Benign Overfitting in Linear Regression

Peter L. Bartlett[a,b,1], Philip M. Long[c], Gábor Lugosi[d,e,f], and Alexander Tsigler[a]

statistical learning theory | overfitting | linear regression | interpolation

Contents

1.Research Background

2.Previous Studies

3.Definitions and Notation

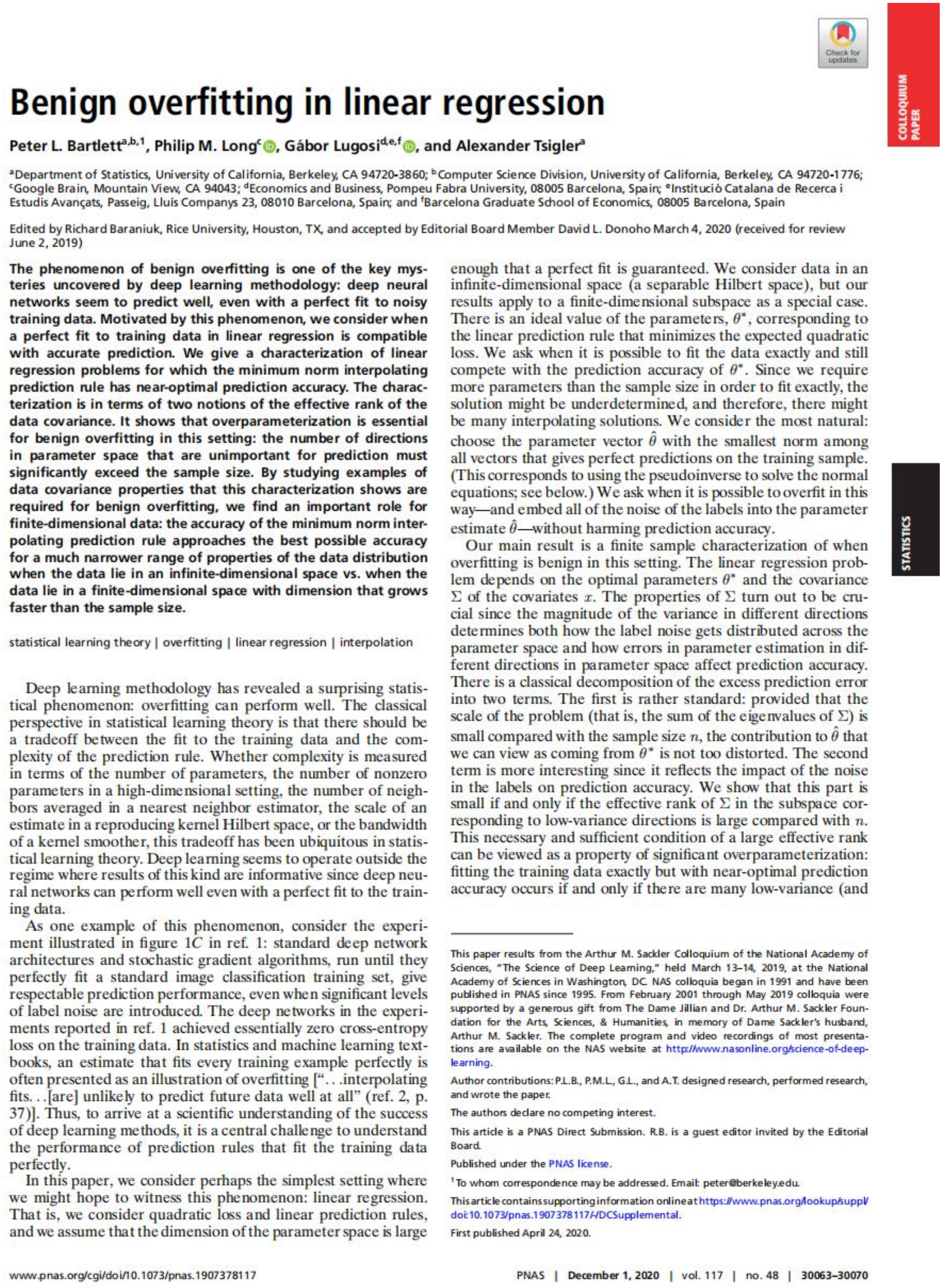
4.Theorems and Conclusions

Th1, Th2

5.Proof（Th1）

Le1~Le11

6.Research Prospects and Future



COLLOQUIUM PAPER

Benign overfitting in linear regression

Peter L. Bartlett[a,b,1], Philip M. Long[c], Gábor Lugosi[d,e,f], and Alexander Tsigler[a]

[a]Department of Statistics, University of California, Berkeley, CA 94720-3860; [b]Computer Science Division, University of California, Berkeley, CA 94720-1776; [c]Google Brain, Mountain View, CA 94043; [d]Economics and Business, Pompeu Fabra University, 08005 Barcelona, Spain; [e]Institució Catalana de Recerca i Estudis Avançats, Passeig, Lluís Companys 23, 08010 Barcelona, Spain; and [f]Barcelona Graduate School of Economics, 08005 Barcelona, Spain

Edited by Richard Baraniuk, Rice University, Houston, TX, and accepted by Editorial Board Member David L. Donoho March 4, 2020 (received for review June 2, 2019)

The phenomenon of benign overfitting is one of the key mysteries uncovered by deep learning methodology: deep neural networks seem to predict well, even with a perfect fit to noisy training data. Motivated by this phenomenon, we consider when a perfect fit to training data in linear regression is compatible with accurate prediction. We give a characterization of linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization is in terms of two notions of the effective rank of the data covariance. It shows that overparameterization is essential for benign overfitting in this setting: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size. By studying examples of data covariance properties that this characterization shows are required for benign overfitting, we find an important role for finite-dimensional data: the accuracy of the minimum norm interpolating prediction rule approaches the best possible accuracy for a much narrower range of properties of the data distribution when the data lie in an infinite-dimensional space vs. when the data lie in a finite-dimensional space with dimension that grows faster than the sample size.

statistical learning theory | overfitting | linear regression | interpolation

STATISTICS

Deep learning methodology has revealed a surprising statistical phenomenon: overfitting can perform well. The classical perspective in statistical learning theory is that there should be a tradeoff between the fit to the training data and the complexity of the prediction rule. Whether complexity is measured in terms of the number of parameters, the number of nonzero parameters in a high-dimensional setting, the number of neighbors averaged in a nearest neighbor estimator, the scale of an estimate in a reproducing kernel Hilbert space, or the bandwidth of a kernel smoother, this tradeoff has been ubiquitous in statistical learning theory. Deep learning seems to operate outside the regime where results of this kind are informative since deep neural networks can perform well even with a perfect fit to the training data.

As one example of this phenomenon, consider the experiment illustrated in figure 1C in ref. 1: standard deep network architectures and stochastic gradient algorithms, run until they perfectly fit a standard image classification training set, give respectable prediction performance, even when significant levels of label noise are introduced. The deep networks in the experiments reported in ref. 1 achieved essentially zero cross-entropy loss on the training data. In statistics and machine learning textbooks, an estimate that fits every training example perfectly is often presented as an illustration of overfitting ["...interpolating fits...[are] unlikely to predict future data well at all" (ref. 2, p. 37)]. Thus, to arrive at a scientific understanding of the success of deep learning methods, it is a central challenge to understand the performance of prediction rules that fit the training data perfectly.

