APTS High-Dimensional Statistics

Yi Yu*

July 16, 2021

There are three principles, roughly expressible in the following terms: Every (measurable) set is nearly a finite sum of intervals; every function (of class L_p) is nearly continuous; every convergent sequence of functions is nearly uniformly convergent.

J. E. Littlewood
Lectures on the Theory of Functions

Every isometry is a direct sum of copies of the unilateral shift and a unitary operator.

The Wold–von Neumann decomposition theorem

Contents

*Department of Statistics, University of Warwick. Email: yi.yu.2@warwick.ac.uk

1 Introduction

Due to the ability of routinely collecting and storing large data sets in numerous application areas, we have witnessed a dramatic surge of interest and activity in high-dimensional analysis in the last two to three decades. Despite that the frenzy has been going on for decades, it is still, arguably, the most important statistics research topic, and has been studied and adopted in most if not all statistics research topics.

First of all, high-dimensional data are usually referred to those data sets having large dimensions, i.e. the dimension of the data are comparable or even larger than the sample size. Formally speaking, we say the dimension is a function of the sample size and is allowed to diverge as the sample size grows unbounded. To deal with such data sets, classical statistics tools often fail and it requires new methodology and theory.

Equally important, high-dimensional analysis also emphasises the non-asymptotic (also known as fixed-sample) viewpoint. This is to say, it is not necessary to have high-dimensional data to conduct high-dimensional analysis. As we shall see clearer throughout this module, the sample size, the dimension, as well as other model parameters, are viewed as fixed, and high-probability statements are made as a function of them (Wainwright, 2019).

Notation

For any matrix $M \in \mathbb{R}^{p\times q}$, let M_i and M^j be the ith row and jth column of M, $i \in \{1,\dots,p\}$, $j \in \{1,\dots,q\}$.

1.1 Linear regression

The running example in this module is the linear regression problem. In this section, we use it to demonstrate i) the limitation of classical results when the dimension is high, ii) what we mean by a non-asymptotic viewpoint and iii) what is needed to deal with the high-dimension feature of the data set.

Model. Let

$$Y_i = X_i^\top \beta^* + \varepsilon_i, \quad i = 1,\dots,n, \tag{1}$$

where

- the coefficient vector $\beta^* \in \mathbb{R}^p$ is unknown,
- the covariates $\{X_i\}_{i=1}^n \subset \mathbb{R}^p$ with $\Sigma_X = n^{-1}\sum_{i=1}^n X_i X_i^\top$ invertible, and
- the noise $\{\varepsilon_i\}_{i=1}^n \subset \mathbb{R}$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

For this model, we have the sample size n, the dimension p and the fluctuation level σ^2.

Remark 1. *Fixed and random designs.*

The limitation of the LSE. Denote the least squares estimator of β^* as $\widehat{\beta}^{\text{LSE}}$, which can be written as

$$\widehat{\beta}^{\text{LSE}} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2. \tag{2}$$

Provided that $\widehat{\Sigma}_X = n^{-1}X^\top X = n^{-1}\sum_{i=1}^n X_i X_i^\top$ is invertible, with $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, we have that

$$\widehat{\beta}^{\text{LSE}} = (X^\top X)^{-1} X^\top Y. \tag{3}$$

Remark 2. *Connections between LSE and MLE.*

When $p > n$, i.e. the dimension is larger than the sample size, $\text{rank}(\widehat{\Sigma}_X) \leq n < p$ and $\widehat{\Sigma}_X$ is not invertible. In fact, the true coefficient β^* plus any p-dimensional vector which is perpendicular to the linear space spanned by $\{X_i\}_{i=1}^n$, is a minimiser in this optimisation problem.

A non-asymptotic viewpoint. Before we proceed to discuss how to deal with the case when $\widehat{\Sigma}_X$ is not invertible, we first have a flavour of what a non-asymptotic viewpoint is. In classical theory, we focus on the asymptotic performances of $\widehat{\beta}^{\text{LSE}}$. Under the assumptions we mentioned, and that X_i's and ε_i's are independent, the dimension p is fixed, we have that

$$\sqrt{n}(\widehat{\beta}^{\text{LSE}} - \beta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 \Sigma^{-1}), \quad n \to \infty,$$

where $n^{-1}XX^\top X \to \Sigma$, as $n \to \infty$. The above statement is asymptotic, in the sense that it holds when the sample size grows unbound. From a non-asymptotic point of view, we can in fact show that for any triplet (n, p, σ^2), if $X^\top X = nI$ (we will discuss this further in the sequel), then for any $t > 0$,

$$\mathbb{P}\{\|\widehat{\beta}^{\text{LSE}} - \beta^*\|_\infty > t\} \leq p \max_{j=1}^p \mathbb{P}\{|\widehat{\beta}_j^{\text{LSE}} - \beta_j^*| > t\} \leq p \max_{j=1}^p \mathbb{P}\{|X^j \varepsilon|/n > t\} \leq 2p \exp\left(-\frac{nt^2}{\sigma^2}\right).$$

What do we need to deal with the high dimensionality. When we allow $p > n$, then the first problem we encounter is that (2) is an ill-defined problem. To overcome this issue, instead of minimising over all $\beta \in \mathbb{R}^p$, we can consider a smaller region, e.g. $\|\beta\|_\infty \leq M$. The second problem we encounter is that $\widehat{\Sigma}_X$ is no longer invertible in (3). To overcome this issue, together with a constraint on the solution region, we should be able to hope that a sub-matrix of $X^\top X$ is invertible.

2 The Lasso in linear regression problems

Recall the model defined in (1). We consider the situation when the dimension p is potentially large and cause problems in classical methods. From the problem we have pointed out in Section 1.1, to have a well-defined problem, we should restrict our solutions into a lower-dimensional space. Arguably, the simplest way to define a low-dimensional space is to assume that β^* satisfies

$$\|\beta^*\|_0 = s < n. \tag{4}$$

This means that β^* lies in an s-dimensional space and can be written as

$$\beta^* \in \mathbb{B}_0(s) = \left\{ \beta \in \mathbb{R}^p; \sum_{j=1}^{p} |\beta_j^*|^0 \leq s \right\},$$

with the convention that $0^0 = 0$. Based on this formulation, there are other types of low-dimensional space, e.g. assuming that $\beta^* \in \mathbb{B}_q(s_q)$, $q \in [0, 1]$, $s_q > 0$. In this module, we focus on the case $q = 0$. Denote

$$\mathcal{O} = \{j : \beta_j^* \neq 0\} \subset \{1, \ldots, p\}.$$

Remark 3. *Plots of different penalties.*

Assuming (1) and (4), we study the Lasso estimator

$$\widehat{\beta} = \widehat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}, \quad \lambda > 0,$$

which is equivalent to the ℓ_∞ constraints due to the Lagrangian duality. The term Lasso is coined in Tibshirani (1996), but the idea is essentially the same as the basis pursuit studied in Chen and Donoho (1994) and the Dantzig selector (Candes and Tao, 2007). The spirit roots in the bias-variance tradeoff, with a connection to the super-efficiency of shrinkage estimators, studied by Hodge, Pinsker, James, Stein, etc. The Lasso estimator is a convex relaxation of the ℓ_0-penalized estimator

$$\widetilde{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X\beta\|^2 + \lambda \|\beta\|_0 \right\},$$

which is an NP-hard optimisation problem.

Remark 4.

- *Spectral clustering in community detection problems in the stochastic block models is also a convex relaxation (e.g. Von Luxburg, 2007).*
- *The ℓ_0 penalisation is not necessarily all NP-hard. We will revisit this in Section 5.*

Lemma 1 (Basic inequality)**.** *We have that*

$$\|X(\widehat{\beta} - \beta^*)\|^2/n + 2\lambda\|\widehat{\beta}\|_1 \leq 2\varepsilon^\top X/n(\widehat{\beta} - \beta^*) + 2\lambda\|\beta^*\|_1.$$

Remark 5. *Lemma 1 is the starting point of analysing the Lasso estimator. More importantly, this type of results is usually the starting point of analysing any penalised estimators.*

Proof of Lemma 1. It follows from the definition of the Lasso estimator that

$$\frac{1}{2n}\|Y - X\widehat{\beta}\|^2 + \lambda\|\widehat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^*\|^2 + \lambda\|\beta^*\|_1,$$

which leads to that

$$\widehat{\beta}^\top X^\top X\widehat{\beta} - 2Y^\top X\widehat{\beta} + 2n\lambda\|\widehat{\beta}\|_1 \leq (\beta^*)^\top X^\top X\beta^* - 2Y^\top X\beta^* + 2n\lambda\|\beta^*\|_1.$$

Plugging in the fact that $Y = X\beta^* + \varepsilon$ concludes the proof. □

In view of the basic inequality, conditional on the design X, the only random part is the term

$$2\varepsilon^\top X(\widehat{\beta} - \beta^*),$$

which can be upper bounded using the inequality

$$|2\varepsilon^\top X(\widehat{\beta} - \beta^*)| \leq 2\|\varepsilon^\top X\|_\infty \|\widehat{\beta} - \beta^*\|_1.$$

(Question: why do we use the ∞-1 inequality? Can we use others, say 2-2? Hanson–Wright inequality.) For any $t > 0$, define the event

$$\mathcal{E}(t) = \left\{\|\varepsilon^\top X\|_\infty < t\right\}.$$

Lemma 2. *If assuming X to be fixed, then for any $t > 0$, we have that*

$$\mathbb{P}\left\{\mathcal{E}(t)\right\} > 1 - 2p\exp\left(-\frac{t^2}{2\sigma^2 \max_{j=1}^p \|X^j\|^2}\right).$$

If assuming $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_X)$, where Σ_X is a positive definite matrix, then for any $t > 0$ and any $t_1 \in (n\max_{j=1}^p(\Sigma_X)_{jj}, 2n\min_{j=1}^p(\Sigma_X)_{jj})$, we have that

$$\mathbb{P}\{\mathcal{E}(t)\} \geq \left\{1 - 2p\exp\left(-\frac{t^2}{2\sigma^2 t_1}\right)\right\}\left[1 - p\max_{j=1}^p \exp\left\{-\frac{n\left\{\frac{t_1}{n(\Sigma_X)_{jj}} - 1\right\}^2}{8}\right\}\right].$$

Remark 6. *This is a usual ingredient in the high-dimensional analysis. First use the sub-Gaussian and/or sub-Exponential tail bounds to control the noise fluctuation, and then discuss the rest of the proof in these large-probability events.*

Proof. Under the fixed design, we have that, for any $t > 0$,

$$\begin{aligned}\mathbb{P}\left\{\mathcal{E}(t)\right\} &= 1 - \mathbb{P}\left\{\max_{j=1}^p |\varepsilon^\top X^j| \geq t\right\} \geq 1 - \sum_{j=1}^p \mathbb{P}\left\{|\varepsilon^\top X^j| \geq t\right\} \\ &\geq 1 - p\max_{j=1}^p \mathbb{P}\left\{|\varepsilon^\top X^j| \geq t\right\} \geq 1 - 2p\exp\left(-\frac{t^2}{2\sigma^2 \max_{j=1}^p \|X^j\|^2}\right), \end{aligned} \tag{5}$$

where the last inequality follows from the Hoeffding inequality (e.g. Proposition 2.5 in Wainwright, 2019).

Under the random design, we have that, for any $t, t_1 > 0$,

$$\begin{aligned}
&\mathbb{P}\{\mathcal{E}(t)\} \geq \mathbb{P}_{\varepsilon|X}\left\{\mathcal{E}(t)|0 < \max_{j=1}^{p}\|X^j\|^2 \leq t_1\right\}\mathbb{P}_X\left\{0 < \max_{j=1}^{p}\|X^j\|^2 \leq t_1\right\} \\
\geq &\left\{1 - 2p\exp\left(-\frac{t^2}{2\sigma^2 t_1}\right)\right\}\mathbb{P}_X\left\{0 < \max_{j=1}^{p}\|X^j\|^2 \leq t_1\right\},
\end{aligned}$$

where the second inequality follows (5).

