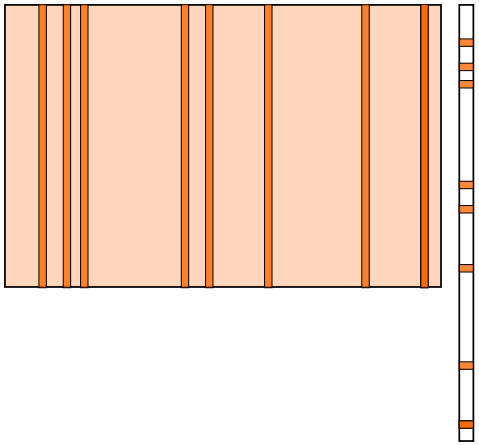
Sparsity

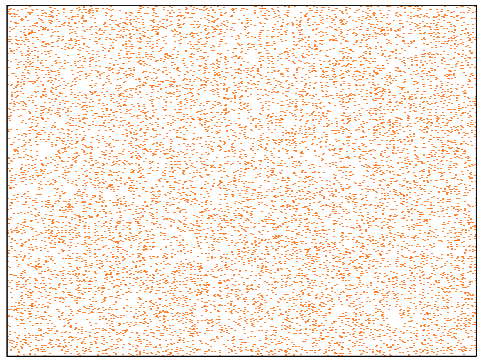
Rajen D. Shah (Statistical Laboratory, University of Cambridge)

Open University
Workshop on Multivariate Analysis Today
18 May 2015

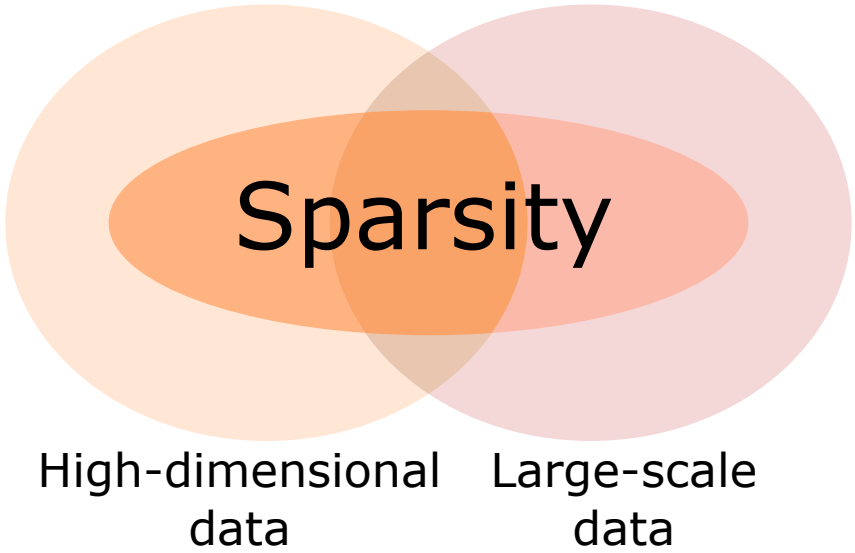
Two types of sparsity



(a) Signal sparsity



(b) Data sparsity



High-dimensional data

- Consider a regression setting:
  - n observations of a response Y_i.
  - p covariates $x_i = (x_{i1}, \ldots, x_{ip})^T$. Let $\mathbf{X}$ be the $n \times p$ design matrix whose i^{th} row is x_i.
- In the classical setting, $p < n$. Here we have in mind $p \gg n$, and p perhaps in the order of thousands or more.
- Such high-dimensional data is becoming increasingly common in many modern statistical applications e.g. gene expression data, GWAS.

Ridge regression (Hoerl and Kennard, 1970)

- $p > n$: OLS coefficients will not be unique. We need to *regularise*.
- Ridge regression solves a penalised optimisation:

$$(\hat{\mu}_\lambda^R, \hat{\beta}_\lambda^R) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \{\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\}.$$

- Equivalent form after centring $\mathbf{Y}$ and $\mathbf{X}$:

$$\begin{aligned}\hat{\beta}_\lambda^R &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\mathbf{Y}.\end{aligned}$$

Optimality of ridge regression

- Linear model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\mathbf{I})$
- Bayesian interpretation of ridge regression:
  - Prior of $N_p(0, \lambda^{-1}/\sigma^2\mathbf{I})$ on $\boldsymbol{\beta}^*$ (σ^2 treated known).
  - $\hat{\boldsymbol{\beta}}^R_\lambda$ is the posterior mean.

Optimality of ridge regression

- Linear model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I})$
- Bayesian interpretation of ridge regression:
  - Prior of $N_p(0, \lambda^{-1}/\sigma^2 \mathbf{I})$ on $\boldsymbol{\beta}^*$ (σ^2 treated known).
  - $\hat{\boldsymbol{\beta}}^R_\lambda$ is the posterior mean.
- Ridge regression is optimal in terms of mean-squared error under a $N_p(0, \lambda^{-1}/\sigma^2)$ prior on $\boldsymbol{\beta}^*$.

The Lasso (Tibshirani, 1996)

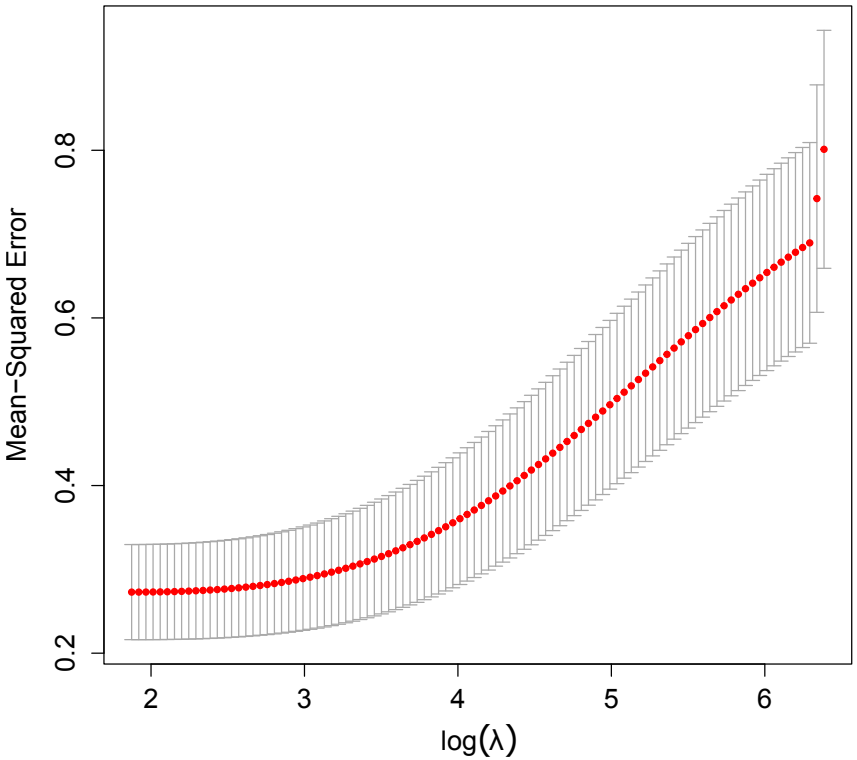
The Lasso solves

$$(\hat{\mu}^L_\lambda, \hat{\boldsymbol{\beta}}^L_\lambda) = \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \{\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^2/(2n) + \lambda\|\boldsymbol{\beta}\|_1\}.$$