In this paper, we consider perhaps the simplest setting where we might hope to witness this phenomenon: linear regression. That is, we consider quadratic loss and linear prediction rules, and we assume that the dimension of the parameter space is large enough that a perfect fit is guaranteed. We consider data in an infinite-dimensional space (a separable Hilbert space), but our results apply to a finite-dimensional subspace as a special case. There is an ideal value of the parameters, θ^*, corresponding to the linear prediction rule that minimizes the expected quadratic loss. We ask when it is possible to fit the data exactly and still compete with the prediction accuracy of θ^*. Since we require more parameters than the sample size in order to fit exactly, the solution might be underdetermined, and therefore, there might be many interpolating solutions. We consider the most natural: choose the parameter vector $\hat{\theta}$ with the smallest norm among all vectors that gives perfect predictions on the training sample. (This corresponds to using the pseudoinverse to solve the normal equations; see below.) We ask when it is possible to overfit in this way—and embed all of the noise of the labels into the parameter estimate $\hat{\theta}$—without harming prediction accuracy.

Our main result is a finite sample characterization of when overfitting is benign in this setting. The linear regression problem depends on the optimal parameters θ^* and the covariance Σ of the covariates x. The properties of Σ turn out to be crucial since the magnitude of the variance in different directions determines both how the label noise gets distributed across the parameter space and how errors in parameter estimation in different directions in parameter space affect prediction accuracy. There is a classical decomposition of the excess prediction error into two terms. The first is rather standard: provided that the scale of the problem (that is, the sum of the eigenvalues of Σ) is small compared with the sample size n, the contribution to $\hat{\theta}$ that we can view as coming from θ^* is not too distorted. The second term is more interesting since it reflects the impact of the noise in the labels on prediction accuracy. We show that this part is small if and only if the effective rank of Σ in the subspace corresponding to low-variance directions is large compared with n. This necessary and sufficient condition of a large effective rank can be viewed as a property of significant overparameterization: fitting the training data exactly but with near-optimal prediction accuracy occurs if and only if there are many low-variance (and

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: P.L.B., P.M.L., G.L., and A.T. designed research, performed research, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.B. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: peter@berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1907378117/-/DCSupplemental.

First published April 24, 2020.

1. Research Background

- A surprising phenomenon in seep learning methodology: **overfitting can perform well**.
- Whether complexity is measured in terms of :
- the number of **parameters**,
- the number of **nonzero parameters** in a h**igh-dimensional setting**,
- the number of **neighbors** averaged in a **nearest neighbor estimator**,
- the scale of an estimate in a **reproducing kernel Hilbert space**,
- the bandwidth of a **kernel smoother**,
- this tradeoff has been ubiquitous in **statistical learning theory**.

1. Research Background

- **Deep learning** seems to operate outside the regime where results of this kind are informative since deep neural networks can **perform well** even with **a perfect fit to the training data**.

- To arrive at a scientific understanding of the success of deep learning methods, it is a central challenge to **understand the performance of prediction rules** that fit the training data perfectly.

1. Research Background

- The simplest setting: **Linear Regression**.

- Consider **quadratic loss** and **linear prediction rules**
- Assume that the **dimension** of the parameter space is **large** enough
- Consider data in an **infinite-dimensional** space

- We ask when it is possible to **fit the data exactly** and still compete with the prediction accuracy of θ^*

1. Research Background

- In an **infinite-dimensional setting**, benign overfitting occurs only for **a narrow range of decay rates** of the eigenvalues.

- On the other hand, it occurs with **any suitably slowly decaying eigenvalue** sequence **in a finite-dimensional space** with **dimension** that grows **faster** than the sample size.

2.Previous Studies

1）. Initial Observations & Experimental Studies

Interpolating prediction rules (fitting noisy data exactly) emerged as a key mystery since 2017 Simons Institute program.

Belkin et al. (3) experimentally showed RKHS interpolants achieve high accuracy despite violating classical generalization bounds.

2）. Interpolating Decision Rules & Kernel Methods

Belkin et al. (4) proposed "simplicial interpolation" with asymptotic consistency in high dimensions.

Belkin et al. (5) studied singular-kernel smoothing methods achieving optimal nonparametric rates while interpolating (building on (6)).

Liang & Rakhlin (7) proved minimum-norm kernel regression (with nonlinear inner-product kernels) has good accuracy under specific sample conditions.

3）. Parameter Space Dimension & Excess Risk

Belkin et al. (8) experimentally analyzed excess risk as a function of parameter space dimension in linear/nonlinear models.

2.Previous Studies

4）. Subsequent Linear Model Analyses

Belkin et al. (11) derived excess risk for linear models (sparse parameters/Fourier features).

Hastie et al. (12) studied linear regression asymptotically ($n, p \to \infty, p/n \to \gamma$):

Assumed convergence of spectral distribution of Σ, used random matrix theory to characterize excess prediction error;

Examined effects of noise variance, eigenvalue distribution, and γ;

Extended to models with random nonlinear features.

5）. Contrast with Present Work

Key contributions of this work:

Provides finite-sample upper/lower bounds for excess prediction error under arbitrary covariances and dimensions;

Requires no asymptotic assumptions or specific data distributions

3. Definitions and Notations

Definition 1 (Linear Regression): A linear regression problem in a separable Hilbert space $\mathbb{H}$ is defined by a random covariate vector $x \in \mathbb{H}$ and outcome $y \in \mathbb{R}$. We define

1) the covariance operator $\Sigma = \mathbb{E}[xx^\top]$ and
2) the optimal parameter vector $\theta^* \in \mathbb{H}$, satisfying $\mathbb{E}(y - x^\top\theta^*)^2 = \min_\theta \mathbb{E}(y - x^\top\theta)^2$.

We assume that

1) x and y are mean zero;
2) $x = V\Lambda^{1/2}z$, where $\Sigma = V\Lambda V^\top$ is the spectral decomposition of Σ and z has components that are independent σ_x^2 sub-Gaussian with σ_x a positive constant: that is, for all $\lambda \in \mathbb{H}$,

$$\mathbb{E}[\exp(\lambda^\top z)] \leq \exp(\sigma_x^2 \|\lambda\|^2/2),$$

where $\|\cdot\|$ is the norm in the Hilbert space $\mathbb{H}$;

3) the conditional noise variance is bounded below by some constant σ^2,

$$\mathbb{E}\left[(y - x^\top\theta^*)^2 \middle| x\right] \geq \sigma^2;$$

4) $y - x^\top\theta^*$ is σ_y^2 sub-Gaussian conditionally on x: that is, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda(y - x^\top\theta^*))|x] \leq \exp(\sigma_y^2\lambda^2/2)$$

(note that this implies $\mathbb{E}[y|x] = x^\top\theta^*$); and

5) almost surely, the projection of the data X on the space orthogonal to any eigenvector of Σ spans a space of dimension n.

Given a training sample $(x_1, y_1), \ldots, (x_n, y_n)$ of n independent pairs with the same distribution as (x, y), an estimator returns a parameter estimate $\theta \in \mathbb{H}$. The excess risk of the estimator is defined as

$$R(\theta) := \mathbb{E}_{x,y}\left[\left(y - x^\top\theta\right)^2 - \left(y - x^\top\theta^*\right)^2\right],$$

3. Definitions and Notations

Definition 2 (Minimum Norm Estimator): Given data $X \in \mathbb{H}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$, the minimum norm estimator $\hat{\theta}$ solves the optimization problem

$$\begin{aligned} \min_{\theta \in \mathbb{H}} \quad & \|\theta\|^2 \\ \text{such that} \quad & \|X\theta - \boldsymbol{y}\|^2 = \min_{\beta} \|X\beta - \boldsymbol{y}\|^2 . \end{aligned}$$

3. Definitions and Notations

- Organize the variables introduced above:

H：随机变量x所在的无限维线性空间

x：H中的随机变量

Σ：协方差矩阵E$[xx']$

θ^*：最优估计值（$y = x'\theta^* + \varepsilon$）

ε：噪声（E$\varepsilon = 0$）

$\hat{\theta}$：最小范数估计值

3. Definitions and Notations

We use $\mu_1(\Sigma) \geq \mu_2(\Sigma) \geq \cdots$ to denote the eigenvalues of Σ in descending order, and we denote the operator norm of Σ by $\|\Sigma\|$. We use I to denote the identity operator on $\mathbb{H}$ and I_n to denote the $n \times n$ identity matrix.

