Some topics in high-dimensional statistical inference

Post-selection inference and controlling the false-discovery rate

Luke Kelly

Department of Statistics
University of Oxford

OxWaSP Applied Statistics
November 6th, 2018

Table of Contents

Problem statement

For input $\mathbf{x} \in \mathbb{R}^p$, we want to predict the associated response $y \in R$ through the model

$$\hat{\mathbf{y}} = f(\mathbf{x}),$$

which we estimate from training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$.

Possible simple approaches for estimating f include

- Linear regression (parametric),
- Nearest neighbour regression (non-parametric).

How does our inference depend on the input dimension p?

Linear regression

If we assume that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}),$$

and $\operatorname{rank} \mathbf{X}^\top \mathbf{X} = p \leq n$, then

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

with $n - p$ degrees of freedom.

The expected prediction error (under the model),

$$\mathbb{E}\, L(y, \hat{y}) = \sigma^2 \left(1 + \frac{p}{n}\right),$$

grows linearly with p.

Nearest neighbours regression

For a choice of k and neighbourhood function $N_k : \mathbb{R}^p \to [n]$ under a suitable metric, we predict

$$\hat{y} = \frac{1}{k} \sum_{i \in N_k(\mathbf{x})} y_i,$$

a locally linear model with n/k effective[1] degrees of freedom.

Bias component of MSE typically increases with k while variance decreases.

Curse of dimensionality as p increases: the sampling density decreases rapidly.

The curse of dimensionality[2]

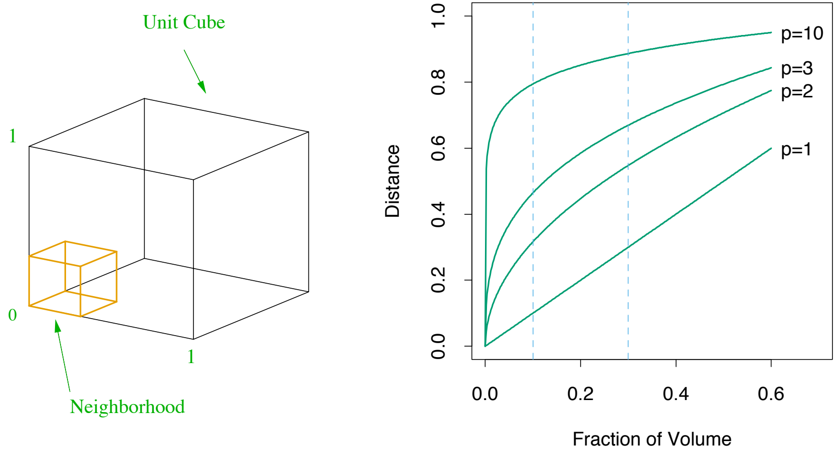


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Problem statement

Questions

Focusing on linear regression, what can we do if

- n is massive
- $p \gg n$?
- β is sparse?

Possible approaches and considerations

- (Random) projections onto lower dimensional subspaces
- Regularisation and variable selection
- Post-selection inference
- Controlling the false-discovery rate

Table of Contents

The Johnson–Lindenstrauss lemma[3]

For vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and constants $\epsilon \in (0, 1)$ and $d = \mathcal{O}(\epsilon^{-2} \log n)$, there exists $\mathbf{S} \in \mathbb{R}^{d \times p}$ such that

$$(1-\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2 \leq \|\mathbf{S}(\mathbf{x}_i - \mathbf{x}_j)\|^2 \leq (1+\epsilon)\|\mathbf{x}_i - \mathbf{x}_j\|^2$$

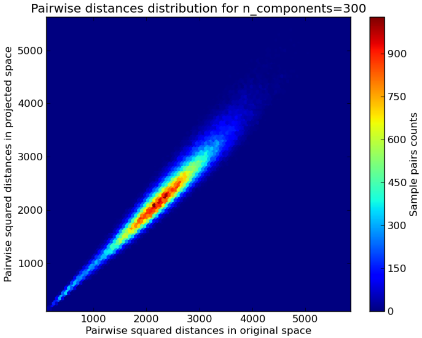
for all $i, j \in [n]$.

