Sparse regression

Optimization-Based Data Analysis

http://www.cims.nyu.edu/~cfgranda/pages/OBDA_spring16

Carlos Fernandez-Granda

3/28/2016

Regression

Aim: Predict the value of a response / dependent variable $y \in \mathbb{R}$ from p predictors / features / independent variables $X_1, X_2, \ldots, X_p \in \mathbb{R}$

Methodology:

1. Fit a model with using n training examples $y_1, y_2, \ldots, y_n$

$$y_i \approx f\left(X_{i1}, X_{i2}, \ldots, X_{ip}\right) \quad 1 \leq i \leq n$$

2. Use learned model f to predict from new data

Linear regression

f is parametrized by an intercept β_0 and a vector of weights $\beta \in \mathbb{R}^p$

$$y_i \approx \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \quad 1 \leq i \leq n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \approx \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ & & \cdots & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{bmatrix}$$

$$y \approx 1\beta_0 + X\beta$$

Aim: Learn β_0 and β from the training data

Linear models

Signal representation (compression, denoising)

$$x = D\ c$$

- **Aim:** Find sparse representation of the signal x
- Columns of D are designed/learned atoms

Linear models

Inverse problems (compressed sensing, super-resolution)

$$y = A\,x$$

- **Aim:** Estimate the signal x from the measurements y
- A models the measurement process

Linear models

Linear regression

$$y = X\,\beta$$

- **Aim:** Learn the relation between the response y and the predictors $X_1, X_2, \ldots, X_p$ to predict from new data
- Each column of X contains a variable that may or may not be related to the response

Preprocessing

1. Center each predictor column X_i subtracting its mean so that

$$\sum_{i=1}^{n} X_{ij} = 0 \quad \text{for all } 1 \leq j \leq p$$

2. Normalize each predictor column X_j so that

$$||X_j||_2 = 1 \quad \text{for all } 1 \leq j \leq p$$

3. Center the response vector y so that

$$\sum_{i=1}^{n} y_i = 0$$

Least-squares fit

Minimize ℓ_2 norm of error over training data

$$\text{minimize} \quad \left|\left| y - X\tilde{\beta} \right|\right|_2$$

After preprocessing $\beta_0 = 0$, so we can ignore it

Geometric interpretation

Decompose y into its projection onto the column space of X and onto the orthogonal complement

$$y = y_X + y_{X^\perp}$$

By Pythagoras's Theorem

$$\left|\left|y - X\tilde{\beta}\right|\right|_2^2 = ||y_{X^\perp}||_2^2 + \left|\left|y_X - X\tilde{\beta}\right|\right|_2^2$$

If $n \geq p$ and X is full rank

$$\beta_{\text{ls}} = \left(X^T X\right)^{-1} X^T y \qquad \text{satisfies} \qquad X\beta_{\text{ls}} = y_X$$

After preprocessing $\beta_0 = 0$

Probabilistic interpretation: Maximum likelihood

Assumption: Linear model is correct but examples are corrupted by additive iid Gaussian noise

$$y = X\beta + z$$

The likelihood is the probability density function of y parametrized by β

$$\mathcal{L}\left(\tilde{\beta}\right) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\left|\left|y - X\tilde{\beta}\right|\right|_2^2\right)$$

The maximum-likelihood estimate of β is

$$\begin{aligned}
\beta_{\mathrm{ML}} &= \arg\max_{\tilde{\beta}} \mathcal{L}\left(\tilde{\beta}\right) \\
&= \arg\max_{\tilde{\beta}} \log \mathcal{L}\left(\tilde{\beta}\right) \\
&= \arg\min_{\tilde{\beta}} \left|\left|y - X\tilde{\beta}\right|\right|_2
\end{aligned}$$

Temperature predictor

A friend tells you:

I found a cool way to predict the temperature in New York: It's just a linear combination of the temperature in every other state. I fit the model on data from the last month and a half and it's perfect!

Overfitting

If a model is very complex, it may overfit the data

To evaluate a model we separate the data into a training and a test set

1. We fit the model using the training set

2. We evaluate the error on the test set

Experiment

X_{train}, X_{test}, z_{train} and β are iid Gaussian with mean 0 and variance 1

$$y_{\text{train}} = X_{\text{train}}\,\beta + z_{\text{train}}$$

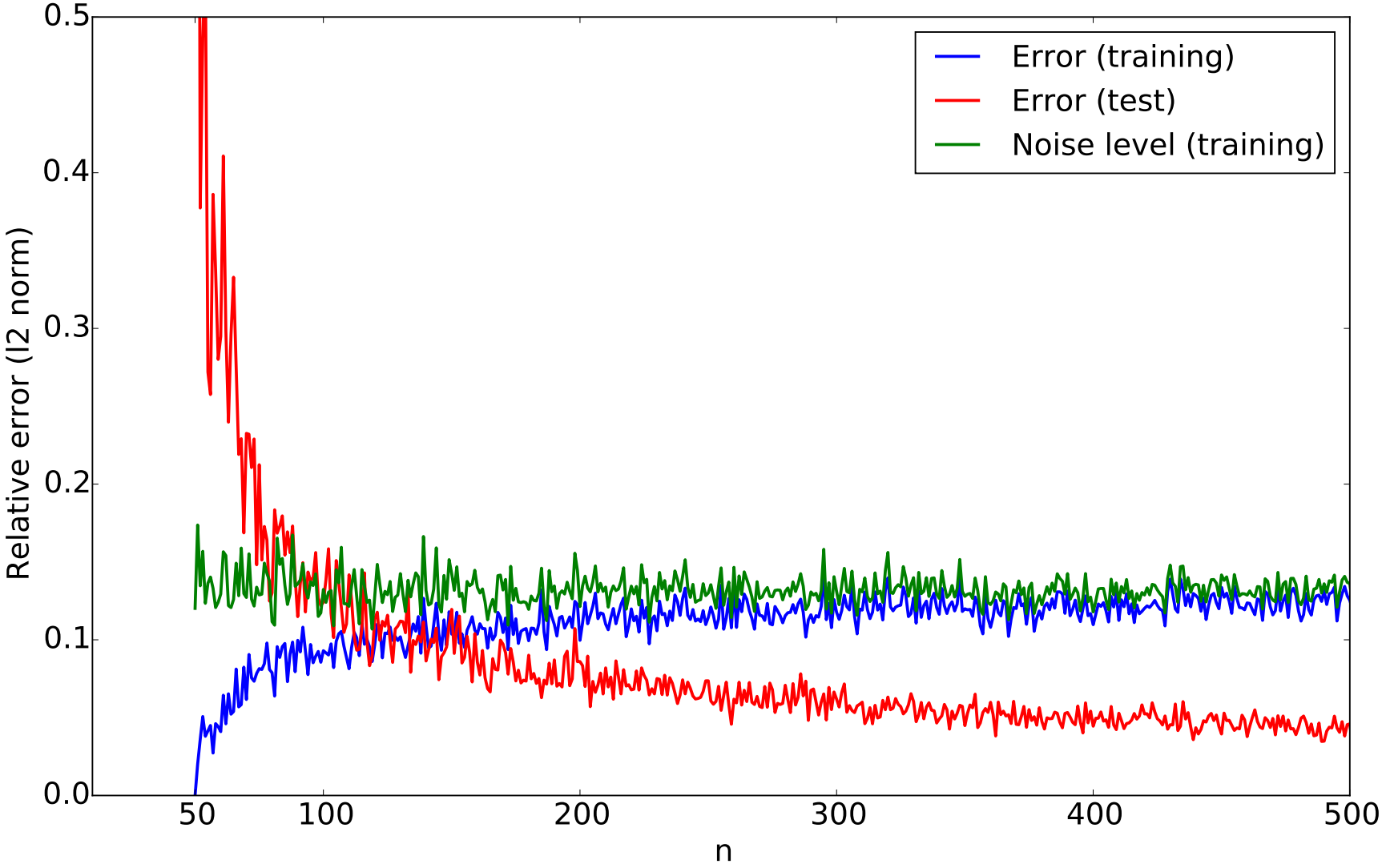
$$y_{\text{test}} = X_{\text{test}}\,\beta$$

We use y_{train} and X_{train} to compute β_{ls}

$$\text{error}_{\text{train}} = \frac{||X_{\text{train}}\,\beta_{\text{ls}} - y_{\text{train}}||_2}{||y_{\text{train}}||_2}$$

$$\text{error}_{\text{test}} = \frac{||X_{\text{test}}\,\beta_{\text{ls}} - y_{\text{test}}||_2}{||y_{\text{test}}||_2}$$

Experiment


Error (training)
Error (test)
Noise level (training)
Relative error (l2 norm)
n
0.0
0.1
0.2
0.3
0.4
0.5
50
100
200
300
400
500

Analysis

Data model, z is Gaussian noise with variance σ_z^2,

$$y = X\beta + z$$

$\sigma_{\text{min}}/\sigma_{\text{max}}$ are the smallest/largest singular values of X

$$\frac{1-\epsilon}{\sigma_{\text{min}}^2} \leq \frac{||\beta - \beta_{\text{ls}}||_2^2}{p\sigma_z^2} \leq \frac{1+\epsilon}{\sigma_{\text{max}}^2}$$

with high probability

Proof

$\beta_{\text{ls}} = V\Sigma^{-1}U^T$ where $X = U\Sigma V^T$

$$||\beta - \beta_{\text{ls}}||_2 = \sqrt{\sum_{j=1}^{p} \left(\frac{U_j^T z}{\sigma_j}\right)^2}$$

$U^T z$ is Gaussian with mean 0 and covariance I

$\left|\left|U^T z\right|\right|_2^2$ is chi-square with p degrees of freedom

$$\left|\left|U^T z\right|\right|_2^2 \approx p$$

Consistency

Assume $\sigma_z^2 = 1$

If X is well conditioned and entries have constant magnitude $\sigma_j \approx \sqrt{n}$

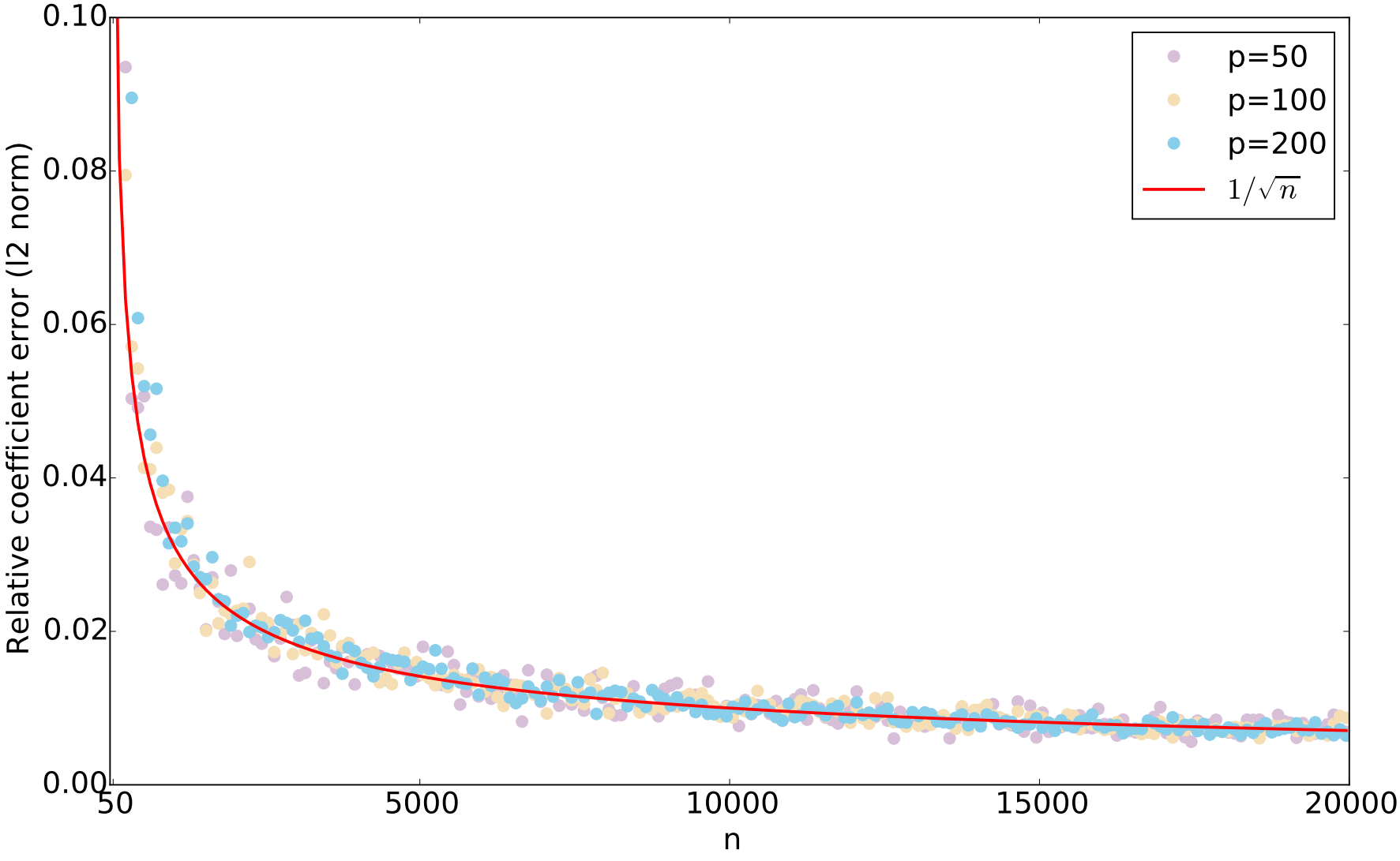
$$||\beta - \beta_{\text{ls}}||_2 \approx \sqrt{\frac{p}{n}}$$