***Definition 3 (Effective Ranks)*:** For the covariance operator Σ, define $\lambda_i = \mu_i(\Sigma)$ for $i = 1, 2, \ldots$. If $\sum_{i=1}^{\infty} \lambda_i < \infty$ and $\lambda_{k+1} > 0$ for $k \geq 0$, define

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i\right)^2}{\sum_{i>k} \lambda_i^2}.$$

4.Theorems and Conclusions

Theorem 1. *For any σ_x, there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem from Definition 1. Define*

$$k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\},$$

where the minimum of the empty set is defined as ∞. Suppose that $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* \geq n/c_1$, then $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$. Otherwise,

$$R(\hat{\theta}) \leq c\left(\|\theta^*\|^2 \|\Sigma\| \max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right)$$
$$+ c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$$

with probability at least $1-\delta$, and

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

Moreover, there are universal constants a_1, a_2, n_0 such that, for all $n \geq n_0$, for all Σ, and for all $t \geq 0$, there is a θ^* with $\|\theta^*\| = t$ such that, for $x \sim \mathcal{N}(0,\Sigma)$ and $y|x \sim \mathcal{N}(x^\top\theta^*, \|\theta^*\|^2\|\Sigma\|)$ with probability at least $1/4$,

$$R(\hat{\theta}) \geq \frac{1}{a_1}\|\theta^*\|^2\|\Sigma\|\mathbb{1}\left[\frac{r_0(\Sigma)}{n\log(1+r_0(\Sigma))} \geq a_2\right].$$

4.Theorems and Conclusions

- **From Theorem 1 we know:**
- $r_0(\Sigma)$ should be small compared with the sample size n (from the first term)
- $r_{k*}(\Sigma)$ and $R_{k*}(\Sigma)$ should be large compared with n.
- the **number** of nonzero eigenvalues should be **large** compared with n
- they should have a **small sum** compared with n
- there should be **many** eigenvalues **no larger** than λ_{k*}
- If the **number** of these small eigenvalues is **not much larger** than n, then they should be roughly **equal**
- they can be more **asymmetric(非对称)** if there are many more of them.

4.Theorems and Conclusions

Theorem 2. (Two Examples)

1) *If* $\mu_k(\Sigma) = k^{-\alpha} \ln^{-\beta}((k+1)e/2)$, *then* Σ *is benign if and only if* $\alpha = 1$ *and* $\beta > 1$.

Theorem 2.1 shows that, for infinite-dimensional data with a fixed covariance, benign overfitting occurs if and only if the eigenvalues of the covariance operator decay just slowly enough for their sum to remain finite.

4.Theorems and Conclusions

Since rescaling X affects the accuracy of the least norm interpolant in an obvious way, we may assume without loss of generality that $\|\Sigma\| = 1$. If we restrict our attention to this case, then informally, Theorem 1 implies that, when the covariance operator for data with n examples is Σ_n, the least norm interpolant converges if $\frac{r_0(\Sigma_n)}{n} \to 0$, $\frac{k_n^*}{n} \to 0$, and $\frac{n}{R_{k_n^*}(\Sigma_n)} \to 0$ and only if $\frac{r_0(\Sigma_n)}{n \log(1+r_0(\Sigma_n))} \to 0$, $\frac{k_n^*}{n} \to 0$, and $\frac{n}{R_{k_n^*}(\Sigma_n)} \to 0$, where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$ for the universal constant b in Theorem 1.

$$R(\hat{\theta}) \leq c\left(\|\theta^*\|^2 \|\Sigma\| \max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\right) + c\log(1/\delta)\sigma_y^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)$$

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

$$R(\hat{\theta}) \geq \frac{1}{a_1}\|\theta^*\|^2 \|\Sigma\| \mathbb{1}\left[\frac{r_0(\Sigma)}{n\log(1+r_0(\Sigma))} \geq a_2\right].$$

***Definition 4*:** A sequence of covariance operators Σ_n with $\|\Sigma_n\| = 1$ is benign if

$$\lim_{n\to\infty} \frac{r_0(\Sigma_n)}{n} = \lim_{n\to\infty} \frac{k_n^*}{n} = \lim_{n\to\infty} \frac{n}{R_{k_n^*}(\Sigma_n)} = 0.$$

4.Theorems and Conclusions

2) *If*

$$\mu_k(\Sigma_n) = \begin{cases} \gamma_k + \epsilon_n & \text{if } k \le p_n, \\ 0 & \text{otherwise} \end{cases}$$

and $\gamma_k = \Theta(\exp(-k/\tau))$, *then* Σ_n *with* $\|\Sigma_n\| = 1$ *is benign if and only if* $p_n = \omega(n)$ *and* $ne^{-o(n)} = \epsilon_n p_n = o(n)$. *Furthermore, for* $p_n = \Omega(n)$ *and* $\epsilon_n p_n = ne^{-o(n)}$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

Theorem 2.2 shows that the situation is very different if the data have finite dimension and a small amount of isotropic noise is added to the covariates. In that case, even if the eigenvalues of the original covariance operator (before the addition of isotropic noise) decay very rapidly, benign overfitting occurs if and only if both the dimension is large compared with the sample size and the isotropic component of the covariance is sufficiently small—but not exponentially small—compared with the sample size.

4.Theorems and Conclusions

- **Tension:**
  - the **slow decay** of eigenvalues that is needed for $\boldsymbol{k/n + n/R_k}$ to be small
  - the **summability** of eigenvalues that is needed for $\boldsymbol{r_0(\Sigma)/n}$ to be small

4.Theorems and Conclusions

- **Explain**
- 1）In the **infinite**-dimensional setting, **slow** decay of the eigenvalues suffices—decay just **fast enough** to ensure **summability**—as shown by Theorem 2.1.
- (Example)

Theorem 31. *Define* $\lambda_{k,n} := \mu_k(\Sigma_n)$ *for all* k, n.

1. *If* $\lambda_{k,n} = k^{-\alpha} \ln^{-\beta}(k+1)$, *then* Σ_n *is benign iff* $\alpha = 1$ *and* $\beta > 1$.
2. *If* $\lambda_{k,n} = k^{-(1+\alpha_n)}$, *then* Σ_n *is benign iff* $\omega(1/n) = \alpha_n = o(1)$. *Furthermore,*

$$R(\hat{\theta}) = \Theta\left(\min\left\{\frac{1}{\alpha_n n} + \alpha_n, 1\right\}\right).$$

4.Theorems and Conclusions

- **Explain**
- 2) Consider a **finite**-dimensional setting (which ensures that the eigenvalues are **summable**), and in this case, arbitrarily slow decay is possible.
- (Example)

3. *If*

$$\lambda_{k,n} = \begin{cases} k^{-\alpha} & \text{if } k \le p_n, \\ \boxed{0 \quad \text{otherwise},} \end{cases}$$

then Σ_n *is benign iff either* $0 < \alpha < 1$, $p_n = \omega(n)$ *and* $p_n = o\left(n^{1/(1-\alpha)}\right)$ *or* $\alpha = 1$, $p_n = e^{\omega(\sqrt{n})}$ *and* $p_n = e^{o(n)}$.