In addition, we have that, for any

$$n\max_{j=1}^{p}(\Sigma_X)_{jj} < t_1 < 2n\min_{j=1}^{p}(\Sigma_X)_{jj},$$

we have that

$$\begin{aligned}
&\mathbb{P}_X\left\{0 < \max_{j=1}^{p}\|X^j\|^2 \leq t_1\right\} \geq 1 - p\max_{j=1}^{p}\mathbb{P}\left\{\sum_{i=1}^{n}X_{ij}^2 - n(\Sigma_X)_{jj} > t_1 - n(\Sigma_X)_{jj}\right\} \\
\geq &1 - p\max_{j=1}^{p}\mathbb{P}\left\{n^{-1}\sum_{i=1}^{n}\frac{X_{ij}^2}{(\Sigma_X)_{jj}} - 1 > \frac{t_1}{n(\Sigma_X)_{jj}} - 1\right\} \\
\geq &1 - p\max_{j=1}^{p}\exp\left\{-\frac{n\left\{\frac{t_1}{n(\Sigma_X)_{jj}} - 1\right\}^2}{8}\right\},
\end{aligned}$$

where the last inequality follows from Bernstein's inequality (e.g. Example 2.11 in Wainwright, 2019). We then conclude the proof. □

We now conduct our analysis in the event

$$\mathcal{E} = \mathcal{E}(n\lambda/2).$$

In the event $\mathcal{E}$, due to Lemma 1, we have that

$$\begin{aligned}
&\|X(\widehat{\beta} - \beta^*)\|^2/n \leq 2\varepsilon^\top X/n(\widehat{\beta} - \beta^*) + 2\lambda(\|\beta^*\|_1 - \|\widehat{\beta}\|_1) \\
\leq &\lambda\|\widehat{\beta} - \beta^*\|_1 + 2\lambda(\|\beta^*\|_1 - \|\widehat{\beta}\|_1),
\end{aligned}$$

which leads to

$$\|X(\widehat{\beta} - \beta^*)\|^2/n + \lambda\|\widehat{\beta} - \beta^*\|_1 \leq 2\lambda\|\widehat{\beta} - \beta^*\|_1 + 2\lambda(\|\beta^*\|_1 - \|\widehat{\beta}\|_1).$$

Note that

$$|\widehat{\beta}_j - \beta_j^*| + |\beta_j^*| - |\widehat{\beta}_j| = 0, \quad j \in \mathcal{O}^c = \{1, \ldots, p\} \setminus \mathcal{O}.$$

We have that

$$\begin{aligned}
&\|X(\widehat{\beta} - \beta^*)\|^2/n + \lambda\|\widehat{\beta} - \beta^*\|_1 \leq 2\lambda\|(\widehat{\beta} - \beta^*)_{\mathcal{O}}\|_1 + 2\lambda(\|\beta_{\mathcal{O}}^*\|_1 - \|\widehat{\beta}_{\mathcal{O}}\|_1) \\
\leq &4\lambda\|(\widehat{\beta} - \beta^*)_{\mathcal{O}}\|_1 \leq 4\lambda\sqrt{|\mathcal{O}|}\|(\widehat{\beta} - \beta^*)_{\mathcal{O}}\| = 4\lambda\sqrt{s}\|(\widehat{\beta} - \beta^*)_{\mathcal{O}}\|.
\end{aligned}$$

Up to now, we have

$$\|X(\widehat{\beta}-\beta^*)\|^2/n \le 4\lambda\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\| \tag{6}$$

and

$$\|\widehat{\beta}-\beta^*\|_1 = \|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\|_1 + \|(\widehat{\beta}-\beta^*)_{\mathcal{O}^c}\|_1 \le 4\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\|_1,$$

which leads to

$$\|(\widehat{\beta}-\beta^*)_{\mathcal{O}^c}\|_1 \le 3\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\|_1 \quad \text{and} \quad \|\widehat{\beta}-\beta^*\|_1 \le 4\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\|. \tag{7}$$

To proceed further, if we have that $X^\top X$ is invertible, then from (6) we have that

$$\mu_{\min}(X^\top X)\|\widehat{\beta}-\beta^*\|^2 \le \|X(\widehat{\beta}-\beta^*)\|^2 \le 4\lambda n\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\| \le 4\lambda n\sqrt{s}\|\widehat{\beta}-\beta^*\|,$$

which leads to that

$$\|\widehat{\beta}-\beta^*\| \le 4\lambda n\sqrt{s}/\mu_{\min}(X^\top X).$$

However, when $p > n$, $X^\top X$ is clearly singular. This prompts us to find new theoretical tools.

2.1 Restricted eigenvalue conditions

In this section, we start from the fixed design and then discuss the random design.

Assumption 1 (Restricted eigenvalue)**.** *There exists a constant $\kappa > 0$ such that, for any*

$$\theta \in \mathcal{C}(\mathcal{O}) = \left\{\theta \in \mathbb{R}^p : \|\theta\|_1 \le 4\sqrt{s}\|\theta_{\mathcal{O}}\|\right\},$$

it holds that

$$\|X\theta\|^2/n \ge \kappa\|\theta_{\mathcal{O}}\|^2.$$

Remark 7. *Assumption 1 is a very strong version of the restricted eigenvalue type conditions. We adopt it here for illustration purpose. There are milder versions of restricted eigenvalue type conditions, under the names of the Reisz condition, the compatibility condition, restricted isometry property, etc.*

Assumption 1 can also be stated in the space $\mathcal{C}(\mathcal{O}) = \{\theta \in \mathbb{R}^p : \|\theta_{\mathcal{O}^c}\|_1 \le 3\|\theta_{\mathcal{O}}\|_1\}$. In the fixed design case, it is assumed that $\mathrm{null}(X) \cap \mathcal{C} = \{0\}$.

A more interesting version of restricted eigenvalue condition is to impose on any arbitrary set $A \subset [p]$, $|A| \le s$, instead of only on a specific set $\mathcal{O}$.

In view of Assumption 1, due to (7), we have that

$$\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\|_1 \le \|\widehat{\beta}-\beta^*\|_1 \le 4\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\|.$$

Under Assumption 1, due to (6), we have that

$$\|X(\widehat{\beta}-\beta^*)\|^2/n \le 4\lambda\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\| \le 4\lambda\sqrt{\frac{s}{n\kappa}}\|X(\widehat{\beta}-\beta^*)\| \le 8\lambda^2 s/\kappa + \|X(\widehat{\beta}-\beta^*)\|^2/(2n),$$

which leads to that

$$\|X(\widehat{\beta}-\beta^*)\|^2/n \le 16\lambda^2 s/\kappa.$$

Moreover, we have that

$$\|\widehat{\beta}-\beta^*\|_1 \le 4\sqrt{s}\|(\widehat{\beta}-\beta^*)_{\mathcal{O}}\| \le 4\sqrt{s}\|X(\widehat{\beta}-\beta^*)\|/\sqrt{n\kappa} \le 16\lambda s/\kappa.$$

To wrap up, we have the following theorem.

Theorem 3. *Assume that* $\|X^j\| = n$, $j = 1, \ldots, n$. *With* $\lambda = C\sigma\sqrt{\log(p)/n}$, $C > 2\sqrt{2}$, *under Assumption 1, we have that, with probability at least*

$$1 - 2p \exp\left(-\frac{C^2 \log(p)}{8}\right),$$

it holds that

$$\|X(\widehat{\beta} - \beta^*)\|^2/n \leq 16C^2\sigma^2 \log(p)s/(n\kappa)$$

and

$$\|\widehat{\beta} - \beta^*\|_1 \leq 16C\sigma s/\kappa\sqrt{\log(p)/n}.$$

We then move on to random designs. When considering X to be random, we cannot assume Assumption 1. One should, instead, prove that with high-probability, Assumption 1 holds. The following theorem is from Raskutti et al. (2010).

Theorem 4. *For any Gaussian random design* $X \in \mathbb{R}^{n\times p}$ *with i.i.d.* $\mathcal{N}(0, \Sigma_X)$ *rows, there are universal positive constants* c, c' *such that*

$$n^{-1/2}\|X\theta\| \geq 4^{-1}\|\Sigma_X^{1/2}\theta\| - 9\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}\|\theta\|_1, \quad \theta \in \mathbb{R}^p,$$

with probability at least $1 - c' \exp(-cn)$.

The implication of Theorem 4 is as follows. For any $A \subset [p]$ with $|A| \leq s$ and any $\theta \in \mathcal{C}(A)$ defined in Assumption 1, it holds that

$$\begin{aligned}
n^{-1/2}\|X\theta\| &\geq 4^{-1}\|\Sigma_X^{1/2}\theta\| - 9\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}\|\theta\|_1 \\
&\geq 4^{-1}\lambda_{\min}^{1/2}(\Sigma_X)\|\theta\| - 9\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}\|\theta\|_1 \\
&= 4^{-1}\lambda_{\min}^{1/2}(\Sigma_X)\|\theta\| - 9\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}(\|\theta_A\|_1 + \|\theta_{A^c}\|_1) \\
&\geq 4^{-1}\lambda_{\min}^{1/2}(\Sigma_X)\|\theta\| - 36\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}\|\theta_A\|_1 \\
&\geq 4^{-1}\lambda_{\min}^{1/2}(\Sigma_X)\|\theta\| - 36\sqrt{s}\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}\|\theta_A\| \\
&\geq \left\{4^{-1}\lambda_{\min}^{1/2}(\Sigma_X) - 36\sqrt{s}\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n}\right\}\|\theta\|.
\end{aligned}$$

Provided that

$$4^{-1}\lambda_{\min}^{1/2}(\Sigma_X) - 36\sqrt{s}\left\{\max_{j=1}^p (\Sigma_X)_{jj}\right\}^{1/2} \sqrt{\log(p)/n} \geq \kappa,$$

we show that with large probability, $n^{-1}XX^\top$ satisfies restricted eigenvalue conditions.

2.2 What we did not cover in this section

- Tuning parameter selection: EBIC, cross-validation, 1se rule, etc.
- Algorithm: LARS, coordinate descent, etc.
- Selection consistency: irrepresentability condition, beta-min condition.
- Other types of penalties: bridge, ridge, group, elastic net, MCP, SCAD, etc.
- Post-processing: restricted MLE, stability selection, etc.
- Beyond linear regression: generalised linear regression, cox regression, etc.
- Minimax lower bound.
- Other methods to quantify the uncertainty of Lasso-type estimators.

3 De-biased Lasso

The Lasso estimators enjoy both theoretical and numerical advantages and have been widely used in many application areas. It, however, still suffers from limitations. Statistical inference based on selection consistency theory typically requires a uniform signal strength condition that all nonzero regression coefficients be greater tin magnitude than an inflated noise level to take model uncertainty into account. This is unfortunately seldom supported by either the data or the underlying science in applications when the presence of weak signals cannot be ruled out (Zhang and Zhang, 2014). In addition, lasso estimators do not have a tractable limiting distribution. Even in the low-dimensional settings, the limiting distribution depends on the unknown parameter and the convergence rate is not uniform (Van de Geer et al., 2014). This means that it is impossible to construct confidence intervals based on the lasso estimators.

The de-biased lasso estimator was developed independently by three groups of researchers, resulting in three papers: Zhang and Zhang (2014), Van de Geer et al. (2014) and Javanmard and Montanari (2014). The differences are subtle and they share the same spirit. In this module, we will use a generic framework.