If entries of β are constant, $||\beta||_2 \approx \sqrt{p}$

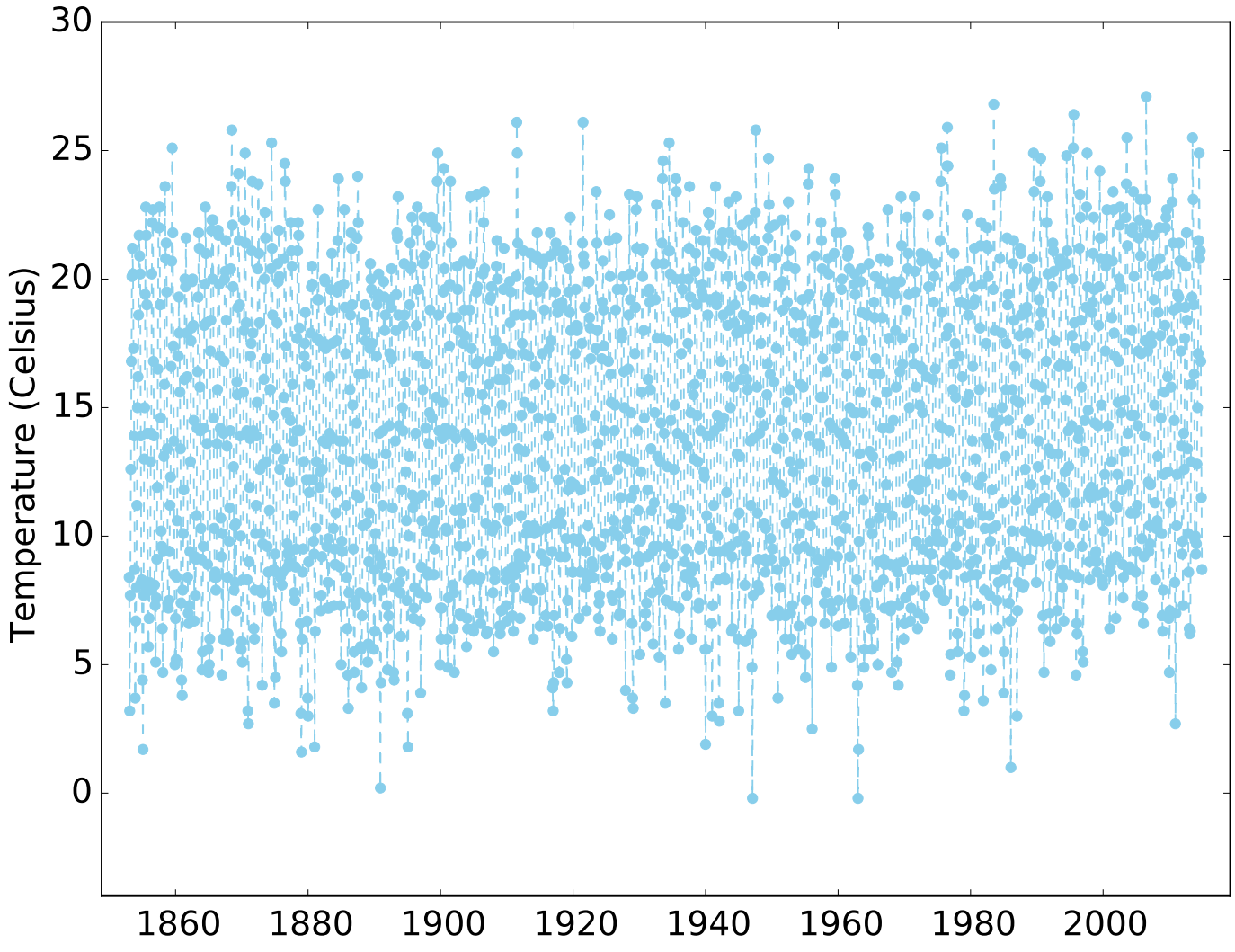
Least-squares estimator is consistent

$$\frac{||\beta - \beta_{\text{ls}}||_2}{||\beta||_2} \approx \frac{1}{\sqrt{n}}$$

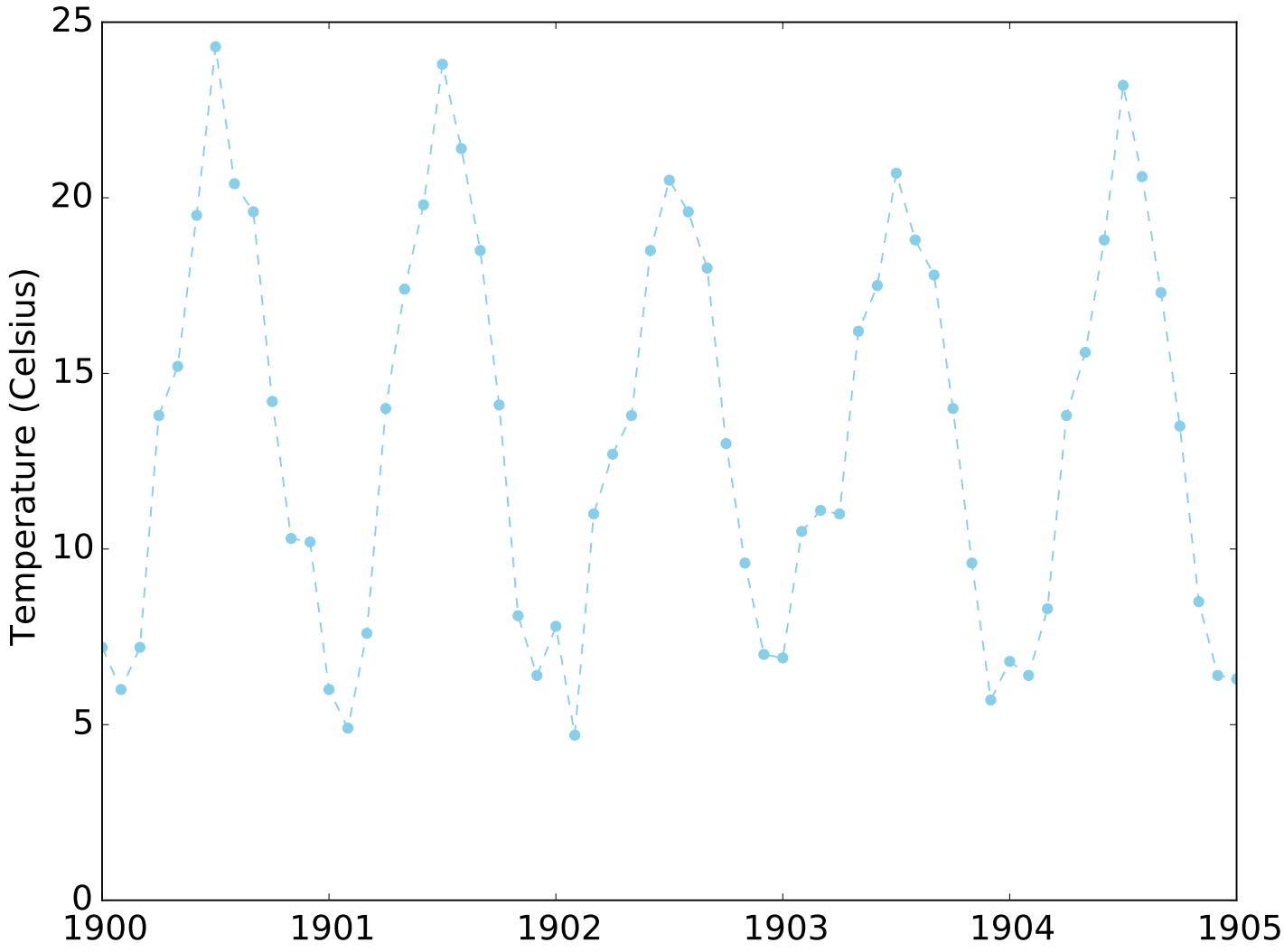
Experiment: $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$, $z \in \mathbb{R}^n$ iid $\mathcal{N}(0,1)$



