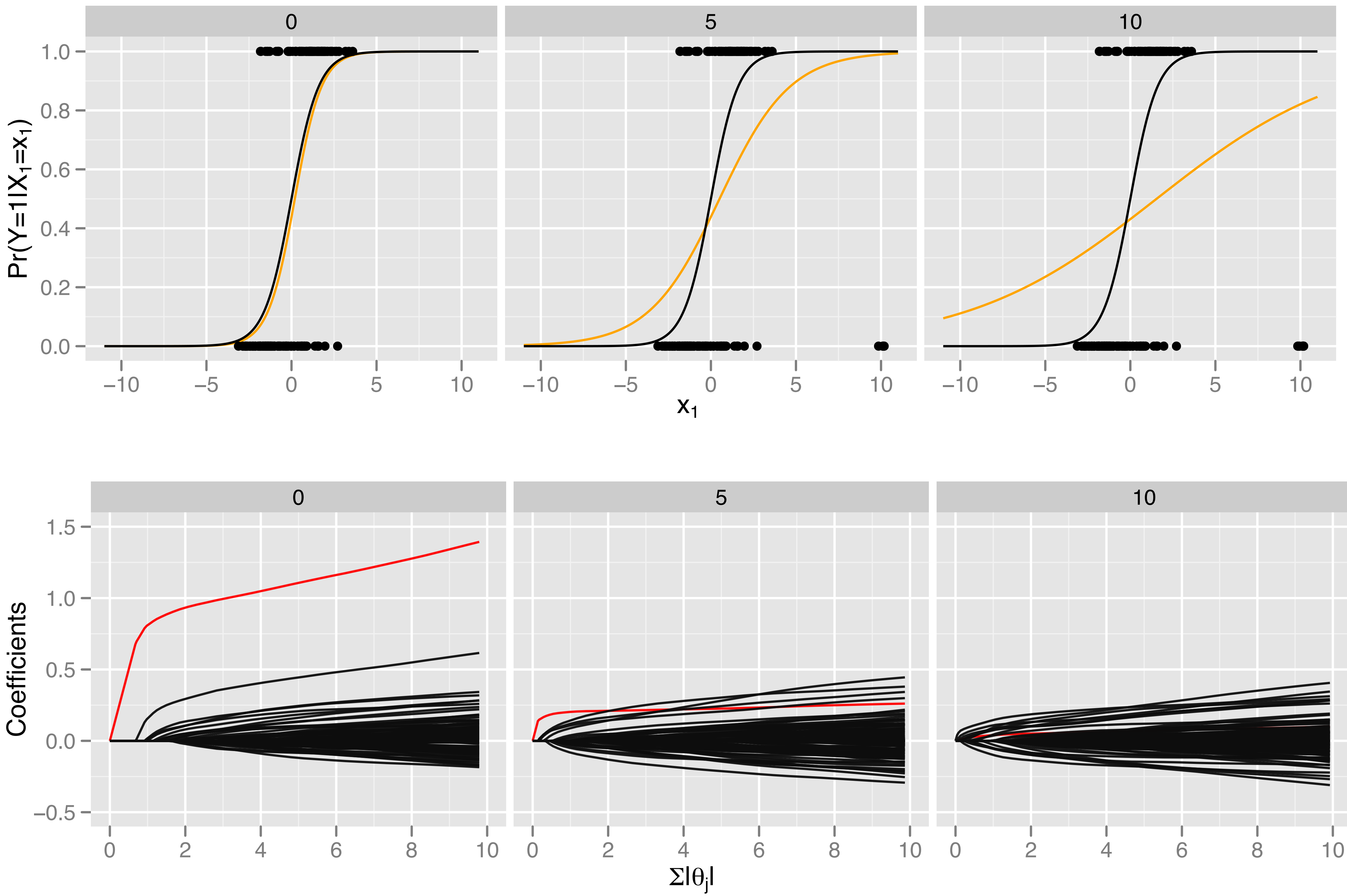


Implosion Breakdown



$$\|\hat{\theta}\|_2 \to 0$$

Outliers + ℓ_1 shrinkage = Unfortunate series of events



Use a different loss function!

β-divergence[2]

- A family of distortion measures.

$$D_\beta(g||f_{\boldsymbol{\theta}}) = \int f_{\boldsymbol{\theta}}^{1+\beta}(z) - \left(1 + \frac{1}{\beta}\right) g(z) f_{\boldsymbol{\theta}}^{\beta}(z) + \frac{1}{\beta} g^{1+\beta}(z) dz.$$

- β trades off robustness for efficiency of the resulting estimator.