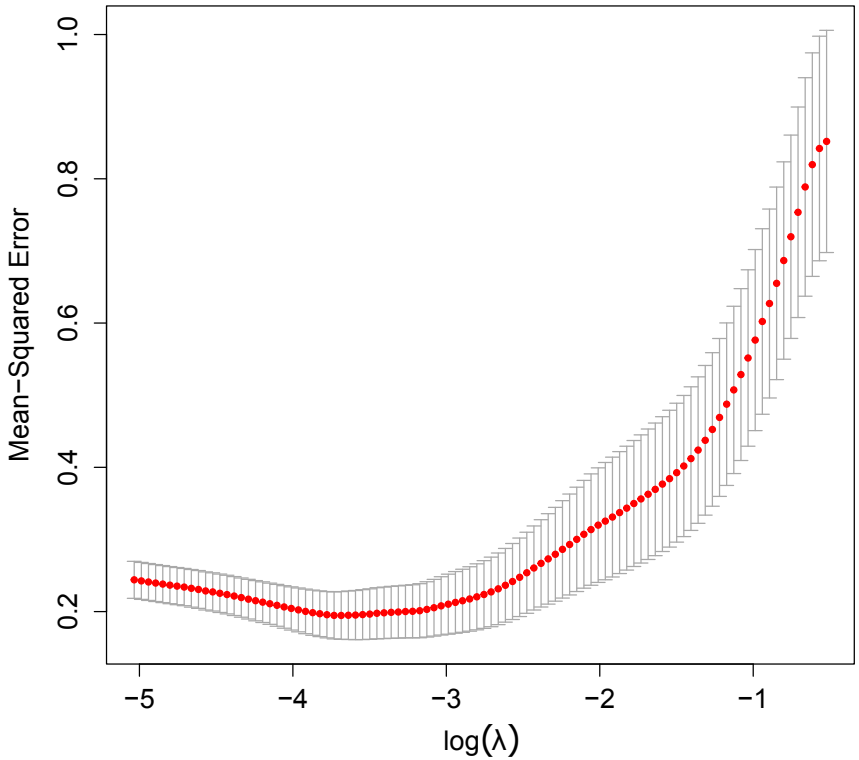
$$\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|.$$

Performance of ridge regression in practice

Gene expression data, $n = 71$ observations of $p = 4088$ predictors.
Response is riboflavin production by *Bacillus subtilis*.



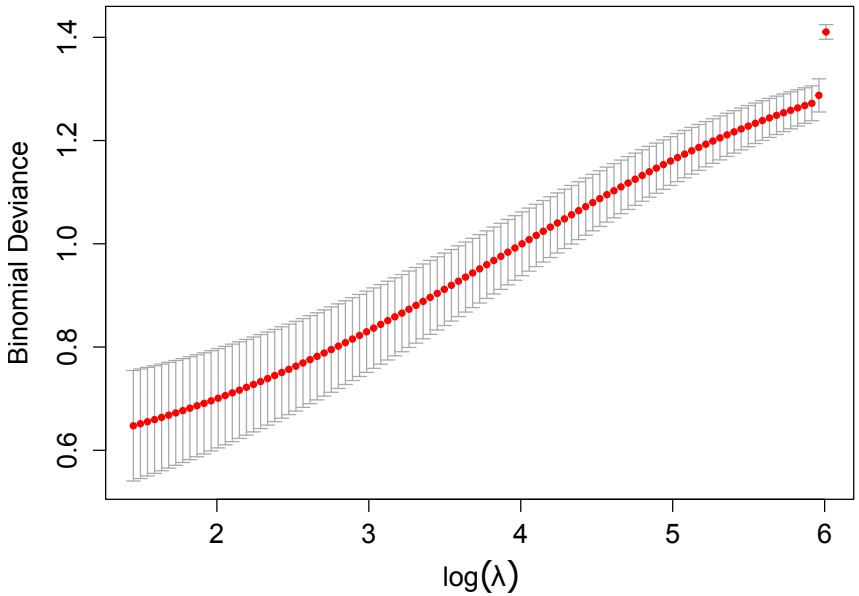
(a) Ridge regression



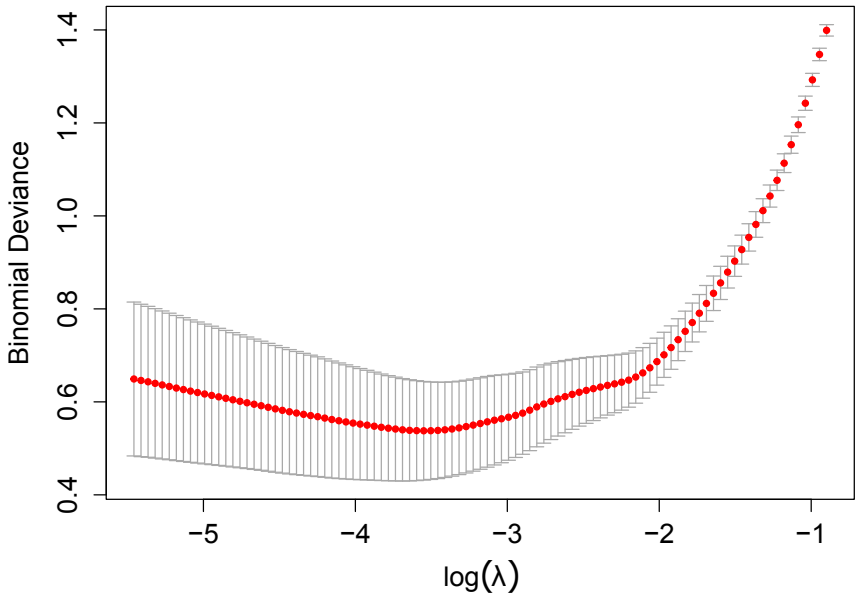
(b) Lasso regression (Tibshirani, 1996)

Performance of ridge regression in practice

Prostate cancer gene expression data. 52 tumour samples, 50 normal samples ($n = 102$) with $p = 6033$ predictors.



(a) Ridge regression



(b) Lasso regression

Signal sparsity

- Typically for high-dimensional data, the Lasso beats ridge regression in terms of prediction error.

Signal sparsity

- Typically for high-dimensional data, the Lasso beats ridge regression in terms of prediction error.
- The normal prior on β^* is often not appropriate.
- Often there is a belief that most of the predictors are irrelevant for determining the response i.e. β^* is sparse.

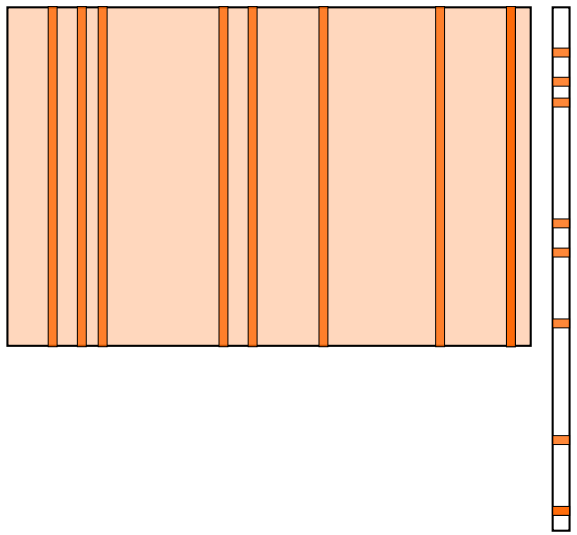


Figure: Schematic of signal $\mathbf{X}\beta^*$

Best subsets regression

If the signal is sparse, best subsets regression may seem natural.

$$\underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \left\{ \|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{k=1}^{p} \mathbb{1}_{\{\beta_k \neq 0\}} \right\}.$$

Best subsets regression

If the signal is sparse, best subsets regression may seem natural.

$$\underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \left\{ \|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{k=1}^{p} \mathbb{1}_{\{\beta_k \neq 0\}} \right\}.$$

Optimisation problem typically infeasible for $p > 50$ as problem is not convex.

Best subsets regression

If the signal is sparse, best subsets regression may seem natural.

$$\underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \left\{ \|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{k=1}^{p} \mathbb{1}_{\{\beta_k \neq 0\}} \right\}.$$

Optimisation problem typically infeasible for $p > 50$ as problem is not convex.

The Lasso solves the closest convex approximation to the objective above.

