高维情形下线性模型的泛化误差研究

第五周 (1.15-1.21) 工作

Yukun Dong

目录

任务:

- 复现 chapter5;
- 复现 chapter6;
- 阅读 chapter7.

5.Misspecified model

5.2 Isotropic

Consider, instead of (2), (3), a data model

$$\big((x_i, w_i), \epsilon_i\big) \sim P_{x,w} \times P_\epsilon, \quad i = 1, \dots, n, \tag{10}$$

$$y_i = x_i^T \beta + w_i^T \theta + \epsilon_i, \quad i = 1, \dots, n, \tag{11}$$

where as before the random draws across $i = 1, \dots, n$ are independent. Here, we partition the features according to $(x_i, w_i) \in \mathbb{R}^{p+d}$, $i = 1, \dots, n$, where the joint distribution $P_{x,w}$ is such that $\mathbb{E}((x_i, w_i)) = 0$ and

$$\mathrm{Cov}\big((x_i, w_i)\big) = \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xw} \\ \Sigma_{xw}^T & \Sigma_w \end{bmatrix}.$$

We collect the features in a block matrix $[\,X\ W\,] \in \mathbb{R}^{n \times (p+d)}$ (which has rows $(x_i, w_i) \in \mathbb{R}^{p+d}$, $i = 1, \dots, n$). We presume that X is observed but W is unobserved, and focus on the min-norm least squares estimator exactly as before in (4), from the regression of y on X (not the full feature matrix $[\,X\ W\,]$).

Given a test point $(x_0, w_0) \sim P_{x,w}$, and an estimator $\hat\beta$ (fit using X, y only, and not W), we define its out-of-sample prediction risk as

$$R_X(\hat\beta; \beta, \theta) = \mathbb{E}\big[\big(x_0^T\hat\beta - \mathbb{E}(y_0|x_0, w_0)\big)^2 \,|\, X\big] = \mathbb{E}\big[\big(x_0^T\hat\beta - x_0^T\beta - w_0^T\theta\big)^2 \,|\, X\big].$$

Note that this definition is conditional on X, and we are integrating over the randomness not only in ϵ (the training errors), but in the unobserved features W, as well. The next lemma decomposes this notion of risk in a useful way.

图 1: 模型设定

Theorem 4. *Assume the misspecified model (10), (11), and assume $(x, w) \sim P_{x,w}$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8 + \eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 + \|\theta\|_2^2 = r^2$ and $\|\beta\|_2^2 / r^2 = \kappa$ for all n, p. Then for the min-norm least squares estimator $\hat\beta$ in (4), as $n, p \to \infty$, with $p/n \to \gamma$, it holds almost surely that*

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

图 2: Theorem 4

定理 4 说明了未参与回归模型的参数 w 的具体维数 d 对极限的性质毫无影响，只有 $\|w\|^2$ 的大小起作用，这里都设 $p = d = 1000$ 进行模拟。这里置 $SNR = r^2/\sigma^2 = 1$.

Theorem 4. *Assume the misspecified model (10), (11), and assume $(x, w) \sim P_{x,w}$ has i.i.d. entries with zero mean, unit variance, and a finite moment of order $8 + \eta$, for some $\eta > 0$. Also assume that $\|\beta\|_2^2 + \|\theta\|_2^2 = r^2$ and $\|\beta\|_2^2 / r^2 = \kappa$ for all n, p. Then for the min-norm least squares estimator $\hat\beta$ in (4), as $n, p \to \infty$, with $p/n \to \gamma$, it holds almost surely that*

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

图 3: 5.1 模拟结果

图中可见极限性质模拟的效果很好，可以说是模拟正确了。

5.3 Polynomial approximation bias

Since adding features should generally improve our approximation capacity, it is reasonable to model $\kappa = \kappa(\gamma)$ as an increasing function of γ. To get an idea of the possible shapes taken by the asymptotic risk curve from Theorem 4, we can inspect different regimes for the approximation bias, i.e., the rate at which $1 - \kappa(\gamma) \to 0$ as $\gamma \to \infty$. For example, we may consider a *polynomial decay* for the approximation bias,

$$1 - \kappa(\gamma) = (1 + \gamma)^{-a}, \tag{13}$$

for some $a > 0$. In this case, the limiting risk in the isotropic setting, from Theorem 4, becomes

$$R_a(\gamma) = \begin{cases} r^2(1+\gamma)^{-a} + (r^2(1+\gamma)^{-a} + \sigma^2)\frac{\gamma}{1-\gamma} & \text{for } \gamma < 1, \\ r^2(1+\gamma)^{-a} + r^2\big(1 - (1+\gamma)^{-a}\big)\big(1 - \frac{1}{\gamma}\big) + (r^2(1+\gamma)^{-a} + \sigma^2)\frac{1}{\gamma-1} & \text{for } \gamma > 1. \end{cases} \tag{14}$$