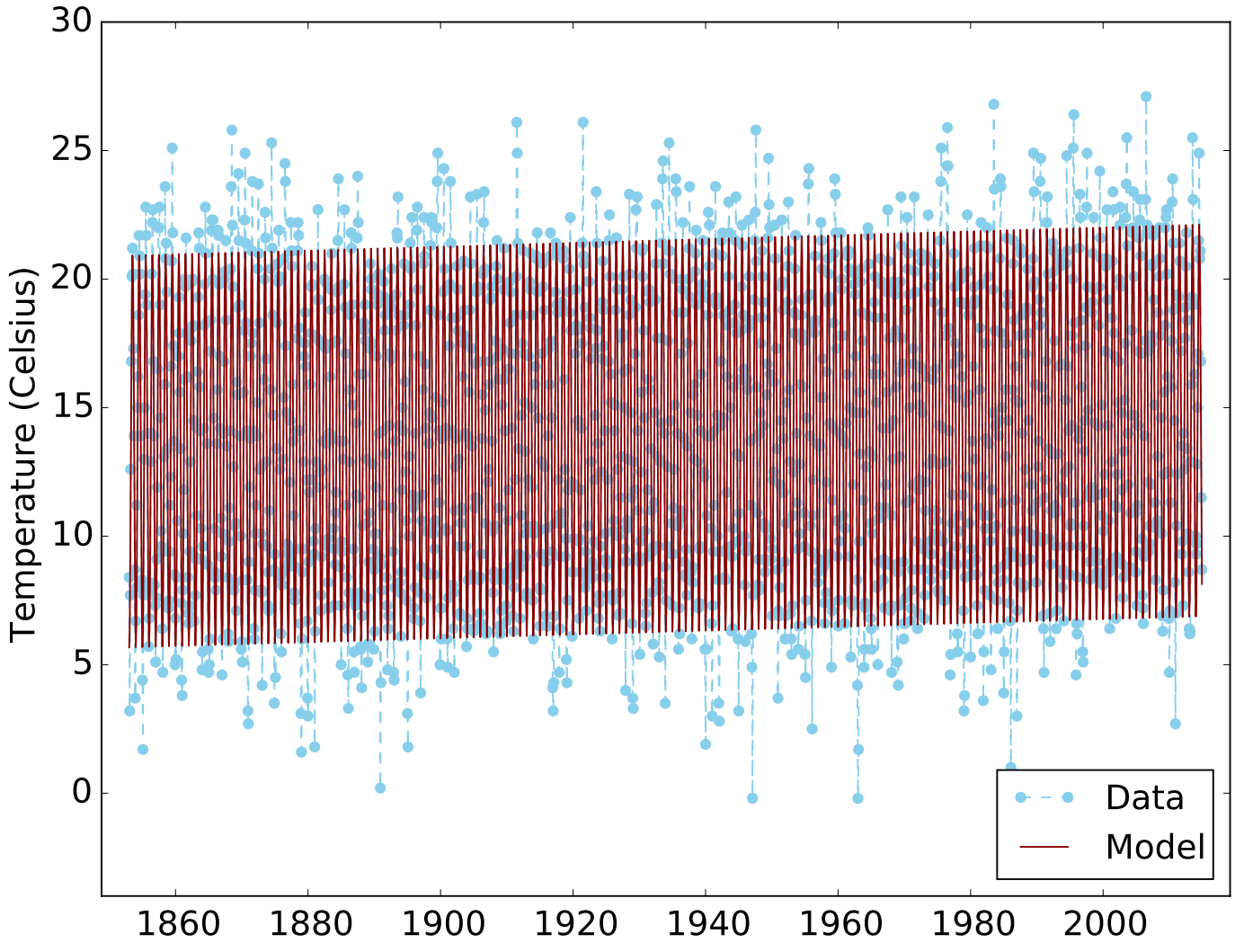
Maximum temperatures in Oxford, UK



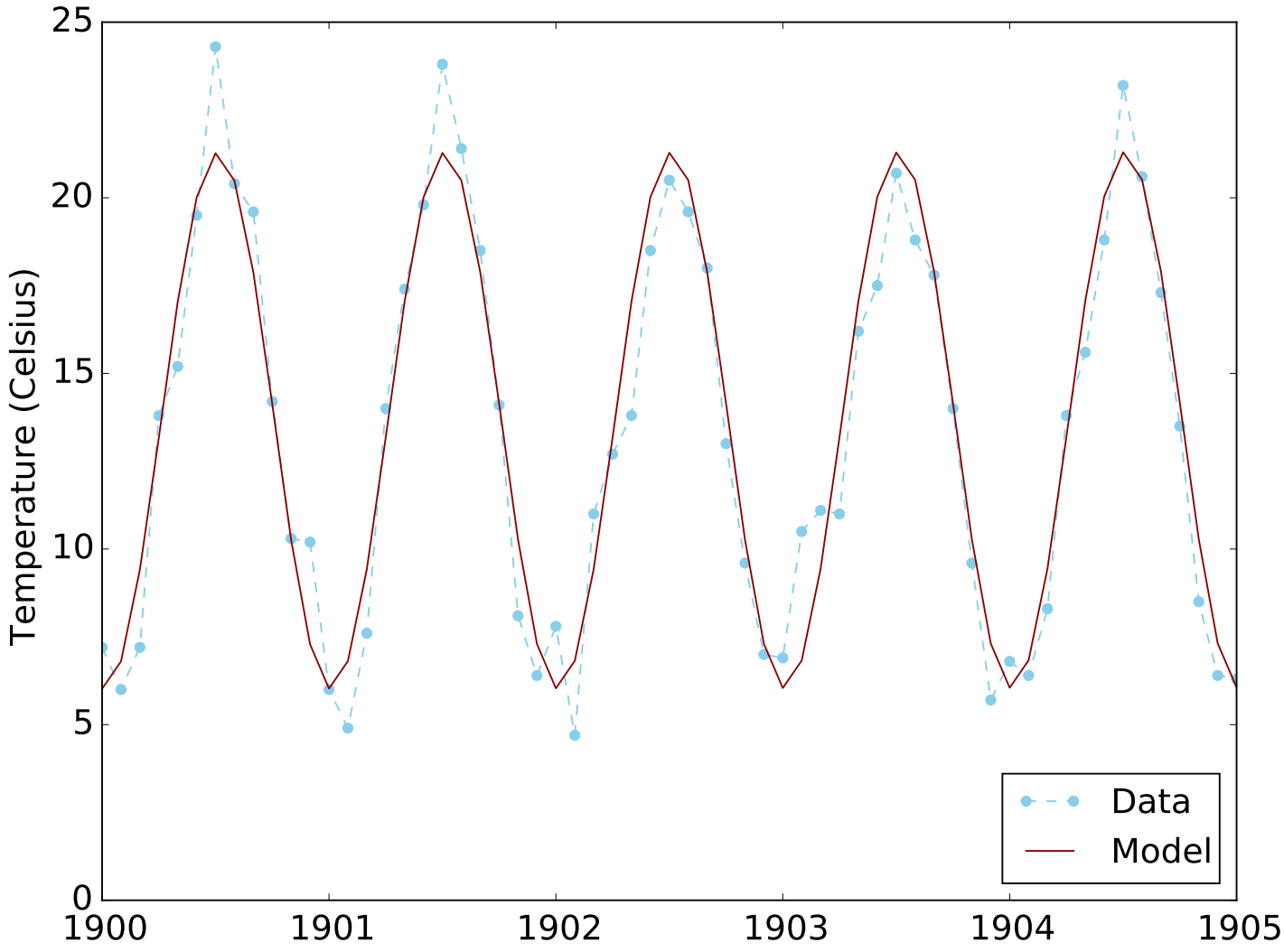
Maximum temperatures in Oxford, UK



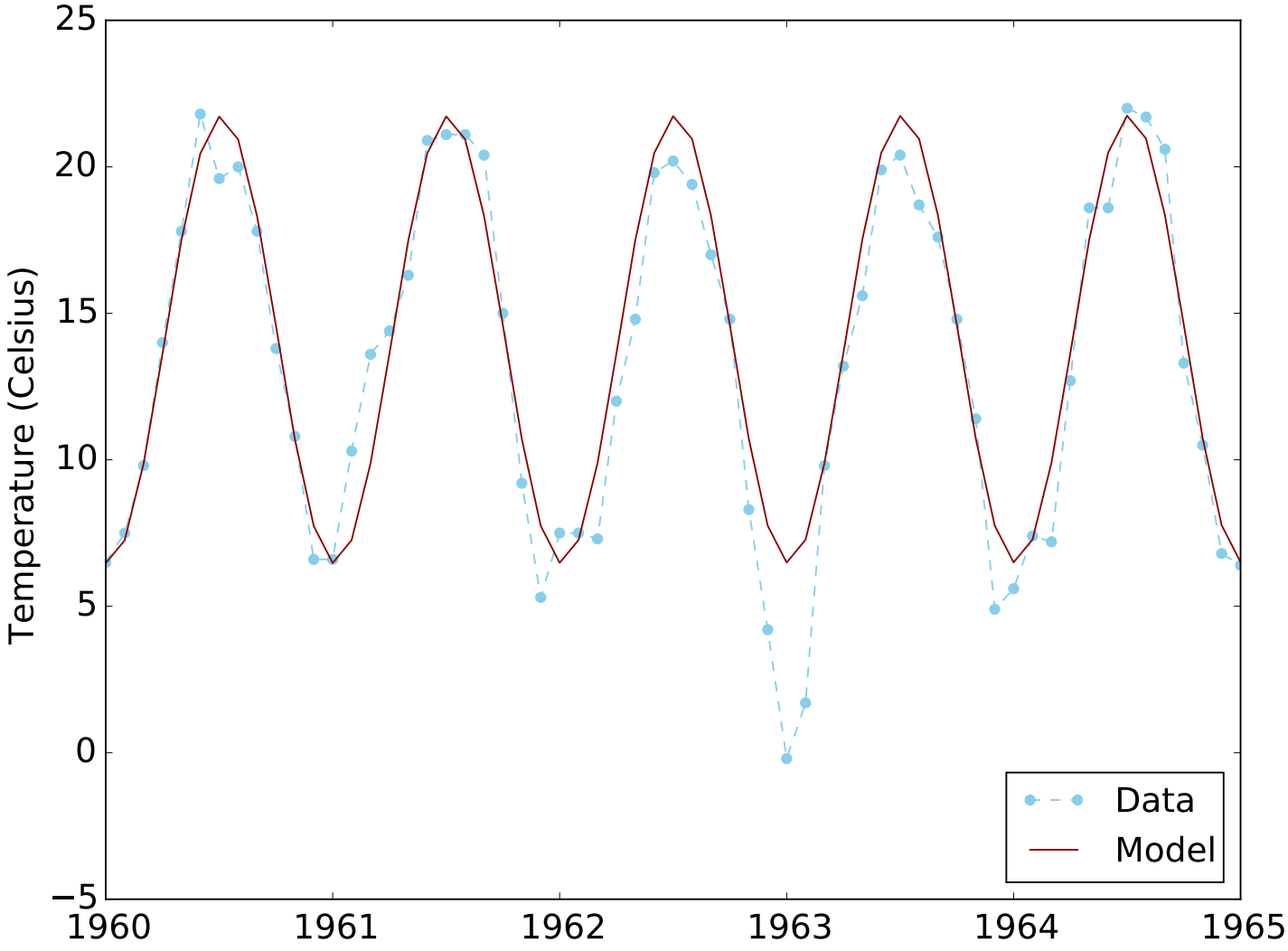
Model fitted by least squares



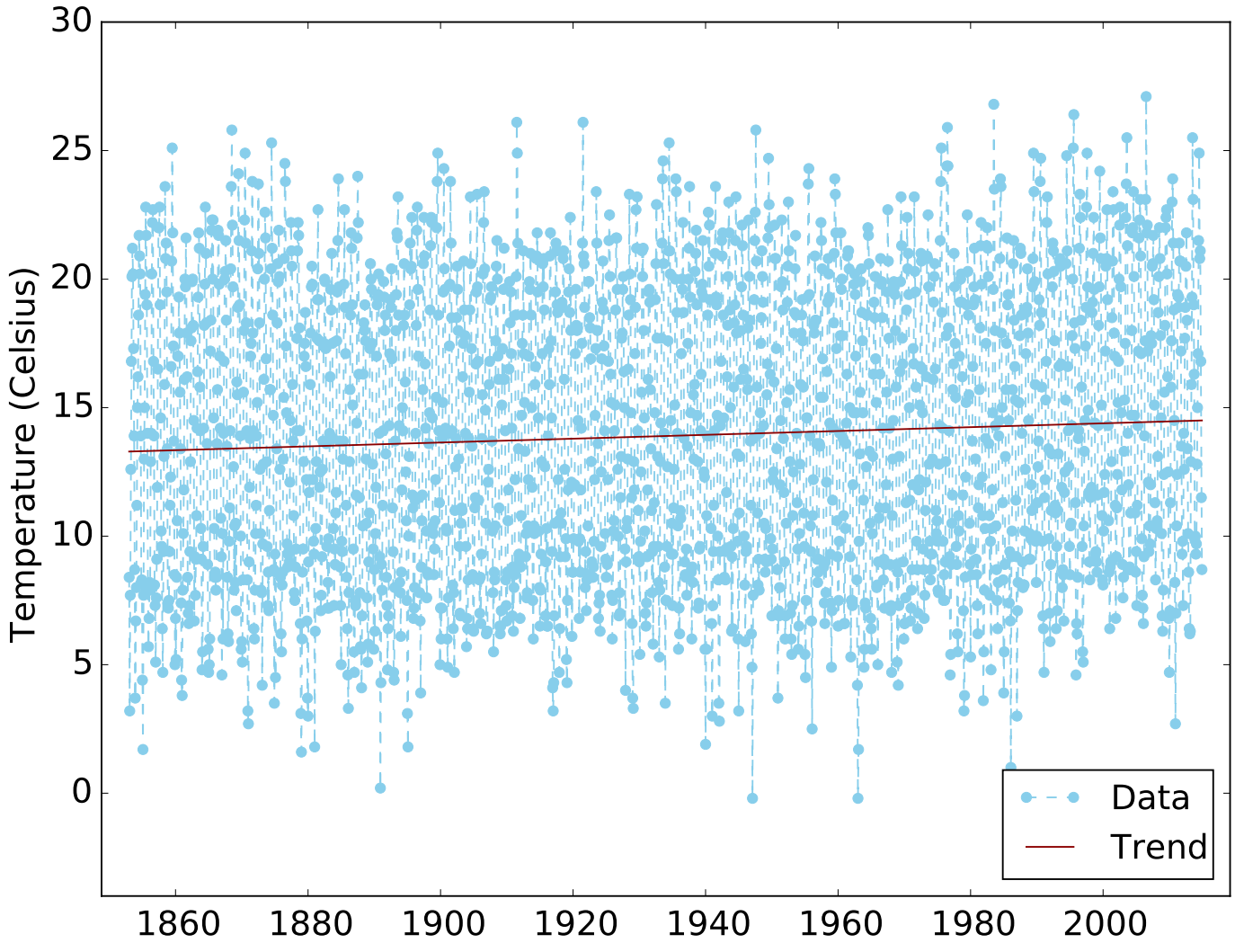
Model fitted by least squares



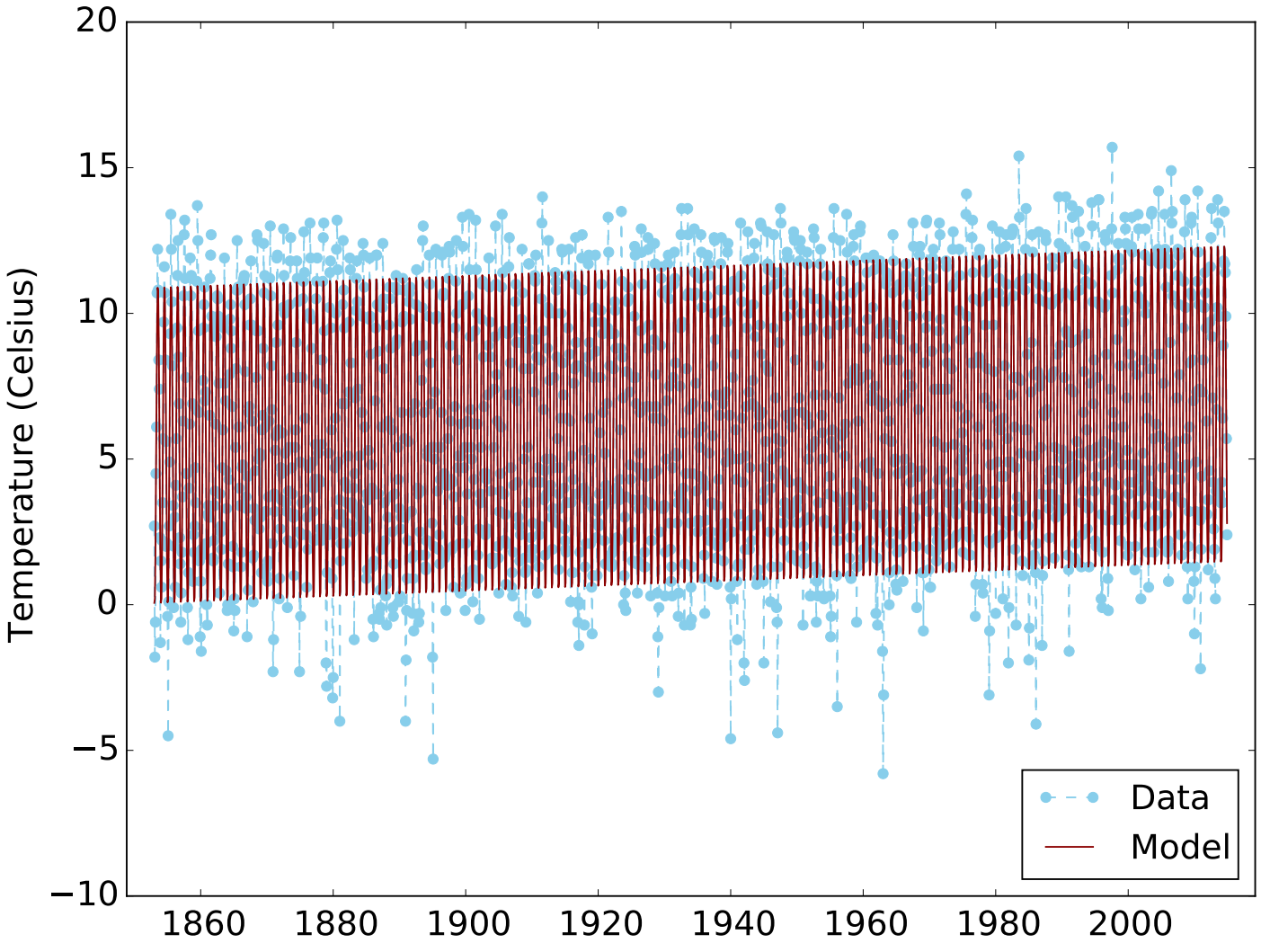
Model fitted by least squares



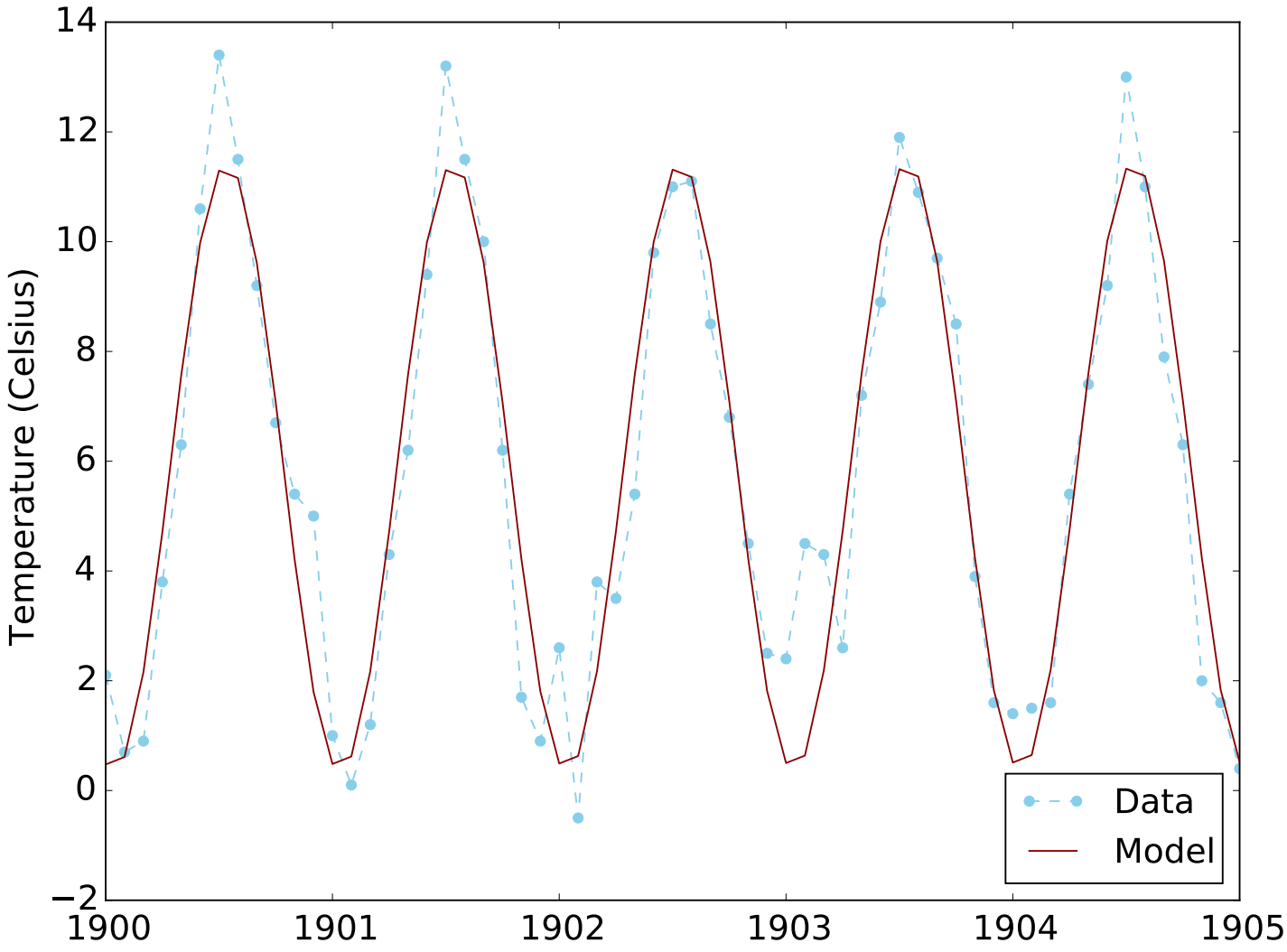
Trend: Increase of 0.75 °C / 100 years (1.35 °F)