Constrained form of Lasso

Note that if

$$(\hat{\mu}_\lambda^L, \hat{\beta}_\lambda^L) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\arg\min} \{\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\beta\|_2^2/(2n) + \lambda\|\beta\|_1\}$$

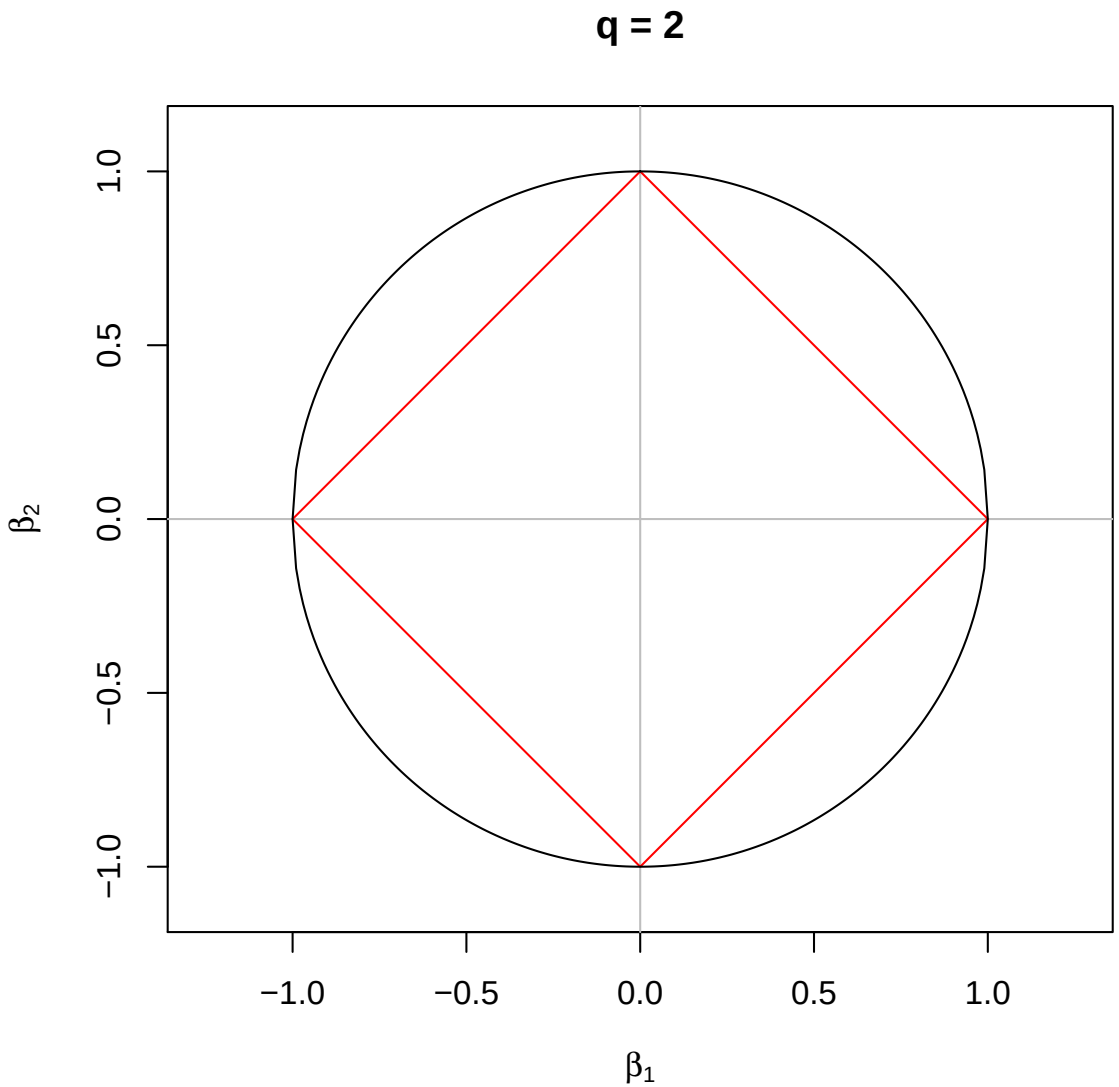
then $(\hat{\mu}_\lambda^L, \hat{\beta}_\lambda^L)$ minimises

$$\|\mathbf{Y} - \mu\mathbf{1} - \mathbf{X}\beta\|_2^2$$

subject to $\|\beta\|_1 \leq \|\hat{\beta}_\lambda^L\|_1$.

ℓ_q balls

Consider penalty functions $\propto \|\boldsymbol{\beta}\|_q = \left(\sum_{k=1}^{p} \beta_k^q\right)^{1/q}$ and $p = 2$.



Lasso coefficients are sparse

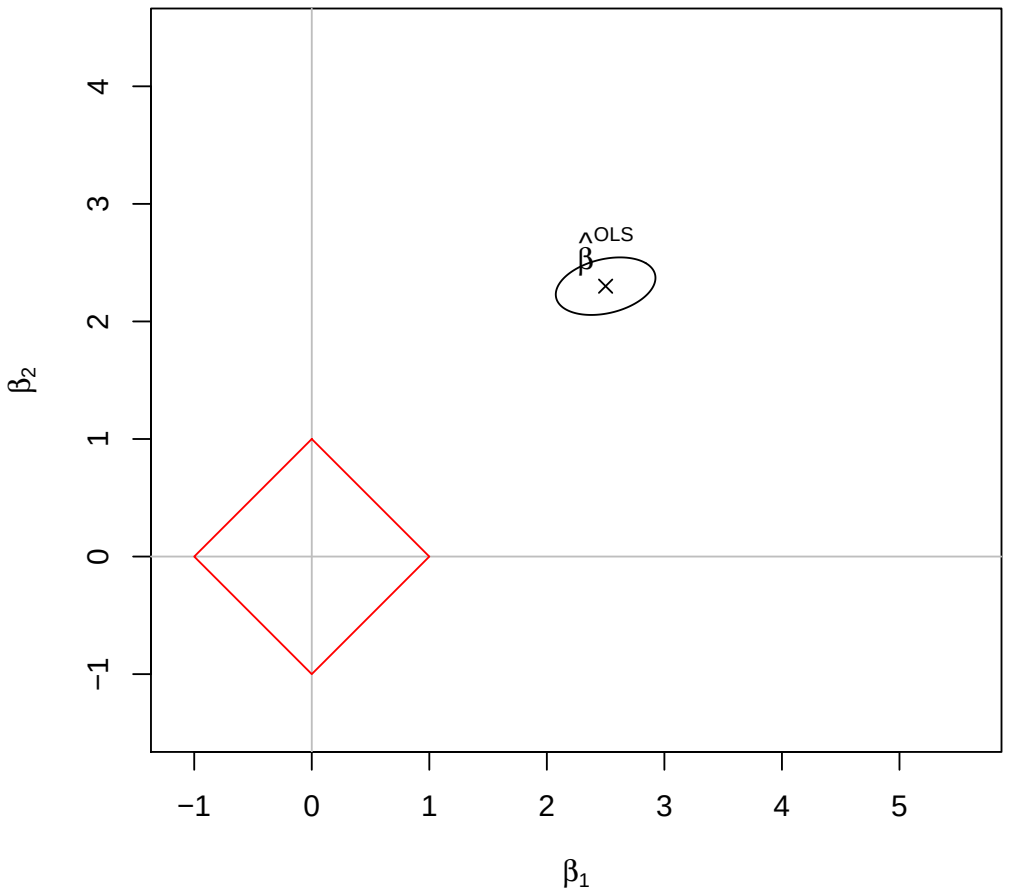


Figure: Contours of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ are ellipses centred at $\hat{\beta}^{OLS}$.

