高维情形下线性模型的泛化误差研究

第三周 (12.18-12.24) 工作

Yukun Dong

目录

任务:

- 开始复现 Hastie 的文章 (Chapter3,4)；
- 阅读其他文献。

Chapter 1~2

模型设定

考虑线性模型。训练集 $(x_i, y_i), i \leq n$，$x_i \in \mathbb{R}^p$ 是输入的特征，$y_i \in \mathbb{R}$ 是相应，他们满足如下的分布：

$$(x_i, \epsilon_i) \sim P_x \times P_\epsilon; y_i = x_i^T \beta + \epsilon_i.$$

其中，P_x 是满足 $\mathbb{E}(x_i) = 0, Cov(x_i) = \Sigma$ 的分布，噪声满足零均值，方差 $Var(\epsilon_i) = \sigma^2$。

在 $p > n$(overparametrized) 情况下，我们使用 L_2 范数最小的估计，即加号逆。加号逆可以用岭回归的极限来逼近。

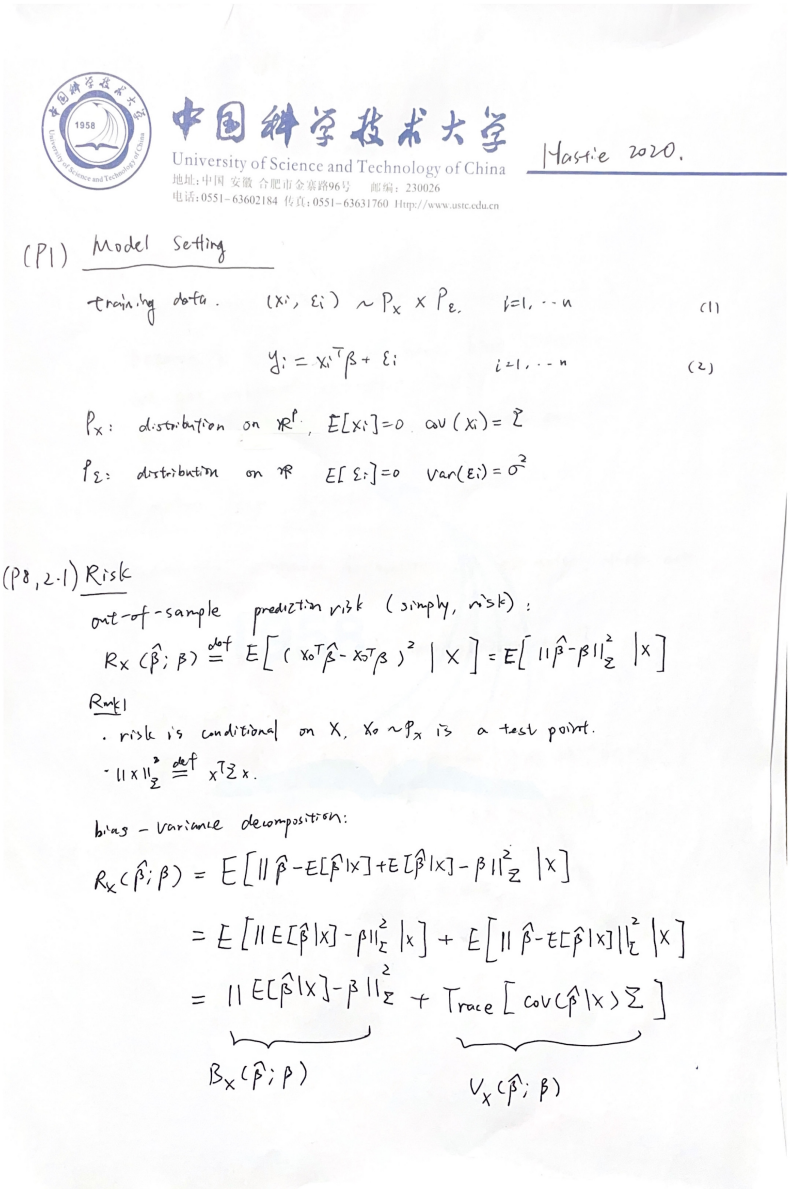

图 1: 模型设定与偏差-方差分解

Chapter 3:Isotropic features

这个情形下先讨论 $\Sigma = I_p$ 的情况。文章指出 $n, p \to \infty, p/n \to \gamma \in (0, \infty)$ 的时候，预报风险的极限仅与 $p/n, \sigma^2, r^2 := \|\beta\|_2^2$ 有关。比例 SNR $:= r^2/\sigma^2$ 被称为信噪比 (signal-to-noise ratio)。下面模拟不同 SNR 下，预报误差关于 γ 的变化。