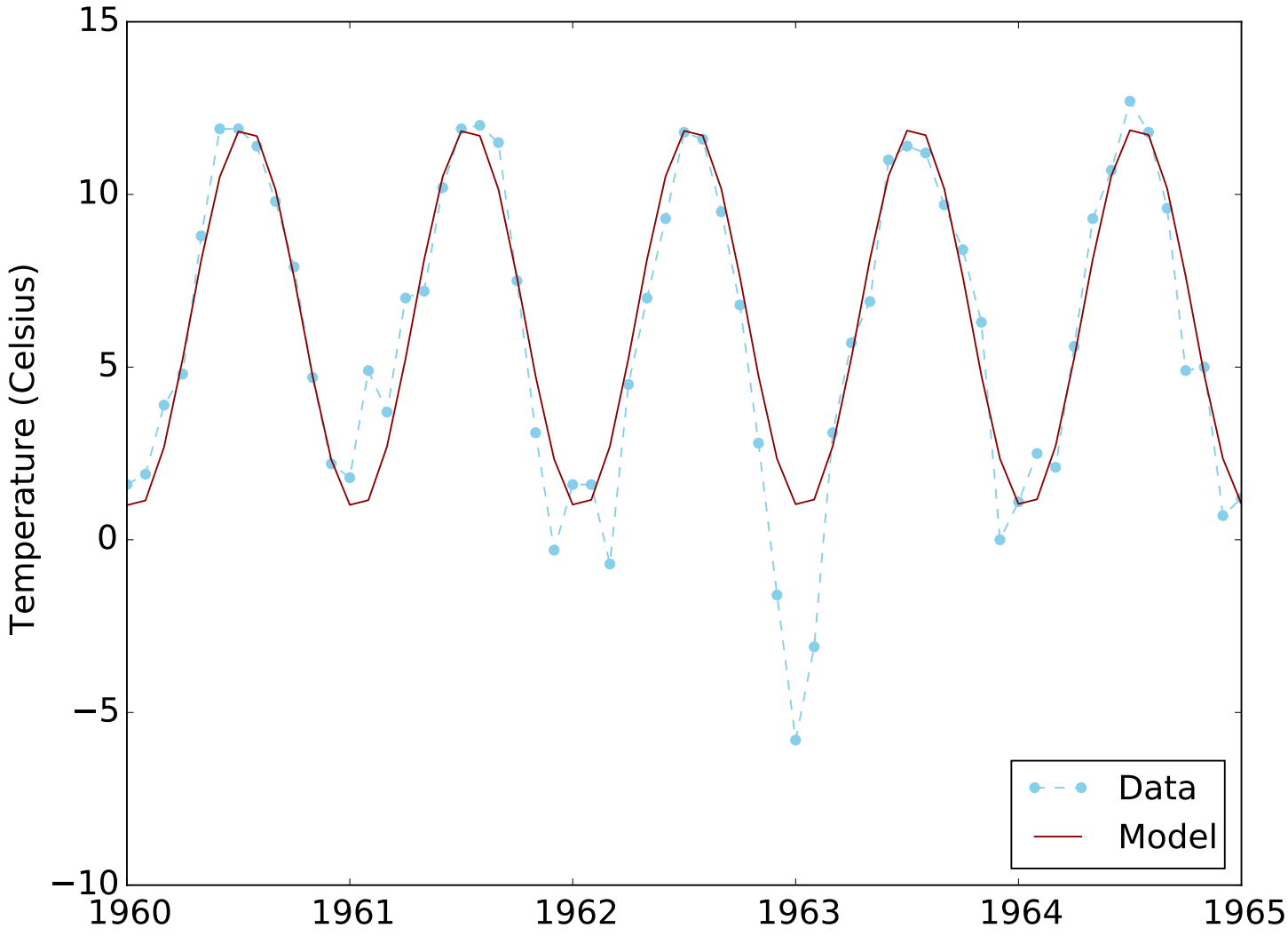
Model for minimum temperatures



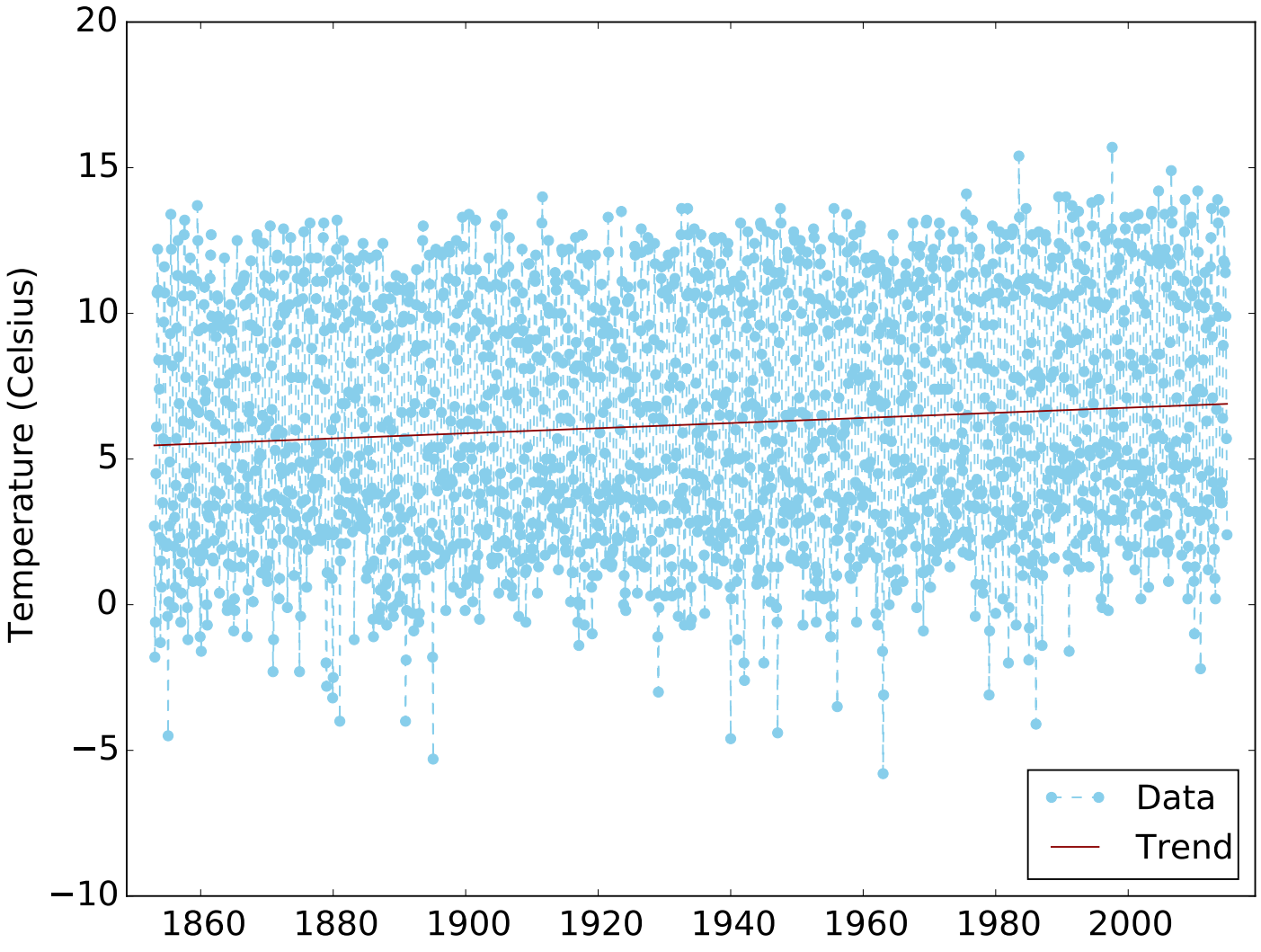
Model for minimum temperatures



Model for minimum temperatures



Trend: Increase of 0.88 °C / 100 years (1.58 °F)



Classification

Aim: Predict the value of a binary response $y \in \{0, 1\}$ from p predictors $X_1, X_2, \ldots, X_p \in \mathbb{R}$

Methodology:

1. Fit a model with using n training examples $y_1, y_2, \ldots, y_n$

$$y_i \approx f\left(X_{i1}, X_{i2}, \ldots, X_{ip}\right) \quad 1 \leq i \leq n$$

2. Use learned model f to predict from new data

Logistic-regression model

We model the probability that $y_i = 1$ by

$$\begin{aligned} \mathrm{P}\left(y_i = 1 | X_{i1}, X_{i2}, \ldots, X_{ip}\right) &:= g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right) \\ &= \frac{1}{1 + \exp\left(-\beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}\right)} \end{aligned}$$

g is the logistic function or sigmoid

$$g\left(t\right) := \frac{1}{1 + \exp -t}$$

Cost function

Likelihood of the data under a Bernouilli model in which

$$\mathrm{P}\left(y_i = 1 | X_{i1}, X_{i2}, \ldots, X_{ip}\right) := g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)$$

$$\mathrm{P}\left(y_i = 0 | X_{i1}, X_{i2}, \ldots, X_{ip}\right) := 1 - g\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)$$

Assuming independence

$$\mathcal{L}\left(\tilde{\beta}_0, \tilde{\beta}\right) = \prod_{i=1}^{n} g\left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j X_{ij}\right)^{y_i} \left(1 - g\left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j X_{ij}\right)\right)^{1-y_i}$$

Cost function

Log likelihood is concave

$$\log \mathcal{L}\left(\tilde{\beta}_0, \tilde{\beta}\right) = \sum_{i=1}^{n} y_i \log g\left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j X_{ij}\right) + (1 - y_i) \log\left(1 - g\left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j X_{ij}\right)\right)$$

Sparse regression

Assumption: Response only depends on a subset $\mathcal{S}$ of $s \ll p$ predictors

$$y \approx 1\beta_0 + X_{\mathcal{S}}\beta_{\mathcal{S}}$$

Regime of interest: $p \approx n$

Model-selection problem: Determine what predictors are relevant

Best-subset selection:

- Choose among all possible sparse models
- Intractable unless s is very small

Forward stepwise regression

Greedy method similar to orthogonal matching pursuit

Initialization:

$$
\begin{aligned}
j_0 &:= \arg\max_j |\langle y, X_j \rangle| \\
\mathcal{J} &:= \{j_0\} \\
\beta_{\text{ls}} &:= \arg\min_{\tilde{\beta}} \left|\left| y - X_{\mathcal{J}}\, \tilde{\beta} \right|\right|_2 \\
r^{(0)} &:= y - X_{\mathcal{J}}\, \beta_{\text{ls}}
\end{aligned}
$$

Forward stepwise regression

Iterations: $k = 1, 2, \ldots$

$$
\begin{aligned}
j_k &:= \arg\max_{j \notin \mathcal{J}} \left| \left\langle y, \mathcal{P}_{\mathrm{col}(X_{\mathcal{J}})^{\perp}} X_j \right\rangle \right| \\
\mathcal{J} &:= \mathcal{J} \cup \{j_k\} \\
\beta_{\mathrm{ls}} &:= \arg\min_{\tilde{\beta}} \left\| y - X_{\mathcal{J}} \, \tilde{\beta} \right\|_2 \\
r^{(k)} &:= r^{(k-1)} - X_{\mathcal{J}} \, \beta_{\mathrm{ls}}
\end{aligned}
$$

Stop when fit does not improve much anymore

The lasso

ℓ_1-norm regularized least-squares regression

$$\text{minimize} \quad \frac{1}{2n} \left|\left| y - X\tilde{\beta} \right|\right|_2^2 + \lambda \left|\left| \tilde{\beta} \right|\right|_1$$

We assume that the data are centered, so $\beta_0 = 0$

Original formulation

$$\begin{aligned} \text{minimize} \quad & \left|\left| y - X\tilde{\beta} \right|\right|_2^2 \\ \text{subject to} \quad & \left|\left| \tilde{\beta} \right|\right|_1 \leq \tau \end{aligned}$$

Equivalent, but relation between λ and τ depends on X and y

Experiment

X_{train} and X_{test} are iid Gaussian $(0, 1)$, z_{train} is iid Gaussian $(0, 0.5)$

β has 10 nonzero entries, $p = 50$, $n = 100$

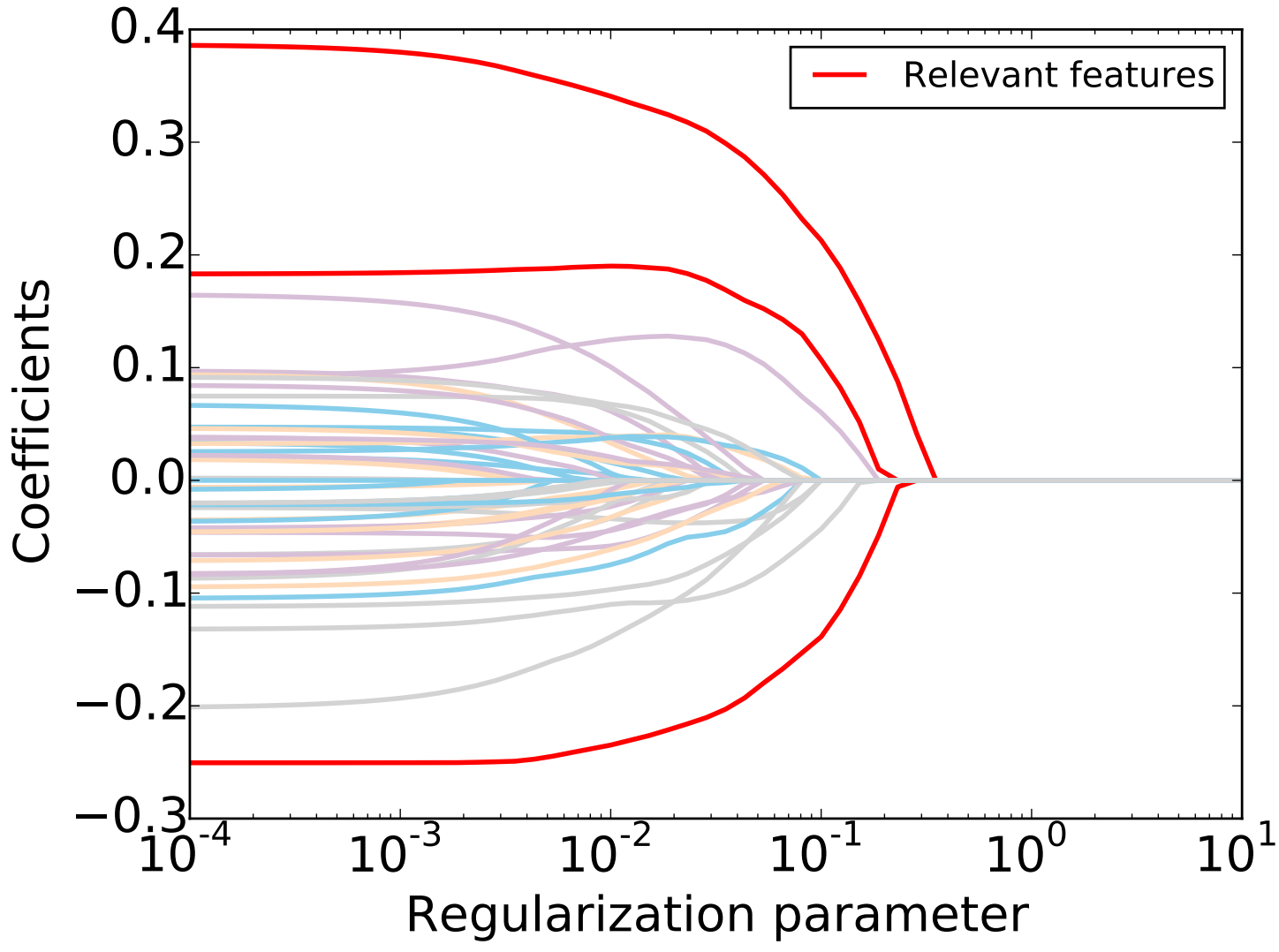
$$y_{\text{train}} = X_{\text{train}}\,\beta + z_{\text{train}} \qquad y_{\text{test}} = X_{\text{test}}\,\beta$$