Lasso coefficients are sparse

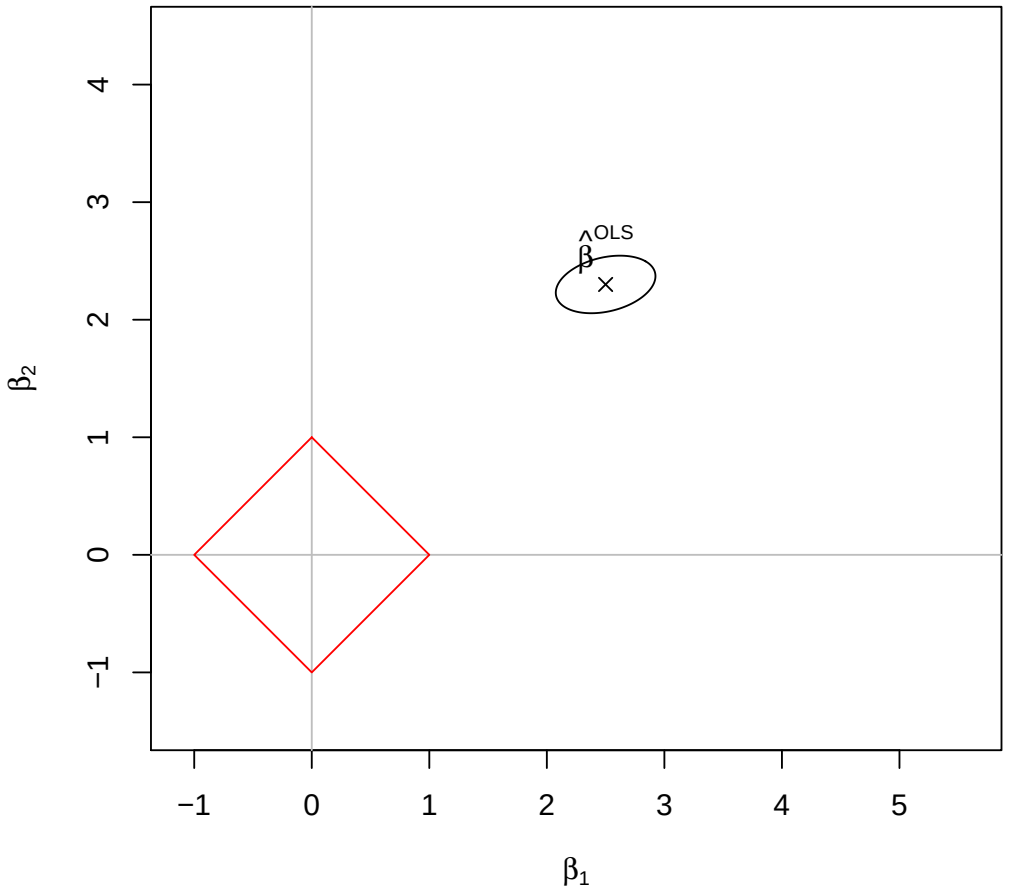
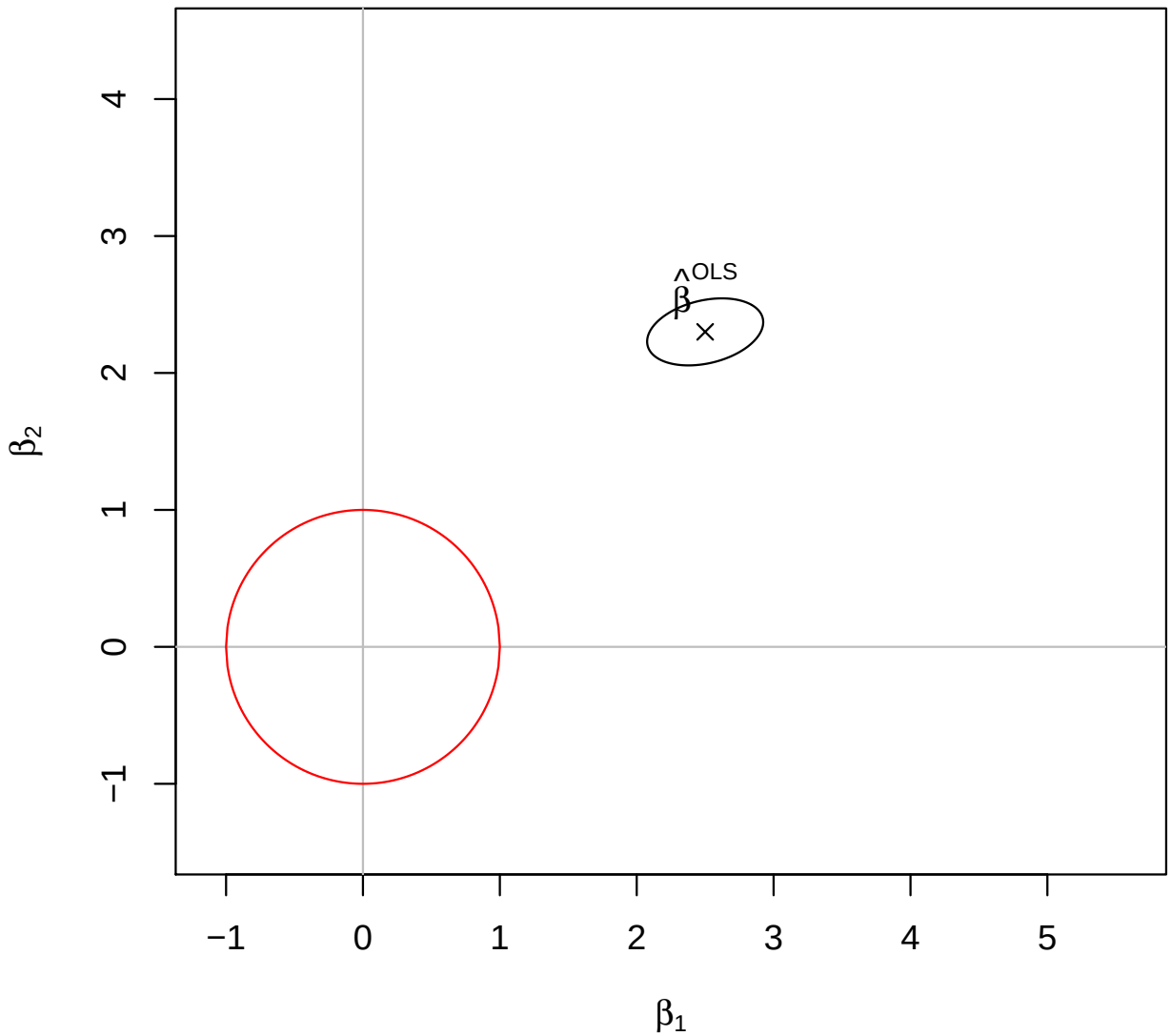


Figure: Contours of $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$ are ellipses centred at $\hat{\beta}^{OLS}$.

Ridge regression coefficients are always non-zero



Benefits of sparse coefficients

- Typically a sparse model fits well for high-dimensional data.

Benefits of sparse coefficients

- Typically a sparse model fits well for high-dimensional data.
- Sparse models can be easier to interpret.

Benefits of sparse coefficients

- Typically a sparse model fits well for high-dimensional data.
- Sparse models can be easier to interpret.
- In order to predict the response for a new observation, we only need measurements of a few covariates.

Benefits of sparse coefficients

- Typically a sparse model fits well for high-dimensional data.
- Sparse models can be easier to interpret.
- In order to predict the response for a new observation, we only need measurements of a few covariates.
- Inner product $\mathbf{x}^T \hat{\beta}$ for new data point $\mathbf{x} \in \mathbb{R}^p$ fast to compute.

Prediction error for the Lasso

Consider the normal linear model

$$\mathbf{Y} = \mathbf{X}\beta^* + \varepsilon \qquad \varepsilon \sim N_n(0, \sigma^2 \mathbf{I}).$$

Theorem

Let $\hat{\beta}$ be the Lasso solution when

$$\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}.$$

With probability at least $1 - p^{-(A^2/2-1)}$

$$\frac{1}{n}\|\mathbf{X}(\beta^* - \hat{\beta})\|_2^2 \leq 4A\sigma\sqrt{\frac{\log(p)}{n}}\|\beta^*\|_1.$$

A faster rate

Under assumptions on $\mathbf{X}$ which in particular prevent columns from being too correlated with each other, we have a stronger result. Suppose β^* has s non-zero components.

Theorem

Let $\hat{\beta}$ be the Lasso solution when

$$\lambda = A\sigma\sqrt{\log(p)/n}$$

With probability at least $1 - p^{-(A^2/8-1)}$,

$$\frac{1}{n}\|\mathbf{X}(\beta^* - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta} - \beta^*\|_1 \leq \frac{16\lambda^2 s}{\phi^2} = \frac{16A^2\log(p)}{\phi^2}\frac{\sigma^2 s}{n},$$

where ϕ^2 is a constant depending on the design.