4. *If*

$$\lambda_{k,n} = \begin{cases} \gamma_k + \epsilon_n & \text{if } k \le p_n, \\ \boxed{0 \quad \text{otherwise},} \end{cases}$$

and $\gamma_k = \Theta(\exp(-k/\tau))$, *then* Σ_n *is benign iff* $p_n = \omega(n)$ *and* $ne^{-o(n)} = \epsilon_n p_n = o(n)$. *Furthermore, for* $p_n = \Omega(n)$ *and* $\epsilon_n p_n = ne^{-o(n)}$,

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{\ln(n/(\epsilon_n p_n))}{n} + \max\left\{\frac{1}{n}, \frac{n}{p_n}\right\}\right).$$

5.Proof

- **5.1.Headline:**
- a standard decomposition of the excessrisk into two pieces:
- 1) a term that corresponds to the distortion that is introduced by viewing $\boldsymbol{\theta}^*$through the lens of **the finite sample**
- 2) a term that corresponds to the distortion introduced by the **noise**

$$\varepsilon = y - X\hat{\theta}$$

5.Proof

- **5.2.可能用到的数学知识**
- 广义逆矩阵

定理 1 设 A 是数域 K 上 $s\times n$ 非零矩阵，则矩阵方程

$$AXA = A \tag{2}$$

一定有解。如果 $\mathrm{rank}(A)=r$，并且

$$A = P\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}Q, \tag{3}$$

其中 P,Q 分别是 K 上 s 级、n 级可逆矩阵，那么矩阵方程(2)的通解为

$$X = Q^{-1}\begin{pmatrix} I_r & B \\ C & D \end{pmatrix}P^{-1}. \tag{4}$$

其中 B,C,D 分别是数域 K 上任意的 $r\times(s-r)$，$(n-r)\times r$，$(n-r)\times(s-r)$矩阵。

5.Proof

定义 1 设 A 是数域 K 上 $s\times n$ 矩阵,矩阵方程 $AXA=A$ 的每一个解都称为 A 的一个**广义逆矩阵**,简称为 A 的**广义逆**,用 A^- 表示 A 的任意一个广义逆。

从定义 1 得出

$$AA^-A = A. \tag{10}$$

从定理 1 得出,当 $A\neq 0$ 时,设 $\mathrm{rank}(A)=r$,且

$$A = P\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}Q.$$

则

$$A^- = Q^{-1}\begin{pmatrix} I_r & B \\ C & D \end{pmatrix}P^{-1}. \tag{11}$$

从定义 1 得出,任意一个 $n\times s$ 矩阵都是 $0_{s\times n}$ 的广义逆。

定理 2 (非齐次线性方程组的相容性定理)非齐次线性方程组 $A\boldsymbol{X}=\boldsymbol{\beta}$ 有解的充分必要条件是

$$\boldsymbol{\beta} = AA^-\boldsymbol{\beta}. \tag{12}$$

证明 必要性。设 $A\boldsymbol{X}=\boldsymbol{\beta}$ 有解 $\boldsymbol{\alpha}$,则

$$\boldsymbol{\beta} = A\boldsymbol{\alpha} = AA^-A\boldsymbol{\alpha} = AA^-\boldsymbol{\beta}.$$

充分性。设 $\boldsymbol{\beta}=AA^-\boldsymbol{\beta}$,则 $A^-\boldsymbol{\beta}$ 是 $A\boldsymbol{X}=\boldsymbol{\beta}$ 的解。 ■

定理 3 (非齐次线性方程组的解的结构定理)非齐次线性方程组 $A\boldsymbol{X}=\boldsymbol{\beta}$ 有解时,它的通解为

$$\boldsymbol{X} = A^-\boldsymbol{\beta}. \tag{13}$$

定理 4 (齐次线性方程组的解的结构定理)数域 K 上 n 元齐次线性方程组 $A\boldsymbol{X}=\boldsymbol{0}$ 的通解为

$$\boldsymbol{X} = (I_n - A^-A)\boldsymbol{Z}, \tag{20}$$

其中 A^- 是 A 的任意给定的一个广义逆,$\boldsymbol{Z}$ 取遍 K^n 中任意列向量。

证明 任取 $\boldsymbol{Z}\in K^n$,有

$$A[(I_n - A^-A)\boldsymbol{Z}] = (A - AA^-A)\boldsymbol{Z} = (A - A)\boldsymbol{Z} = \boldsymbol{0},$$

因此 $\boldsymbol{X}=(I_n-A^-A)\boldsymbol{Z}$ 是齐次线性方程组 $A\boldsymbol{X}=\boldsymbol{0}$ 的解。

反之,设 $\boldsymbol{\eta}$ 是 $A\boldsymbol{X}=\boldsymbol{0}$ 的一个解,则

$$(I_n - A^-A)\boldsymbol{\eta} = \boldsymbol{\eta} - A^-A\boldsymbol{\eta} = \boldsymbol{\eta}.$$

综上所述,$\boldsymbol{X}=(I_n-A^-A)\boldsymbol{Z}$ 是齐次线性方程组 $A\boldsymbol{X}=0$ 的通解。 ■

推论 1 设数域 K 上 n 元非齐次线性方程组 $A\boldsymbol{X}=\boldsymbol{\beta}$ 有解,则它的通解为

$$\boldsymbol{X} = A^-\boldsymbol{\beta} + (I_n - A^-A)\boldsymbol{Z}, \tag{21}$$

其中 A^- 是 A 的任意给定的一个广义逆,$\boldsymbol{Z}$ 取遍 K^n 中任意列向量。

证明 由于 $A^-\boldsymbol{\beta}$ 是 $A\boldsymbol{X}=\boldsymbol{\beta}$ 的一个解,且 $(I_n-A^-A)\boldsymbol{Z}$ 是导出方程组 $A\boldsymbol{X}=0$ 的通解,因此 $\boldsymbol{X}=A^-\boldsymbol{\beta}+(I_n-A^-A)\boldsymbol{Z}$ 是 $A\boldsymbol{X}=\boldsymbol{\beta}$ 的通解。 ■

5.Proof

定义 2 设 A 是复数域上 $s\times n$ 矩阵，矩阵方程组

$$\begin{cases} AXA = A, \\ XAX = X, \\ (AX)^* = AX, \\ (XA)^* = XA, \end{cases} \tag{22}$$

称为 A 的 **Penrose 方程组**，它的解称为 A 的 **Moore-Penrose 广义逆**，记作 A^+。(22)式中 $(AX)^*$ 表示把 AX 的每个元素取共轭复数得到的矩阵再转置。

定理 5 如果 A 是复数域上 $s\times n$ 非零矩阵，A 的 Penrose 方程组总是有解，并且它的解唯一。设 $A=BC$，其中 B、C 分别是列满秩与行满秩矩阵，则 Penrose 方程组的唯一解是

$$X = C^*(CC^*)^{-1}(B^*B)^{-1}B^*. \tag{23}$$

证明 把(23)式代入 Penrose 方程组的每一个方程，验证每一个方程都变成恒等式：

$$AXA = (BC)C^*(CC^*)^{-1}(B^*B)^{-1}B^*(BC) = BC = A,$$

$$XAX = C^*(CC^*)^{-1}(B^*B)^{-1}B^*(BC)C^*(CC^*)^{-1}(B^*B)^{-1}B^*$$
$$= C^*(CC^*)^{-1}(B^*B)^{-1}B^* = X,$$

$$(AX)^* = X^*A^* = B(B^*B)^{-1}(CC^*)^{-1}CC^*B^*$$
$$= B(B^*B)^{-1}B^* = B(CC^*)(CC^*)^{-1}(B^*B)^{-1}B^* = AX,$$

$$(XA)^* = A^*X^* = C^*B^*B(B^*B)^{-1}(CC^*)^{-1}C$$
$$= C^*(CC^*)^{-1}C = C^*(CC^*)^{-1}(B^*B)^{-1}(B^*B)C = XA.$$

因此(12)式的确是 Penrose 方程组的解。

下面证解的唯一性。设 X_1 和 X_2 都是 Penrose 方程组的解。则

$$X_1 = X_1AX_1 = X_1(AX_2A)X_1 = X_1(AX_2)(AX_1)$$

$$= X_1(AX_2)^*(AX_1)^* = X_1(AX_1AX_2)^* = X_1X_2^*(AX_1A)^*$$
$$= X_1X_2^*A^* = X_1(AX_2)^* = X_1AX_2 = X_1(AX_2A)X_2$$
$$= (X_1A)(X_2A)X_2 = (X_1A)^*(X_2A)^*X_2 = (X_2AX_1A)^*X_2$$
$$= (X_2A)^*X_2 = X_2AX_2 = X_2.$$