For an estimate $\hat{\beta}$ learnt from the training data

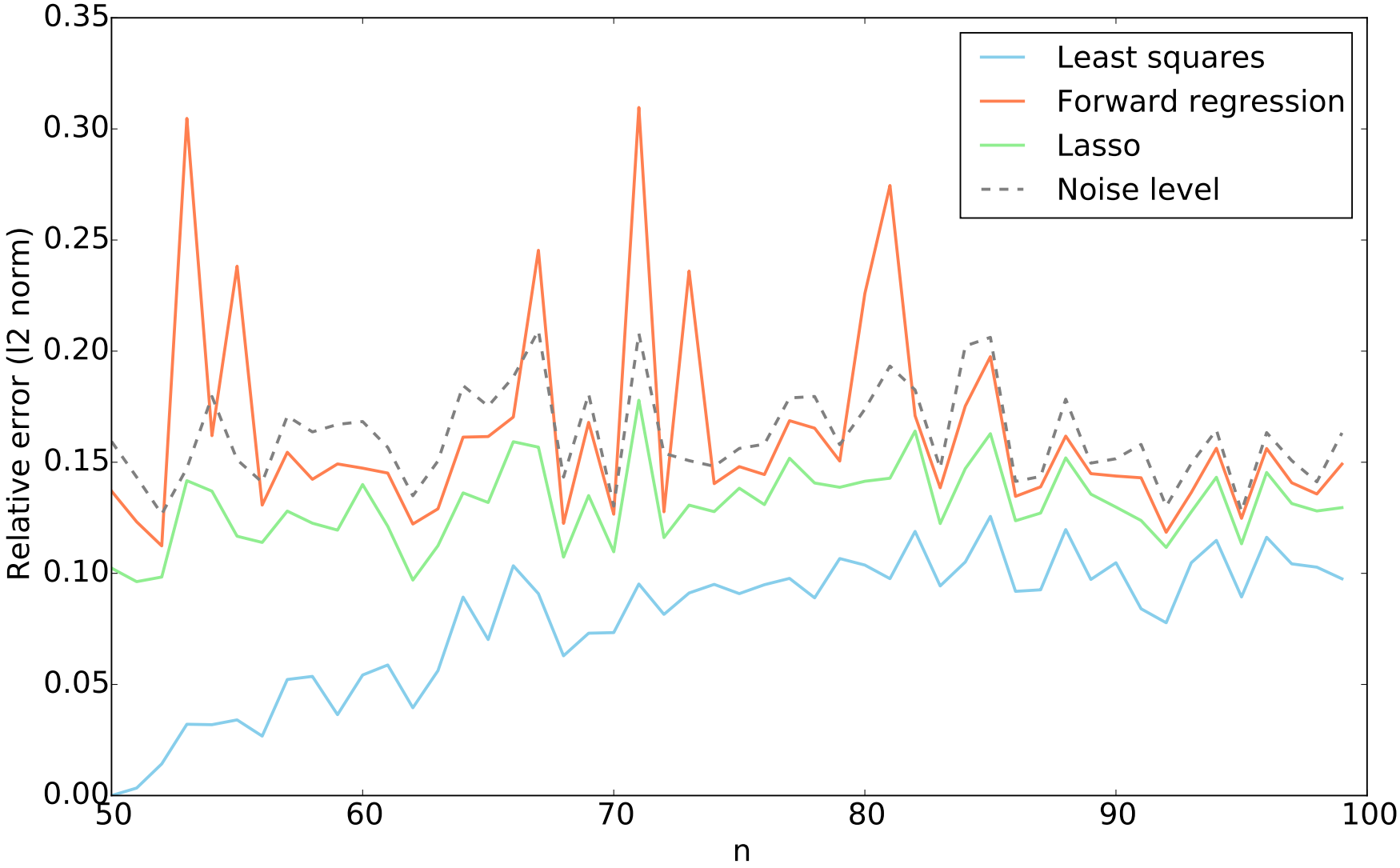
$$\text{error}_{\text{train}} = \frac{\left|\left|X_{\text{train}}\,\hat{\beta} - y_{\text{train}}\right|\right|_2}{||y_{\text{train}}||_2}$$

$$\text{error}_{\text{test}} = \frac{\left|\left|X_{\text{test}}\,\hat{\beta} - y_{\text{test}}\right|\right|_2}{||y_{\text{test}}||_2}$$

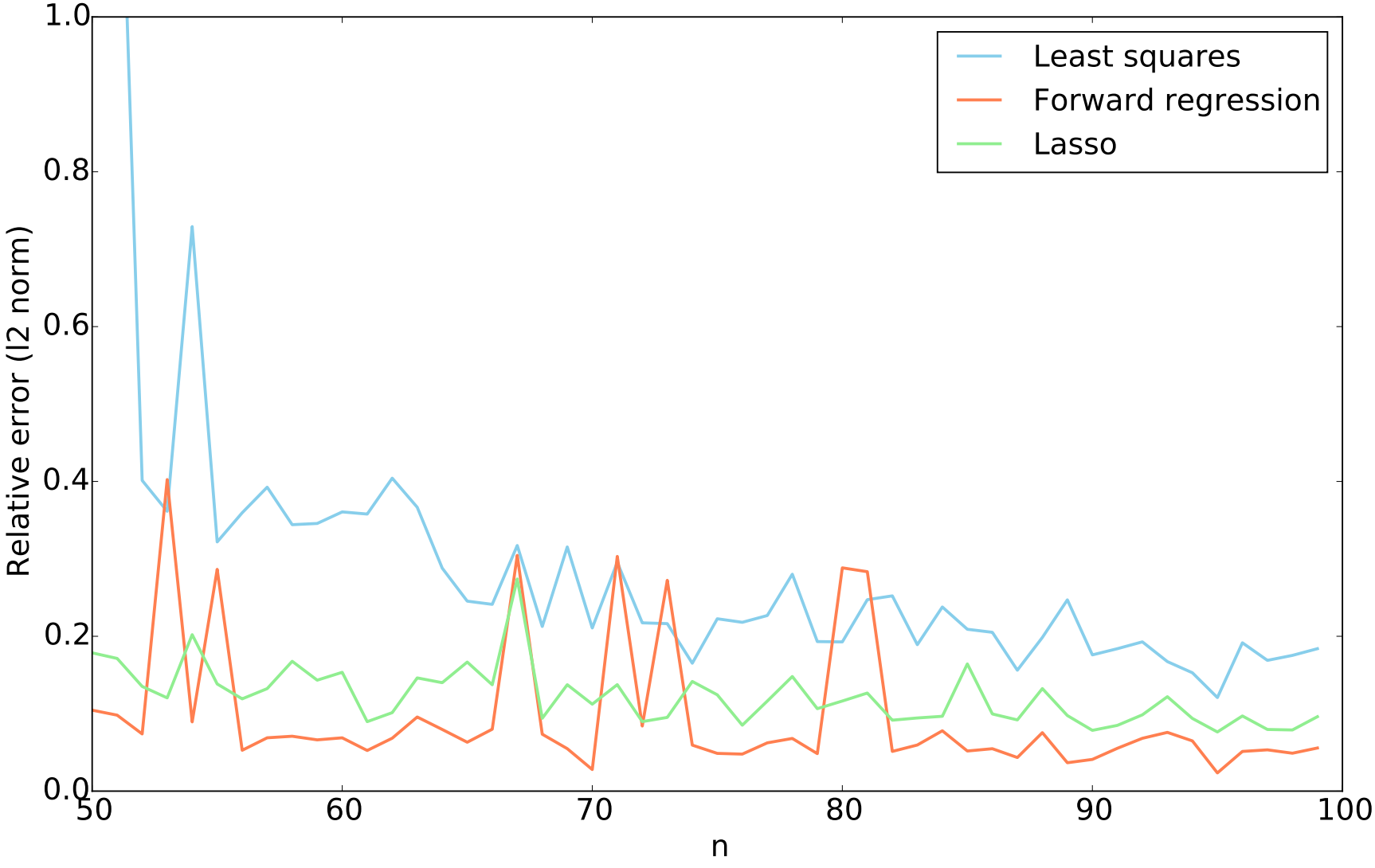
Lasso path


Relevant features
Coefficients
Regularization parameter
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
10-4
10-3
10-2
10-1
100
101

Training error



Test error


Least squares
Forward regression
Lasso
Relative error (l2 norm)
n
0.0
0.2
0.4
0.6
0.8
1.0
50
60
70
80
90
100

Logistic regression

Add ℓ_1-norm regularization term to log likelihood, we minimize

$$
\begin{aligned}
&- \sum_{i=1}^{n} y_i \log g\left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j X_{ij}\right) \\
&- (1 - y_i) \log\left(1 - g\left(\tilde{\beta}_0 + \sum_{j=1}^{p} \tilde{\beta}_j X_{ij}\right)\right) \\
&+ \lambda \left|\left|\tilde{\beta}\right|\right|_1
\end{aligned}
$$

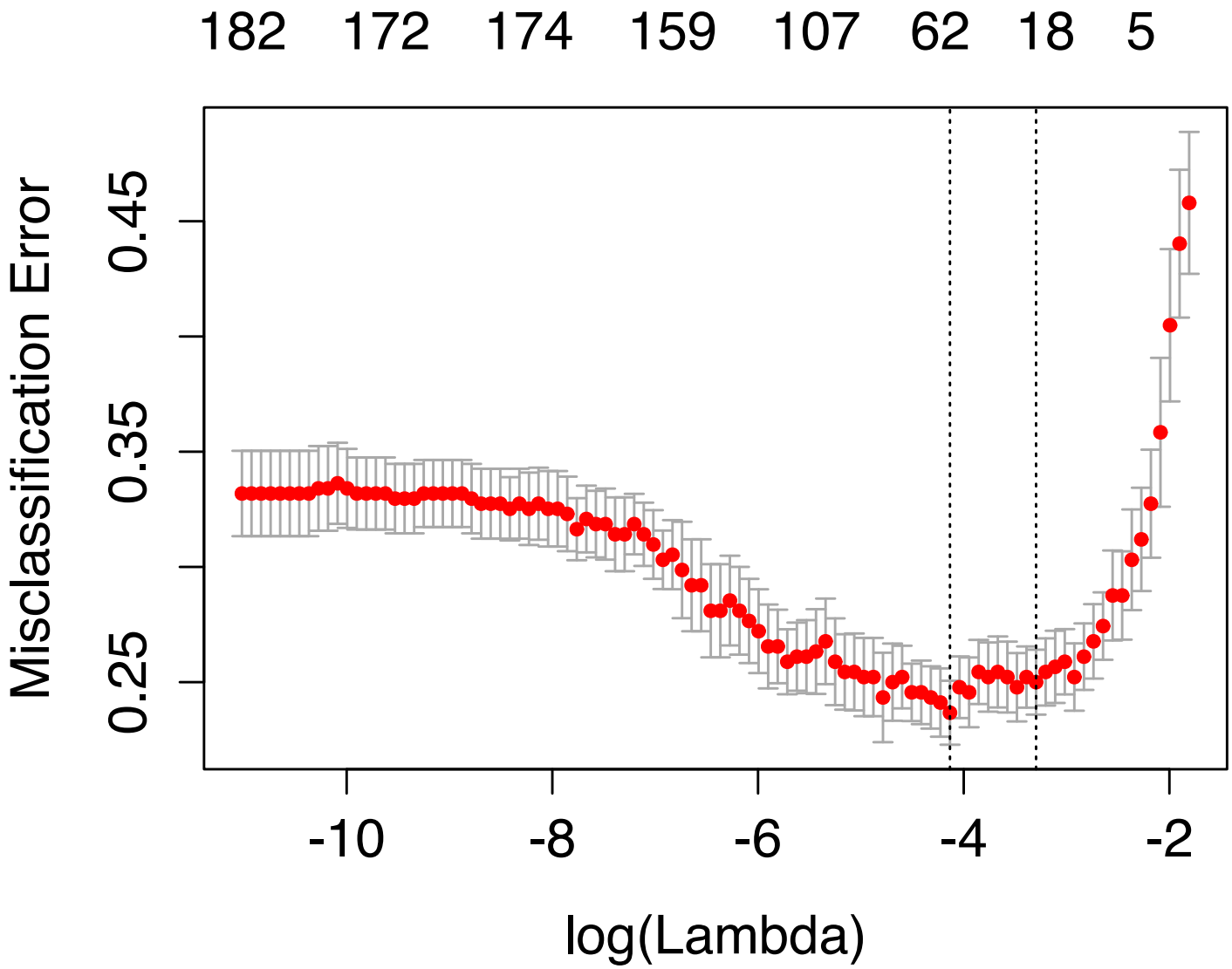
Arrhythmia prediction

Predict whether patient has arrhythmia from $n = 271$ examples and $p = 182$ predictors

- Age, sex, height, weight
- Features obtained from electrocardiogram recordings

Best sparse model uses around 60 predictors

Prediction accuracy



Least-squares analysis

Data model, z is Gaussian noise with variance σ_z^2 and β is s sparse

$$y = X\beta + z$$

Assumption: X is well conditioned, entries have constant magnitude

- $\sigma_{\min} \approx \sqrt{n}$
- Columns of X have ℓ_2 norm $\approx \sqrt{n}$

For the least-squares estimator we have

$$||\beta - \beta_{\text{ls}}||_2 \approx \sigma_z \sqrt{\frac{p}{n}}$$

If $p \approx n$, but $s \ll p$ we should do better!

Restricted eigenvalue property

There exists $\gamma > 0$ such that for any $v \in \mathbb{R}^p$ if

$$||v_{T^c}||_1 \leq ||v_T||_1$$

for any subset T, $|T| \leq s$, then

$$\frac{1}{n} ||M\ v||_2^2 \geq \gamma\, ||v||_2^2$$

Similar to restricted isometry property, may hold even if $p > n$!