图 4: polynomial approximation

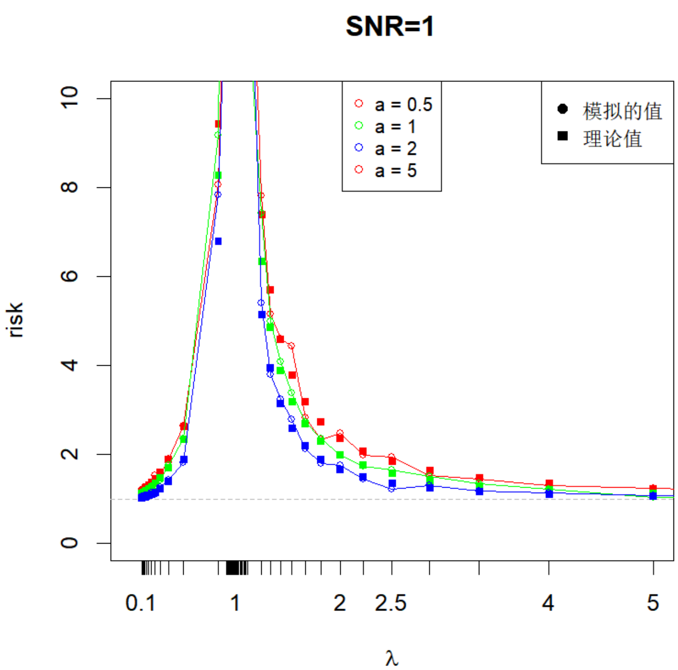

图 5: SNR=1, 多项式衰减模拟

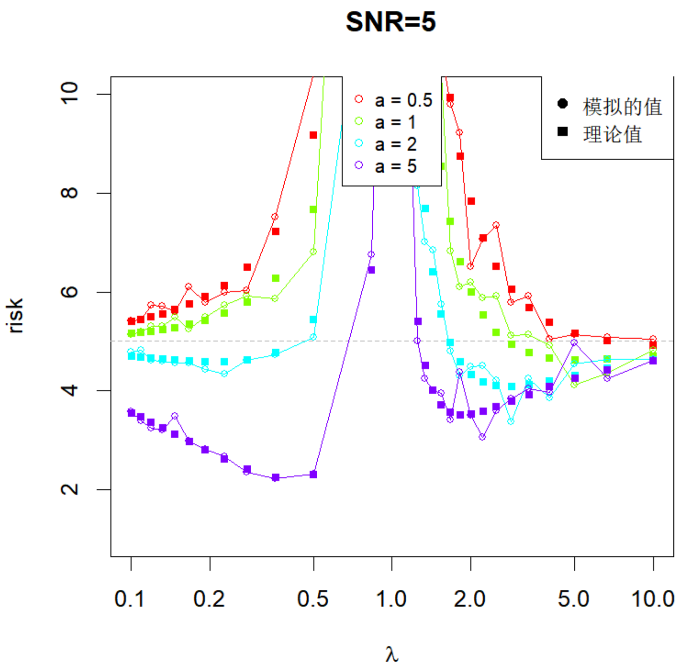

图 6: SNR=5, 多项式衰减模拟

6. Ridge regularization

Theorem 5. *Assume the conditions of Theorem 2 (well-specified model, isotropic features). Then for ridge regression in (5) with $\lambda > 0$, as $n, p \to \infty$, such that $p/n \to \gamma \in (0, \infty)$, it holds almost surely that*

$$R_X(\hat{\beta}_\lambda; \beta) \to \sigma^2 \gamma \int \frac{\alpha \lambda^2 + s}{(s+\lambda)^2} \, dF_\gamma,$$