5.Proof

●5.3. Lemma

Lemma 1. $r_k(\Sigma) \geq 1$, $r_k^2(\Sigma) = r_k(\Sigma^2) R_k(\Sigma)$, *and*

$$r_k(\Sigma^2) \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$

Notice that $r_0(I_p) = R_0(I_p) = p$. More generally, if all of the nonzero eigenvalues of Σ are identical, then $r_0(\Sigma) = R_0(\Sigma) = \operatorname{rank}(\Sigma)$. For Σ with finite rank, we can express both $r_0(\Sigma)$ and $R_0(\Sigma)$ as a product of the rank and a notion of symmetry. In particular, for $\operatorname{rank}(\Sigma) = p$, we can write

$$r_0(\Sigma) = \operatorname{rank}(\Sigma) s(\Sigma), \qquad R_0(\Sigma) = \operatorname{rank}(\Sigma) S(\Sigma),$$

$$\text{with } s(\Sigma) = \frac{\frac{1}{p}\sum_{i=1}^{p} \lambda_i}{\lambda_1}, \qquad S(\Sigma) = \frac{\left(\frac{1}{p}\sum_{i=1}^{p} \lambda_i\right)^2}{\frac{1}{p}\sum_{i=1}^{p} \lambda_i^2}.$$

Both notions of symmetry s and S lie between $1/p$ (when $\lambda_2 \to 0$) and 1 (when the λ_i are all equal).

5.Proof

Lemma 2. *The excess risk of the minimum norm estimator satisfies* $R(\hat{\theta}) \leq 2\theta^{*\top} B\theta^* + c\sigma^2 \log(1/\delta)\,\mathrm{tr}(C)$ *with probability at least* $1-\delta$ *over* ϵ*, and* $\mathbb{E}_\varepsilon R(\hat{\theta}) \geq \theta^{*\top} B\theta^* + \sigma^2\,\mathrm{tr}(C)$*, where*

$$B = \left(I - X^\top \left(XX^\top\right)^{-1} X\right)\Sigma\left(I - X^\top \left(XX^\top\right)^{-1} X\right),$$

$$C = \left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}.$$

Proof

Le2 $R(\hat{\theta}) = \mathbb{E}_{x,y}(y - x'\hat{\theta})^2 - \mathbb{E}(y - x'\theta^*)^2$

$= \mathbb{E}_{x,y}\left(y - x'\theta^* + x'(\theta^* - \hat{\theta})\right)^2 - \mathbb{E}(y - x'\theta^*)^2$

$= \mathbb{E}_x\left(x'(\theta^* - \hat{\theta})\right)^2$

$y = X\theta^* + \varepsilon = X\hat{\theta}$ 得到

$\hat{\theta} = X'(XX')^{-1}X\theta^* + X'(XX')^{-1}\varepsilon$

$\mathbb{E}_x\left(x'(\theta^* - \hat{\theta})\right)^2 = \mathbb{E}_x\left(x'\left((I - X'(XX')^{-1}X)\theta^* - X'(XX')^{-1}\varepsilon\right)\right)^2$

$\leq 2\mathbb{E}_x\left(x'(I - X'(XX')^{-1}X)\theta^*\right)^2 + 2\mathbb{E}_x\left(x'X'(XX')^{-1}\varepsilon\right)^2$

$= 2\mathbb{E}_x\left(\theta^{*\prime}(I - X'(XX')^{-1}X)\underbrace{xx'}_{\Sigma}(I - X'(XX')^{-1}X)\theta^*\right)$

$+ 2\mathbb{E}_x\left(\varepsilon'(XX')^{-1}X\underbrace{xx'}_{\Sigma}X'(XX')^{-1}\varepsilon\right)$

$= 2\theta^{*\prime}B\theta^* + 2\varepsilon'C\varepsilon$

ε是σ^2次高斯的，即$\forall\lambda\in R$

$\mathbb{E}[e^{\lambda\varepsilon}|X] \leq e^{\frac{\lambda^2\sigma^2}{2}}$

以至少$1-e^{-t}$的概率

$\varepsilon'C\varepsilon \leq \sigma^2\mathrm{tr}(C) + 2\sigma^2 t|C| + 2\sigma^2\sqrt{|C|t^2 + \mathrm{tr}(C^2)t}$ (1)

5.Proof

Lemma 2. *The excess risk of the minimum norm estimator satisfies* $R(\hat{\theta}) \le 2\theta^{*\top} B\theta^* + c\sigma^2 \log(1/\delta)\,\mathrm{tr}(C)$ *with probability at least* $1-\delta$ *over* ϵ*, and* $\mathbb{E}_{\varepsilon} R(\hat{\theta}) \ge \theta^{*\top} B\theta^* + \sigma^2\,\mathrm{tr}(C)$*, where*

$$B = \left(I - X^\top \left(XX^\top\right)^{-1} X\right)\Sigma\left(I - X^\top \left(XX^\top\right)^{-1} X\right),$$

$$C = \left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}.$$

$\forall \alpha \in k^n$,

$$\frac{|C\alpha|}{|\alpha|} = \sqrt{\frac{(C\alpha, C\alpha)}{(\alpha, \alpha)}} = \sqrt{\frac{\alpha' C' C \alpha}{\alpha' \alpha}}$$

C是半正定矩阵，因此，存在正交矩阵P s.t.

$$P C P^{-1} = \mathrm{diag}\{\mu_1, \mu_2 \dots \mu_n\} := D$$

$$\frac{|C\alpha|}{|\alpha|} = \sqrt{\frac{\alpha' P' D^2 P\alpha}{\alpha'\alpha}} = \sqrt{\frac{(\mu_1\theta_1)^2 + (\mu_2\theta_2)^2 + \dots + (\mu_n\theta_n)^2}{\alpha_1^2 + \alpha_2^2 + \dots + \alpha_n^2}}$$

$$\le \sqrt{\frac{\mathrm{tr}(C^2)(\theta_1^2 + \dots + \theta_n^2)}{\alpha_1^2 + \dots + \alpha_n^2}} = \sqrt{\mathrm{tr}(C^2)}\,\frac{|P\alpha|}{|\alpha|} = \sqrt{\mathrm{tr}(C^2)}$$

$\mathrm{tr}(C^2) = \mu_1^2 + \mu_2^2 + \dots + \mu_n^2 \le (\mu_1 + \mu_2 + \dots + \mu_n)^2$

$= (\mathrm{tr}C)^2$，因此

$\frac{|C\alpha|}{|\alpha|} \le \mathrm{tr}C$. 结合 $|C| = \sup\limits_{\alpha \in k^n} \frac{|C\alpha|}{|\alpha|}$，因此

$|C| \le \mathrm{tr}C$

代入(1)式有，以至少 $1-e^{-t}$ 的概率

$$\varepsilon' C \varepsilon \le \sigma^2 \mathrm{tr}(C)(2t+1) + 2\sigma^2 \mathrm{tr}(C)\sqrt{t^2+t}$$

$$\le (4t+2)\sigma^2 \mathrm{tr}(C)$$

取 $t = \log(\frac{1}{\delta})$，有，以至少 $1-\delta$ 的概率

$$R(\hat{\theta}) \le 2\varepsilon' B \varepsilon + 2\varepsilon' C \varepsilon$$

$$\le 2\varepsilon' B\varepsilon + 4(2\log(\tfrac{1}{\delta}) + 1)\sigma^2 \mathrm{tr}(C)$$