Guarantees

If X satisfies the RE property and $\tau = ||\beta||_1$, the solution β_{lasso} to

$$\begin{aligned} \text{minimize} \quad & \left|\left| y - X\tilde{\beta} \right|\right|_2^2 \\ \text{subject to} \quad & \left|\left| \tilde{\beta} \right|\right|_1 \leq \tau \end{aligned}$$

satisfies

$$||\beta - \beta_{\text{lasso}}||_2 \leq \frac{\sigma_z}{\gamma} \sqrt{\frac{32\ \alpha\ s \log p}{n}}$$

with probability $1 - 2\exp\left(-\left(\alpha - 1\right)\log p\right)$ for any $\alpha > 1$

Close to least-squares error if we know which predictors are relevant

Proof

The error

$$h := \beta - \beta_{\text{lasso}}$$

satisfies

$$||h_{T^c}||_1 \leq ||h_T||_1$$

so

$$||\beta - \beta_{\text{lasso}}||_2^2 \leq \frac{1}{\gamma n} ||X\ h||_2^2$$

Proof

By optimality of β_{lasso}

$$||Xh||_2^2 \leq \left|2\, z^T Xh\right| \leq 2 \left|\left|X^T z\right|\right|_\infty ||h||_1$$

Since $||h_{T^c}||_1 \leq ||h_T||_1$

$$||h||_1 \leq 2\sqrt{s}\, ||h||_2$$

so

$$||\beta - \beta_{\text{lasso}}||_2^2 \leq \frac{4\sqrt{s}}{\gamma n} ||\beta - \beta_{\text{lasso}}||_2 \left|\left|X^T z\right|\right|_\infty$$

Proof

$X_i^T z$ is Gaussian with variance $\sigma_z^2 ||X_i||_2^2$, so for $t > 0$

$$\mathrm{P}\left(\left|X_i^T z\right| > t\sigma_z ||X_i||_2\right) \leq 2\exp\left(-\frac{t^2}{2}\right)$$

By the union bound,

$$\begin{aligned}\mathrm{P}\left(\left|\left|X^T z\right|\right|_\infty > t\sigma_z \max_i ||X_i||_2\right) &\leq 2\ p\exp\left(-\frac{t^2}{2}\right) \\ &= 2\exp\left(-\frac{t^2}{2} + \log p\right)\end{aligned}$$

Proof

Choosing $t = \sqrt{2\alpha \log p}$, $\alpha > 2$, if $\max_i ||X_i||_2 = \sqrt{n}$ we have

$$\mathrm{P}\left(\left|\left|X^T z\right|\right|_\infty > \sigma_z \sqrt{2\alpha\, n \log p}\right) \leq 2 \exp\left(-(\alpha - 1) \log p\right)$$

We conclude

$$||\beta - \beta_{\text{lasso}}||_2 \leq \frac{\sigma_z}{\gamma} \sqrt{\frac{32\ \alpha\ s \log p}{n}}$$

with probability $1 - 2 \exp\left(-(\alpha - 1) \log p\right)$

Correlated predictors

Error of least-squares estimator

$$||\beta - \beta_{\text{ls}}||_2 = \sqrt{\sum_{j=1}^{p} \left(\frac{U_j^T z}{\sigma_j} \right)^2}$$

If predictors are strongly correlated singular values can be very small

Estimator suffers from noise amplification

Ridge regression

Aim: Control norm of weights

$$\text{minimize} \quad \left|\left|y - X\tilde{\beta}\right|\right|_2^2 + \lambda \left|\left|\tilde{\beta}\right|\right|_2^2$$

Equivalent to regression problem

$$\text{minimize} \quad \left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \lambda I \end{bmatrix} \tilde{\beta} \right|\right|_2^2$$

Ridge regression

If $y = X\beta + z$

$$\beta_{\text{ridge}} = V \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda^2} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_2^2}{\sigma_2^2+\lambda^2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \frac{\sigma_p^2}{\sigma_p^2+\lambda^2} \end{bmatrix} V^T \beta$$

$$+ V \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\lambda^2} & 0 & \cdots & 0 \\ 0 & \frac{\sigma_2}{\sigma_2^2+\lambda^2} & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \frac{\sigma_p}{\sigma_p^2+\lambda^2} \end{bmatrix} U^T z$$

Decreases noise amplification if singular values are small

Correlated predictors in model selection

Lasso controls norm amplification but tends to choose a small subset of the relevant predictors

Aim: Selecting all relevant predictors

Why? Interpretability + improving prediction

Identical predictors

If two predictors are the same $X_i = X_j$ then β_i and β_j should be the same

True for any cost function of the form

$$\text{minimize} \quad \frac{1}{2n} \left|\left| y - X\tilde{\beta} \right|\right|_2^2 + \lambda \, \mathcal{R}\left(\tilde{\beta}\right)$$

as long as the regularizer $\mathcal{R}$ is strictly convex

Not true for the lasso

Experiment

The entries of z_{train}, Z_{train} and Z_{test} are iid Gaussian $(0, 1)$

$$y_{\text{train}} := X_{A,\text{train}}\,\beta_A + X_{B,\text{train}}\,\beta_B + z_{\text{train}}$$

$$y_{\text{test}} := X_{A,\text{test}}\,\beta_A + X_{B,\text{test}}\,\beta_B$$

$$X_{\text{train}} := \begin{bmatrix} X_{A,\text{train}} & X_{B,\text{train}} & Z_{\text{train}} \end{bmatrix}$$

$$X_{\text{test}} := \begin{bmatrix} X_{A,\text{test}} & X_{B,\text{test}} & Z_{\text{test}} \end{bmatrix}$$

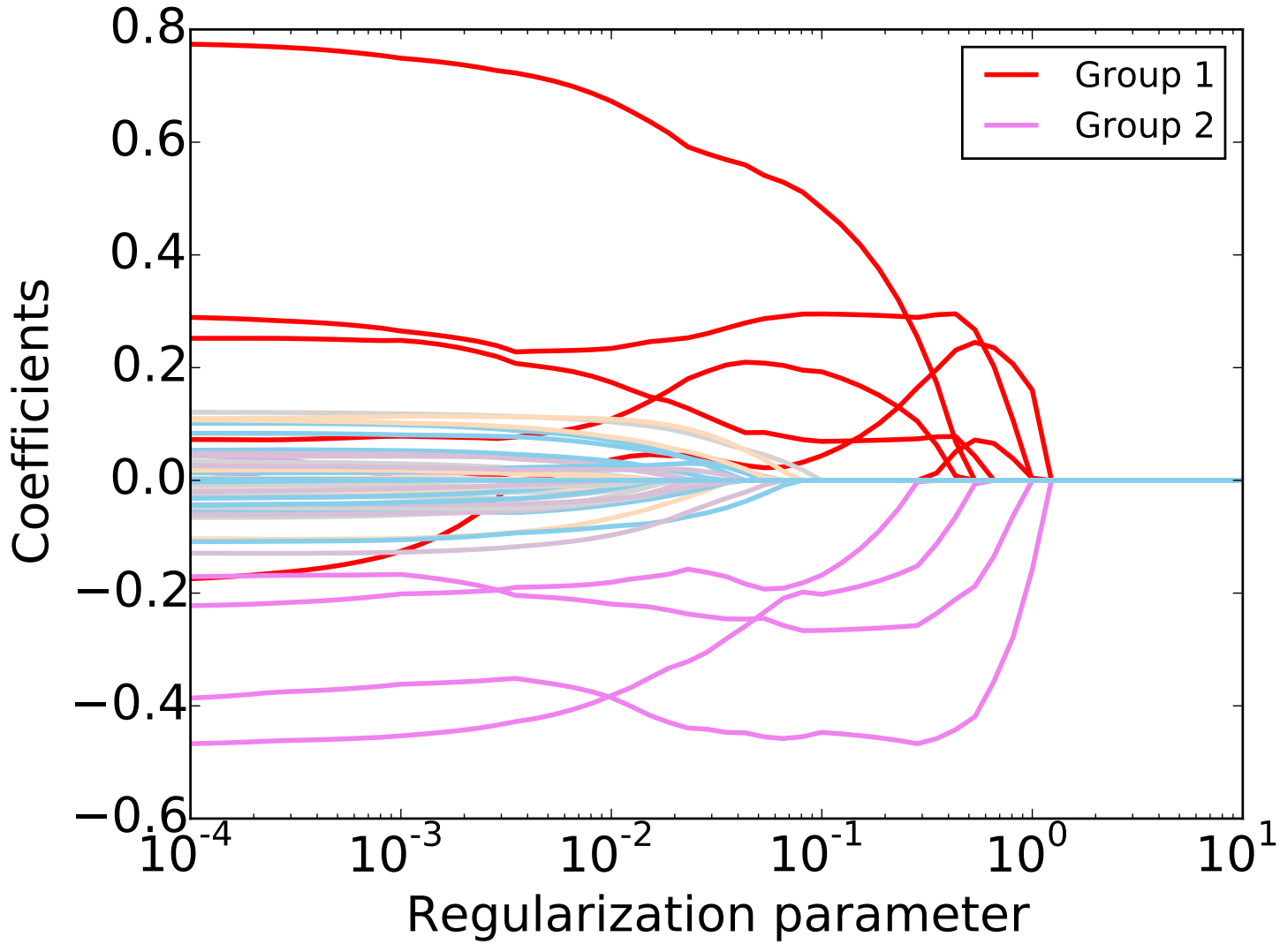
Experiment

Rows of $X_{A,\text{train}}$, $X_{B,\text{train}}$, $X_{A,\text{train}}$, $X_{B,\text{train}}$ are independent Gaussian random vectors with mean zero and covariance matrix

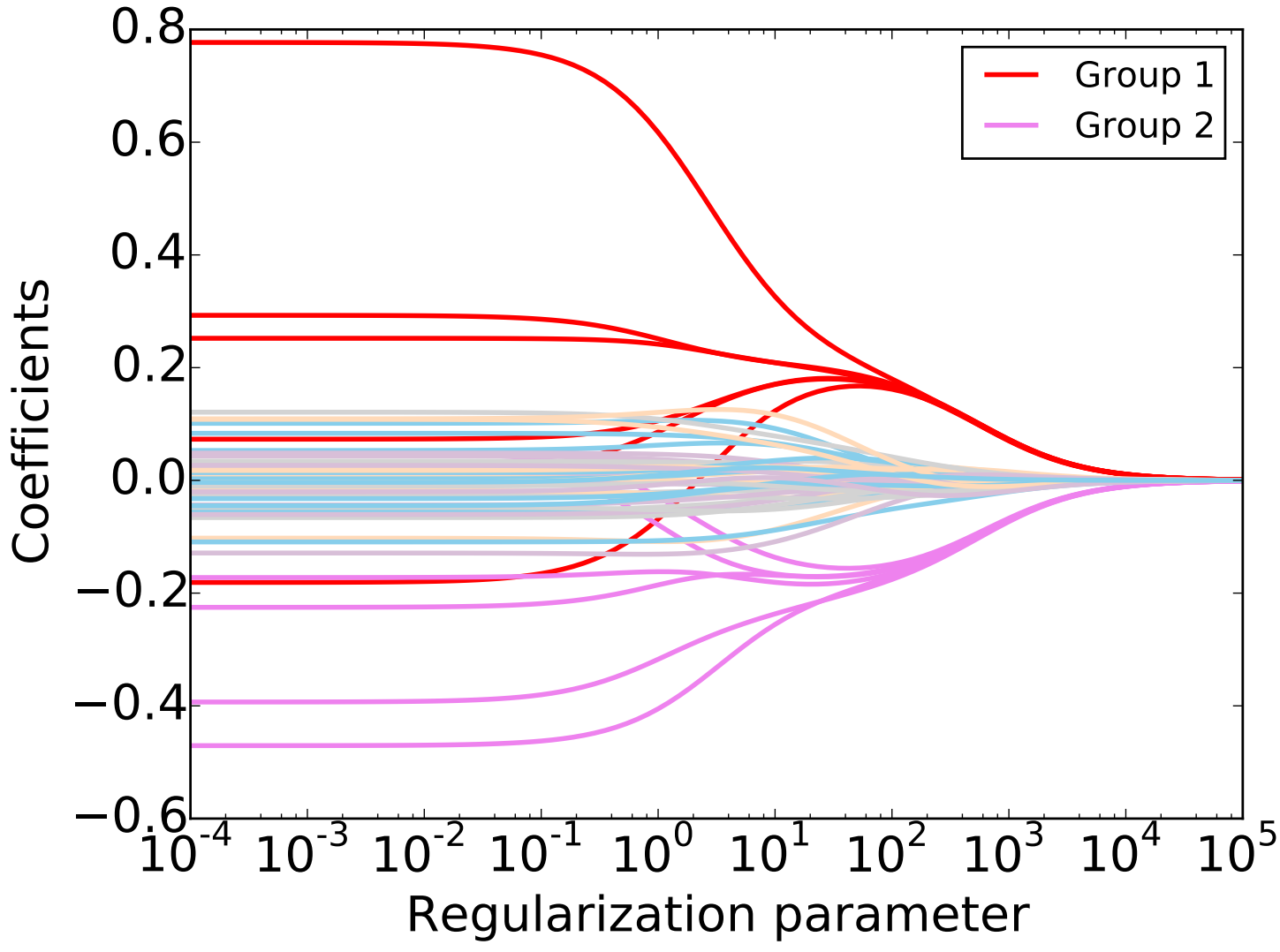
$$\Sigma := \begin{bmatrix} 1 & 0.95 & 0.95 & 0.95 \\ 0.95 & 1 & 0.95 & 0.95 \\ 0.95 & 0.95 & 1 & 0.95 \\ 0.95 & 0.95 & 0.95 & 1 \end{bmatrix}$$

β_A and β_B have 4 nonzero entries each

Lasso path


Group 1
Group 2
Coefficients
Regularization parameter
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
10^-4
10^-3
10^-2
10^-1
10^0
10^1