3.2 Limiting risk

这一小节探讨预报误差的渐进性质。

取定 $n = 200, \sigma^2 = 1$，SNR $= 1, 2, 3, 4, 5$，$\gamma \in (0.1, 10)$。由于根据这一章假设，x 各个分量之间的生成是独立且满足一定正则条件的，我们生成每个样本 $x \sim N(0, I_p)$。

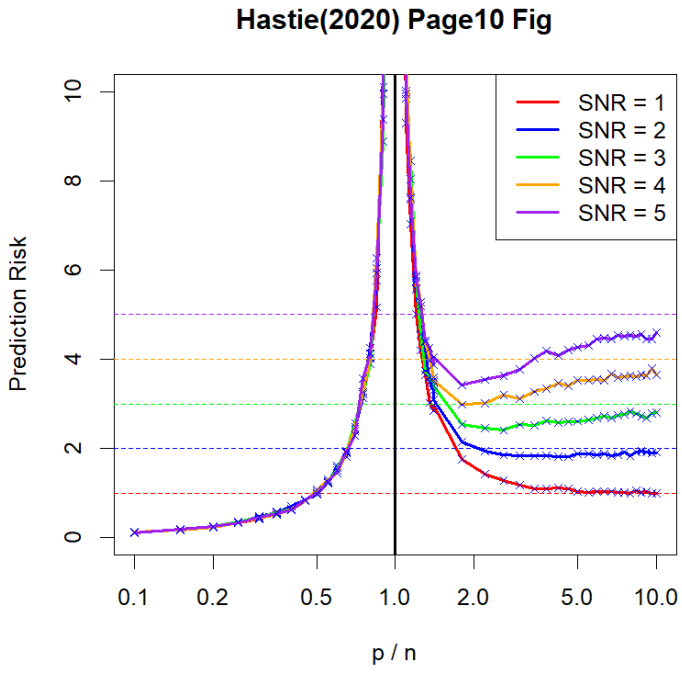

图 2: Page10 Fig

这个结果跟理论的结果非常相符，$\gamma < 1$ 的时候极限分布与 SNR 无关，这五条线也几乎重合；虚线代表零误差，即 $\hat{\beta} = 0$ 的时候对应的预报风险。$\gamma < 1$ 的时候 SNR 越大的时候右半部分 U 形曲线越明显，理论上该最小值在 $\gamma = \frac{\sqrt{\text{SNR}}}{\sqrt{\text{SNR}}-1}$ 处取到，最终收敛值如下：

Theorem 1. *Assume the model (1), (2), where $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $4+\eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 = r^2$ for all n, p. Then for the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, such that $p/n \to \gamma \in (0, \infty)$, it holds almost surely that*

$$B_X(\hat{\beta}; \beta) \to r^2\left(1 - \frac{1}{\gamma}\right), \tag{6}$$

$$V_X(\hat{\beta}; \beta) \to \sigma^2 \frac{1}{\gamma - 1}. \tag{7}$$

Hence, summarizing with Proposition 2, we have

$$R_X(\hat{\beta}; \beta) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\left(1 - \frac{1}{\gamma}\right) + \sigma^2 \frac{1}{\gamma - 1} & \text{for } \gamma > 1. \end{cases} \tag{8}$$