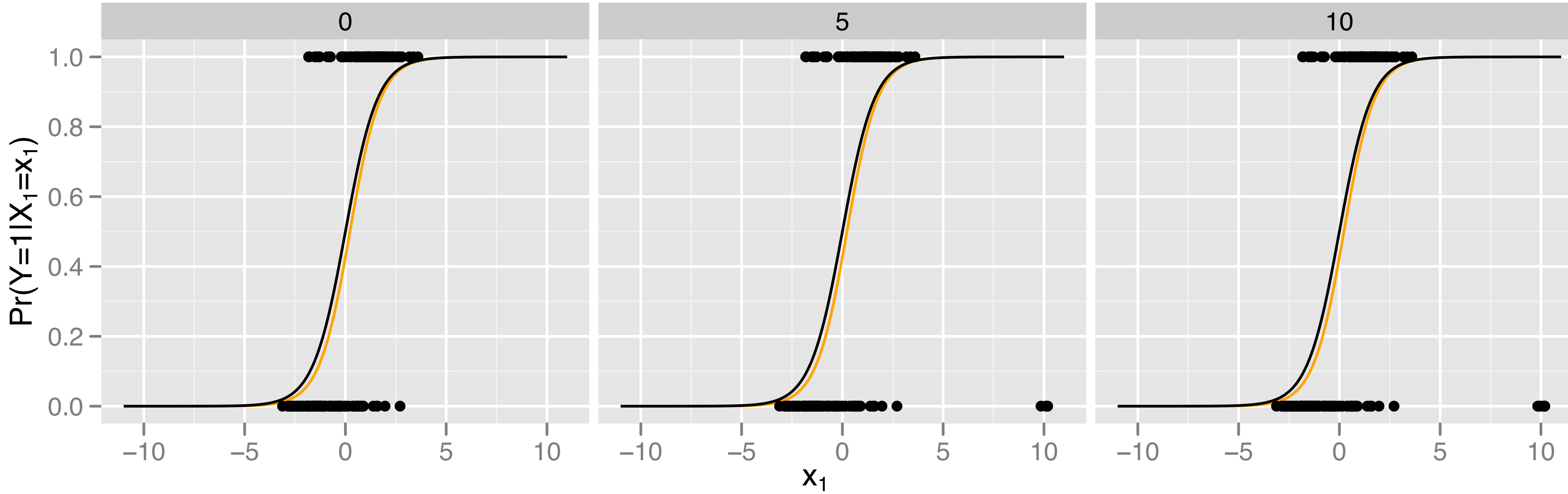
$$\hat{\theta} = \arg\min_{\boldsymbol{\theta}} \hat{D}_\beta(\mathbf{y}||f_{\boldsymbol{\theta}})$$