where F_γ is the Marchenko-Pastur law, and $\alpha = r^2/(\sigma^2\gamma)$. The limiting risk can be alternatively written as

$$\sigma^2\gamma\big(m(-\lambda) - \lambda(1-\alpha\lambda)m'(-\lambda)\big).$$

where we abbreviate $m = m_{F_\gamma}$ for the Stieltjes transform of the Marchenko-Pastur law F_γ. Furthermore, the limiting ridge risk is minimized at $\lambda^* = 1/\alpha$, in which case the optimal limiting risk can be written explicitly as

$$\sigma^2\gamma \cdot m(-1/\alpha) = \sigma^2 \frac{-(1-(1+\sigma^2/r^2)\gamma) + \sqrt{(1-(1+\sigma^2/r^2)\gamma)^2 - 4\sigma^2\gamma^2/r^2}}{2\gamma},$$

where we have used the closed-form for the Stieltjes transform of the Marchenko-Pastur law, see (7).

图 7: Theorem 5

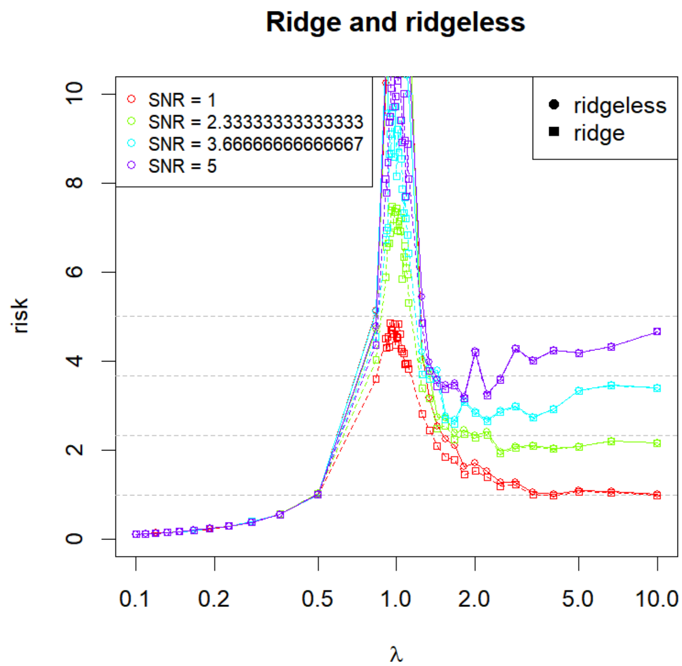

图 8: 模拟岭回归，全参数模型，会得到比 ridgeless 更优的 risk

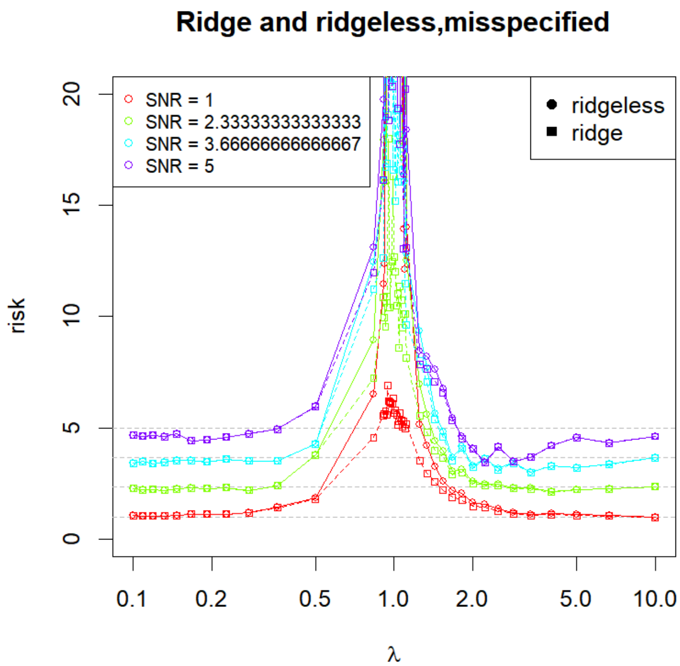

图 9: 模拟岭回归，部分参数模型，会得到比 ridgeless 更优的 risk