图 3: Theorem 1

3.3 Limiting L_2 norm

这一小节探讨 $\|\hat{\beta}\|_2^2$ 渐进性质。文章没有给出模拟，这里简要模拟一下。

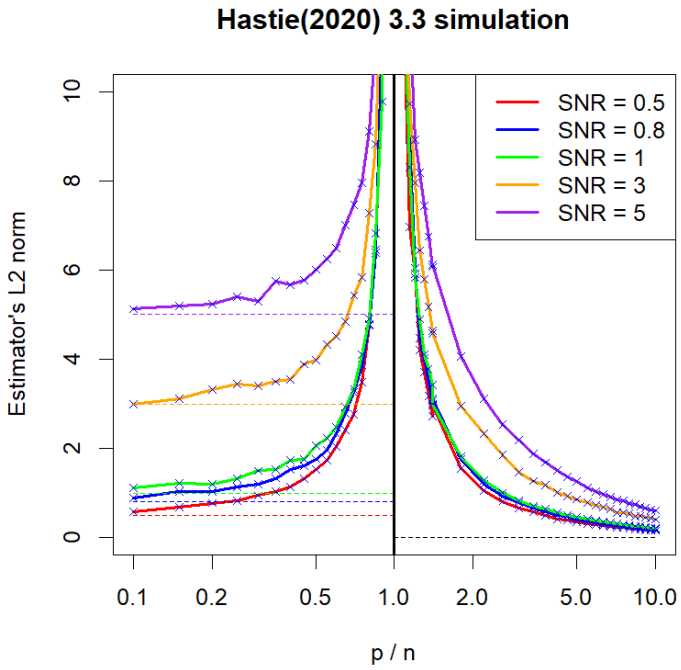

图 4: 3.3 simulation

Theorem 1. *Assume the model (1), (2), where $x \sim P_x$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $4+\eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 = r^2$ for all n, p. Then for the min-norm least squares estimator $\hat{\beta}$ in (4), as $n, p \to \infty$, such that $p/n \to \gamma \in (0, \infty)$, it holds almost surely that*

$$B_X(\hat{\beta};\beta) \to r^2\big(1-\frac{1}{\gamma}\big)\,, \tag{6}$$

$$V_X(\hat{\beta};\beta) \to \sigma^2\frac{1}{\gamma-1}\,. \tag{7}$$

Hence, summarizing with Proposition 2, we have

$$R_X(\hat{\beta};\beta) \to \begin{cases} \sigma^2\frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2\big(1-\frac{1}{\gamma}\big) + \sigma^2\frac{1}{\gamma-1} & \text{for } \gamma > 1. \end{cases} \tag{8}$$

图 5: Corollary 1

Chapter 4:Correlated features

这个情形下先讨论 $\Sigma \neq I_p$ 的情况。这时候产生训练样本 x 的办法是先生成 p 维向量 z，满足 z_i iid $\sim (0,1)$，再做变换 $x = \Sigma^{\frac{1}{2}} z$。

最小范数回归的风险取决于 Σ 和 β 的几何形态。用 $\Sigma = \sum_{i=1}^{p} s_i v_i v_i^T$ 表示 Σ 的特征值分解，其中 $s_1 \geq s_2 \geq ... \geq s_p \geq 0$。问题的几何形态由特征值序列 $(s_1, ..., s_p)$ 和在特征向量基 $((v_1^T\beta), ..., (v_p^T\beta))$ 中的 β 的系数决定。

文章通过在 $\mathbb{R}_{\geq 0}$ 上的两个概率分布来编码这些：

$$\hat{H}_n(s) := \frac{1}{p}\sum_{i=1}^{p} 1_{\{s \geq s_i\}}, \quad \hat{G}_n(s) = \frac{1}{\|\beta\|_2^2}\sum_{i=1}^{p} \langle \beta, v_i \rangle^2 1_{\{s \geq s_i\}}.$$

4.1 Prediction risk

基于一些假设，给出了 Bias 和 Variance 的置信域。

Assumption 1. The covariates vector $x \sim P_x$ is of the form $x = \Sigma^{1/2} z$, where, defining $\hat{H}_n$ as per Eq. (9), we have

(a) The vector $z = (z_1, \ldots, z_p)$ has independent (not necessarily identically distributed) entries with $\mathbb{E}\{z_i\} = 0$, $\mathbb{E}\{z_i^2\} = 1$, and $\mathbb{E}\{|z_i|^k\} \leq C_k < \infty$ for all $k \geq 2$.

(b) $s_1 = \|\Sigma\|_{op} \leq M$, $\int s^{-1} d\hat{H}_n(s) < M$.

(c) $|1 - (p/n)| \geq 1/M$, $1/M \leq p/n \leq M$.

Condition (a) bounds the tail probabilities on the covariates. Requiring finite moment of all order is useful to get strong bounds on the deviations of the risk from its predicted value. As discussed below, bounds on the first few moments are sufficient if we are satisfied in weaker probability bounds.

Conditions (b) requires the eigenvalues of Σ to be bounded, and not to accumulate[4] near 0. For the analysis of min-norm interpolation, we will add the additional assumption that the minimum eigenvalue of Σ is bounded away from zero. However condition (b) is sufficient for the analysis of ridge regression in Section 6.

Finally, as our statements are non-asymptotic, we do not assume p/n to converge to to a value. However condition (c) requires p/n to be bounded and bounded away from the interpolation threshold $p/n = 1$.