We first claim that Lasso is a biased estimator, for β_j^*, $j \in \mathcal{O}$. To see this, we consider a very simple toy example with orthonormal design, i.e. $(X^j)^\top X^k = n\delta_{jk}$, where $\delta_{jk} = \mathbb{1}\{j = k\}$. Therefore, the jth Lasso estimator can be written as

$$\widehat{\beta}_j = \underset{\beta_j \in \mathbb{R}}{\arg\min} \left\{ \frac{1}{2n} \|Y - X^j \beta_j\|^2 + \lambda |\beta_j| \right\}.$$

The least squares estimator of β_j^* is

$$\widehat{\beta}_j^{\mathrm{LSE}} = Y^\top X^j / n.$$

Some calculation shows that

$$\widehat{\beta}_j = \underset{\beta_j \in \mathbb{R}}{\arg\min} \left\{ \frac{1}{2} \left(\widehat{\beta}_j^{\mathrm{LSE}} - \beta_j \right)^2 + \lambda |\beta_j| \right\} = \begin{cases} \widehat{\beta}_j^{\mathrm{LSE}} - \lambda, & \widehat{\beta}_j^{\mathrm{LSE}} > \lambda, \\ 0, & |\widehat{\beta}_j^{\mathrm{LSE}}| \le \lambda, \\ \widehat{\beta}_j^{\mathrm{LSE}} + \lambda, & \widehat{\beta}_j^{\mathrm{LSE}} < -\lambda. \end{cases}$$

From this toy example, we can see that the bias comes from the shrinkage λ and if we can debias Lasso, then we should be able to wish for an unbiased estimator with a tractable limiting distribution.

3.1 Bias corrected linear estimators

We can write the least squares estimator as

$$\widehat{\beta}_j^{\mathrm{LSE}} = Y^\top (X^j)^\perp / (X^j)^\top (X^j)^\perp, \tag{8}$$

where $(X^j)^\perp$ is the projection of X^j to the orthogonal complement of the column space of $X^{-j} = (X^k, k \neq j)$. In the case when $p > n$, $\mathrm{rank}(X^{-j}) = n$, for all $j \in \{1, \ldots, p\}$. As a consequence, $(X^j)^\perp = 0$ and (8) is undefined.

We notice that $\widehat{\beta}_j^{\mathrm{LSE}}$ defined in (8) is equivalent to solving the equation

$$((X^j)^\perp)^\top (Y - \beta_j X^j) = ((X^j)^\perp)^\top X^k = 0, \quad \forall k \neq j.$$

For the case when $p > n$, we retain the main equation $z_j^\top (Y - \beta_j X^j) = 0$, but relax the constraint $z_j^\top X^k = 0$ for all $k \neq j$. Based on this relaxation, we consider a linear estimator

$$\widehat{\beta}_j^{\mathrm{linear}} = \frac{Y^\top z_j}{z_j^\top X^j} = \beta_j^* + \frac{z_j^\top \varepsilon}{z_j^\top X^j} + \sum_{k \neq j} \frac{z_j^\top X^k \beta_k}{z_j^\top X^j}.$$

To correct the bias, we can start with a nonlinear initial estimator $\widehat{\beta}^{\mathrm{initial}}$ and have the final estimator as

$$\widehat{b}_j = \widehat{\beta}_j^{\mathrm{initial}} - \sum_{k \neq j} \frac{z_j^\top X^k \widehat{\beta}_k^{\mathrm{initial}}}{z_j^\top X^j} = \frac{Y^\top z_j}{z_j^\top X^j} - \sum_{k \neq j} \frac{z_j^\top X^k \widehat{\beta}_k^{\mathrm{initial}}}{z_j^\top X^j},$$

which can be interpreted as a one-step correction from the initial estimator. We can rewrite it as

$$\widehat{b}_j = \widehat{\beta}_j^{\mathrm{initial}} + \frac{z_j^\top \{Y - X\widehat{\beta}^{\mathrm{initial}}\}}{z_j^\top X^j}.$$

We have the decomposition that

$$\widehat{b}_j - \beta_j^* = \frac{z_j^\top \varepsilon}{z_j^\top X^j} + \sum_{k \neq j} \frac{z_j^\top X^k (\beta_k - \widehat{\beta}_k^{\mathrm{initial}})}{z_j^\top X^j},$$

where the first term can be shown to by asymptotically normal and the second term can be expected to be negligible provided the initial estimator $\widehat{\beta}^{\mathrm{initial}}$ has a small ℓ_1 error.

3.2 Another viewpoint

Recall the Lasso estimator

$$\widehat{\beta} = \widehat{\beta}(\lambda) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}.$$

From convex optimisation theory, we know that $\widehat{\beta}$ satisfies the Karush–Kuhn–Tucker (KKT) conditions

$$-X^\top(Y - X\widehat{\beta}) + \lambda n \widehat{s} = 0,$$
$$\|\widehat{s}\|_\infty \leq 1 \quad \text{and} \quad \widehat{s}_j = \text{sign}(\widehat{\beta}_j) \text{ if } \widehat{\beta}_j \neq 0.$$

The vector $\widehat{s}$ is the weak derivative of $\|\beta\|_1$. We therefore have

$$\lambda \widehat{s} = X^\top(Y - X\widehat{\beta})/n.$$

The KKT conditions can be rewritten with the notation $\widehat{\Sigma} = X^\top X/n$ that

$$\widehat{\Sigma}(\widehat{\beta} - \beta^*) + \lambda \widehat{s} = X^\top \varepsilon/n.$$

Let $\widehat{\Theta}$ be a reasonable approximation of a relaxed form of an inverse of $\widehat{\Sigma}$, in the sense that $\widehat{\Theta}\widehat{\Sigma} \approx I$. We then have that

$$\widehat{\beta} - \beta^* + \widehat{\Theta}\lambda\widehat{s} = \widehat{\Theta}X^\top\varepsilon/n - (\widehat{\Theta}\widehat{\Sigma} - I)(\widehat{\beta} - \beta^*).$$

Let

$$\widehat{b} = \widehat{\beta} + \widehat{\Theta}\lambda\widehat{s} = \widehat{\beta} + \widehat{\Theta}X^\top(Y - X\widehat{\beta})/n.$$

We have that

$$\begin{aligned}\sqrt{n}(\widehat{b} - \beta^*) &= n^{-1/2}\widehat{\Theta}X^\top\varepsilon - \sqrt{n}(\widehat{\Theta}\widehat{\Sigma} - I)(\widehat{\beta} - \beta^*) \\ &= n^{-1/2}\Sigma X^\top\varepsilon + n^{-1/2}(\widehat{\Theta} - \Sigma)X^\top\varepsilon - \sqrt{n}(\widehat{\Theta}\widehat{\Sigma} - I)(\widehat{\beta} - \beta^*).\end{aligned}$$

For each $j \in \{1, \ldots, p\}$, if we can show that

- $n^{-1/2}\Sigma_j X^\top\varepsilon$ converges to a normal distribution, using CLT;
- $\|\widehat{\Theta} - \Sigma\|_1 \|n^{-1/2}X^\top\varepsilon\|_\infty \to 0$; and
- $\|\sqrt{n}(\widehat{\beta} - \beta^*)\|_1 \|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty \to 0$,

then we can show that $\sqrt{n}(\widehat{b} - \beta^*)_j$ converges to a normal distribution.

3.3 What we did not cover in this section

- The estimator $\widehat{\Theta}$, which we will discuss in Section 4.
- Tuning parameter selection.
- Practical issues.
- Convergence rates.
- Other models.

4 Graphical Lasso and graphical models

In Section 3, we discuss the de-biased Lasso but we do not discuss the estimator $\widehat{\Theta}$. To thoroughly understand the choice of $\widehat{\Theta}$ used in the de-biased Lasso, we start with Gaussian graphical models.

4.1 Gaussian graphical models

Let $X \in \mathbb{R}^p \sim \mathcal{N}(\mu, \Sigma)$, with density

$$(2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-2^{-1}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$

Denote $\Theta = \Sigma^{-1}$ as the precision matrix or the concentration matrix. In terms of Θ, the density function of $\mathcal{N}(\mu, \Sigma)$ can be equivalently formulated as

$$\exp\left\{\mu^\top \Theta x - \langle \Theta, 2^{-1} x x^\top \rangle - p/2 \log(2\pi) + 2^{-1}\log(|\Theta|) - 2^{-1}\mu^\top \Theta \mu\right\},$$

which implies that the Gaussian distribution is an exponential family with canonical parameters $(-\mu^\top \Theta, \Theta)$.

Partition the random vector X into two components $X_A \in \mathbb{R}^a$ and $X_B \in \mathbb{R}^b$ such that $A \sqcup B = \{1, \ldots, p\}$. Let μ and Σ be partitioned accordingly, i.e.

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix}.$$

For any $x_A \in \mathbb{R}^a$ and $x_B \in \mathbb{R}^b$, the conditional density is

$$f(x_A|x_B) \propto \exp\Bigg[-\frac{1}{2}\{x_A - \mu_A - \Theta_{A,A}^{-1}\Theta_{A,B}(x_B - \mu_B)\}^\top \Theta_{A,A} \times \{x_A - \mu_A - \Theta_{A,A}^{-1}\Theta_{A,B}(x_B - \mu_B)\}\Bigg],$$

where $\Theta_{A,A}^{-1} = \Sigma_{A,A} - \Sigma_{A,B}(\Sigma_{B,B})^{-1}\Sigma_{B,A}$. We therefore have that

$$X_A | X_B = x_B \sim \mathcal{N}(\mu_{A|B}, \Sigma_{A|B}),$$

where

$$\mu_{A|B} = \mu_A + \Sigma_{A,B}(\Sigma_{B,B})^{-1}(x_B - \mu_B) \quad \text{and} \quad \Sigma_{A|B} = \Sigma_{A,A} - \Sigma_{A,B}(\Sigma_{B,B})^{-1}\Sigma_{B,A}.$$

For any $i, j \in \{1, \ldots, p\}$, $i \neq j$, it holds that $X_i \perp X_j$ if and only if $\Sigma_{i,j} = 0$. Based on the formula above, we have that $X_i \perp X_j | X_{-(i,j)}$ if and only if $(\Sigma_{(i,j)|-(i,j)})_{1,2} = 0$. Based on the above derivation, it means that $X_i \perp X_j | X_{-(i,j)}$ if and only if $(\Sigma^{-1})_{i,j} = \Theta_{i,j} = 0$.

Let $G = (V, E)$ be an undirected graph with vertices $V = \{1, \ldots, p\}$ and edges E. A random vector $X \in \mathbb{R}^p$ is said to satisfy the Gaussian graphical model with graph G, if $X \sim \mathcal{N}(\mu, \Sigma)$ with

$$\Theta_{i,j} = (\Sigma^{-1})_{i,j} = 0 \quad \text{if and only if} \quad (i,j) \notin E.$$

Gaussian graphical models only model the pairwise interactions between nodes. Gaussian graphical models are the continuous counterpart to Ising models.

Remark 8.

- *Markov random fields. Undirected graphical models are also known as Markov random fields (MRF). An MRF is specified by an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$. The structure of this graph encodes certain conditional independence assumptions among subsets of the p-dimensional discrete random vector $X = (X_1, \ldots, X_p)^\top$, where X_i is associated with vertex $i \in V$. One important problem for such models is to estimate the underlying graph from n i.i.d. samples $\{x^{(i)}\}_{i=1}^n$ drawn from the distribution specified by some MRF.*

- *Pairwise MRF. Let $X = (X_1, \ldots, X_p)$ denote a random vector with each variable $X_s \in \mathcal{X}_s$. Given an undirected graph $G = (V, E)$, each X_s is associated with $s \in V$. The pairwise Markov random field associated with the graph G over the random vector X is the family of distributions of X which factorise as*

$$\mathbb{P}(x) \propto \exp\left\{\sum_{(s,t)\in E} \phi_{st}(x_s, x_t)\right\},$$

where for each edge $(s, t) \in E$, ϕ_{st} is a mapping from pairs $(x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t$ to the real line.

For models involving discrete random variables, the pairwise assumption involves no loss of generality since any Markov random field with higher-order interactions can be converted (by introducing additional variables) to an equivalent Markov random field with purely pairwise interactions.