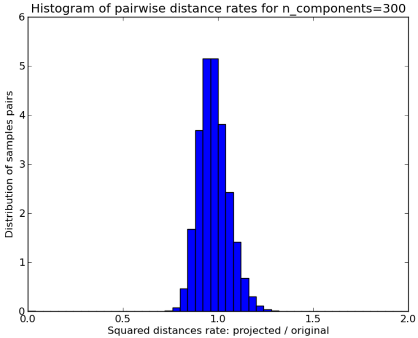
- Subspace dimension d does not depend on p.
- Simple proof using Markov's inequality and the union bound.
- The projection $\mathbf{S}$ can be found in randomised polynomial time through random projections.

Cannings and Samworth derive error bounds in d for the k-NN classifier.

The Johnson–Lindenstrauss lemma[4]

Projecting from $p = 100,000$ features down to 300.





Sketched regression[5]

Ahfock et al. apply the JL lemma to reduce the data dimension from n to k and analyse the regression estimators

$$\hat{\beta}_{\text{complete}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{y},$$
$$\hat{\beta}_{\text{partial}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \mathbf{X}^\top \mathbf{y},$$

where $\tilde{\mathbf{X}} = \mathbf{S}\mathbf{X}$ and $\tilde{\mathbf{y}} = \mathbf{S}\mathbf{y}$ and the sketch $\mathbf{S}$ is

- Gaussian with $\mathcal{N}(0, 1/k)$ entries
- Hadamard with Rademacher noise
- Clarkson–Woodruff with a sparse structure.

By drawing repeated sketches, the authors develop a CLT in n for the sketched data and corresponding estimators.

Sketched regression[5]

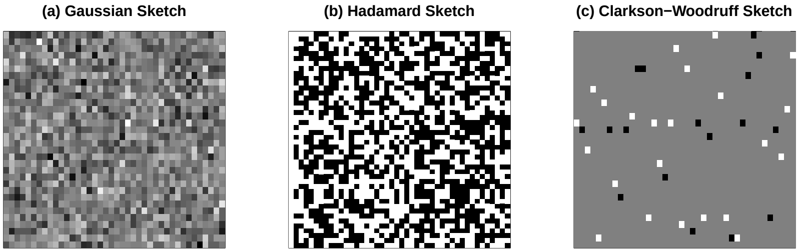


Figure 1: Sampled sketching matrices $\boldsymbol{S}$ for $k = 32, n = 36$. Elements in the sketching matrix are coloured based on the value. One and negative one are coloured as black and white respectively. Intermediate values are in shades of grey.

- The computational cost of the sketches varies
- The best choice of sketch depends on the signal-to-noise ratio in the data.

Table of Contents

Ridge regression

(The columns of $\mathbf{X}$ are scaled and centred and $\mathbf{y}$ is centred.)

We place an ℓ_2 penalty on the regression coefficients so

$$\begin{aligned}\hat{\beta}_{\mathsf{ridge}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_2^2\\ &= (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y},\end{aligned}$$

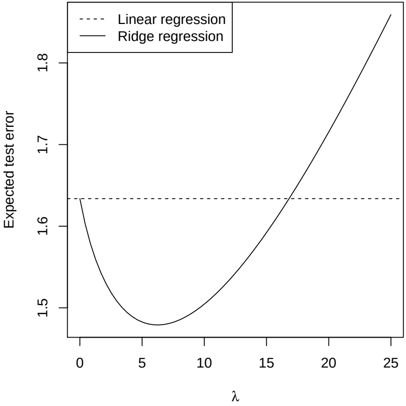
where λ controls the level of shrinkage.

- OLS solution for $\lambda \downarrow 0$ and null model for $\lambda \uparrow \infty$
- Problem is non-singular even if $p > n$.

As λ increases, bias increases while variance decreases.

Ridge regression[6]

We can estimate λ from the data.