Ridge-regression path

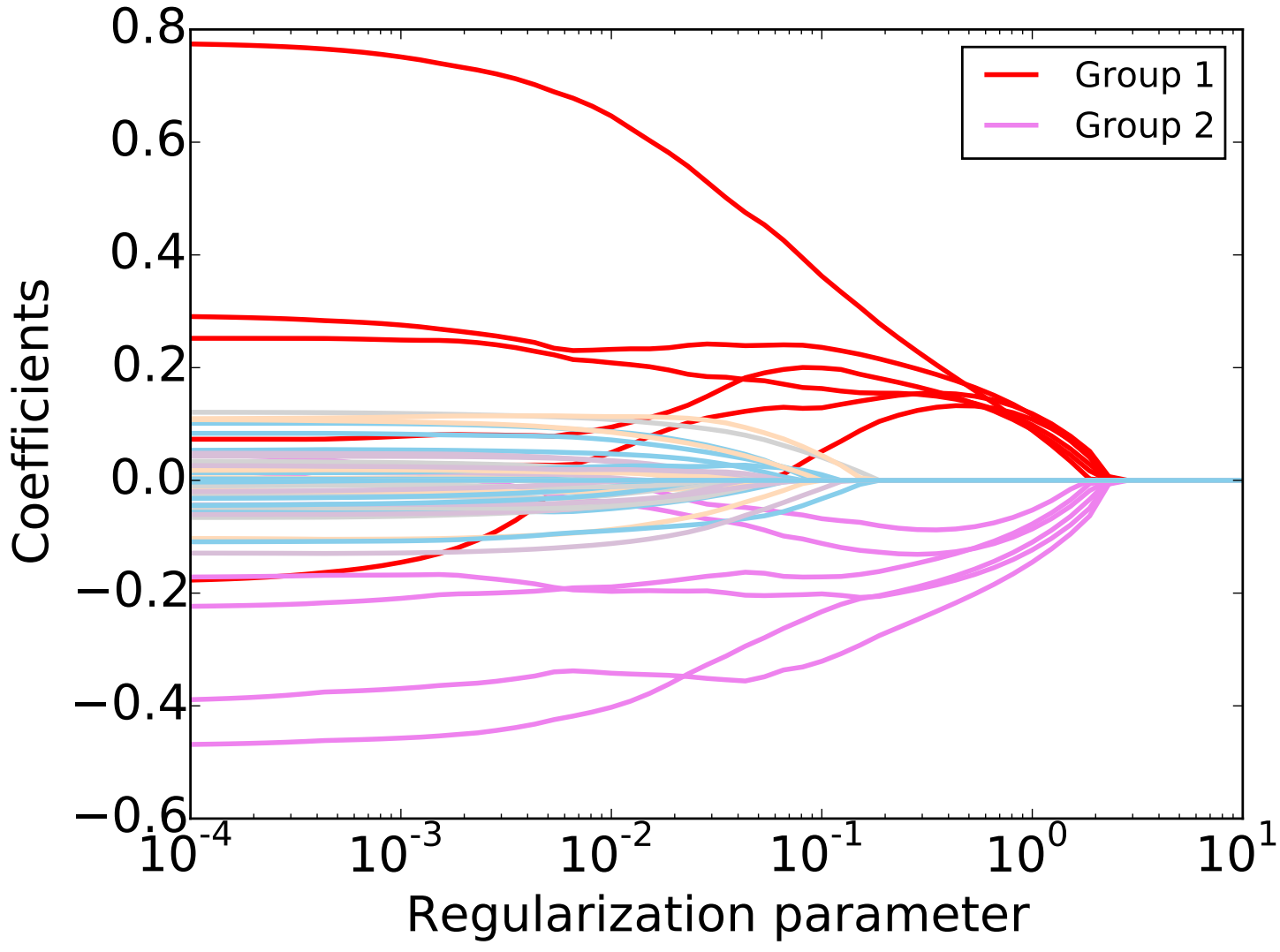


Elastic net

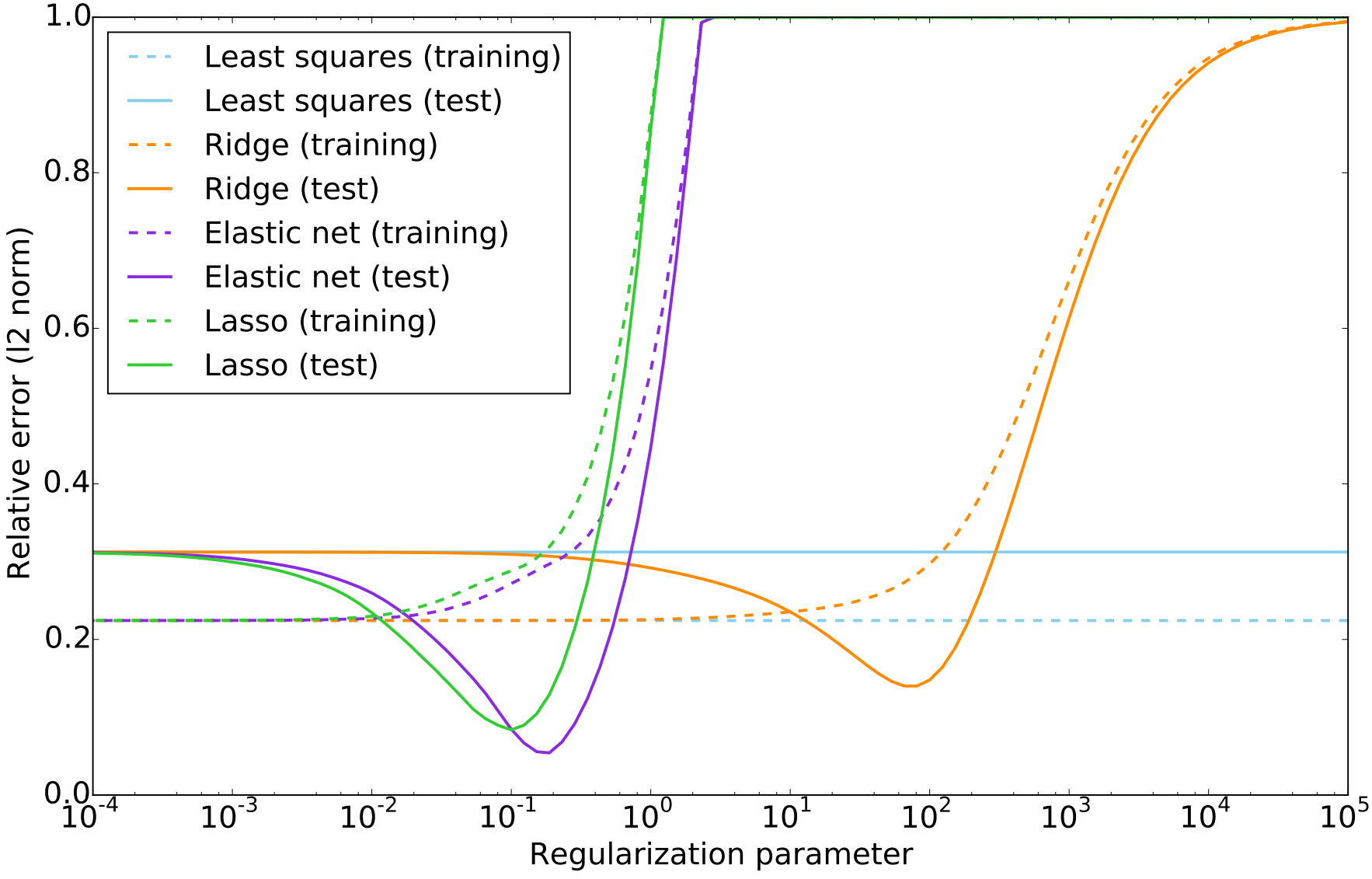
Combines lasso and ridge-regression penalties

$$\text{minimize} \quad \left|\left|y - X\tilde{\beta}\right|\right|_2^2 + \lambda \left(\frac{1-\alpha}{2} \left|\left|\tilde{\beta}\right|\right|_2^2 + \frac{\alpha}{2} \left|\left|\tilde{\beta}\right|\right|_1 \right)$$

Elastic-net path, $\alpha = 0.5$



Errors


Least squares (training)
Least squares (test)
Ridge (training)
Ridge (test)
Elastic net (training)
Elastic net (test)
Lasso (training)
Lasso (test)
Relative error (l2 norm)
Regularization parameter
1.0
0.8
0.6
0.4
0.2
0.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}

Group sparse structure

Predictors may be partitioned into k groups $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_m$

$$y \approx \beta_0 + X\beta = \beta_0 + \begin{bmatrix} X_{\mathcal{G}_1} & X_{\mathcal{G}_2} & \cdots & X_{\mathcal{G}_m} \end{bmatrix} \begin{bmatrix} \beta_{\mathcal{G}_1} \\ \beta_{\mathcal{G}_2} \\ \cdots \\ \beta_{\mathcal{G}_m} \end{bmatrix}$$

Aim: Select groups that are relevant

Multi-task learning

Several responses y_1, y_2, $\ldots$, y_k modeled with the same predictors

$$Y = \begin{bmatrix} y_1 & y_2 & \ldots & y_k \end{bmatrix} \approx B_0 + XB = B_0 + X \begin{bmatrix} \beta_1 & \beta_2 & \cdots & \beta_k \end{bmatrix}$$

Assumption: Responses depend on the same subset of predictors

Aim: Learn a group-sparse model

Mixed ℓ_1/ℓ_2 norm

Mixed ℓ_1/ℓ_2 norm

$$\left|\left|\tilde{\beta}\right|\right|_{1,2} := \sum_{i=1}^{k} \left|\left|\tilde{\beta}_{\mathcal{G}_i}\right|\right|_2$$

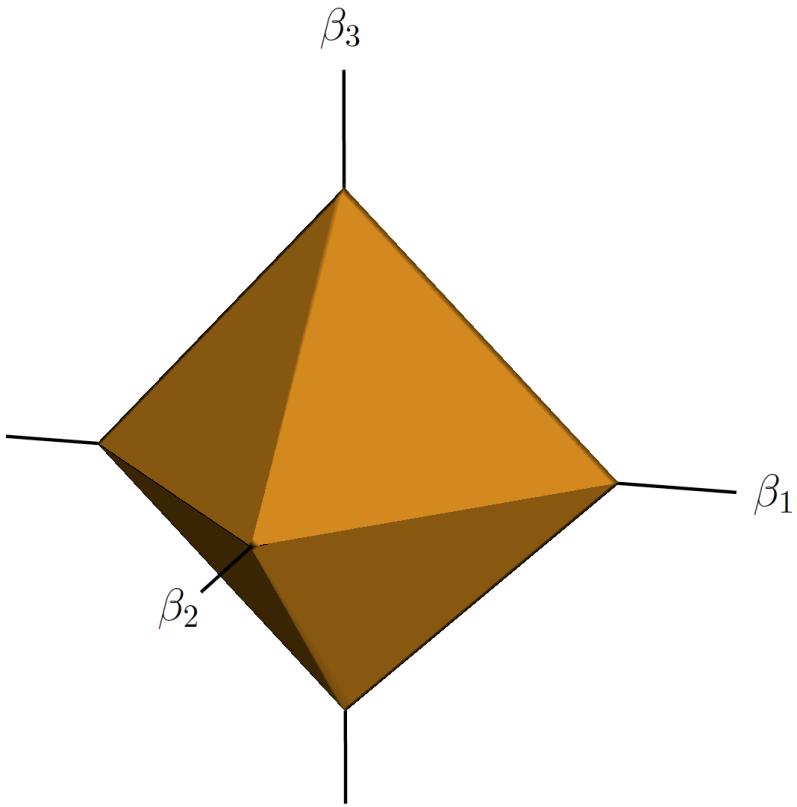
promotes group-sparse structure

For multitask learning

$$\left|\left|\tilde{B}\right|\right|_{1,2} := \sum_{i=1}^{k} \left|\left|\tilde{B}_{:i}\right|\right|_2$$

Geometric intuition $\mathcal{G}_1 = \{1, 2\}$ $\mathcal{G}_2 = \{3\}$

ℓ_1-norm ball



ℓ_1/ℓ_2-norm ball

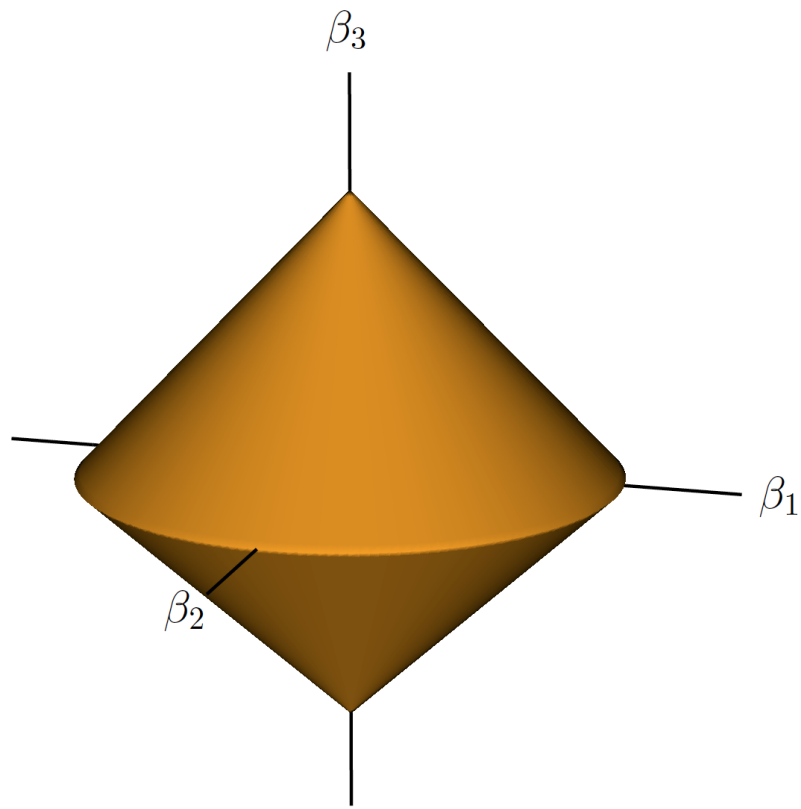


Figure taken from *Statistical Learning with Sparsity The Lasso and Generalizations* by Hastie, Tibshirani and Wainwright

Group and multi-task lasso

Group lasso

$$\text{minimize} \quad \left|\left|y - \tilde{\beta}_0 - X\tilde{\beta}\right|\right|_2^2 + \lambda \left|\left|\tilde{\beta}\right|\right|_{1,2}$$

Multi-task lasso

$$\text{minimize} \quad \left|\left|Y - \tilde{B}_0 - X\tilde{B}\right|\right|_{\mathsf{F}}^2 + \lambda \left|\left|\tilde{B}\right|\right|_{1,2}$$

Proximal gradient method

Method to solve the optimization problem

$$\text{minimize} \quad f(x) + g(x),$$

where f is differentiable and prox_g is tractable

Proximal-gradient iteration:

$$x^{(0)} = \text{arbitrary initialization}$$

$$x^{(k+1)} = \text{prox}_{\alpha_k g}\left(x^{(k)} - \alpha_k \nabla f\left(x^{(k)}\right)\right)$$

Subdifferential of ℓ_1/ℓ_2 norm

$q \in \mathbb{R}^p$ is a subgradient of the ℓ_1/ℓ_2 norm at $\beta \in \mathbb{R}^p$ if

$$q_{\mathcal{G}_i} = \frac{\beta_{\mathcal{G}_i}}{||\beta_{\mathcal{G}_i}||_2} \quad \text{if } \beta_{\mathcal{G}_i} \neq 0,$$

$$||q_{\mathcal{G}_i}||_2 \leq 1 \qquad \text{if } \beta_{\mathcal{G}_i} = 0$$

Proximal operator of ℓ_1/ℓ_2 norm

Proximal operator of ℓ_1/ℓ_2 norm is block soft-thresholding

$$\text{prox}_{\alpha\,||\cdot||_{1,2}}(\beta) = \mathcal{BS}_\alpha(\beta)$$

where $\alpha > 0$ and

$$\mathcal{BS}_\alpha(\beta)_{\mathcal{G}_i} := \begin{cases} \beta_{\mathcal{G}_i} - \alpha \frac{\beta_{\mathcal{G}_i}}{||\beta_{\mathcal{G}_i}||_2} & \text{if } ||\beta_{\mathcal{G}_i}||_2 \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