- *Ising model is a special case that $X_s \in \{\pm 1\}$ and $\phi_{st}(x_s, x_t) = \theta^* x_s x_t$. The distribution takes the form*

$$\mathbb{P}_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp\left\{\sum_{(s,t)\in E} \theta^*_{st} x_s x_t\right\}, \quad x \in \{\pm 1\}^{\otimes p}.$$

A composite likelihood approach works on

$$\prod_{j=1}^p \mathbb{P}_\theta(x_r | x_{-r}) = \prod_{j=1}^p \left\{\frac{\exp\left(2x_r \sum_{t\in V\setminus\{r\}} \theta_{rt} x_t\right)}{\exp\left(2x_r \sum_{t\in V\setminus\{r\}} \theta_{rt} x_t\right) + 1}\right\}.$$

4.2 Graphical Lasso

To estimate the Gaussian graphical model, from 2006-2011, a series of papers have proposed similar methods, in the spirit of utilising ℓ_1 penalisation. These papers include Meinshausen and Bühlmann (2006), Yuan and Lin (2007), Banerjee et al. (2008), Cai et al. (2011), etc.

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$. The likelihood for μ and $\Theta = \Sigma^{-1}$, based on $\{X_1, \ldots, X_n\}$, is

$$\frac{n}{2} \log(|\Theta|) - \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^\top \Theta (X_i - \mu),$$

up to a constant not depending on μ or Θ. The maximum likelihood estimator of (μ, Σ) is $(\overline{X}, S)$, with

$$\overline{X} = n^{-1} \sum_{i=1}^n X_i \quad \text{and} \quad S = n^{-1} \sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^\top.$$

The precision (concentration) matrix can be naturally estimated by S^{-1}, but this is not a good estimator even if S is invertible when p is moderate. Before we consider further, we first investigate the likelihood.

Since the MLE of μ is $\overline{X}$, if we further assume the observations are centred, then the likelihood can be written as (up to constants)

$$\log(|\Theta|) - \frac{1}{n}\sum_{i=1}^{n} X_i^\top \Theta X_i = \log(|\Theta|) - \mathrm{tr}(\Theta S).$$

If we know that the graph is sparse, then we can adapt the Lasso estimator and study

$$\widehat{\Theta} = \underset{M \succ 0}{\arg\max} \left\{ \log(|M|) - \mathrm{tr}(MS) - \lambda \sum_{i,j=1}^{p} |M_{ij}| \right\}.$$

There are also variants of the above, where the diagonals are not penalised.

Remark 9. *Computational issues: the maxdet problem and quadratic approximations. Relationship with Van de Geer et al. (2014).*

4.3 Node-wise regression

It is also called neighbourhood selection. Let $X \sim \mathcal{N}(\mu, \Sigma)$. For any $j \in \{1, \ldots, p\}$, consider optimal prediction of X^j, given all the remaining variables. Let $\theta^j \in \mathbb{R}^p$ be the vector of coefficients for optimal prediction that

$$\theta^j = \underset{\theta \in \mathbb{R}^p : \theta_j = 0}{\arg\min}\ \mathbb{E}\left(X^j - \sum_{k \neq j} \theta_k X^k \right)^2.$$

The elements of θ^j are determined by the precision matrix. For $k \in \{1, \ldots, p\} \setminus \{j\}$ and $\Theta = \Sigma^{-1}$, it holds that $\theta_k^j = -\Theta_{jk}/\Theta_{kk}$.

Given data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$, considering the high-dimensionality, we encourage the sparsity by a Lasso penalty that

$$\widehat{\theta}^j = \underset{\theta \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} \|X^j - X^{-j}\theta\|^2 + \lambda \|\theta\|_1 \right\}.$$

We now come back to the de-biased lasso. For simplicity, we consider fixed design here. Recall that

$$\sqrt{n}(\widehat{b} - \beta^*) = n^{-1/2}\Sigma X^\top \varepsilon + n^{-1/2}(\widehat{\Theta} - \Sigma)X^\top \varepsilon - \sqrt{n}(\widehat{\Theta}\widehat{\Sigma} - I)(\widehat{\beta} - \beta^*).$$

In order to show the asymptotic normality, we need

- $\|\widehat{\Theta} - \Sigma\|_1 \|n^{-1/2} X^\top \varepsilon\|_\infty \to 0$; and
- $\|\sqrt{n}(\widehat{\beta} - \beta^*)\|_1 \|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty \to 0$.

This is to say that we need

$$\|\widehat{\Theta} - \Sigma\|_1 = o(\sqrt{\log(p)}) \quad \text{and} \quad \|\widehat{\Theta}\widehat{\Sigma} - I\|_\infty = o(\sqrt{\log(p)}).$$

With these goals in mind, we adopt the node-wise regression estimator to construct $\widehat{\Theta}$. For each $j \in \{1, \ldots, p\}$, let

$$\widehat{\theta}_j = \underset{\theta \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} \|X^j - X^{-j}\theta\|^2 + \lambda_j \|\theta\|_1 \right\},$$

with components of $\widehat{\theta}_j = \{\widehat{\theta}_{jk} : k = 1, \ldots, p,\ k \neq j\}$. Denote by

$$\widehat{C} = \begin{pmatrix} 1 & -\widehat{\theta}_{12} & \cdots & -\widehat{\theta}_{1p} \\ -\widehat{\theta}_{21} & 1 & \cdots & -\widehat{\theta}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ -\widehat{\theta}_{p1} & -\widehat{\theta}_{p2} & \cdots & 1 \end{pmatrix}$$

and write

$$\widehat{T}^2 = \mathrm{diag}(\widehat{\tau}_1^2, \ldots, \widehat{\tau}_p^2),$$

where for $j = 1, \ldots, p$,

$$\widehat{\tau}_j^2 = \|X^j - X^{-j}\widehat{\theta}_j\|^2/n + \lambda_j \|\widehat{\theta}_j\|_1.$$

Then define

$$\widehat{\Theta} = \widehat{T}^{-2}\widehat{C}.$$

It follows from the KKT conditions that

$$-(X^{-j})^\top (X^j - X^{-j}\widehat{\theta}_j)/n + \lambda_j \widehat{s}_j = 0,$$

where $\widehat{s}_j$ is the weak derivative. We therefore have that

$$\begin{aligned} \widehat{\tau}_j^2 &= \|X^j - X^{-j}\widehat{\theta}_j\|^2/n + \lambda_j \|\widehat{\theta}_j\|_1 \\ &= (X^j)^\top (X^j - X^{-j}\widehat{\theta}_j)/n - (\widehat{\theta}_j)^\top (X^{-j})^\top (X^j - X^{-j}\widehat{\theta}_j)/n + \lambda_j \|\widehat{\theta}_j\|_1 \\ &= (X^j)^\top (X^j - X^{-j}\widehat{\theta}_j)/n - \lambda_j (\widehat{\theta}_j)^\top \widehat{s}_j + \lambda_j (\widehat{\theta}_j)^\top \widehat{s}_j \\ &= (X^j)^\top (X^j - X^{-j}\widehat{\theta}_j)/n, \end{aligned}$$

which leads to that

$$X^j X \widehat{\Theta}_j / n = 1.$$

The KKT conditions also imply that

$$\|(X^{-j})^\top (X^j - X^{-j}\widehat{\theta}_j)/n\|_\infty \leq \lambda_j,$$

which is equivalent to

$$\|(X^{-j})^\top X \widehat{\Theta}_j\|_\infty / n \leq \lambda_j / \widehat{\tau}_j^2.$$

We thus have

$$\|\widehat{\Sigma}\widehat{\Theta}_j^\top - e_j\|_\infty \leq \lambda_j / \widehat{\tau}_j^2.$$

Remark 10. *Note that $\widehat{\Theta}$ is not necessarily a symmetric matrix. Relationship with Zhang and Zhang (2014).*

4.4 What we did not cover in this section

- Tuning parameter selection.
- CLIME
- Theoretical results
- Other large matrix estimation procedures: low rank, sparse, banded.
- Ising model, Markov random fields, pseudo-likelihood.
- Networks
- High-dimensional matrix estimation: banded, low-rank, etc.
- Factor analysis.

5 Fused Lasso and change point detection problems

In this section, we move on to consider the case that there exists a linear ordering among $\beta_1, \ldots, \beta_p$. The methods we are discussing are under a number of different names: total variation penalisation, fused lasso, generalised fused lasso, trend filtering, etc. At the end of this section, we will make connections with change point detection problems.

5.1 Fused Lasso

The term fused Lasso is coined in Tibshirani et al. (2005). The idea was exploited earlier in Rudin et al. (1992), Steidl et al. (2006) and others. Tibshirani et al. (2005) studied a high-dimensional regression problem, where the coefficients β^* are not only sparse, but also piece-wise constant. For simplicity, in this module, we start with a much simpler problem

$$y_i = \theta_i^0 + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i, $i = 1, \ldots, n$, are i.i.d. sub-Gaussian random variables, and $\theta^0 = (\theta_1^0, \ldots, \theta_n^0)^\top \in \mathbb{R}^n$ is a piece-wise constant mean sequence, having a set of change points

$$S_0 = \{i \in \{2, \ldots, n\} : \theta_i^0 \neq \theta_{i-1}^0\} = \{\eta_1, \ldots, \eta_K\}.$$

Let $\eta_0 = 1$ and $\eta_{K+1} = n + 1$. The fused-lasso estimator is defined as

$$\widehat{\theta} = \mathop{\arg\min}_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n-1} |\theta_i - \theta_{i+1}| \right\},$$

where $\lambda > 0$ is a tuning parameter. Let

$$\Delta = \min_{i=1,\ldots,K+1} (\eta_i - \eta_{i-1}) \quad \text{and} \quad \kappa = \min_{i=1,\ldots,K} |\theta_{\eta_i}^0 - \theta_{\eta_i - 1}^0|$$

be the minimal spacing and jump size. Let the incidence matrix be

$$D = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \cdots & -1 & 1 \end{pmatrix}.$$

We can rewrite the fused lasso estimator as

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \|D\theta\|_1 \right\}.$$

Remark 11. *Incidence matrix. General graphs. Generalised fused lasso. Higher order trend filtering.*

We then to show the de-noising property of the fused Lasso estimator (Lin et al., 2017), that, with $\lambda = (n\Delta)^{1/4}$, with probability at least $1 - \exp(-C\gamma)$, $\gamma > 1$, it holds that

$$\|\widehat{\theta} - \theta^0\|^2 \lesssim \gamma^2 \sigma^2 \frac{K}{n} \left\{ (\log(K) + \log\log(n)) \log(n) + \sqrt{\frac{n}{\Delta}} \right\}$$

and

$$\mathbb{E}\left(\|\widehat{\theta} - \theta^0\|^2 \right) \lesssim \sigma^2 \frac{K}{n} \left\{ (\log(K) + \log\log(n)) \log(n) + \sqrt{\frac{n}{\Delta}} \right\}.$$

We first remark that the minimax lower bound is

$$\inf_{\widehat{\theta}} \sup_{\|D\theta^0\|_0 \leq K} \mathbb{E}\left(\|\widehat{\theta} - \theta^0\|^2 \right) \gtrsim \sigma^2 \frac{K}{n} \log(n/K).$$

This means that when $n/\Delta \lesssim \log^2(n)$, then the fused Lasso estimator is optimal up to logarithmic factors.