5.Proof

Lemma 2. *The excess risk of the minimum norm estimator satisfies* $R(\hat{\theta}) \leq 2\theta^{*\top} B\theta^{*} + c\sigma^{2} \log(1/\delta) \operatorname{tr}(C)$ *with probability at least* $1-\delta$ *over* ϵ, *and* $\mathbb{E}_{\varepsilon} R(\hat{\theta}) \geq \theta^{*\top} B\theta^{*} + \sigma^{2} \operatorname{tr}(C)$, *where*

$$B = \left(I - X^{\top}\left(XX^{\top}\right)^{-1}X\right)\Sigma\left(I - X^{\top}\left(XX^{\top}\right)^{-1}X\right),$$

$$C = \left(XX^{\top}\right)^{-1}X\Sigma X^{\top}\left(XX^{\top}\right)^{-1}.$$

$$\mathbb{E}R(\hat{\theta}) = \mathbb{E}_{x,\varepsilon}\left[(x'B\theta^{*})^{2} + (x'X'(XX')^{-1}\varepsilon)^{2}\right] \quad (\text{交叉项是关于}\varepsilon\text{的线性函数均值为}0)$$

$$= \mathbb{E}_{x,\varepsilon}\,\theta^{*\prime}B'xx'B\theta^{*} + \mathbb{E}_{x,\varepsilon}\,\varepsilon'(XX')^{-1}X\,xx'\,X'(XX')^{-1}\varepsilon$$

$$= \theta^{*\prime}\;B\;\theta^{*} + \mathbb{E}_{\varepsilon}\,\varepsilon'(XX')^{-1}X\Sigma X'(XX')^{-1}\varepsilon$$

$$= \theta^{*\prime}\;B\;\theta^{*} + \operatorname{tr}\left((XX')^{-1}X\Sigma X'(XX')^{-1}\mathbb{E}[\varepsilon\varepsilon'|X]\right)$$

$$= \theta^{*\prime}B\theta^{*} + \operatorname{tr}(C\,\mathbb{E}(\varepsilon\varepsilon'|X)) \leq \theta^{*\prime}B\theta^{*} + \sigma^{2}\operatorname{tr}(C)$$

5.Proof

Lemma 3. *Consider a covariance operator* Σ *with* $\lambda_i = \mu_i(\Sigma)$ *and* $\lambda_n > 0$. *Write its spectral decomposition* $\Sigma = \sum_j \lambda_j v_j v_j^\top$, *where the orthonormal* $v_j \in \mathbb{H}$ *are the eigenvectors corresponding to the* λ_j. *For* i *with* $\lambda_i > 0$, *define* $z_i = Xv_i/\sqrt{\lambda_i}$. *Then,*

$$\operatorname{tr}(C) = \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i \right],$$

and these $z_i \in \mathbb{R}^n$ *are independent* σ_x^2 *sub-Gaussian. Furthermore, for any* i *with* $\lambda_i > 0$, *we have*

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where $A_{-i} = \sum_{i \neq i} \lambda_j z_j z_j^\top$.

Le3 由Def 1中的Assumption 2可知，$\frac{x' v_i}{\sqrt{\lambda_i}}$ 为 σ_x^2 次高斯变量

注意到 $X v_i = \sqrt{\lambda_i} z_i$

$\{v_j\}$ 为 $\mathbb{H}$ 上的一组标准正交基，因此，X的一行可写作

$$x_k' = \sum_j (x_k' v_j) v_j'$$

$$(z_i)_k = \frac{x_k' v_i}{\sqrt{\lambda_i}}$$

所以 $x_k' = \sum_j \sqrt{\lambda_j} (z_j)_k v_j'$

$$X = \begin{pmatrix} \sum_j \sqrt{\lambda_j} (z_j)_1 v_j' \\ \cdots \\ \sum_j \sqrt{\lambda_j} (z_j)_n v_j' \end{pmatrix} = \sum_j \sqrt{\lambda_j} z_j v_j'$$

5.Proof

Lemma 3. *Consider a covariance operator Σ with $\lambda_i = \mu_i(\Sigma)$ and $\lambda_n > 0$. Write its spectral decomposition $\Sigma = \sum_j \lambda_j v_j v_j^\top$, where the orthonormal $v_j \in \mathbb{H}$ are the eigenvectors corresponding to the λ_j. For i with $\lambda_i > 0$, define $z_i = X v_i / \sqrt{\lambda_i}$. Then,*

$$\operatorname{tr}(C) = \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i \right],$$

and these $z_i \in \mathbb{R}^n$ are independent σ_x^2 sub-Gaussian. Furthermore, for any i with $\lambda_i > 0$, we have

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where $A_{-i} = \sum_{i \neq i} \lambda_j z_j z_j^\top$.

$$XX' = \sum_i \sum_j \sqrt{\lambda_i}\sqrt{\lambda_j}\, z_j v_j' v_i z_j'$$

$$= \sum_j \lambda_j z_j z_j'$$

$$X\Sigma X' = \Big(\sum_i \sqrt{\lambda_i}\, z_i v_i'\Big)\Big(\sum_j \lambda_j v_j v_j'\Big)\Big(\sum_k \sqrt{\lambda_k}\, v_k z_k'\Big)$$

$$= \sum_i \sum_j \sum_k \sqrt{\lambda_i}\sqrt{\lambda_k}\,\lambda_j\, z_i (v_i' v_j)(v_j' v_k) z_k'$$

$$= \sum_j \lambda_j^2 z_j z_j'$$

$$\operatorname{tr}(C) = \operatorname{tr}\big((XX')^{-1} X\Sigma X' (XX')^{-1} \big)$$

$$= \operatorname{tr}\big((XX')^{-2} X\Sigma X' \big)$$

$$= \operatorname{tr}\Big(\big(\sum_j \lambda_j z_j z_j'\big)^{-2} \big(\sum_j \lambda_j^2 z_j z_j'\big) \Big)$$

$$= \sum_j \lambda_j^2 \operatorname{tr}\Big(\big(\sum_i \lambda_i z_i z_i'\big)^{-2} z_j z_j' \Big)$$

$$= \sum_j \lambda_j^2 \operatorname{tr}\Big(z_j' \big(\sum_i \lambda_i z_i z_i'\big)^{-2} z_j \Big)$$

$$= \sum_j \lambda_j^2 z_j' \big(\sum_i \lambda_i z_i z_i'\big)^{-2} z_j$$

5.Proof

Lemma 3. *Consider a covariance operator Σ with $\lambda_i = \mu_i(\Sigma)$ and $\lambda_n > 0$. Write its spectral decomposition $\Sigma = \sum_j \lambda_j v_j v_j^\top$, where the orthonormal $v_j \in \mathbb{H}$ are the eigenvectors corresponding to the λ_j. For i with $\lambda_i > 0$, define $z_i = Xv_i/\sqrt{\lambda_i}$. Then,*

$$\operatorname{tr}(C) = \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i \right],$$

and these $z_i \in \mathbb{R}^n$ are independent σ_x^2 sub-Gaussian. Furthermore, for any i with $\lambda_i > 0$, we have

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where $A_{-i} = \sum_{i \neq i} \lambda_j z_j z_j^\top$.

根据 Sherman-Morrison公式: $(A+xx')^{-1} = A^{-1} - \frac{A^{-1}xx'A^{-1}}{1+x'A^{-1}x}$

$$\left(\sum_j \lambda_j z_j z_j'\right)^{-1} = \left(A_{-i} + \lambda_i z_i z_i'\right)^{-1}$$

$$= A_{-i}^{-1} - \frac{\lambda_i A_{-i}^{-1} z_i z_i' A_{-i}^{-1}}{1+\lambda_i z_i' A_{-i}^{-1} z_i}$$

通过计算可得 (过程省去)

$$\lambda_i^2 z_i' \left(\sum_j \lambda_j z_j z_j'\right) z_i = \frac{\lambda_i^2 z_i' A_{-i}^{-2} z_i}{(1+\lambda_i z_i' A_{-i}^{-1} z_i)^2}$$

整合过后:

$$\operatorname{tr}(C) = \sum_j \frac{\lambda_j^2 z_j' A_{-j}^{-2} z_j}{(1+\lambda_j z_j' A_{-j}^{-1} z_j)^2}$$

5.Proof

- **Define:**

$$A = \sum_i \lambda_i z_i z_i^\top, \quad A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top, \quad A_k = \sum_{i > k} \lambda_i z_i z_i^\top,$$

- The next step is to show that eigenvalues of A, A_{-i}, and A_k are **concentrated**.