Linear regression:
Squared bias ≈ 0.006
Variance ≈ 0.627
Test error $\approx 1 + 0.006 + 0.627 = 1.633$

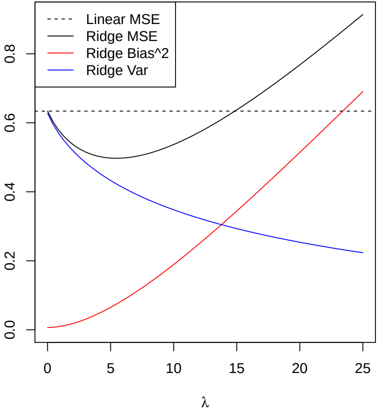
Ridge regression, at its best:
Squared bias ≈ 0.077
Variance ≈ 0.403
Test error $\approx 1 + 0.077 + 0.403 = 1.480$

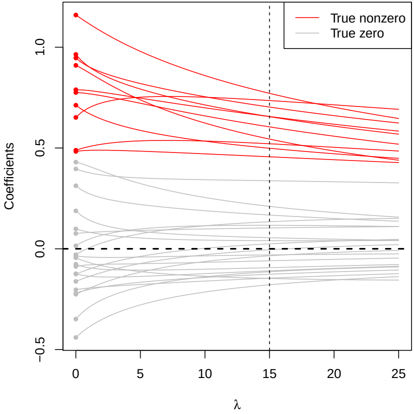
What if β is truly sparse?[7]

The ℓ_2 penalty in ridge regression

- Shrinks coefficients towards 0 but never exactly

so does not perform variable selection





Lasso regression

The lasso estimator is

$$\hat{\beta}_{\text{lasso}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

an ℓ_1-penalised regression.

Although the optimisation problem is similar to ridge regression, the lasso ℓ_1 penalty

- Shrinks coefficients exactly to zero.

Lasso regression[8]

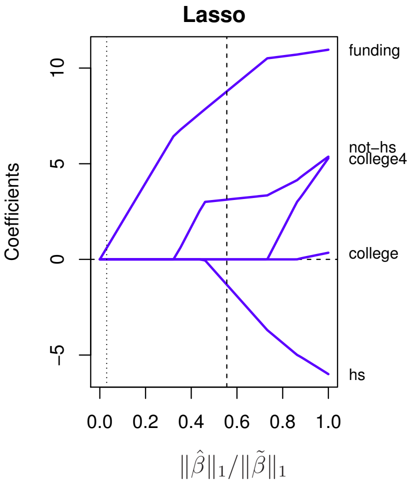


Figure 2.1 *Left: Coefficient path for the lasso, plotted versus the ℓ_1 norm of the coefficient vector, relative to the norm of the unrestricted least-squares estimate $\tilde{\beta}$. Right: Same for ridge regression, plotted against the relative ℓ_2 norm.*

Lasso regression[8]

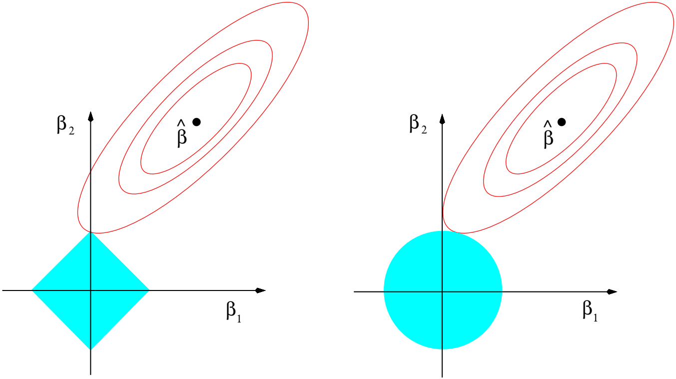


Figure 2.2 *Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions* $|\beta_1|+|\beta_2| \leq t$ *and* $\beta_1^2+\beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point* $\widehat{\beta}$ *depicts the usual (unconstrained) least-squares estimate.*

Table of Contents

Hypothesis testing after variable selection

In performing variable selection, we use the data to estimate

- The penalty term λ
- The subset of true (non-zero) coefficients and their corresponding values.