Extensions

- ℓ_1-penalised generalised linear models.

Extensions

- ℓ_1-penalised generalised linear models.
- Structural penalties e.g. the group Lasso (Yuan & Lin, 2006): $G_1 \cup \cdots \cup G_q = \{1, \ldots, p\}$, multipliers $m_1, \ldots, m_q$,

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{G_j}\|_2.$$

Extensions

- ℓ_1-penalised generalised linear models.
- Structural penalties e.g. the group Lasso (Yuan & Lin, 2006): $G_1 \cup \cdots \cup G_q = \{1, \ldots, p\}$, multipliers $m_1, \ldots, m_q$,

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{G_j}\|_2.$$

- 'De-biasing' the Lasso e.g. using non-convex penalty functions.

Extensions

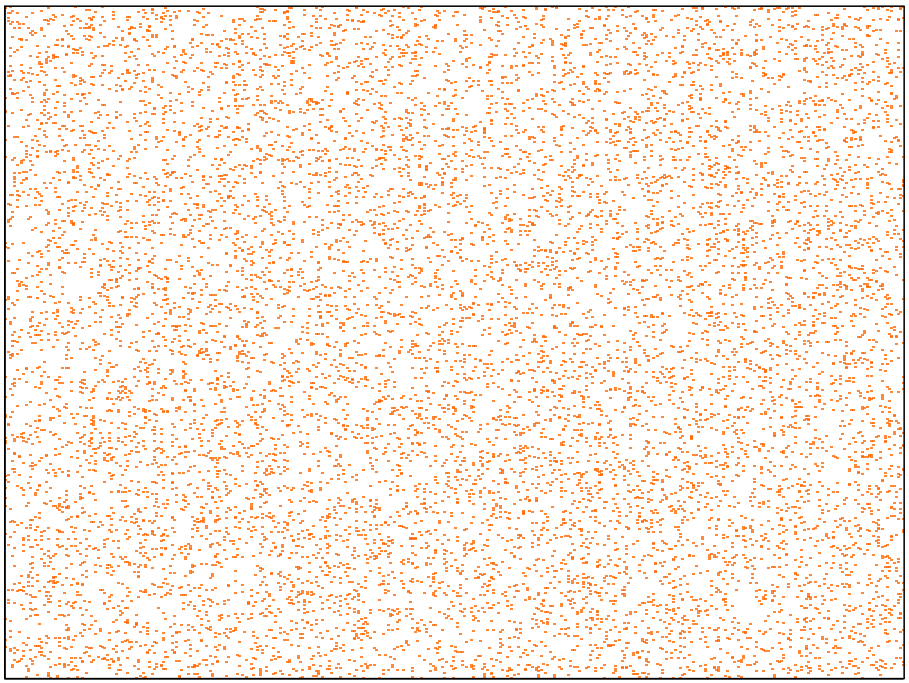
- ℓ_1-penalised generalised linear models.
- Structural penalties e.g. the group Lasso (Yuan & Lin, 2006): $G_1 \cup \cdots \cup G_q = \{1, \ldots, p\}$, multipliers $m_1, \ldots, m_q$,

$$\lambda \sum_{j=1}^{q} m_j \|\beta_{G_j}\|_2.$$

- 'De-biasing' the Lasso e.g. using non-convex penalty functions.
- Inference.

Large-scale data

- Large p, large n.
- Data is not the only relevant resource to consider. The computational budget is also an issue (both memory and computing power).
- In many large-scale applications, the design matrix $\mathbf{X}$ is sparse.



Text analysis

Given a collection of documents, construct variables which count the number of occurrences of different words. Can add variables giving the frequency of consecutive pairs of words (bigrams) or consecutive triples of words (trigrams).

	"statistics"	"multivariate analysis"	"Big Data"	$\cdots$
Doc 1	4	0	4	$\cdots$
Doc 2	3	2	4	$\cdots$
Doc 3	0	0	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Dimension reduction

- Our computational budget may mean that OLS and Ridge regression are computationally infeasible.

Dimension reduction

- Our computational budget may mean that OLS and Ridge regression are computationally infeasible.
- The computer science literature has a variety of algorithms to form a low-dimensional "sketch" of the design matrix i.e. a mapping

$$\underset{n \times p}{\mathbf{X}} \mapsto \underset{n \times L,}{\mathbf{S}} \qquad L \ll p.$$

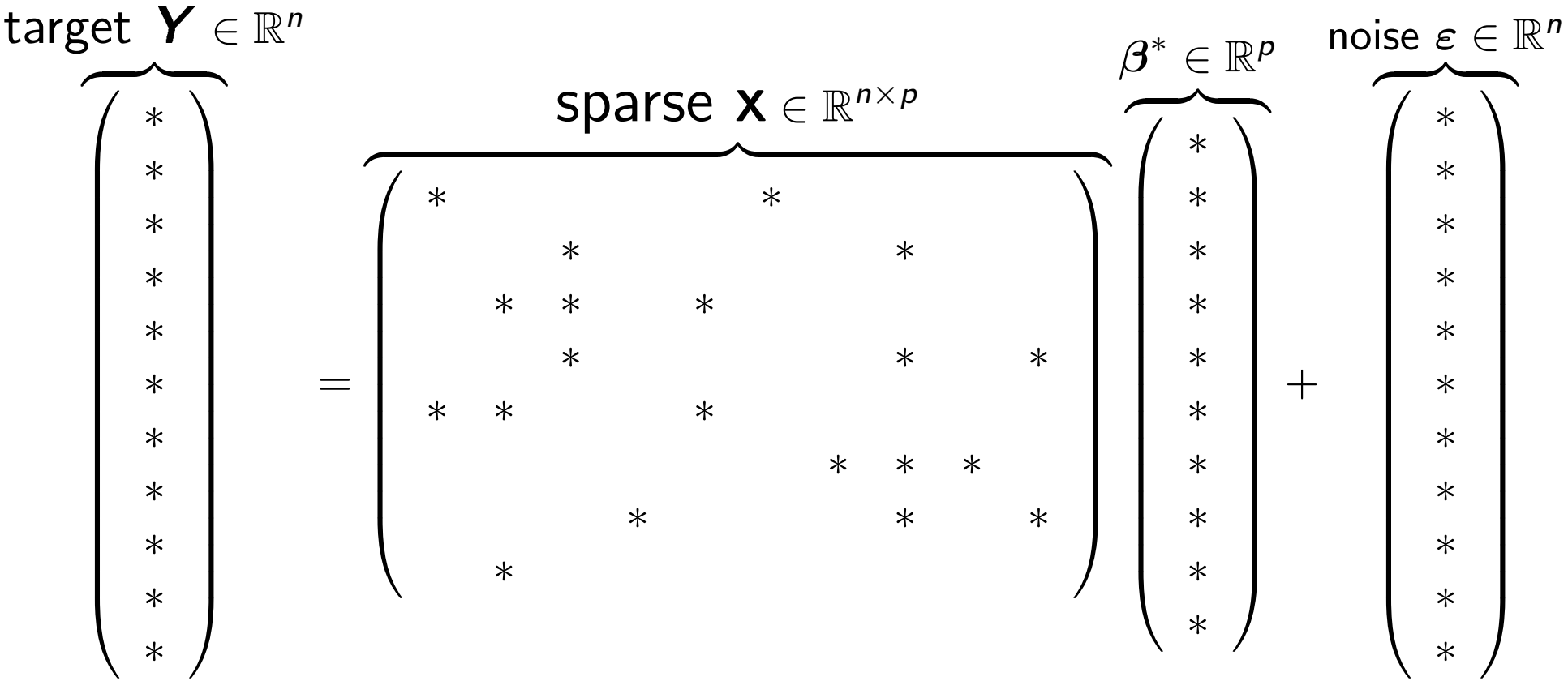
Dimension reduction

- Our computational budget may mean that OLS and Ridge regression are computationally infeasible.
- The computer science literature has a variety of algorithms to form a low-dimensional "sketch" of the design matrix i.e. a mapping

$$\underset{n \times p}{\mathbf{X}} \mapsto \underset{n \times L,}{\mathbf{S}} \qquad L \ll p.$$

- The idea is then to perform the regression on $\mathbf{S}$ rather than the larger $\mathbf{X}$.