The proof also starts with the basic inequality. It follows from the definition that

$$\|Y - \widehat{\theta}\|^2 + 2\lambda \|D\widehat{\theta}\|_1 \leq \|Y - \theta^0\|^2 + 2\lambda \|D\theta^0\|_1,$$

which leads to that

$$\|\widehat{\theta} - \theta^0\|^2 \leq 2\varepsilon^\top (\widehat{\theta} - \theta^0) + 2\lambda (\|D\theta^0\|_1 - \|D\widehat{\theta}\|_1).$$

Different from the Lasso case, the fluctuation process $\varepsilon^\top(\widehat{\theta} - \theta^0)$ needs more sophisticated control. To be specific, we consider the decomposition

$$\widehat{\theta} - \theta^0 = \widehat{\delta} + \widehat{x},$$

where $\widehat{\delta} = P_0(\widehat{\theta} - \theta^0)$ and $\widehat{x} = P_1\widehat{\theta}$. The matrix P_0 denotes the projection matrix onto the piecewise constant structure inherent in θ^0 and $P_1 = I - P_0$. With such notation, we have that

$$\begin{aligned} &\|\widehat{\theta} - \theta^0\|^2 = \|\widehat{\delta}\|^2 + \|\widehat{x}\|^2 \leq 2\varepsilon^\top \widehat{\delta} + 2\varepsilon^\top \widehat{x} + 2\lambda(\|D\theta^0\|_1 - \|D\widehat{\theta}\|_1) \\ =&2\varepsilon^\top \widehat{\delta} + 2\varepsilon^\top \widehat{x} + 2\lambda(\|D_{S_0}\theta^0\|_1 - \|D_{S_0}\widehat{\theta}\|_1 - \|D_{-S_0}\widehat{\theta}\|_1) \\ \leq&2\varepsilon^\top \widehat{\delta} + 2\varepsilon^\top \widehat{x} + 2\lambda(\|D_{S_0}(\theta^0 - \widehat{\theta})\|_1 - \|D_{-S_0}\widehat{\theta}\|_1) \end{aligned}$$

$$
\begin{aligned}
&\leq 2\varepsilon^\top \widehat{\delta} + 2\varepsilon^\top \widehat{x} + 2\lambda(\|D_{S_0}\widehat{\delta}\|_1 + \|D_{S_0}\widehat{x}\|_1 - \|D_{-S_0}\widehat{x}\|_1) \\
&= \underbrace{2\varepsilon^\top \widehat{\delta} + 2\lambda\|D_{S_0}\widehat{\delta}\|_1}_{A_0} + \underbrace{2\varepsilon^\top \widehat{x} + 2\lambda(\|D_{S_0}\widehat{x}\|_1 - \|D_{-S_0}\widehat{x}\|_1)}_{B_0}.
\end{aligned}
$$

We then bound the terms A_0 and B_0 separately. As for the term A_0, which only involves $\widehat{\delta}$. The quantity $\widehat{\delta}$ lies in a low-dimensional subspace. The term B_0 involves $\widehat{x}$, which requires more intricate argument.

Bounding A_0. Note that

$$
A_0 = 2\left(\frac{|\varepsilon^\top \widehat{\delta}|}{\|\widehat{\delta}\|} + \lambda \frac{\|D_{S_0}\widehat{\delta}\|_1}{\|\widehat{\delta}\|}\right)\|\widehat{\delta}\|
$$

and

$$
\begin{aligned}
\|D_{S_0}\widehat{\delta}\|_1 = \sum_{i=1}^{K+1} |\widehat{\delta}_{\eta_i} - \widehat{\delta}_{\eta_{i-1}}| &\leq 2\sum_{i=1}^{K+1} |\widehat{\delta}_{\eta_i}| \leq 2\sqrt{(K+1)\sum_{i=1}^{K+1} \widehat{\delta}^2_{\eta_i}} \\
&\leq 4\sqrt{K\sum_{i=1}^{K+1} \frac{\eta_i - \eta_{i-1}}{\Delta}\widehat{\delta}^2_{\eta_i}} \leq 4\sqrt{\frac{K}{\Delta}}\|\widehat{\delta}\|.
\end{aligned}
$$

Consider the large probability event

$$
\Omega_1 = \left\{\sup_{\delta \in \mathcal{R}} \frac{\varepsilon^\top \delta}{\|\delta\|} \leq \gamma C\sqrt{K}\right\},
$$

where $\mathcal{R} = \text{span}\{\mathbb{1}_{B_0}, \ldots, \mathbb{1}_{B_K}\}$, $B_j = \{\eta_j, \ldots, \eta_{j+1} - 1\}$. In the event Ω_1, we have that

$$
A_0 \leq 2(\gamma c\sqrt{K} + 4\lambda\sqrt{K/\Delta})\|\widehat{\delta}\|.
$$

Bounding B_0. In order to bound B_0, we consider a lower interpolant $\widehat{z}$ to $\widehat{x}$. This interpolant approximates $\widehat{x}$ using $2K+2$ monotonic segments, and the corresponding fluctuation $\varepsilon^\top \widehat{z}$ can be controlled. The residual from the interpolant approximation, denoted $\widehat{w} = \widehat{x} - \widehat{z}$, can also be controlled. Putting the results together, we bound the term B_0.

We first define the class of vectors containing the lower interpolant. Let $\mathcal{M}$ be the set of piecewise monotonic vectors $z \in \mathbb{R}^n$, with the following properties, for each $i \in \{0, \ldots, K\}$:

(i) there exists a point $t \in [t_i, t_{i+1})$ such that the absolute value $|z_j|$ is non-increasing over the segment $j \in [t_i, t]$ and non-decreasing over the segment $j \in (t, t_{i+1})$;

(ii) the signs remain constant on the monotone pieces that

$$
\text{sign}(z_{\eta_i})\text{sign}(z_j) \geq 0, \quad j \in [\eta_i, t]
$$

and

$$
\text{sign}(z_{\eta_{i+1}})\text{sign}(z_j) \geq 0, \quad j \in (t, \eta_{i+1}).
$$

With this notation, we first notice that, for any $x \in \mathbb{R}^n$, there exists $z \in \mathcal{M}$ (not necessarily unique), such that the following hold:

$$\begin{aligned}
&\|D_{-S_0}x\|_1 = \|D_{-S_0}z\|_1 + \|D_{-S_0}(x-z)\|_1,\\
&\|D_{S_0}x\|_1 = \|D_{S_0}z\|_1 \le \|D_{-S_0}z\|_1 + \frac{4\sqrt{K}}{\sqrt{\Delta}}\|z\|,\\
&\|z\| \le \|x\| \quad \text{and} \quad \|x-z\| \le \|x\|.
\end{aligned}$$

In order to control $\varepsilon^\top \widehat{z}$, we have the following result

$$\mathbb{P}\left\{\sup_{z\in\mathcal{M}} \frac{\varepsilon^\top z}{\|z\|} > \gamma c\sqrt{(\log(K)+\operatorname{loglog}(n))K\log(n)}\right\} \le 2\exp\left\{-C\gamma^2(\log(K)+\log\log(n))\right\}.$$

In order to control $\varepsilon^\top \widehat{w}$, we have the following result

$$\mathbb{P}\left\{\sup_{w\in\mathcal{R}^\perp} \frac{|\varepsilon^\top w|}{\sqrt{\|D_{-S_0}w\|_1\|w\|}} > \gamma(nK)^{1/4}\right\} \le 2\exp(-C\gamma^2\sqrt{K}).$$

Remark 12.

- *Computational cost.*
- *Optimality.*
- *General designs.*

5.2 The ℓ_0-penalisation and dynamic programming

In the Lasso estimation, we mention that the ℓ_1 penalisation is a convex relaxation of the ℓ_0 penalisation. In the linear regression problem we studied thereof, there is no ordering among the covariates, therefore, finding a non-zero subset of $\{1,\dots,p\}$, is an NP-hard problem. In what we discuss in this section, there is a linear ordering among $\theta_1^0,\dots,\theta_n^0$, and therefore a polynomial time algorithm is reachable.

Formally speaking, instead of studying the fused Lasso estimator, we study the ℓ_0-penalised estimator, also known as the Potts estimator, which is defined as

$$\widetilde{\theta} = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^n} H(\theta) = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^n}\left\{\frac{1}{2}\|Y-\theta\|^2 + \lambda\|D\theta\|_0\right\}.$$

We first introduce the dynamic programming algorithm, which solves the above optimisation problem (Friedrich et al., 2008). It starts with $t=2$ and $t_0=0$. At every time point $t\ge 2$, it computes

$$\widehat{t} = \operatorname*{arg\,max}_{s\in\{t_0+1,\dots,t-1\}}\left\{\sum_{i=t_0+1}^{t}(Y_i-\overline{Y}_{[t_0+1,t]})^2 - \sum_{i=t_0+1}^{s}(Y_i-\overline{Y}_{[t_0+1,s]})^2 - \sum_{i=s+1}^{t}(Y_i-\overline{Y}_{[s+1,t]})^2\right\}.$$

If

$$\sum_{i=t_0+1}^{t}(Y_i-\overline{Y}_{[t_0+1,t]})^2 - \sum_{i=t_0+1}^{\widehat{t}}(Y_i-\overline{Y}_{[t_0+1,\widehat{t}]})^2 - \sum_{i=\widehat{t}+1}^{t}(Y_i-\overline{Y}_{[\widehat{t}+1,t]})^2 \ge \lambda,$$

then we let $t_0 = \widehat{t}$, $t = t+1$ and include $\widehat{t}$ as a new change point; otherwise, we just update $t = t+1$.

The computational cost of this dynamic programming is of order $O(n^2)$. There are many different variants of the dynamic programming, aiming to accelerate the algorithm. We remark that without stronger model assumptions, all these variants in fact operate with the same computational cost $O(n^2)$. Despite the larger computational cost, the ℓ_0-penalised estimator is optimal in terms of both the required condition and its change point estimator.

In the change point analysis problem, our goal is to obtain consistent change point estimators $\{\widehat{\eta}_1, \ldots, \widehat{\eta}_{\widehat{K}}\}$, with $\widehat{\eta}_1 < \ldots \widehat{\eta}_{\widehat{K}}$, such that

$$\widehat{K} = K \quad \text{and} \quad \max_{k=1,\ldots,\widehat{K}} |\widehat{\eta}_k - \eta_k| = \epsilon,$$

where $\epsilon/n \to 0$, with probability tending to 1 as $n \to \infty$.

In order to quantify the difficulty of the problem, we rely on the quantity $\kappa\sqrt{\Delta}/\sigma$, which is a signal-to-noise ratio. It is established that, if

$$\kappa\sqrt{\Delta}/\sigma < \sqrt{\log(n)},$$

then in the minimax sense, there is no consistent estimator of change points. Based on this minimax lower bound, in order to claim that $\widetilde{\theta}$ is optimal in terms of condition, then we need to show that $\widetilde{\theta}$ is consistent provided that

$$\kappa\sqrt{\Delta}/\sigma \gtrsim \log^{1/2}(n).$$

In addition, it is also established that

$$\inf_{\widehat{\eta}} \sup_{P:\kappa\sqrt{\Delta}/\sigma \gtrsim \log^{1/2}(n)} \mathbb{E}_P\{H(\widehat{\eta}, \eta)\} \geq c\sigma^2\kappa^{-2},$$

where $H(\cdot, \cdot)$ denotes the two-sided Hausdorff distance. Based on this minimax lower bound, in order to claim that $\widetilde{\theta}$ is optimal in terms of estimation, then we need to show that the change points induced by θ has the localisation error rate of order $\sigma^2\kappa^{-2}$.

To show the optimality of $\widetilde{\theta}$, we show provided that

$$\kappa\sqrt{\Delta}/\sigma \geq C\sqrt{\log(n)}a_n,$$

where a_n is any arbitrarily diverging sequence, with $\lambda = C_\lambda\sigma^2\log(n)$, it holds that

$$\mathbb{P}\left\{\widetilde{K} = K \quad \text{and} \quad |\widetilde{\eta}_k - \eta_k| \leq C\sigma^2\log(n)/\kappa_k^2, \, \forall k\right\} \geq 1 - n^{-c},$$

where

- $\{\widetilde{\eta}_k\}_{k=1}^{\widetilde{K}}$ is the collection of change points induced by $\widetilde{\theta}$ and
- $\kappa_k = |\theta_{\eta_k} - \theta_{\eta_k - 1}|$.