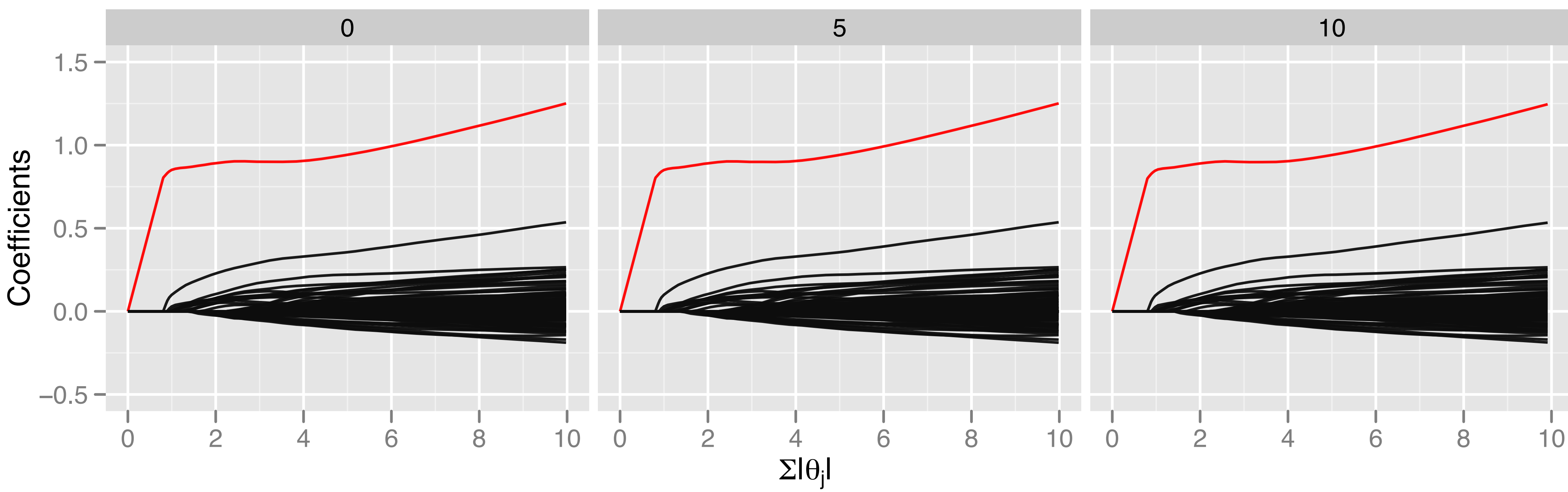
- Optimality conditions

max likelihood	min β-div
$\sum_{i=1}^n u_{\boldsymbol{\theta}}(y_i) = \mathbf{0}$	$\sum_{i=1}^n u_{\boldsymbol{\theta}}(y_i) f_{\boldsymbol{\theta}}^{\beta}(y_i) = \mathbf{0}$

[2] Basu et al. 1998

Rescue by min β-div



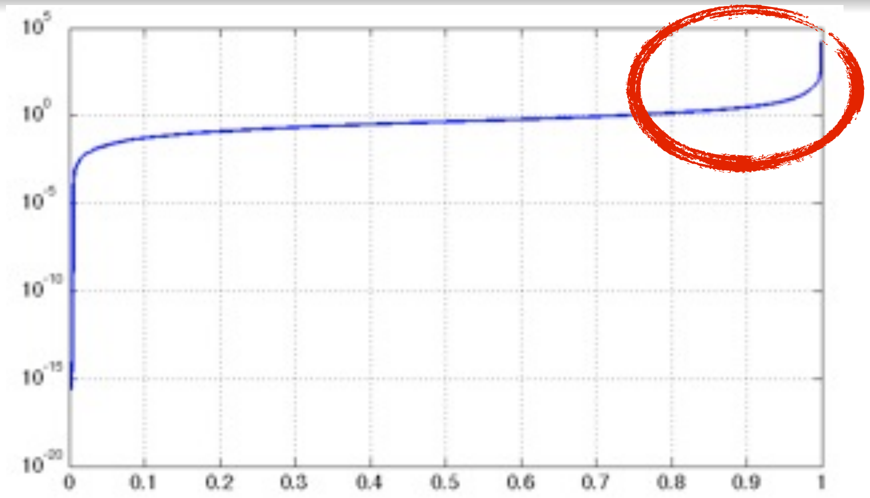


The blood, sweat, and tears

Computation

- The optimization problem is *not* convex.
- Solve the problem as a series of convex approximations (Majorizations/Auxiliary functions).
  - Convex + ℓ_1-penalties well studied.
- Guarantees on convergence to stationary points.
- Some heuristics on choosing starting points.

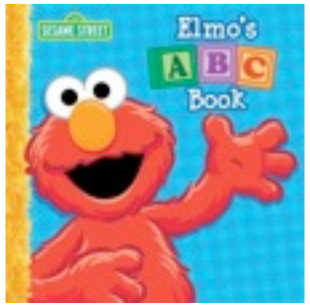
Summary



- **Ss** is for **S**parsity.
  - Haystack = all possible sources of variation.
  - Needle = minority of sources that explain majority of systematic variation.
- **Ee** is for "**e**ll-one"-penalized regression.

$$\min_{\boldsymbol{\theta}} L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta}\|_1$$

- **Bb** is for **B**ias.
  - ℓ_1-penalization **b**ias + implosion breakdown = missed detections.
  - Fight **b**ias with a robust loss function.



$$\min_{\boldsymbol{\theta}} \hat{D}(\mathbf{y} || f_{\boldsymbol{\theta}}) + \lambda \|\boldsymbol{\theta}\|_1$$

Acknowledgments

- Rice University
  - David Scott
  - Dennis Cox
  - Yin Zhang
  - Hadley Wickham
- LBNL/Berkeley/Sandia
  - Paul Spellman
  - Elizabeth Purdom
  - Tammy Kolda
  - David Gleich
- DOE CSGF & Krell Institute
- The letters **Ss**, **Ee**, and **Bb**