Any conclusions we draw about the resulting model using classical tools will be biased as

- We have used the data to generate the hypotheses!

Can we correct for the biases in our inference without splitting the data? Surprisingly, yes!

Coverage[9]

Although the set of possible models is

$$\{\beta_j^M : j \in M \subset [p]\},$$

we only perform inference on $\beta^{\hat{M}}$ for the selected model, $\hat{M}$.

A confidence interval $C_j^{\hat{M}}$ for $\beta_j^{\hat{M}}$ satisfying

$$\mathbb{P}(\beta_j^{\hat{M}} \in C_j^{\hat{M}}) \geq 1 - \alpha,$$

is not well-defined when $j \notin M$ so we focus on conditional coverage instead,

$$\mathbb{P}(\beta_j^M \in C_j^M \,|\, M = \hat{M}) \geq 1 - \alpha,$$

by characterising $\boldsymbol{\eta}^\top \mathbf{y} \,|\, \{\hat{M}(\mathbf{y}) = M\}$.

Example[10]

For example, with $p = 3$, the forward stagewise approach

- Selects variable 3, and
- Assigns it a positive coefficient

after one step if and only if both

$$\frac{\mathbf{X}_3^\top \mathbf{y}}{\|\mathbf{X}_3\|_2} \geq \frac{|\mathbf{X}_1^\top \mathbf{y}|}{\|\mathbf{X}_1\|_2} \qquad \text{and} \qquad \frac{\mathbf{X}_3^\top \mathbf{y}}{\|\mathbf{X}_3\|_2} \geq \frac{|\mathbf{X}_2^\top \mathbf{y}|}{\|\mathbf{X}_2\|_2}.$$

We can represent this event as $\{\mathbf{A}\mathbf{y} \leq \mathbf{b}\}$, a polyhedron.

Polyhedra[9]

The event $\{\hat{M} = M\}$ for the lasso is a union of polyhedra.

Denoting by $\mathbf{s}_M$ the signs of selected variables, the event $\{\hat{M} = M, \hat{\mathbf{s}}_M = \mathbf{s}\}$ is a polyhedron,

$$\{\mathbf{y} \in \mathbb{R}^n : \mathbf{A}(M, \mathbf{s}_M)\mathbf{y} \leq \mathbf{b}(M, \mathbf{s}_M)\},$$

so it suffices to study $\boldsymbol{\eta}^\top \mathbf{y} \,|\, \{\hat{M}(\mathbf{y}) = M\}$.

One can then derive a statistic $F(\boldsymbol{\eta}^\top \mathbf{y})$ such that

$$F(\boldsymbol{\eta}^\top \mathbf{y}) \,|\, \{\mathbf{A}\mathbf{y} \leq \mathbf{b}\} \sim \mathrm{Unif}(0, 1),$$

where F is a truncated Gaussian CDF with computable terms and, for example, $\boldsymbol{\eta} = \mathbf{e}_j^\top (\mathbf{X}_M^\top \mathbf{X}_M)^{-1} \mathbf{X}_M^\top$ returns variable j.

Polyhedral lemma[9]

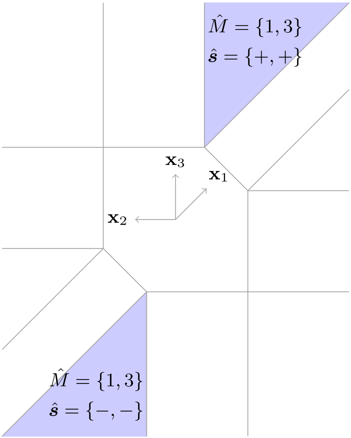


FIG. 1. *A geometric picture illustrating Theorem 4.3 for $n = 2$ and $p = 3$. The lasso partitions $\mathbb{R}^n$ into polyhedra according to the selected model and signs.*

Polyhedral lemma[8]