图 6: Assumption 1

Definition 1 (Predicted bias and variance: min-norm regression)**.** Let $\widehat{H}_n$ be the empirical distribution of eigenvalues of Σ, and $\widehat{G}_n$ the reweighted distribution as per Eq. (9). For $\gamma \in \mathbb{R}_{>0}$, define $c_0 = c_0(\gamma, \widehat{H}_n) \in \mathbb{R}_{>0}$ to be the unique non-negative solution of

$$1 - \frac{1}{\gamma} = \int \frac{1}{1 + c_0 \gamma s} \, d\widehat{H}_n(s) \,, \tag{10}$$

We then define the predicted bias and variance by

$$\mathscr{B}(\widehat{H}_n, \widehat{G}_n, \gamma) := \|\beta\|_2^2 \left\{ 1 + \gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} \, d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} \, d\widehat{H}_n(s)} \right\} \cdot \int \frac{s}{(1 + c_0 \gamma s)^2} \, d\widehat{G}_n(s) \,, \tag{11}$$

$$\mathscr{V}(\widehat{H}_n, \gamma) := \sigma^2 \gamma \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} \, d\widehat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} \, d\widehat{H}_n(s)} \,. \tag{12}$$

图 7: Definition 1

Theorem 2. *Assume the data model (1), (2) and that the covariates distribution satisfies Assumption 1. Further assume $s_p = \lambda_{\min}(\Sigma) > 1/M$. Define $\gamma = p/n$ and let $\hat{\beta}$ be the min-norm least squares estimator in Eq. (4).*

Then for any constants $D > 0$ (arbitrarily large) there exist $C = C(M, D)$ such that, with probability at least $1 - Cn^{-D}$ the following hold

$$R_X(\hat{\beta}; \beta) = B_X(\hat{\beta}; \beta) + V_X(\hat{\beta}; \beta) \,, \tag{13}$$

$$\left| B_X(\hat{\beta}; \beta) - \mathscr{B}(\widehat{H}_n, \widehat{G}_n, \gamma) \right| \le \frac{C \|\beta\|_2^2}{n^{1/7}} \,, \tag{14}$$

$$\left| V_X(\hat{\beta}; \beta) - \mathscr{V}(\widehat{H}_n, \gamma) \right| \le \frac{C}{n^{1/7}} \,, \tag{15}$$

where $\mathscr{B}$ and $\mathscr{V}$ are given in Definition 2.

图 8: Theorem 2

在渐进意义下，给出 $\widehat{H}_n, \widehat{G}_n$ 的弱收敛假设，得到 Bias 和 Variance 几乎处处收敛的结论：

Theorem 3. *Consider the setting of Theorem 2 but, instead of Assumption 1 (a), assume that $(z_i)_{i \le p}$ are identically distributed and satisfy the conditions $\mathbb{E} z_i = 0$, $\mathbb{E}(z_i^2) = 1$, $\mathbb{E}(|z_i|^{4+\delta}) \le C < \infty$. Further assume $p/n \to \gamma \in (0, \infty)$, $\widehat{H}_n \Rightarrow H$, $\widehat{G}_n \Rightarrow G$. Then, almost surely $B_X(\hat{\beta}; \beta)/\|\beta\|_2^2 \to \mathscr{B}(H, G, \gamma)$, $V_X(\hat{\beta}; \beta) \to \mathscr{V}(H, \gamma)$.*

图 9: Theorem 3

下面试着模拟这一点。目标是先固定 n，换不同的 p，看偏差和方差的蒙特卡洛估计是否收敛于理论值。步骤如下：

- 选取合适的 Σ 和 z 的分布，使得弱收敛条件成立；
- 对于每个 p 生成 $\beta_p \in \mathbb{R}^p; \Sigma_p \in \mathbb{R}^{p\times p}$，计算出偏差和方差的理论值 $\mathcal{B}(H,G,\lambda), \mathcal{V}(H,\lambda)$；
- 生成设计阵 $X \in \mathbb{R}^{n\times p}$，注意这之后 X 不能变化，因为我们所说的偏差和方差是以设计阵为条件的；
- 蒙特卡罗模拟 (即对于刚刚固定好的 X，生成 m 次 y，得到 m 个 $\hat{\beta}$)，得到偏差和方差的估计；
- 横轴为 γ，纵轴为偏差或者方差，对比估计和理论值。