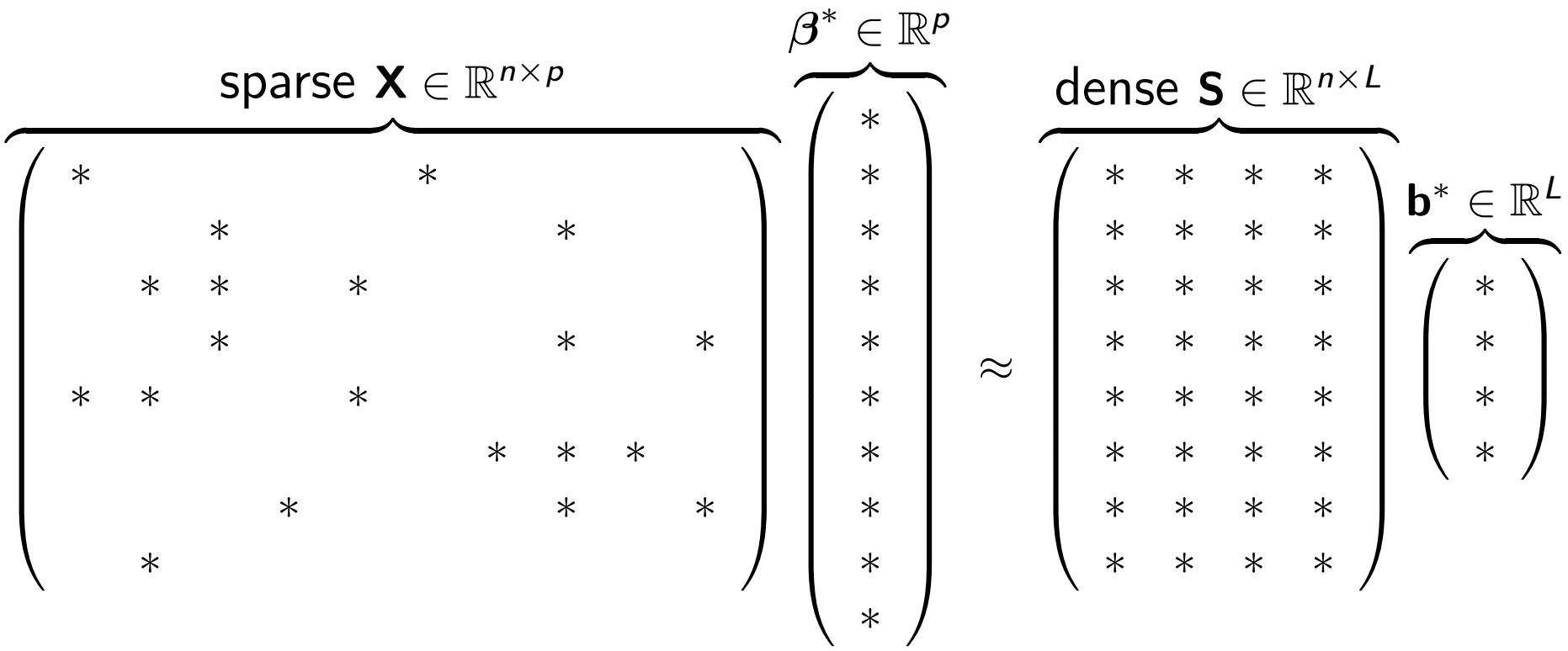
Linear model with sparse design



Non-zero entries are marked with $*$.

Linear model with sparse design

Can we safely reduce our sparse p-dimensional problem to a (possibly dense) L-dimensional one with $L \ll p$?



Sketching methods

- PCA may be too expensive to compute.
- Random projections e.g. $\mathbf{S} = \mathbf{XA}$, $\mathbf{A}$ $p \times L$ with i.i.d. Gaussian entries.

Sketching methods

- PCA may be too expensive to compute.
- Random projections e.g. $\mathbf{S} = \mathbf{XA}$, $\mathbf{A}$ $p \times L$ with i.i.d. Gaussian entries.
- *b-bit min-wise hashing* (Li and König, 2011).
  - Dimension reduction for a binary $\mathbf{X}$
  - Based on earlier technique of *min-wise hashing* (Broder, 1997).
  - Impressive empirical results.
- Shah & Meinshausen (2015) study a variant, *random-sign hashing* that also deals with continuous data.

Random-sign hashing

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} & 1 & & 3 \\ & & 6 & 2 \\ 3 & & 1 & \\ & 2 & 5 & \\ 2 & 4 & & \end{pmatrix} \end{array}$$

Random-sign hashing

$$\mathbf{X} = \begin{array}{c} \pi_1 \quad 2 \quad 3 \quad 1 \quad 4 \\ \begin{pmatrix} 1 & & & 3 \\ & 6 & & 2 \\ & 1 & 3 & \\ 2 & 5 & & \\ 4 & & 2 & \end{pmatrix} \end{array} \mapsto \mathbf{H} = \begin{pmatrix} 2 \\ 3 \\ 3 \\ 2 \\ 2 \end{pmatrix} \quad \mathbf{S}' = \begin{pmatrix} 1 \\ 6 \\ 1 \\ 2 \\ 4 \end{pmatrix}$$

First columns of $\mathbf{H}$ and $\mathbf{S}'$ generated by the random permutation π_1 of the variables.

Random-sign hashing

$$\mathbf{X} = \begin{array}{c} \pi_2 \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} 3 & 1 & 4 & 2 \end{array} \\ \begin{pmatrix} & & 3 & 1 \\ 6 & & 2 & \\ 1 & 3 & & \\ 5 & & & 2 \\ & 2 & & 4 \end{pmatrix} \end{array} \mapsto \mathbf{H} = \begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 2 & 3 \\ 1 & 3 \\ 1 & 1 \end{pmatrix} \quad \mathbf{S}' = \begin{pmatrix} 1 & 3 \\ 6 & 6 \\ 1 & 1 \\ 2 & 5 \\ 4 & 2 \end{pmatrix}$$

Random-sign hashing

Choose **random sign assignments** $\{1, \ldots, p\} \to \{-1, 1\} : k \mapsto \Psi_{kl}$ independently for all columns $l = 1, \ldots, L$.

Suppose $\Psi_{11} = +$, $\Psi_{21} = -$ and $\Psi_{42} = -$, $\Psi_{32} = +$, $\Psi_{12} = -$.

$$\mathbf{H} = \begin{pmatrix} 1 & 4 \\ 2 & 3 \\ 2 & 3 \\ 1 & 3 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{S}' = \begin{pmatrix} 1 & 3 \\ 6 & 6 \\ 1 & 1 \\ 2 & 5 \\ 4 & 2 \end{pmatrix} \mapsto \mathbf{S} = \begin{pmatrix} 1 & -3 \\ -6 & 6 \\ -1 & 1 \\ 2 & 5 \\ 4 & -2 \end{pmatrix}$$