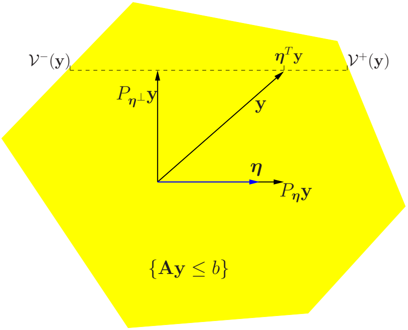


Figure 6.9 *Schematic illustrating the polyhedral lemma (6.7), for the case* $N = 2$ *and* $\|\boldsymbol{\eta}\|_2 = 1$*. The yellow region is the selection event* $\{\mathbf{Ay} \leq b\}$*. We decompose* $\mathbf{y}$ *as the sum of two terms: its projection* $P_{\boldsymbol{\eta}}\mathbf{y}$ *onto* $\boldsymbol{\eta}$ *(with coordinate* $\boldsymbol{\eta}^T\mathbf{y}$*) and its projection onto the* $(N-1)$*-dimensional subspace orthogonal to* $\boldsymbol{\eta}$*:* $\mathbf{y} = P_{\boldsymbol{\eta}}\mathbf{y} + P_{\boldsymbol{\eta}^\perp}\mathbf{y}$*. Conditioning on* $P_{\boldsymbol{\eta}^\perp}\mathbf{y}$*, we see that the event* $\{\mathbf{Ay} \leq b\}$ *is equivalent to the event* $\{\mathcal{V}^-(\mathbf{y}) \leq \boldsymbol{\eta}^T\mathbf{y} \leq \mathcal{V}^+(\mathbf{y})\}$*. Furthermore* $\mathcal{V}^+(\mathbf{y})$ *and* $\mathcal{V}^-(\mathbf{y})$ *are independent of* $\boldsymbol{\eta}^T\mathbf{y}$ *since they are functions of* $P_{\boldsymbol{\eta}^\perp}\mathbf{y}$ *only, which is independent of* $\mathbf{y}$*.*

Table of Contents

False discovery rate (FDR)

How many variables in the selected model are truly associated with the response?

- The FDR is the expected fraction of false variables returned by the selection procedure,

$$\mathrm{FDR} = \mathbb{E}\,\frac{|\hat{M} \cap \overline{M}|}{|\hat{M}| \vee 1}.$$

Bounding the FDR is important for replicability but we only have a finite amount of data.

Provided $p \leq n$, we can bound the lasso FDR exactly with only a finite amount of data using knockoffs .

Construct knockoff features[11]

Rescale columns of $\mathbf{X}$ so that $\boldsymbol{\Sigma} = \mathbf{X}^\top \mathbf{X}$ has $\operatorname{diag} \boldsymbol{\Sigma} = 1$.

Construct knockoff features $\tilde{\mathbf{X}}$ such that

- $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \boldsymbol{\Sigma}$
  - Same covariance structure as $\mathbf{X}$.
- $\mathbf{X}^\top \tilde{\mathbf{X}} = \boldsymbol{\Sigma} - \operatorname{diag} \mathbf{s}$ for choice of $\mathbf{s}$
  - Same correlations between distinct originals and knockoffs
  - We minimise correlation between a feature j and its knockoff: $\mathbf{X}_j^\top \tilde{\mathbf{X}}_j = 1 - s_j$.

If $\mathbf{X}_j$ is a true variable then we want it to enter the model before its knockoff.

Proportion of knockoffs entering model estimates the FDR.

Construct knockoffs and compute statistics[11]

Choose $\mathbf{s} \in \mathbb{R}_+^p$ satisfying $\operatorname{diag} \mathbf{s} \preceq 2\boldsymbol{\Sigma}$ and form

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\Sigma}^{-1} \operatorname{diag} \mathbf{s}) + \tilde{\mathbf{U}}\mathbf{C},$$

where $n \times p$ orthonormal $\tilde{\mathbf{U}}$ orthogonal to $\operatorname{span} \mathbf{X}$ and $\mathbf{C}^\top \mathbf{C} = 2(\operatorname{diag} \mathbf{s}) - (\operatorname{diag} \mathbf{s})\boldsymbol{\Sigma}^{-1}(\operatorname{diag} \mathbf{s})$.