Remark 13. *Optimality.*

The proof of the above result is conducted in the large probability event defined below

$$\mathcal{A} = \left\{ \max_{0 \leq a < b < c \leq n} \sqrt{\frac{(b-a)(c-b)}{c-a}} \left| \overline{Y}_{[a+1,b]} - \overline{\theta^*}_{[a+1,b]} + \overline{Y}_{[b+1,c]} - \overline{\theta^*}_{[b+1,c]} \right| \leq \sigma\sqrt{C_\lambda \log(n)} \right\}.$$

Let $\widetilde{P}$ be the partition induced by $\widetilde{\theta}$. In order to show the performances of $\widetilde{\theta}$, we let $[s, e]$ be any member of $\widetilde{P}$ and complete the proof in the following four steps.

S1. The interval $[s, e]$ contains no more than two true change points.

S2. If $[s, e]$ contains exactly two true change points, say η_k and η_{k+1}, then

$$\eta_k - s \leq C\sigma^2 \log(n)/\kappa_k^2 \quad \text{and} \quad e - \eta_{k+1} \leq C\sigma^2 \log(n)/\kappa_{k+1}^2.$$

S3. If $[s, e]$ contains only one true change point, say η_k, without loss of generality, letting $\eta_k - s \leq e - \eta_k$, then it must hold that

$$s \leq \eta_k \leq e \leq \eta_{k+1}, \quad \eta_k - s \leq C\sigma^2 \log(n)/\kappa_k^2 \quad \text{and} \quad e - \eta_{k+1} \leq C\sigma^2 \log(n)/\kappa_{k+1}^2.$$

S4. If $[s, e]$ contains no true change point, then there must exist two true change points, say η_k and η_{k+1}, satisfying that

$$\eta_k < s < e < \eta_{k+1}, \quad s - \eta_k \leq C\sigma^2 \log(n)/\kappa_k^2 \quad \text{and} \quad \eta_{k+1} - e \leq C\sigma^2 \log(n)/\kappa_{k+1}^2.$$

See detailed proofs in Wang et al. (2020). We prove S1. here for an illustration. We first observe that, for any two disjoint intervals $I_1, I_2 \subset \{1, \ldots, n\}$, $I = I_1 \cup I_2$, it holds that

$$\sum_{i \in I} (Y_i - \overline{Y}_I)^2 = \sum_{i \in I_1} (Y_i - \overline{Y}_{I_1})^2 + \sum_{i \in I_2} (Y_i - \overline{Y}_{I_2})^2 + \frac{|I_1||I_2|}{|I_1| + |I_2|} (\overline{Y}_{I_1} - \overline{Y}_{I_2})^2.$$

To show S1., we prove by contradiction and assume that there exists $I \in \widetilde{P}$ containing at least three true change points. This implies that there exists $\eta_k \in I$ satisfying that

$$\min\{e - \eta_k, \, \eta_k - s\} > \Delta.$$

Denote $I_1 = [s, \eta_k - \Delta/3]$, $I_2 = (\eta_k - \Delta/3, \eta_k)$, $I_3 = [\eta_k, \eta_k + \Delta/3]$ and $I_4 = (\eta_k + \Delta/3, e]$. Let $\widetilde{P}_1$ be such that

$$\widetilde{P}_1 = \widetilde{P} \cup \{I_1, I_2, I_3, I_4\} \setminus \{I\}$$

and u be the piecewise constant vector induced by $\widetilde{P}_1$. By the definition of $\widetilde{\theta}$, it holds that

$$\begin{aligned}
0 \geq H(\widehat{\theta}) - H(u) &= -3\lambda + \sum_{i \in I} (Y_i - \overline{Y}_I)^2 - \sum_{i \in I_1} (Y_i - \overline{Y}_{I_1})^2 \\
&\quad - \sum_{i \in I_2} (Y_i - \overline{Y}_{I_2})^2 - \sum_{i \in I_3} (Y_i - \overline{Y}_{I_3})^2 - \sum_{i \in I_4} (Y_i - \overline{Y}_{I_4})^2 \\
&\geq -3\lambda + \frac{|I_2||I_3|}{|I_2| + |I_3|} (\overline{Y}_{I_2} - \overline{Y}_{I_3})^2 \\
&= -3\lambda + \frac{|I_2||I_3|}{|I_2| + |I_3|} \{(\overline{Y}_{I_2} - \theta_{\eta_k}) - (\overline{Y}_{I_3} - \theta_{\eta_{k+1}}) + (\theta_{\eta_k} - \theta_{\eta_{k+1}})\}^2 \\
&\geq -3\lambda + \frac{\Delta}{12} (\theta_{\eta_k} - \theta_{\eta_{k+1}})^2 - \frac{|I_2||I_3|}{|I_2| + |I_3|} \{(\overline{Y}_{I_2} - \theta_{\eta_k}) - (\overline{Y}_{I_3} - \theta_{\eta_{k+1}})\}^2 \\
&\geq -4\lambda + \frac{\Delta}{12} \kappa_k^2 > 0,
\end{aligned}$$

which leads to the contradiction.

5.3 Change point detection in high-dimensional linear regression models

Now we consider a more challenging problem, which will bring what we have learnt together. Let the data $\{(x_t, y_t)\}_{t=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ satisfy the model

$$y_t = x_t^\top \beta_t^* + \varepsilon_t, \quad t = 1, \ldots, n,$$

where $\{\beta_t^*\}_{t=1}^n \subset \mathbb{R}^p$ are the unknown coefficient vectors, $\{x_t\}_{t=1}^n$ are i.i.d. mean-zero sub-Gaussian random vectors with $\mathbb{E}(x_t x_t^\top) = \Sigma$, and $\{\varepsilon_t\}_{t=1}^n$ are independent mean-zero sub-Gaussian random variables with sub-Gaussian parameters upper bounded by σ^2 and independent of $\{x_t\}_{t=1}^n$. In addition, there exists a sequence of change points $1 = \eta_0 < \eta_1 < \ldots \eta_K \leq n < \eta_{K+1} = n + 1$ such that $\beta_t^* \neq \beta_{t+1}^*$, if and only if $t \in \{\eta_k\}_{k=1}^K$.

We can solve this problem using an ℓ_0-penalisation framework with ℓ_1-penalisation sub-routine. To be specific, we let

$$\widehat{\mathcal{P}} \in \arg\min_{\mathcal{P}} \left\{ \sum_{I \in \mathcal{P}} L(I) + \gamma |\mathcal{P}| \right\},$$

where

- the minimisation is over all possible interval partitions of $\{1, \ldots, n\}$,
- the loss function
$$L(I) = \sum_{t \in I} (y_t - x_t^\top \widehat{\beta}_I^\lambda)^2,$$
with
$$\widehat{\beta}_I^\lambda = \arg\min_{v \in \mathbb{R}^p} \left\{ \sum_{t \in I} (y_t - x_t^\top v)^2 + \lambda \sqrt{\max\{|I|, \log(n \vee p)\}} \|v\|_1 \right\}.$$

Remark 14. *Optimality.*

To refine $\widehat{P}$, we can have a further step to prompt the optimality, that is

$$(\widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\eta}_k) = \arg\min_{\substack{\eta \in \{s_k+1, \ldots, e_k-1\} \\ \beta_1, \beta_2 \in \mathbb{R}^p, \beta_1 \neq \beta_2}} \left\{ \sum_{t=s_k+1}^{\eta} \|y_t - \beta_1^\top x_t\|^2 + \sum_{t=\eta+1}^{e_k} \|y_t - \beta_2 x_t\|^2 \right.$$
$$\left. + \zeta \sum_{i=1}^{p} \sqrt{(\eta - s_k)(\beta_1)_i^2 + (e_k - \eta)(\beta_2)_i^2} \right\},$$

where $s_k = 2\widehat{\eta}_{k-1}/3 + \widehat{\eta}_k/3$ and $e_k = \widehat{\eta}_k/3 + 2\widehat{\eta}_{k+1}/3$, $\zeta > 0$ is a tuning parameter.

Remark 15. *Group lasso penalty.*

5.4 What we did not cover in this section

- Dyadic CART
- Binary segmentation
- Minimax lower bounds

- Tuning parameter selection
- General graphs, lattices
- Dependence in data
- Testing results

6 Functional linear regression and reproducing kernel Hilbert spaces

In terms of the types of linear regression, we have already seen the case where both the predictors and responses are vectors/scalars. More generally speaking, there are more complicated linear regressions, where the predictors and/or responses are functions.

In terms of the types of penalisation, we have already seen ℓ_0 and ℓ_1 penalisations. Although we have seen that they can be solved in some scenarios, in the whole real line, neither is differentiable. Considering the computations, ℓ_2 penalisation, also known as ridge regression, is a more natural choice.

In this section, we consider a functional linear regression problem, utilising both ℓ_2 and ℓ_1 penalisations, also requiring knowledge in reproducing kernel Hilbert spaces (RKHS).

The model we consider in this model is

$$Y_t(r) = \int_{[0,1]} A^*(r,s) X_t(s)\,\mathrm{d}s + \sum_{j=1}^{p} Z_{tj}\beta_j^*(r) + \epsilon_t(r), \quad r \in [0,1], \quad t \in \{1,\ldots,T\},\, r \in [0,1],$$

where

- $Y_t(\cdot) : [0,1] \to \mathbb{R}$ is the functional response,
- $X_t(\cdot) : [0,1] \to \mathbb{R}$ is the functional covariate,
- $Z_t = (Z_{tj})_{j=1}^p \in \mathbb{R}^p$ is the vector covariate,
- $\epsilon_t(\cdot) : [0,1] \to \mathbb{R}$ is the functional noise,
- $A^*(\cdot,\cdot) : [0,1] \times [0,1] \to \mathbb{R}$ is a bivariate coefficient function, and
- $\{\beta_j^*(\cdot) : [0,1] \to \mathbb{R}\}_{j=1}^p$ is a collection of p univariate coefficient functions.

We consider a discretised observations of the functions. To be specific, assume the data are

$$\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=1,i=1,j=1}^{T,n_1,n_2}.$$

In order to provide theoretical guarantees, we need some regularity conditions on the behaviours of the functions.

6.1 Reproducing kernel Hilbert spaces (RKHS)

Let $\mathcal{L}^2 = \mathcal{L}^2([0,1])$ be the space of all square integrable functions with respect to the uniform distribution on $[0,1]$, i.e. $\mathcal{L}^2 = \{f : [0,1] \to \mathbb{R},\ \|f\|^2_{\mathcal{L}^2} = \int_{[0,1]} f^2(s)\,\mathrm{d}s < \infty\}$. We consider a Hilbert space $\mathcal{H} \subset \mathcal{L}^2$ and an associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, under which $\mathcal{H}$ is complete. We assume that there exists a continuous symmetric nonnegative-definite kernel functions $\mathbb{K} : [0,1] \times [0,1] \to \mathbb{R}_+$ such that the space $\mathcal{H}$ is an RKHS, in the sense that for each $s \in [0,1]$, the function $\mathbb{K}(\cdot, s) \in \mathcal{H}$ and $g(s) = \langle g(\cdot), \mathbb{K}(\cdot, s) \rangle_{\mathcal{H}}$, for all $g \in \mathcal{H}$.

It follows from Mercer's theorem (Mercer, 1909) that there exists an orthonormal basis of $\mathcal{L}^2$, $\{\phi_k\}_{k=1}^{\infty} \subset \mathcal{L}^2$, such that $\mathbb{K}(\cdot, \cdot)$ has the representation $\mathbb{K}(s,t) = \sum_{k=1}^{\infty} \mu_k \phi_k(s)\phi_k(t),\ s,t \in [0,1]$, where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are the eigenvalues of $\mathbb{K}$ and $\{\phi_k\}_{k=1}^{\infty}$ are the corresponding eigen-functions. We further denote $\Phi_k = \sqrt{\mu_k}\phi_k$ and note that $\|\Phi_k\|_{\mathcal{H}} = 1$, for $k \in \mathbb{N}$.

Any function $f \in \mathcal{H}$ can be then written as

$$f(s) = \sum_{k=1}^{\infty} \left\{ \int_{[0,1]} f(s)\phi_k(s)\,\mathrm{d}s \right\} \phi_k(s) = \sum_{k=1}^{\infty} a_k \phi_k(s), \quad s \in [0,1].$$

Its RKHS norm is defined as $\|f\|_{\mathcal{H}} = \sqrt{\sum_{k=1}^{\infty} a_k^2/\mu_k}$. For the eigen-functions, we have $\|\phi_k\|^2_{\mathcal{H}} = \mu_k^{-1}$.