Random-sign hashing: summary

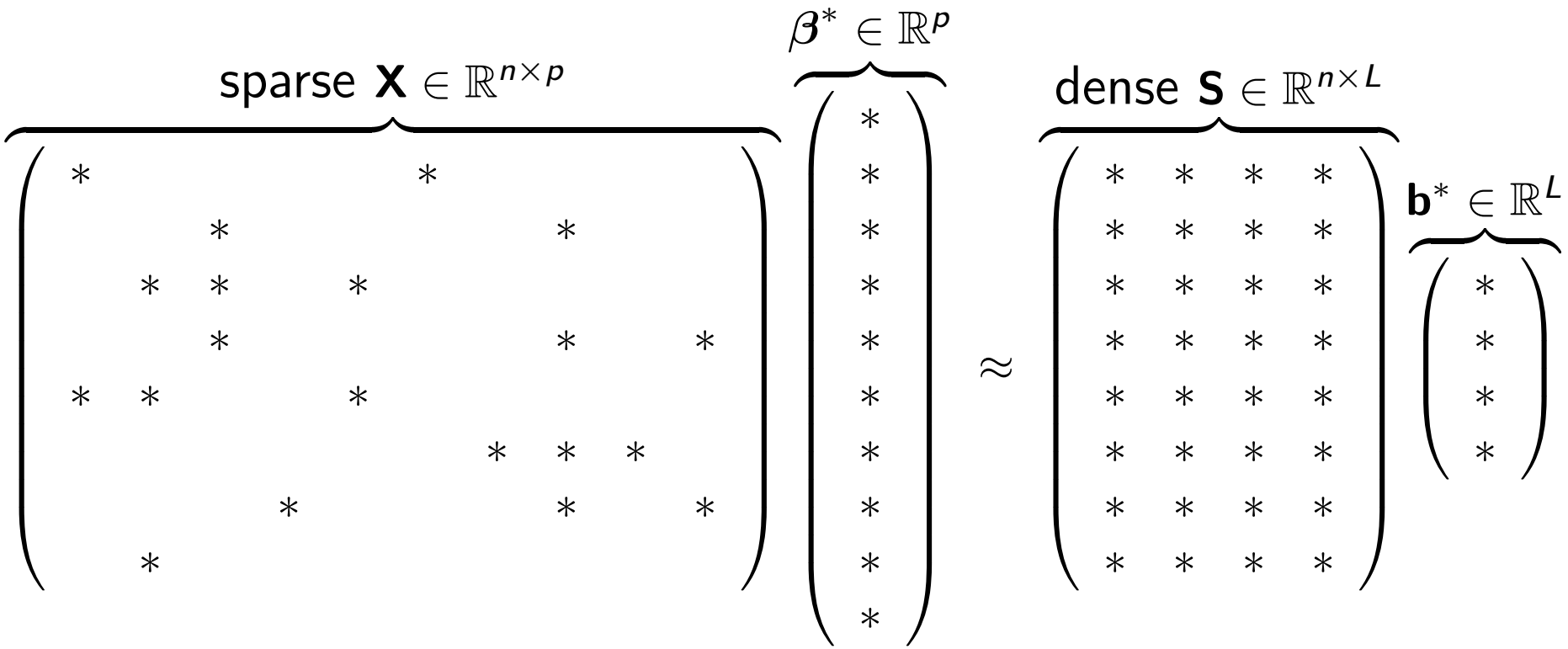
We get $n \times L$ matrices $\mathbf{H}$, and $\mathbf{S}$ given by

$$H_{il} = \arg\min_{k \in \mathbf{z}_i} \pi_l(k)$$
$$S_{il} = \Psi_{H_{il} l} X_{i H_{il}},$$

where Ψ_{hl} is the random sign of the h^{th} variable in the l^{th} permutation.

Approximation error

Can we *construct* a $\mathbf{b}^* \in \mathbb{R}^L$ such that $\mathbf{X}\beta^*$ is close to $\mathbf{S}\mathbf{b}^*$ on average i.e. such that $\mathbb{E}\|\mathbf{X}\beta^* - \mathbf{S}\mathbf{b}^*\|_2^2$ is small?



Approximation error

Is there a $\mathbf{b}^*$ such that we have unbiasedness $\mathbb{E}(\mathbf{S}_l b_l^*) = \mathbf{X}\beta^*/L$?

$$\overbrace{\begin{pmatrix} * & & & & * & & & \\ & & * & & & * & & \\ & * & * & * & & & & \\ & & * & & & * & & * \\ * & * & & * & & & & \\ & & & & * & * & * & \\ & & & * & & * & & * \\ & * & & & & & & \end{pmatrix}}^{\text{sparse } \mathbf{X} \in \mathbb{R}^{n\times p}} \overbrace{\begin{pmatrix} * \\ * \\ * \\ * \\ * \\ * \\ * \\ * \\ * \\ * \end{pmatrix}}^{\beta^* \in \mathbb{R}^p} \overset{?}{=} \frac{1}{L}\mathbb{E}_{\pi,\psi} \left[\overbrace{\begin{pmatrix} * \\ * \\ * \\ * \\ * \\ * \\ * \\ * \end{pmatrix}}^{\mathbf{S}_l \in \mathbb{R}^{n\times 1}} \overbrace{\begin{pmatrix} * \end{pmatrix}}^{b^*{}_l \in \mathbb{R}} \right]$$

Approximation error

- Assume for now that there are $q \leq p$ non-zero entries in each row of $\mathbf{X}$. Unequal row sparsity can also be dealt with.

Approximation error

- Assume for now that there are $q \leq p$ non-zero entries in each row of $\mathbf{X}$. Unequal row sparsity can also be dealt with.
- Consider one permutation with min-hash value H_i for $i = 1, \ldots, n$ and random signs ψ_k, $k = 1, \ldots, p$.

$$\mathbb{E}_{\pi,\psi}\left[\overbrace{\begin{pmatrix} \psi_{H_1} X_{1H_1} \\ \psi_{H_2} X_{2H_2} \\ \ldots \\ \ldots \\ \ldots \end{pmatrix}}^{\mathbf{S}\in\mathbb{R}^{n\times 1}} \overbrace{\left(q \sum_{k=1}^{p} \beta_k^* \psi_k\right)}^{=:b^*\in\mathbb{R}^1}\right] =$$

Approximation error

- Assume for now that there are $q \leq p$ non-zero entries in each row of $\mathbf{X}$. Unequal row sparsity can also be dealt with.
- Consider one permutation with min-hash value H_i for $i = 1, \ldots, n$ and random signs ψ_k, $k = 1, \ldots, p$.

$$\mathbb{E}_{\pi,\psi}\left[\underbrace{\begin{pmatrix} \psi_{H_1} X_{1H_1} \\ \psi_{H_2} X_{2H_2} \\ \cdots \\ \cdots \\ \cdots \end{pmatrix}}_{\mathbf{S}} \underbrace{\left(q \sum_{k=1}^{p} \beta_k^* \psi_k\right)}_{=:b^*}\right] = \begin{pmatrix} \sum_{k=1}^{p} X_{1k} \beta_k^* q \mathbb{P}(H_1 = k) \\ \sum_{k=1}^{p} X_{2k} \beta_k^* q \mathbb{P}(H_2 = k) \\ \cdots \\ \cdots \\ \cdots \end{pmatrix}$$

Approximation error

- Assume for now that there are $q \leq p$ non-zero entries in each row of $\mathbf{X}$. Unequal row sparsity can also be dealt with.
- Consider one permutation with min-hash value H_i for $i = 1, \ldots, n$ and random signs ψ_k, $k = 1, \ldots, p$.

$$\mathbb{E}_{\pi,\psi}\left[\underbrace{\begin{pmatrix} \psi_{H_1} X_{1H_1} \\ \psi_{H_2} X_{2H_2} \\ \cdots \\ \cdots \\ \cdots \end{pmatrix}}_{\mathbf{S}} \underbrace{\left(q \sum_{k=1}^{p} \beta_k^* \psi_k\right)}_{=:b^*}\right] = \begin{pmatrix} \sum_{k=1}^{p} X_{1k} \beta_k^* q \mathbb{P}(H_1 = k) \\ \sum_{k=1}^{p} X_{2k} \beta_k^* q \mathbb{P}(H_2 = k) \\ \cdots \\ \cdots \\ \cdots \end{pmatrix}$$

$$= \mathbf{X}\beta^* \text{ (unbiased).}$$

Approximation error

Theorem

Let $\mathbf{b}^* \in \mathbb{R}^L$ *be defined by*

$$b_l^* = \frac{q}{L} \sum_{k=1}^{p} \beta_k^* \Psi_{kl} w_{\pi_l(k)},$$

where $\mathbf{w}$ *is a vector of weights. Then there is a choice of* $\mathbf{w}$*, such that:*

(i) *The approximation is unbiased:* $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{S}\mathbf{b}^*) = \mathbf{X}\boldsymbol{\beta}^*$.