Run lasso on augmented $n \times 2p$ design matrix $[\mathbf{X}\ \tilde{\mathbf{X}}]$ and compute

$$W_j = (Z_j \vee \tilde{Z}_j) \cdot \operatorname{sign}(Z_j - \tilde{Z}_j), \quad j \in [p],$$

where $Z_j = \sup\{\lambda : \hat{\beta}_j^\lambda \neq 0\}$ and $\tilde{Z}_j = \sup\{\lambda : \hat{\beta}_{j+p}^\lambda \neq 0\}$.

- $Z_j \gg 0$ evidence against null that $\beta_j = 0$.

Select variables[11]

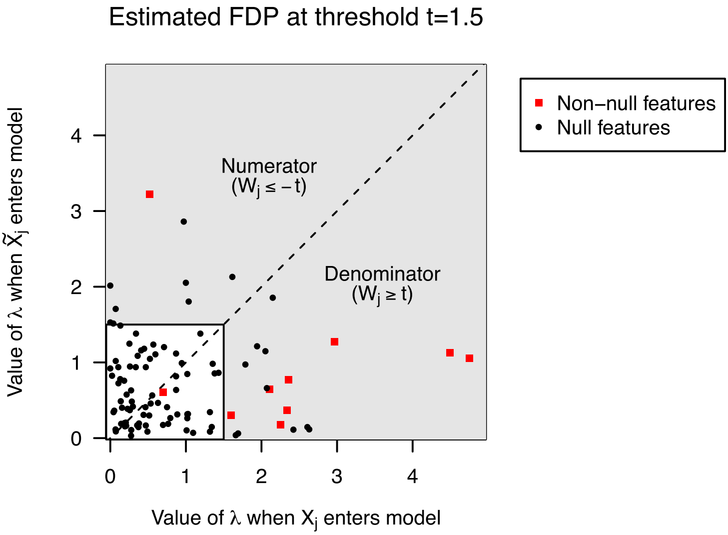
For the target FDR q we compute the threshold

$$T = \min\left\{ t \in \mathcal{W} : \frac{|\{j : W_j \leq -t\}|}{|\{j : W_j \geq t\}| \vee 1} \leq q \right\},$$

where $\mathcal{W} = \{|W_j| : j \in [p]\} \setminus \{0\}$.

The selected model $\hat{M} = \{j : W_j \geq T\}$ has an expected FDR bounded by q.

Knockoff filter[11]



Concluding remarks

Regression is an active area of statistical research.

We have described some recent methods to

- Select variables
- Account for biases in adaptively chosen hypothesis tests
- Control the false-discovery rate.

There are many others!

References / source material

[1] L. Janson, W. Fithian, and T.J. Hastie. Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485, 2015.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, USA, 2nd edition, 2009.

[3] T.I. Cannings and R.J. Samworth. Random-projection ensemble classification. *J. Roy. Stat. Soc. Ser. B.*, 79(4):959–1035.

[4] D. Lopez-Paz and D Duvenaud. Random projections, 2013.

[5] D. Ahfock, W. J. Astle, and S. Richardson. Statistical properties of sketching algorithms. *ArXiv 1706.03665*, 2017.

[6] R.J. Tibshirani. High-dimensional regression, 2014.

[7] R.J. Tibshirani. Modern regression 2: The lasso, 2013.

[8] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. Chapman and Hall/CRC, New York, USA, 1st edition, 2015.

[9] J.D. Lee, D.L. Sun, Y. Sun, and J.E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 06 2016.

[10] R.J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *J. Am. Stat. Assoc.*, 111(514):600–620, 2016.

[11] R.F. Barber and E.J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 10 2015.

Table of Contents

Practical

Generate synthetic data sets with varying numbers of data points n, feature dimensions p and true variables M.

- Compare ordinary least squares, ridge regression and Lasso (`glmnet`).
- Correct for biases using post-selection inference (`selectiveInference`).
- Control the false-discovery rate using knock-offs (`knockoff`).