Iterative Shrinkage Thresholding for the ℓ_1/ℓ_2 norm

Proximal gradient method for the problem

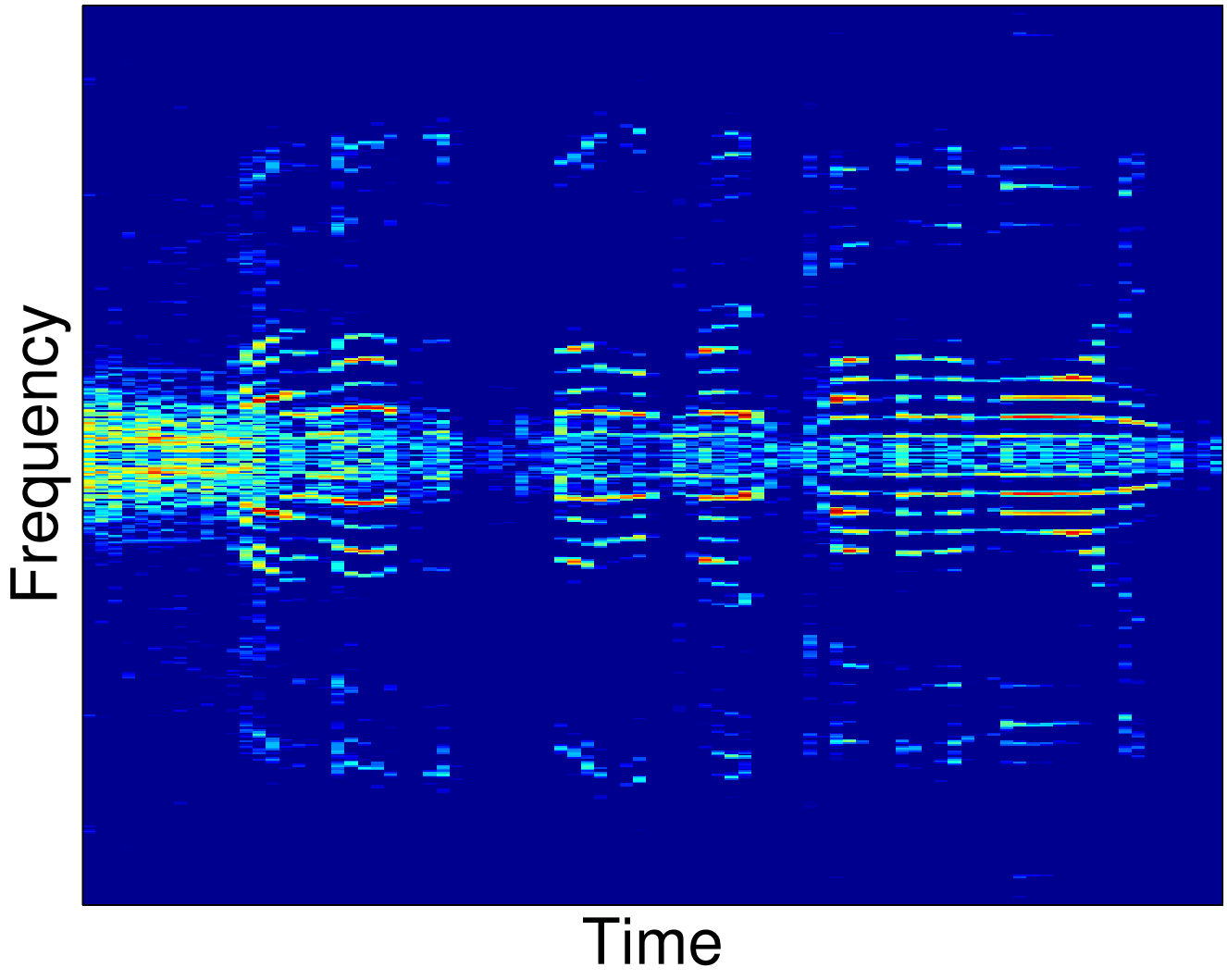
$$\text{minimize} \quad \frac{1}{2}\,||Ax - y||_2^2 + \lambda\ ||x||_{1,2}$$

ISTA iteration:

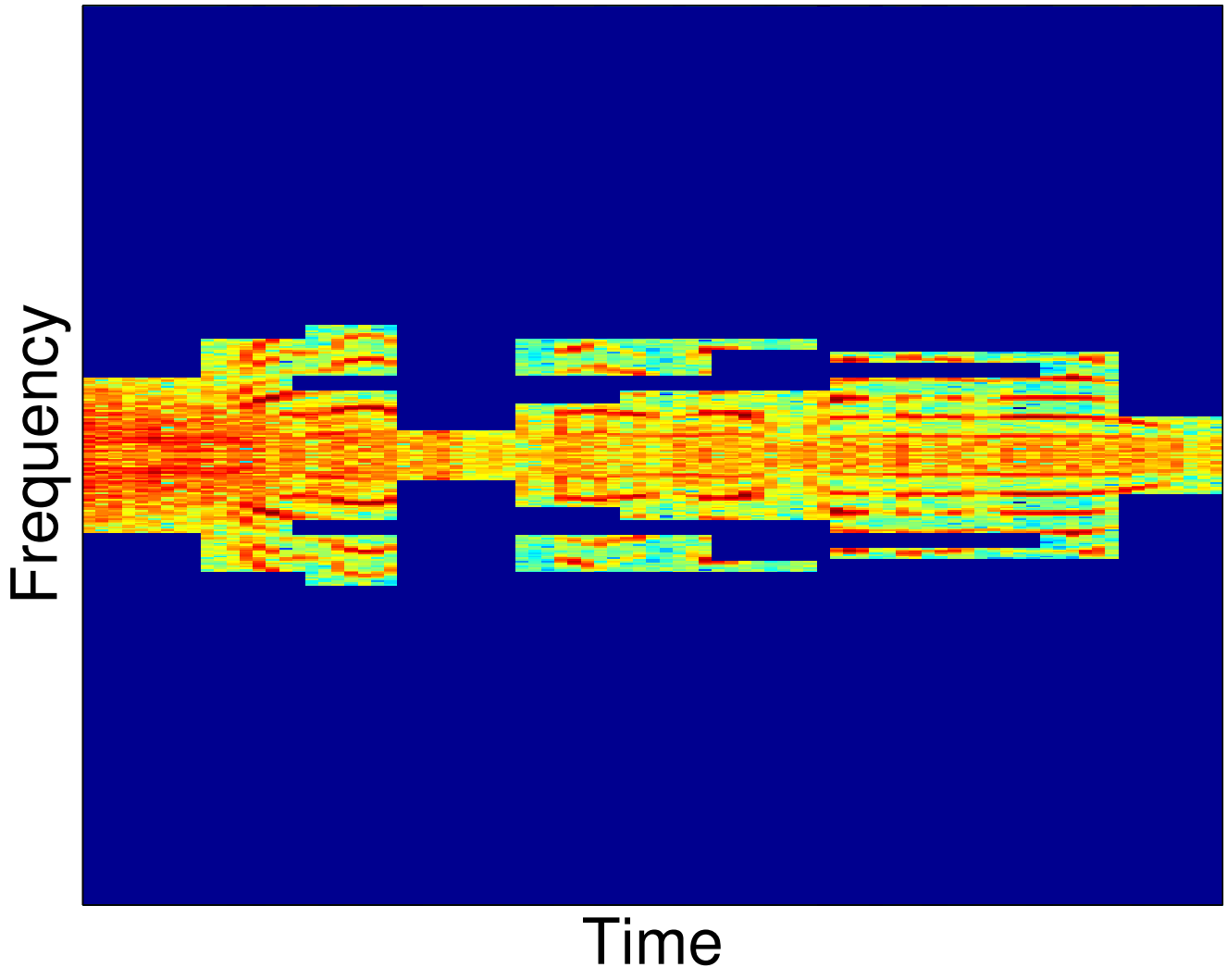
$$x^{(0)} = \text{arbitrary initialization}$$

$$x^{(k+1)} = \mathcal{BS}_{\alpha_k\,\lambda}\left(x^{(k)} - \alpha_k\,A^T\left(Ax^{(k)} - y\right)\right)$$

Speech denoising: STFT thresholding



Speech denoising: STFT block thresholding



Experiment

X_{train} and X_{test} are iid Gaussian $(0, 1)$, Z_{train} is iid Gaussian $(0, 0.5)$

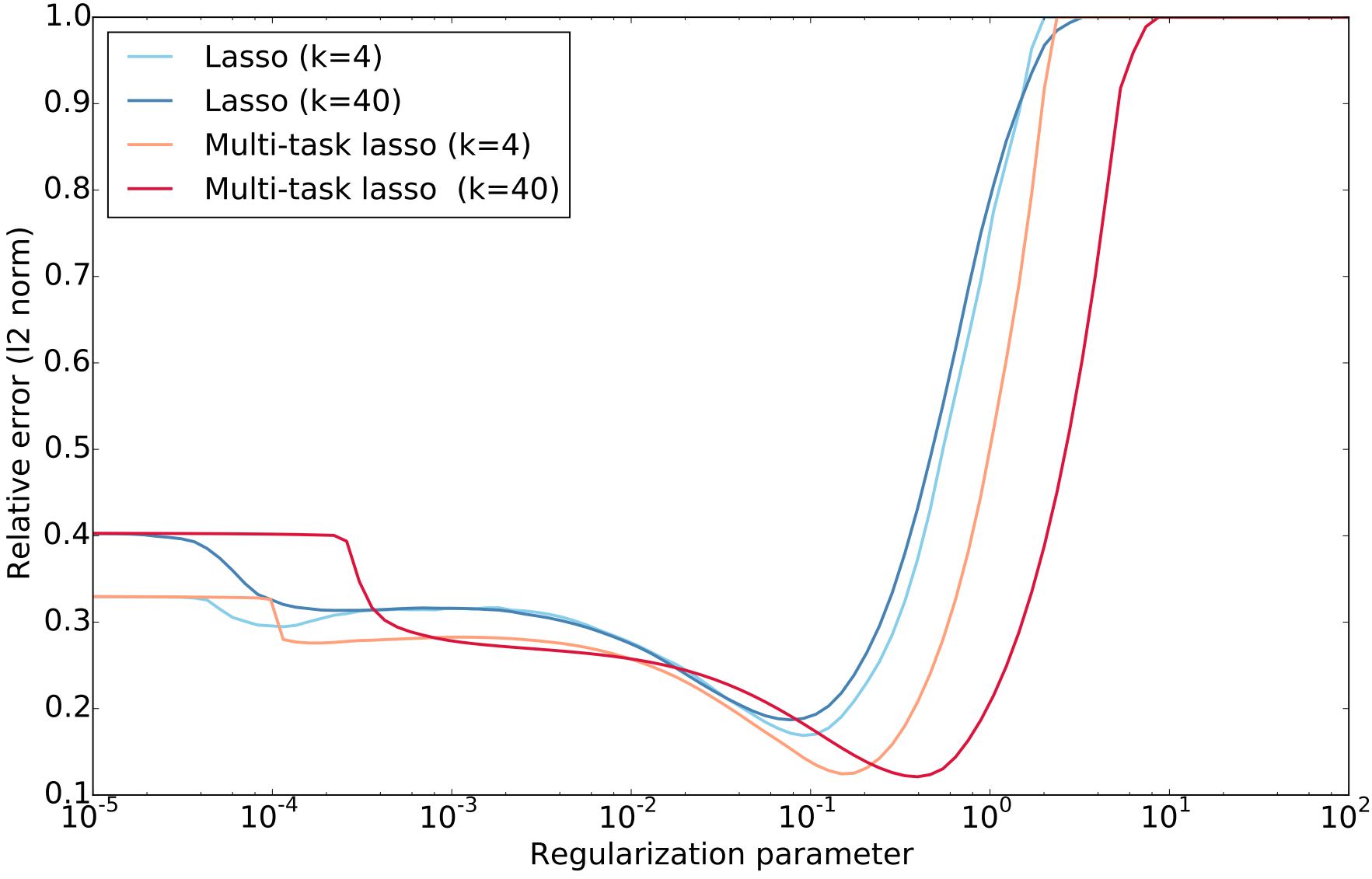
$B \in \mathbb{R}^{p \times k}$ has s nonzero rows, $p = 50$, $n = 100$

$$Y_{\text{train}} = X_{\text{train}}\, B + Z_{\text{train}} \qquad Y_{\text{test}} = X_{\text{test}}\, B$$

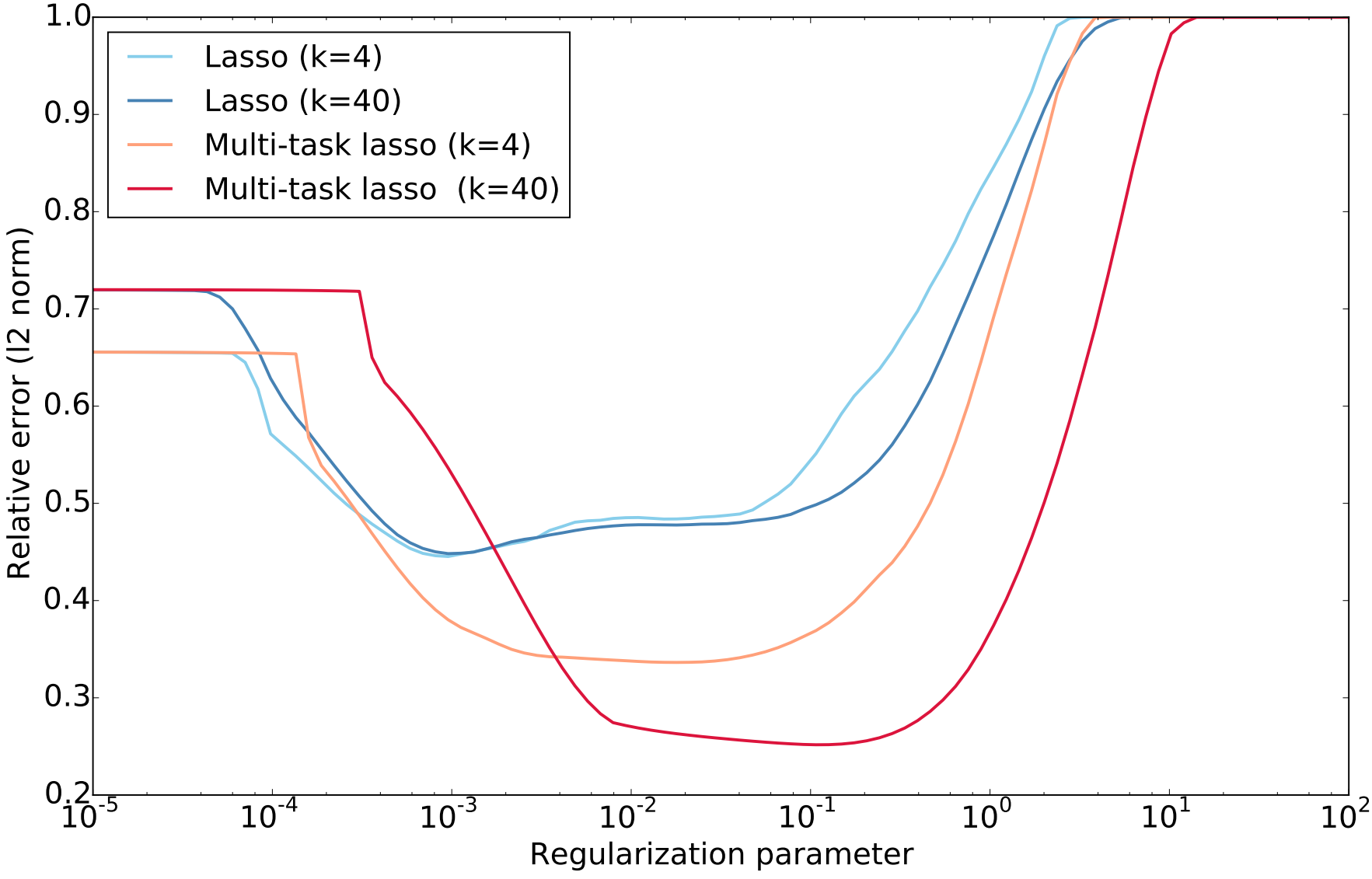
For an estimate $\hat{B}$ learnt from the training data

$$\text{error}_{\text{test}} = \frac{\left|\left| X_{\text{test}}\, \hat{B} - Y_{\text{test}} \right|\right|_2}{||Y_{\text{test}}||_2}$$

Test error, $s = 4$



Test error, $s = 30$



Magnitude of coefficients, $s = 30$ $k = 40$