5.Proof

Lemma 4. *There is a constant c such that, for any $k \geq 0$ with probability at least $1-2e^{-n/c}$,*

$$\frac{1}{c}\sum_{i>k}\lambda_i - c\lambda_{k+1}n \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\left(\sum_{i>k}\lambda_i + \lambda_{k+1}n\right).$$

Lemma 25 (ϵ-net argument). *Suppose $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix, and $\mathcal{N}_\epsilon$ is an ϵ-net on the unit sphere $\mathcal{S}^{n-1}$ in the Euclidean norm, where $\epsilon < \frac{1}{2}$. Then*

$$\|A\| \leq (1-\epsilon)^{-2}\max_{x\in\mathcal{N}_\epsilon}|x^\top Ax|.$$

Proof. Denote the eigenvalues of A as $\lambda_1, \ldots, \lambda_n$ and assume $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Denote the first eigenvector of A as $v \in \mathcal{S}^{n-1}$, and take $\Delta v \in \mathbb{R}^n$ such that $v + \Delta v \in \mathcal{N}_\epsilon$ and $\|\Delta v\| \leq \epsilon$. Denote the coordinates of Δv in the eigenbasis of A as $\Delta v_1, \ldots, \Delta v_n$. Now we can write

$$\begin{aligned}
\left|(v+\Delta v)^\top A(v+\Delta v)\right| &= \left|\lambda_1 + 2\lambda_1\Delta v_1 + \sum_{i=1}^n \lambda_i \Delta v_i^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 + \sum_{i=2}^n \frac{\lambda_i}{\lambda_1}\Delta v_i^2\right| \\
&\geq |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \sum_{i=2}^n \Delta v_i^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\Delta v_1 + \Delta v_1^2 - \|\Delta v\|^2 + \Delta v_1^2\right| \\
&= |\lambda_1| \cdot \left|1 + 2\left(\Delta v_1 + \Delta v_1^2\right) - \|\Delta v\|^2\right| \\
&\geq |\lambda_1| \cdot \left|1 + 2\left(-\|\Delta v\| + (-\|\Delta v\|)^2\right) - \|\Delta v\|^2\right| \\
&= |\lambda_1| \cdot \left|1 - 2\|\Delta v\| + \|\Delta v\|^2\right| \\
&\geq |\lambda_1| \cdot \left|1 - 2\epsilon + \epsilon^2\right| \\
&= \|A\|(1-\epsilon)^2,
\end{aligned}$$

where the first inequality holds because the λ_is are decreasing in magnitude, and the last two inequalities hold since the functions $x + x^2$ and $2x + x^2$ are both increasing on $(-\frac{1}{2}, \infty)$ and $\Delta v_1 \geq -\|\Delta v\| \geq -\epsilon \geq -\frac{1}{2}$. $\square$

5.Proof

Lemma 4. *There is a constant c such that, for any $k \geq 0$ with probability at least $1-2e^{-n/c}$,*

$$\frac{1}{c}\sum_{i>k}\lambda_i - c\lambda_{k+1}n \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\left(\sum_{i>k}\lambda_i + \lambda_{k+1}n\right).$$

Le4

对于一个$u\in\mathbb{R}^n$ ($|u|=1$)，$u'Au=\sum\lambda_i(u'z_i)^2$，存在$C_2>0$，以至少$1-2e^{-t}$的概率

$$\left|u'Au-\sum\lambda_i\right| \leq C_2\sigma_x^2\max\left(\lambda_1 t, \sqrt{t\sum\lambda_i^2}\right)$$

设N为S^{n-1}的$\frac{1}{4}$-网且$|N|\leq 9^n$，以至少$1-2e^{-t}$的概率

$$\left|u'Au-\sum\lambda_i\right| \leq C_2\sigma_x^2\max\left\{\lambda_1(t+n\ln 9), \sqrt{(t+n\ln 9)\sum_i\lambda_i^2}\right\}$$

设$\hat{v}=\lambda_1(t+n\ln 9)+\sqrt{(t+n\ln 9)\sum\lambda_i^2}$. 存在$C_3'>0$ s.t.

$$\left|A-\left(\sum_i\lambda_i\right)I_n\right| \leq C_3'\max_{\alpha\in N}\left|\alpha'\left(A-\left(\sum_i\lambda_i\right)I_n\right)\alpha\right| \leq C_3'\left|u'Au-\sum\lambda_i\right|$$

$$\leq C_3\sigma_x^2\hat{v}$$

以至少$1-2e^{-t}$概率成立.

$\exists C_4, C_5$ s.t. $t+n\ln 9\leq C_5 n$ as $t\leq\frac{n}{C_4}$，此时

$$\hat{v}\leq C_5\left(\lambda_1 n+\sqrt{n\sum_i\lambda_i^2}\right)$$

由AM-GM不等式

$$\sqrt{\lambda_1 n\sum\lambda_i} \leq \frac{1}{2\delta}\lambda_1 n+\frac{\delta}{2}\sum\lambda_i \quad (\forall\delta>0)$$

取$\delta=C_3\sigma_x^2$，适当整合系数

$$\hat{v}\leq C_6\sigma_x^2\lambda_1 n+\frac{1}{2C_3\sigma_x^2}\sum\lambda_i$$

$$\left|A-\left(\sum\lambda_i\right)I_n\right| = \max\left\{\left|\mu_1(A)-\sum\lambda_i\right|, \left|\mu_n(A)-\sum\lambda_i\right|\right\}$$

$$\leq\frac{1}{2}\sum\lambda_i+C_6C_3\sigma_x^4\lambda_1 n$$

故 $\mu_1(A)-\sum\lambda_i\leq\frac{1}{2}\sum\lambda_i+C_6C_3\sigma_x^4\lambda_1 n$

$$\sum\lambda_i-\mu_n(A)\leq\frac{1}{2}\sum\lambda_i+C_6C_3\sigma_x^4\lambda_1 n$$

适当调整$C_1, C_2, \ldots, C_6$，得到Lemma 4

5.Proof

Lemma 5 *There are constants $b, c \geq 1$ such that for any $k \geq 0$, with probability at least $1-2e^{-n/c}$,*

1. *for all $i \geq 1$,*

$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c\left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right),$$

2. *for all $1 \leq i \leq k$,*

$$\mu_n(A) \geq \mu_n(A_{-i}) \geq \mu_n\left(A_k\right) \geq \frac{1}{c} \sum_{j>k} \lambda_j - c\lambda_{k+1} n,$$

3. *if $r_k(\Sigma) \geq bn$, then*

$$\frac{1}{c}\lambda_{k+1} r_k(\Sigma) \leq \mu_n\left(A_k\right) \leq \mu_1(A_k) \leq c\lambda_{k+1} r_k(\Sigma).$$

5.Proof

Le5 $\mathrm{rank}(A-A_k)\le k$，因此 $\dim\ker(A-A_k)\geq n-k$.