(ii) *If* $\|\mathbf{X}\|_\infty \leq 1$, *then* $\frac{1}{n}\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{X}\boldsymbol{\beta}^*\|_2^2) \leq 2q\|\boldsymbol{\beta}^*\|_2^2/L$.

Row scaling

- In some applications e.g. text analysis, may assume the signal is

$$\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$$

where δ_i is the row sparsity, and κ is a scaling function.

Row scaling

- In some applications e.g. text analysis, may assume the signal is

$$\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$$

  where δ_i is the row sparsity, and κ is a scaling function.
- Example: When $\mathbf{X}$ is binary $\kappa(\delta) = 1/\sqrt{\delta}$ effectively scales all $\mathbf{x}_i$ to have same ℓ_2-norm.

Row scaling

- In some applications e.g. text analysis, may assume the signal is

$$\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$$

  where δ_i is the row sparsity, and κ is a scaling function.
- Example: When $\mathbf{X}$ is binary $\kappa(\delta) = 1/\sqrt{\delta}$ effectively scales all $\mathbf{x}_i$ to have same ℓ_2-norm.
- With a different $\mathbf{b}^*$, it is possible to approximate $\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$.

Row scaling

- In some applications e.g. text analysis, may assume the signal is

$$\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$$

where δ_i is the row sparsity, and κ is a scaling function.

- Example: When $\mathbf{X}$ is binary $\kappa(\delta) = 1/\sqrt{\delta}$ effectively scales all $\mathbf{x}_i$ to have same ℓ_2-norm.
- With a different $\mathbf{b}^*$, it is possible to approximate $\kappa(\delta_i)\mathbf{x}_i^T\boldsymbol{\beta}^*$.
- Example

$$\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}[\{\sqrt{\delta_{\min}/\delta_i}\mathbf{x}_i^T\boldsymbol{\beta}^* - \mathbf{s}_i^T\mathbf{b}^*\}^2] \leq \frac{q_{\min}\|\boldsymbol{\beta}^*\|^2}{L}\log\{4\log(L)/\delta_{\min}\}$$

where $q_{\min}$ is the minimal number of non-zeroes in $\mathbf{x}_i$; $\delta_{\min}$ is the minimal row sparsity.

Interaction models

Let $\mathbf{f}^* \in \mathbb{R}^n$ be given by

$$f_i^* = \sum_{k=1}^{p} X_{ik}\theta_k^{*,(1)} + \sum_{k,k_1=1}^{p} X_{ik}\mathbb{1}_{\{X_{ik_1}=0\}}\Theta_{k,k_1}^{*,(2)}, \quad i = 1, \ldots, n.$$

Interaction models

Let $\mathbf{f}^* \in \mathbb{R}^n$ be given by

$$f_i^* = \sum_{k=1}^{p} X_{ik}\theta_k^{*,(1)} + \sum_{k,k_1=1}^{p} X_{ik}\mathbb{1}_{\{X_{ik_1}=0\}}\Theta_{k,k_1}^{*,(2)}, \quad i = 1, \ldots, n.$$

Assume $\|\mathbf{X}\|_\infty \leq 1$. Previous results hold if $\|\boldsymbol{\beta}^*\|_2$ is replaced by

$$\ell(\boldsymbol{\Theta}^*) := \|\boldsymbol{\theta}^{*,(1)}\|_2 + 2\left(q \sum_{k,k_1,k_2} \left|\Theta_{kk_1}^{*,(2)}\Theta_{kk_2}^{*,(2)}\right|\right)^{1/2}.$$

Theorem

There exists $\mathbf{b}^* \in \mathbb{R}^L$ *such that*

(i) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{S}\mathbf{b}^*) = \mathbf{f}^*$;

(ii) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{f}^*\|_2^2)/n \leq 2q\ell^2(\boldsymbol{\Theta}^*)/L$.

Interaction models

Let $\mathbf{f}^* \in \mathbb{R}^n$ be given by

$$f_i^* = \sum_{k=1}^{p} X_{ik}\theta_k^{*,(1)} + \sum_{k,k_1=1}^{p} X_{ik}\mathbb{1}_{\{X_{ik_1}=0\}}\Theta_{k,k_1}^{*,(2)}, \quad i = 1, \ldots, n.$$

Assume $\|\mathbf{X}\|_\infty \leq 1$. Previous results hold if $\|\boldsymbol{\beta}^*\|_2$ is replaced by

$$\ell(\boldsymbol{\Theta}^*) := \|\boldsymbol{\theta}^{*,(1)}\|_2 + 2\left(q \sum_{k,k_1,k_2} \left|\Theta_{kk_1}^{*,(2)}\Theta_{kk_2}^{*,(2)}\right|\right)^{1/2}.$$

Theorem

There exists $\mathbf{b}^* \in \mathbb{R}^L$ *such that*

(i) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\mathbf{S}\mathbf{b}^*) = \mathbf{f}^*$;

(ii) $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\Psi}}(\|\mathbf{S}\mathbf{b}^* - \mathbf{f}^*\|_2^2)/n \leq 2q\ell^2(\boldsymbol{\Theta}^*)/L$.

If there are a finite number of non-zero interaction terms with finite value, the approximation error becomes very small if $L \gg q^2$.

Linear model

- Assume model
$$\mathbf{Y} = \alpha^* \mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}.$$
- Random noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ satisfies $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

Linear model

- Assume model
$$\mathbf{Y} = \alpha^* \mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}.$$
- Random noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ satisfies $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}(\varepsilon_i^2) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.
- We give bounds on a mean-squared prediction error (MSPE) of the form
$$\mathrm{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) := \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\pi},\boldsymbol{\Psi}}\left(\|\alpha^* \mathbf{1} + \mathbf{X}\boldsymbol{\beta}^* - \hat{\alpha}\mathbf{1} - \mathbf{S}\hat{\mathbf{b}}\|_2^2\right)/n.$$

Ordinary least squares

Theorem

Let $(\hat{\alpha}, \hat{\mathbf{b}})$ *be the least squares estimator and let* $L^* = \sqrt{2qn}\|\beta^*\|_2/\sigma$. *We have*

$$\text{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \leq 2 \max\left\{\frac{L}{L^*}, \frac{L^*}{L}\right\} \sigma \sqrt{\frac{2q}{n}} \|\beta^*\|_2 + \frac{\sigma^2}{n}.$$

Ordinary least squares

Theorem

Let $(\hat{\alpha}, \hat{\mathbf{b}})$ *be the least squares estimator and let* $L^* = \sqrt{2qn}\|\beta^*\|_2/\sigma$. *We have*

$$\text{MSPE}((\hat{\alpha}, \hat{\mathbf{b}})) \leq 2 \max\left\{\frac{L}{L^*}, \frac{L^*}{L}\right\} \sigma \sqrt{\frac{2q}{n}} \|\beta^*\|_2 + \frac{\sigma^2}{n}.$$

- Suppose more predictors are added to the design matrix but their associated coefficients are all 0, so $\|\beta^*\|_2 = O(1)$. The MSPE only increases like $\sqrt{q}$ compared to the factor of p we would see if OLS were used.
- Additionally assume $n = O(q)$ (ensures MSPE is bounded asymptotically). Then $L^* = O(q)$. This could be a substantial reduction over p.

Discussion

- Signal sparsity: useful assumption for high-dimensional data.
- The Lasso (Tibshirani, 1996) has been central to developments in the field.
- Remaining challenges: different sorts of structural sparsity; computation; inference.

Discussion

- Signal sparsity: useful assumption for high-dimensional data.
- The Lasso (Tibshirani, 1996) has been central to developments in the field.
- Remaining challenges: different sorts of structural sparsity; computation; inference.
- Data sparsity: often present in large-scale data.
- Several CS algorithms for dimension reduction that may be of interest to statisticians.

Discussion

- Signal sparsity: useful assumption for high-dimensional data.
- The Lasso (Tibshirani, 1996) has been central to developments in the field.
- Remaining challenges: different sorts of structural sparsity; computation; inference.
- Data sparsity: often present in large-scale data.
- Several CS algorithms for dimension reduction that may be of interest to statisticians.

Thank you for listening