Remark 16.

- *RKHS vs. Functional PCA.*
- *SVD*
- *Choice of kernel functions.*

We now consider the class of Hilbert–Schmidt operators, which is an important subclass of compact linear operators.

For any compact linear operator $A_2 : \mathcal{H} \to \mathcal{H}$, denote

$$A_2[f,g] = \langle A_2[g], f \rangle_{\mathcal{H}}, \quad f,g \in \mathcal{H}.$$

Note that $A_2[f,g]$ is well defined for any $f,g \in \mathcal{H}$ due to the compactness of A_2. Define $a_{ij} = A_2[\Phi_i, \Phi_j] = \langle A_2[\Phi_j], \Phi_i \rangle_{\mathcal{H}}$, $i,j \in \mathbb{N}$. We thus have for any $f,g \in \mathcal{H}$, it holds that

$$\begin{aligned} A_2[f,g] = \langle A_2[g], f \rangle_{\mathcal{H}} &= \sum_{i=1}^{\infty} \langle f, \Phi_i \rangle_{\mathcal{H}} \langle A_2[g], \Phi_i \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \langle f, \Phi_i \rangle_{\mathcal{H}} \langle A_2 \left[\sum_{j=1}^{\infty} \langle g, \Phi_j \rangle_{\mathcal{H}} \Phi_j \right], \Phi_i \rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{\infty} \langle f, \Phi_i \rangle_{\mathcal{H}} \langle g, \Phi_j \rangle_{\mathcal{H}} \langle A_2[\Phi_j], \Phi_i \rangle_{\mathcal{H}} = \sum_{i,j=1}^{\infty} a_{ij} \langle f, \Phi_i \rangle_{\mathcal{H}} \langle g, \Phi_j \rangle_{\mathcal{H}}. \end{aligned}$$

Therefore, we can define a bivariate function $A_1(r,s)$, $r,s \in [0,1]$, affiliated with the compact linear operator A_2. Plugging $f = \mathbb{K}(r, \cdot)$ and $g = \mathbb{K}(s, \cdot)$ into the above, we have that

$$A_1(r,s) = \sum_{i,j=1}^{\infty} a_{ij} \langle \mathbb{K}(r, \cdot), \Phi_i \rangle_{\mathcal{H}} \langle \mathbb{K}(s, \cdot), \Phi_j \rangle_{\mathcal{H}} = \sum_{i,j=1}^{\infty} a_{ij} \Phi_i(r) \Phi_j(s).$$

Furthermore, it is straightforward to verify that for any $v \in \mathcal{H}$, we have that

$$A_2[v](r) = \langle A_1(r, \cdot), v(\cdot) \rangle_{\mathcal{H}}, \quad r \in [0, 1].$$

Thus, we have established an equivalence between a compact linear operator A_2 and its corresponding bivariate function $A_1(r, s)$, therefore any compact linear operator $A_2 : \mathcal{H} \to \mathcal{H}$ can be viewed as a bivariate function $A_1 : [0, 1] \times [0, 1] \to \mathbb{R}$.

6.2 The penalised estimator and the representer theorem

Given absolute constants $C_A, C_\beta > 0$ and two RKHS's $\mathcal{H}$ and $\mathcal{H}_\beta$, we define

$$\mathcal{C}_A = \{A : \mathcal{H} \to \mathcal{H}, \, \|A\|_{\mathrm{F}} \leq C_A\} \quad \text{and} \quad \mathcal{C}_\beta = \left\{ \{\beta_j\}_{j=1}^p \subset \mathcal{H}_\beta : \sum_{j=1}^p \|\beta_j\|_{\mathcal{H}_\beta}^2 \leq C_\beta \right\},$$

where $\|A\|_{\mathrm{F}}^2 = \sum_{i,j=1}^\infty \langle A[\Phi_j], \Phi_i \rangle_{\mathcal{H}}^2$.

The constrained/penalised least squares estimator is

$$(\widehat{A}, \widehat{\beta}) = \underset{\substack{A \in \mathcal{C}_A \\ \{\beta_j\}_{j=1}^p \in \mathcal{C}_\beta}}{\arg\min} \left[\frac{1}{T n_2} \sum_{t=1}^T \sum_{j=1}^{n_2} \left\{ Y_t(r_j) - \frac{1}{n_1} \sum_{i=1}^{n_1} A(r_j, s_i) X_t(s_i) - \langle \beta(r_j), Z_t \rangle_p \right\}^2 + \lambda \sum_{j=1}^p \|\beta_j\|_{n_2} \right],$$

where

- $\langle \cdot, \cdot \rangle_p$ denotes the p-dimensional vector inner product;
- $\|\cdot\|_n$ denotes the discretised ℓ_2-norm; to be specific, given an observation grid $\{t_i\}_{i=1}^n \subset [0, 1]$, $\|f\|_n = \sqrt{n^{-1} \sum_{i=1}^n f^2(t_i)}$.

This is an infinite dimensional optimisation problem. Fortunately, a form of the representer theorem states that the estimator $(\widehat{A}, \widehat{\beta})$ can in fact be written as linear combinations of their corresponding kernel functions evaluated at the discrete grids $\{s_i\}_{i=1}^{n_1}$ and $\{r_j\}_{j=1}^{n_2}$.

Denote $\mathbb{K}$ and $\mathbb{K}_\beta$ as the RKHS kernels of $\mathcal{H}$ and $\mathcal{H}_\beta$ respectively. There always exists a minimiser $(\widehat{A}, \widehat{\beta})$ such that

$$\widehat{A}(r, s) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \widehat{a}_{ij} \mathbb{K}(r, r_j) \mathbb{K}(s, s_i), \quad (r, s) \in [0, 1] \times [0, 1], \, \{\widehat{a}_{ij}\}_{i,j=1}^{n_1, n_2} \subset \mathbb{R},$$

and

$$\widehat{\beta}_l(r) = \sum_{j=1}^{n_2} \widehat{b}_{lj} \mathbb{K}_\beta(r_j, r), \quad r \in [0, 1], \, l \in \{1, \ldots, p\}, \, \{\widehat{b}_{lj}\}_{l=1,j=1}^{p,n_2} \subset \mathbb{R}.$$

Due to the equivalence between constrained and penalised optimisation, we can formulate the above into a penalised optimisation such that

$$(\widehat{A}, \widehat{\beta}) = \underset{A, \beta}{\arg\min} \left[\sum_{t=1}^T \sum_{j=1}^{n_2} \left\{ Y_t(r_j) - \frac{1}{n_1} \sum_{i=1}^{n_1} A(r_j, s_i) X_t(s_i) - \langle \beta(r_j), Z_t \rangle_p \right\}^2 \right.$$

$$+ \lambda_1 \|A\|_{\mathrm{F}}^2 + \lambda_2 \sum_{l=1}^{p} \|\beta_l\|_{\mathcal{H}}^2 + \lambda_3 \sum_{l=1}^{p} \|\beta_l\|_{n_2} \Bigg].$$

The representer theorem is a very powerful tool, which is used almost everywhere when the kernels are used. We present a proof here.

For any linear subspaces $R, S \subset \mathcal{H}$, let $\mathcal{P}_R$ and $\mathcal{P}_S$ be the projection mappings of the spaces R and S, respectively, with respect to $\|\cdot\|_{\mathcal{H}}$. For any $f, g \in \mathcal{H}$ and any compact linear operator $A : \mathcal{H} \to \mathcal{H}$, denote

$$A|_{R\times S}[f, g] = A[\mathcal{P}_R f, \mathcal{P}_S g].$$

Let $(\widehat{B}, \widehat{\alpha})$ be any solution to the optimisation problem. Let $R_1 = \mathrm{span}\{\mathbb{K}_\beta(r_j, \cdot)\}_{j=1}^n \subset \mathcal{H}_\beta$, $R = \mathrm{span}\{\mathbb{K}(r_j, \cdot)\}_{j=1}^n \subset \mathcal{H}$ and $S = \mathrm{span}\{\mathbb{K}(s_j, \cdot)\}_{j=1}^n \subset \mathcal{H}$. Denote $\widehat{\beta}_l = \mathcal{P}_R(\widehat{\alpha}_l)$, $l \in \{1, \ldots, p\}$ and $\widehat{A}[\cdot, \cdot] = \widehat{B}|_{R\times S}[\cdot, \cdot] = \widehat{B}[\mathcal{P}_R \cdot, \mathcal{P}_S \cdot]$.

Let $S^\perp$ and $R^\perp$ be the orthogonal complements of S and R in $\mathcal{H}$. Then for any compact linear operator A, we have the decomposition

$$A = A|_{R\times S} + A|_{R\times S^\perp} + A|_{R^\perp \times S} + A|_{R^\perp \times S^\perp}.$$

We can show that there exist $\{a_{ij}\}_{i,j=1}^n \subset \mathbb{R}$ such that

$$A|_{R\times S}[f, g] = \sum_{i,j=1}^{n} a_{ij} \langle \mathbb{K}(r_i, \cdot), f\rangle_{\mathcal{H}} \langle \mathbb{K}(s_j, \cdot), g\rangle_{\mathcal{H}}.$$

To complete the proof, we proceed in four steps.

S1. In this step, we are to show that for any compact linear operator A, its associate bivariate function $A(\cdot, \cdot)$ satisfies that $\{A(r_i, s_j)\}_{i,j=1}^n$ only depend on $A|_{R\times S}$. This is because

$$A|_{R\times S^\perp}(r_i, s_j) = A|_{R\times S^\perp}[\mathbb{K}(r_i, \cdot), \mathbb{K}(s_j, \cdot)] = A[\mathcal{P}_R \mathbb{K}(r_i, \cdot), \mathcal{P}_{S^\perp} \mathbb{K}(s_j, \cdot)] = 0.$$

Similar arguments also lead to that $A|_{R^\perp\times S}(r_i, s_j) = A|_{R^\perp\times S^\perp}(r_i, s_j) = 0$.

S2. By S1. we have that $\widehat{A}(r_i, s_j) = \widehat{B}(r_i, s_j)$, $i, j = 1, \ldots, n$, which implies that, there exists $\{a_{ij}\}_{i,j=1}^n \subset \mathbb{R}$ such that for any $f, g \in \mathcal{H}$,

$$\widehat{A}[f, g] = \sum_{i,j=1}^{n} \widehat{a}_{ij} \langle \mathbb{K}(r_i, \cdot), f\rangle_{\mathcal{H}} \langle \mathbb{K}(s_j, \cdot), g\rangle_{\mathcal{H}}.$$

Therefore, the associated bivariate function satisfies that, for all $r, s \in [0, 1]$,

$$\begin{aligned}\widehat{A}(r, s) &= \sum_{i,j=1}^{\infty} \widehat{A}[\Phi_i, \Phi_j]\Phi_i(r)\Phi_j(s) = \sum_{i,j=1}^{\infty} \sum_{k,l=1}^{n} \widehat{a}_{kl}\Phi_i(r_k)\Phi_j(s_l)\Phi_i(r)\Phi_j(s) \\ &= \sum_{k,l=1}^{n} \widehat{a}_{kl} \left\{\sum_{i=1}^{\infty} \Phi_i(r)\Phi_i(r_k)\right\} \left\{\sum_{i=1}^{\infty} \Phi_i(s)\Phi_i(s_l)\right\} = \sum_{k,l=1}^{n} \widehat{a}_{kl} \mathbb{K}(r, r_k)\mathbb{K}(s, s_l).\end{aligned}$$

S3. Similar arguments lead to that, for any $j \in \{1, \ldots, n\}$ and $l \in \{1, \ldots, p\}$, we have that

$$\widehat{\alpha}_l(r_j) = \langle \widehat{\alpha}_l, \mathbb{K}_\beta(r_j, \cdot)\rangle_{\mathcal{H}_\beta} = \widehat{\beta}_l(r_j).$$

S4. We now have that

$$\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{Y_t(r_i)-\frac{1}{n}\sum_{j=1}^{n}\widehat{A}(r_i,s_j)X_t(s_j)-\langle\widehat{\beta}(r_i),Z_t\rangle_p\right\}^2+\lambda\sum_{l=1}^{p}\|\widehat{\beta}_l\|_n$$
$$=\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{Y_t(r_i)-\frac{1}{n}\sum_{j=1}^{n}\widehat{B}(r_i,s_j)X_t(s_j)-\langle\widehat{\alpha}(r_i),Z_t\rangle_p\right\}^2+\lambda\sum_{l=1}^{p}\|\widehat{\alpha}_l\|_n,$$

$$\|\widehat{A}\|_{\mathrm{F}}\le\|\widehat{B}\|_{\mathrm{F}}\le C_A\quad\text{and}\quad\sum_{l=1}^{p}\|\widehat{\beta}_l\|_{\mathcal{H}_\beta}\le\sum_{l=1}^{p}\|\widehat{\alpha}_l\|_{\mathcal{H}_\beta}\le C_\beta.$$

We now discuss the numerical issues. For any $r,s\in[0,1]$, denote the RKHS kernels as

$k_1(r)=[\mathbb{K}(r,r_1),\mathbb{K}(r,r_2),\cdots,\mathbb{K}(r,r_{n_2})]^\top\in\mathbb{R}^{n_2}$ and $k_2(s)=[\mathbb{K}(s,s_1),\mathbb{K}(s,s_2),\cdots,\mathbb{K}(s,s_{n_1})]^\top\in\mathbb{R}^{n_1}$.