∃线性空间 L ($\dim L=n-k$)，$\forall v\in L$

$$v'Av-v'A_kv=0$$

$$v'Av=v'A_kv\le \mu_1(A_k)|v|^2$$

取 v_0 为 $\mu_{k+1}(A)$ 的单位特征向量，有

$$\mu_{k+1}(A)=v_0'Av_0=v_0'A_kv_0\le\mu_1(A_k)$$

$A-A_{-T}=\lambda_i z_i z_i'$，从而 $A-A_{-T}$ 正定，因此

$$\mu_{k+1}(A)\geq\mu_{k+1}(A_{-T})$$

同理 $$\mu_n(A)\geq\mu_n(A_{-T})$$

当 $1\le i\le k$ 时

$$\mu_n(A_{-T})\geq\mu_n(A_k)$$

当 $r_k(\Sigma)\geq bn$ 时

$$\sum_{j>k}\lambda_j+\lambda_{k+1}n=\lambda_{k+1}r_k(\Sigma)+\lambda_{k+1}n\le\left(1+\frac{1}{b}\right)\lambda_{k+1}r_k(\Sigma)$$

$$\frac{1}{c_1}\sum_{j>k}\lambda_j-c_1\lambda_{k+1}n=\frac{1}{c_1}\lambda_{k+1}r_k(\Sigma)-c_1\lambda_{k+1}n\geq\left(\frac{1}{c_1}-\frac{c_1}{b}\right)\lambda_{k+1}r_k(\Sigma)$$

5.Proof

Lemma 6 *Suppose $\{\lambda_i\}_i^\infty$ is a non-increasing sequence of non-negative numbers such that $\sum_{i=1}^\infty \lambda_i < \infty$, and $\{\xi_i\}_{i=1}^\infty$ are independent centered σ-subexponential random variables. Then for some universal constant a for any $t > 0$ with probability at least $1 - 2e^{-t}$*

$$\left|\sum_i \lambda_i \xi_i\right| \le a\sigma \max\left(t\lambda_1, \sqrt{t\sum_i \lambda_i^2}\right).$$

Corollary 1. *Suppose that $z \in \mathbb{R}^n$ is a centered random vector with independent σ^2 sub-Gaussian coordinates with unit variances, $\mathscr{L}$ is a random subspace of $\mathbb{R}^n$ of codimension k, and $\mathscr{L}$ is independent of z. Then, for some universal constant a and any $t > 0$, with probability at least $1 - 3e^{-t}$,*

$$\|z\|^2 \le n + a\sigma^2(t + \sqrt{nt}),$$
$$\|\Pi_{\mathscr{L}} z\|^2 \ge n - a\sigma^2(k + t + \sqrt{nt}),$$

where $\Pi_{\mathscr{L}}$ is the orthogonal projection on $\mathscr{L}$.

5.Proof

Upper Bound on the Trace Term. Lemma 7 *There are constants* $b, c \geq 1$ *such that, if* $0 \leq k \leq n/c$, $r_k(\Sigma) \geq bn$, *and* $l \leq k$, *then with probability at least* $1 - 7e^{-n/c}$,

$$\operatorname{tr}(C) \leq c\left(\frac{l}{n} + n\frac{\sum_{i>l}\lambda_i^2}{(\sum_{i>k}\lambda_i)^2}\right).$$

5.Proof

Lower bound on tr(C)

Lemma 8. *There is a constant c such that, for any $i \geq 1$ with $\lambda_i > 0$ and any $0 \leq k \leq n/c$, with probability at least $1 - 5e^{-n/c}$,*

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \geq \frac{1}{cn}\left(1 + \frac{\sum_{j>k}\lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}.$$

Lemma 9. *Suppose that $n \leq \infty$, $\{\eta_i\}_{i=1}^n$ is a sequence of non-negative random variables, and that $\{t_i\}_{i=1}^n$ is a sequence of nonnegative real numbers (at least one of which is strictly positive) such that, for some $\delta \in (0,1)$ and any $i \leq n$, $\Pr(\eta_i > t_i) \geq 1 - \delta$. Then,*

$$\Pr\left(\sum_{i=1}^n \eta_i \geq \frac{1}{2}\sum_{i=1}^n t_i\right) \geq 1 - 2\delta.$$

Lemma 10. *There are constants c such that, for any $0 \leq k \leq n/c$ and any $b > 1$ with probability at least $1 - 10e^{-n/c}$,*

1) *if $r_k(\Sigma) < bn$, then $\operatorname{tr}(C) \geq \frac{k+1}{cb^2 n}$*; and
2) *if $r_k(\Sigma) \geq bn$, then*

$$\operatorname{tr}(C) \geq \frac{1}{cb^2}\min_{l \leq k}\left(\frac{l}{n} + \frac{b^2 n \sum_{i>l}\lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right).$$

In particular, if all choices of $k \leq n/c$ give $r_k(\Sigma) < bn$, then $r_{n/c}(\Sigma) < bn$ implies that, with probability at least $1 - 10e^{-n/c}$, $\operatorname{tr}(C) = \Omega_{\sigma_x}(1)$.

5.Proof

Lemma 11. *For any* $b \geq 1$ *and* $k^* := \min\{k : r_k(\Sigma) \geq bn\}$, *if* $k^* < \infty$, *we have*

$$\min_{l \leq k^*}\left(\frac{l}{bn} + \frac{bn\sum_{i>l}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*}(\Sigma))^2}\right)$$

$$= \frac{k^*}{bn} + \frac{bn\sum_{i>k^*}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*}(\Sigma))^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}.$$

5.Proof

• 5.4 Prove

Lemma 10. *There are constants c such that, for any $0 \le k \le n/c$ and any $b > 1$ with probability at least $1 - 10e^{-n/c}$,*

1) *if $r_k(\Sigma) < bn$, then* $\operatorname{tr}(C) \ge \frac{k+1}{cb^2 n}$; and
2) *if $r_k(\Sigma) \ge bn$, then*

$$\operatorname{tr}(C) \ge \frac{1}{cb^2} \min_{l \le k} \left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2} \right).$$

Upper Bound on the Trace Term. Lemma 7 *There are constants $b, c \ge 1$ such that, if $0 \le k \le n/c$, $r_k(\Sigma) \ge bn$, and $l \le k$, then with probability at least $1 - 7e^{-n/c}$,*

$$\operatorname{tr}(C) \le c \left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{(\sum_{i>k} \lambda_i)^2} \right).$$

$$\boxed{R(\hat{\theta}) \le c \left(\|\theta^*\|^2 \|\Sigma\| \max\left\{ \sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}} \right\} \right) + c \log(1/\delta) \sigma_y^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right)}$$

$$R(\hat{\theta}) \le 2\theta^{*\top} B \theta^* + c\sigma^2 \log(1/\delta) \operatorname{tr}(C)$$

$$\boxed{\mathbb{E} R(\hat{\theta}) \ge \frac{\sigma^2}{c} \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right).}$$

$$\mathbb{E}_{\varepsilon} R(\hat{\theta}) \ge \theta^{*\top} B \theta^* + \sigma^2 \operatorname{tr}(C)$$

Lemma 11. *For any $b \ge 1$ and $k^* := \min\{k : r_k(\Sigma) \ge bn\}$, if $k^* < \infty$, we have*

$$\min_{l \le k^*} \left(\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} \right) = \frac{k^*}{bn} + \frac{bn \sum_{i>k^*} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}.$$

5.Proof

Lemma 35. *There is a constant c, that depends only on σ_x, such that for any $1 < t < n$, with probability at least $1 - e^{-t}$,*

$$\theta^{*\top} B \theta^* \leq c \|\theta^*\|^2 \|\Sigma\| \max\left\{ \sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{t}{n}} \right\}.$$

6.Research Prospects and Future

6.1 Conclusion:

1)We give finite sample excessrisk bounds that reveal the covariance structure that ensuresthat the minimum norm interpolating prediction rule has nearoptimal prediction accuracy.

2)Overparameterization (that is, the existence of many **low-variance and hence**, unimportant directions in parameter space) is essential for benign overfitting and that data thatlie in a large but finite-dimensional space exhibit the benignoverfitting phenomenon with a much wider range of covariance properties than data that lie in an infinite-dimensionalspace.

6.Research Prospects and Future

6.2 Future

1）条件期望E(y|x)不是x的一个线性函数

2）放宽“协变量作为独立随机变量向量的线性函数分布”这一条件

3）扩展损失函数

4）其他非线性参数化的函数类