先简单画一条没有理论值的：

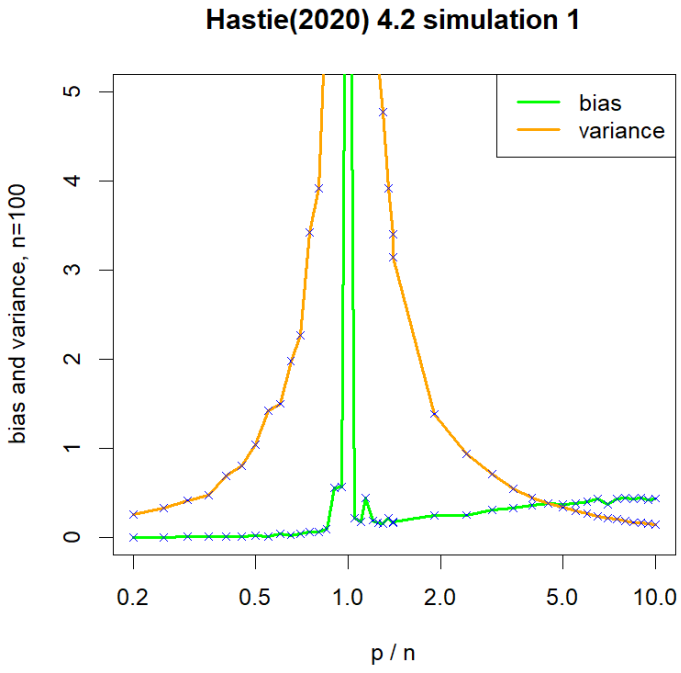

图 10: 蒙特卡洛模拟效果，没有附加理论值

目前偏差模拟与理论符合，方差部分仍有问题。下图中方框代表理论值。

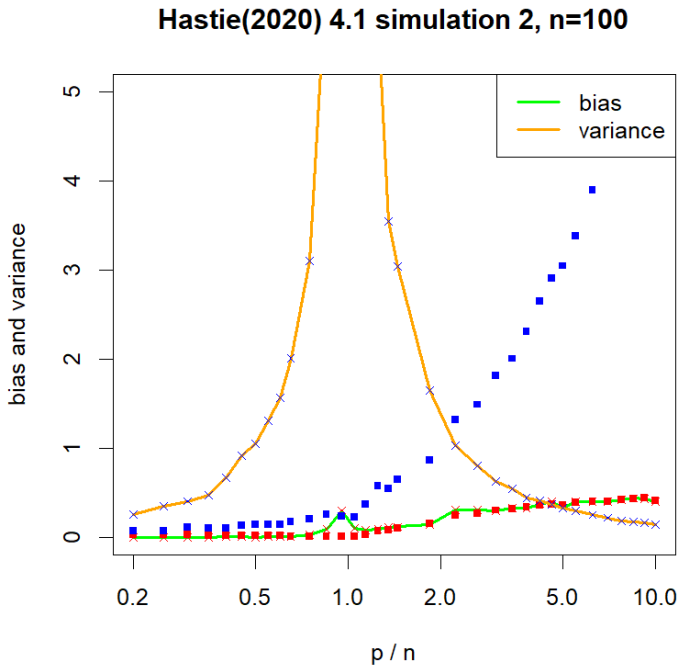

图 11: 方差模拟失败的结果，但偏差模拟成功了

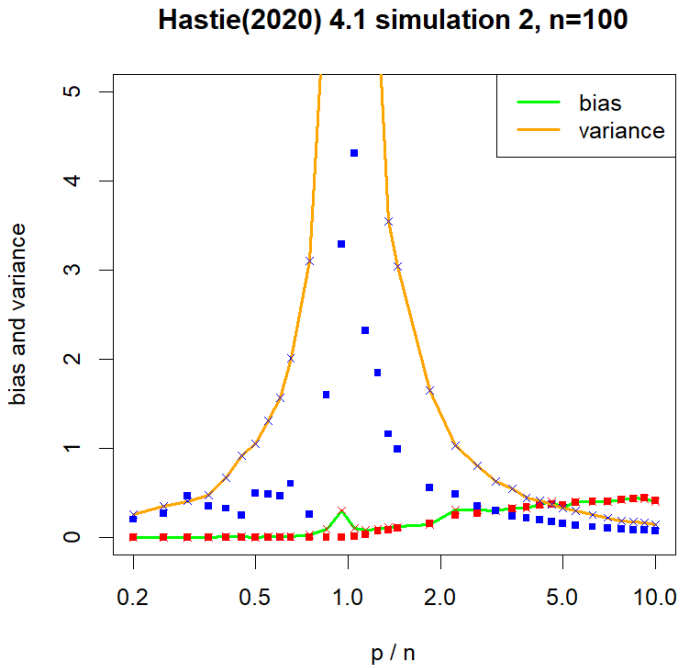

图 12: 稍作改进，还是有问题