Denote $K_1=[k_1(r_1),k_1(r_2),\cdots,k_1(r_{n_2})]\in\mathbb{R}^{n_2\times n_2}$ and $K_2=[k_2(s_1),k_2(s_2),\cdots,k_2(s_{n_1})]\in\mathbb{R}^{n_1\times n_1}$. Note that $K_1=\langle k_1(r),k_1(r)^\top\rangle_{\mathcal{H}}$ and $K_2=\langle k_2(s),k_2(s)^\top\rangle_{\mathcal{H}}$, thus both are symmetric and positive definite matrices.

By the representer theorem, we have the minimier taking the form

$$A(r,s)=k_1(r)^\top Rk_2(s)\text{ and }\beta_l(r)=k_1(r)^\top\mathbf{b}_l,\ l=1,\cdots,p,$$

where $R\in\mathbb{R}^{n_2\times n_1}$ is an $n_2\times n_1$ matrix and $\mathbf{b}_l=[b_{l1},b_{l2},\cdots,b_{ln_2}]^\top\in\mathbb{R}^{n_2}$ is an n_2-dimensional vector for $l=1,\cdots,p$. Denote $\beta(r)=[\beta_1(r),\beta_2(r),\cdots,\beta_p(r)]^\top=[\mathbf{b}_1,\mathbf{b}_2,\cdots,\mathbf{b}_p]^\top k_1(r)=B^\top k_1(r)$, where $B=[\mathbf{b}_1,\mathbf{b}_2,\cdots,\mathbf{b}_p]\in\mathbb{R}^{n_2\times p}$. For $t=1,\ldots,T$, denote

$$Y_t=[Y_t(r_1),Y_t(r_2),\cdots,Y_t(r_{n_2})]^\top\in\mathbb{R}^{n_2}\quad\text{and}\quad X_t=[X_t(s_1),X_t(s_2),\cdots,X_t(s_{n_1})]^\top\in\mathbb{R}^{n_1}.$$

Define $Y=[Y_1,\cdots,Y_T]\in\mathbb{R}^{n_2\times T}$, $X=[X_1,\cdots,X_T]\in\mathbb{R}^{n_1\times T}$ and $Z=[Z_1,\cdots,Z_T]\in\mathbb{R}^{p\times T}$.

With the notation defined above, in the following, we rewrite the penalized optimization problem as a function of R and $B=[\mathbf{b}_1,\mathbf{b}_2,\cdots,\mathbf{b}_p]$.

- The squared loss can be rewritten as

$$\sum_{t=1}^{T}\sum_{j=1}^{n_2}\left(Y_t(r_j)-\frac{1}{n_1}\sum_{i=1}^{n_1}A(r_j,s_i)X_t(s_i)-\langle\beta(r_j),Z_t\rangle_p\right)^2$$
$$=\sum_{t=1}^{T}\left\|Y_t-\frac{1}{n_1}K_1^\top RK_2X_t-K_1^\top BZ_t\right\|_2^2=\left\|Y-\frac{1}{n_1}K_1RK_2X-K_1BZ\right\|_{\mathrm{F}}^2,$$

  where $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ are the Euclidean norm of a vector and the Frobenius norm of a matrix.

- As for the Frobenius norm penalty $\|A\|_F^2$, first, it is easy to see $A^\top(r,s)=k_1(s)^\top Rk_2(r)$, where $A^\top$ is the adjoint operator of A. Let $u(s)=k_2(s)^\top c$, for any $c\in\mathbb{R}^{n_1}$. We have that

$$\begin{aligned}&A^\top A[u](s)=\langle A^\top(s,r),A[u](r)\rangle_{\mathcal{H}}=\langle k_1(r)^\top Rk_2(s),A[u](r)\rangle_{\mathcal{H}}\\=&\langle k_2(s)^\top R^\top k_1(r),\langle A(r,s),u(s)\rangle_{\mathcal{H}}\rangle_{\mathcal{H}}=k_2(s)^\top R^\top\langle k_1(r),k_1(r)^\top\rangle_{\mathcal{H}}R\langle k_2(s),u(s)\rangle_{\mathcal{H}}\\=&k_2(s)^\top R^\top K_1RK_2c.\end{aligned}$$

  Thus, the eigenvalues of $A^\top A$ are the same as those of $R^\top K_1RK_2$ and $\|A\|_{\mathrm{F}}^2=\mathrm{tr}(R^\top K_1RK_2)$.

- As for the $\mathcal{H}$-norm penalty $\|\beta_l\|_{\mathcal{H}}^2$ and the group Lasso-type penalty $\|\beta_l\|_{n_2}$, we have $\|\beta_l\|_{\mathcal{H}}^2 = \langle \beta_l(r), \beta_l(r) \rangle_{\mathcal{H}} = \mathbf{b}_l^\top K_1 \mathbf{b}_l$ and

$$\|\beta_l\|_{n_2} = \sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2} \beta_l^2(r_j)} = \sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{b}_l^\top k_1(r_j) k_1(r_j)^\top \mathbf{b}_l}$$
$$= \sqrt{\mathbf{b}_l^\top \frac{1}{n_2}\sum_{j=1}^{n_2} k_1(r_j) k_1(r_j)^\top \mathbf{b}_l} = \sqrt{\frac{1}{n_2}\mathbf{b}_l^\top K_1^2 \mathbf{b}_l}, \text{ for } l = 1, \ldots, p.$$

Combining all the components, the optimization can be written as

$$\left\| Y - \frac{1}{n_1} K_1 R K_2 X - K_1 B Z \right\|_{\mathrm{F}}^2 + \lambda_1 \mathrm{tr}(R^\top K_1 R K_2) + \lambda_2 \sum_{l=1}^{p} \mathbf{b}_l^\top K_1 \mathbf{b}_l + \lambda_3 \sum_{l=1}^{p} \sqrt{\frac{1}{n_2}\mathbf{b}_l^\top K_1^2 \mathbf{b}_l}.$$

Elementary algebra shows that the above is a convex function of R and $B = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_p]$. Note that the first three terms of the above are quadratic functions and can be handled easily, while the main difficulty of the optimisation lies in the group Lasso-type penalty $\lambda_3 \sum_{l=1}^{p} \sqrt{\frac{1}{n_2}\mathbf{b}_l^\top K_1^2 \mathbf{b}_l}$.

Remark 17. *Ridge regression. Iterative coordinate descent.*

6.3 Different measurements of space complexity

In order to provide theoretical guarantees of the penalised estimators, we need more regularity conditions, which are down to the complexity of the spaces. There are many different ways to characterise the complexity. Roughly speaking, we can name three categories:

- distribution-free notions, e.g. packing number , covering number, Vapnik–Chervonenkis dimension;
- distribution-dependent notions, e.g. packing number, covering number, Vapnik–Chervonenkis dimension in the $L_2(P)$ distance;
- data-dependent notions, e.g. Rademacher complexity, Gaussian complexity.

In this section, we introduce the Rademacher complexity, but we remark that in order to provide sharp analysis in functional linear regression, one in fact needs the notion of local Rademacher complexity. The Rademacher complexity provides global estimates of the complexity of the function class, that is, they do not reflect the fact that the algorithm will likely pick functions that have a small error, and in particular, only a small subset of the function class will be used. We refer to Bartlett et al. (2005), Mendelson (2002) and Mendelson and Vershynin (2002), for more detailed explanations.

Let $(\mathcal{X}, P)$ be a probability space. Denote by $\mathcal{F}$ a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$, and set $X_1, \ldots, X_n$ to be independent random variables distributed according to P. Let $\sigma_1, \ldots, \sigma_n$ be n independent Rademacher random variables, that is $\mathbb{P}\{\sigma_1 = 1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$.

For a function $f : \mathcal{X} \to \mathbb{R}$, define

$$P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad Pf = \mathbb{E}\{f(X)\} \quad \text{and} \quad R_n f = \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i).$$

For a class $\mathcal{F}$, set

$$R_n\mathcal{F} = \sup_{f\in\mathcal{F}} R_n f.$$

Define $\mathbb{E}_\sigma$ to be the expectation with respect to the random variables $\{\sigma_i\}_{i=1}^n$, conditional on all of the other random variables. The Rademacher average of $\mathcal{F}$ is $\mathbb{E}(R_n\mathcal{F})$ and the empirical (or conditional) Rademacher averages of $\mathcal{F}$ are

$$\mathbb{E}_\sigma R_n\mathcal{F} = \frac{1}{n}\mathbb{E}\left\{\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(X_i)|X_1,\ldots,X_n\right\}.$$

A standard fact is that the expected deviation of the empirical means from the actual ones can be controlled by the Rademacher averages of the class. That is, for any class of functions $\mathcal{F}$, it holds that

$$\max\left\{\mathbb{E}\sup_{f\in\mathcal{F}}(Pf - P_n f), \mathbb{E}\sup_{f\in\mathcal{F}}(P_n f - Pf)\right\} \leq 2\mathbb{E}R_n\mathcal{F}.$$

6.4 What we did not cover in this section

- Tuning parameter selection
- Kernel selection
- Theoretical results
- Testing
- Estimation
- Different observation grids
- Different RKHS's

7 What we did not cover in this module

- Robustness
- Privacy
- Missing-ness
- Testing & confidence region
- Nonparametric statistics
- Causal inference
- Topological data anlysis
- and many many other topics...

References

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

Shaobing Chen and David Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44. IEEE, 1994.

Felix Friedrich, Angela Kempe, Volkmar Liebscher, and Gerhard Winkler. Complexity penalized m-estimation: fast computation. *Journal of Computational and Graphical Statistics*, 17(1):201–224, 2008.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Kevin Lin, James L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused lasso, with application to approximate changepoint screening. *Advances in neural information processing systems*, 30, 2017.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2006.

Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002.

Shahar Mendelson and Roman Vershynin. Entropy, combinatorial dimensions and random averages. In *International Conference on Computational Learning Theory*, pages 14–28. Springer, 2002.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Gabriele Steidl, Stephan Didas, and Julia Neumann. Splines in higher order tv regularization. *International journal of computer vision*, 70(3):241–255, 2006.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Sara Van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Univariate mean change point detection: Penalization, cusum and optimality. *Electronic Journal of Statistics*, 14(1):1917–1961, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 217–242, 2014.