SSE-DP-2026-1

過適合と二重降下現象
– Part I –

国友直人・望月泰博

統計数理研究所・早稲田大学

2026 年 1 月

SSE-DP(ディスカッションペーパー・シリーズ)は以下のサイトから無料で入手可能です。
https://stat-expert.ism.ac.jp/training/discussionpaper/

SSE-DP-2026-1

Over-parametrization and Double Decent Phenomena

-Part I -

by

Naoto Kunitomo and Yasuhiro Mochizuki

The Institute of Statistical Mathematics・Waseda University

January 2026

(Summary)

In overparametrized linear regression models, some recent studies report the double decent phenomena, that is, the predictive mean squared errors become small when the number of coefficient parameters increases. We examine the phenomena claimed in recent studies by Hastie et al. (2022, Annals of Statistics), and Kelly et al. (2024, Journal of Finance). We conducted two simulations, which correspond to their studies, and report the preliminary results.

過適合と二重降下現象
- Part I - [1]

国友直人[2]

望月泰博[3]

2025 年 12 月 31 日

鍵言葉 (Key Words)

線形回帰モデル, 最小二乗推定, 過適合 (over-fitting) と二重降下 (double-decent) 現象, 深層学習 (deep learning), シミュレーション実験

概要 (Summary)

線形回帰モデルと最小二乗法は応用上で最も利用されている統計的方法の一つである。「過適合と二重降下現象-Part I-」では線形回帰モデルを利用して説明変数の階数に関する通常の仮定が成り立たず, データ数よりも説明変数の数が多い高次元の場合に起きうる過適合 (overfitting), 二重降下 (double decent) 現象というデータ分析における最近の話題について考察する。Hastie et al. (2022, Annals of Statistics) による結果の一部を紹介, Kelly et al. (2024, Journal of Finance) に類似したシミュレーションを含む二つの実験により二重降下現象を分析した結果を報告する。

1　はじめに

統計検定 2 級の教科書「統計学基礎」5 章では多くの教科書と同様に線形回帰モデル (linear regression models) と最小二乗法 (least squares method) の標準的な説明にかなりのスペースを割いている。今でも最小二乗法は

[1] 2025-12-31. 統計エキスパート養成事業の一環として行われた共同研究をまとめた未完成な原稿であり、コメントを歓迎する。

[2] 統計数理研究所, 〒 190-8562 東京都立川市緑町 10-3

[3] 早稲田大学

統計的データ分析では様々な統計ソフトウエアでは標準的に装備されている最もよく利用されている統計的方法の一つである。「統計学基礎」ではさらに最小二乗法の計算方法、幾つかの統計的性質、様々な利用法について解説している。

この研究ノートでは「統計エキスパート」にとり基本的な線形回帰モデルを利用して, 最近のデータ分析を巡る話題の中からデータ数よりも説明変数が多い場合に生じうる過適合 (overfitting), 二重降下 (double decent) 現象をとりあげる。これまでの良く用いられている線形回帰モデルによる統計分析では母数の数がデータ数よりも小さいことを仮定するのが一般的である。この状況における回帰分析については標準的な議論が統計学の教科書では展開されている[4]。

これに対して、近年ではデータ分析において母数の数がデータ数よりも大きくなり得る場合として深層学習 (deep learning) モデルと呼ばれる学習理論が登場し、その応用が注目されている。通常の統計的分析では母数の数がデータ数より大きい場合は分析の初めから想定されていないのでこうした過適合 (overparametrization と呼ばれる) 何が起きるか、自明な問題とは言えない。したがって, 過適合な統計モデルで何が起きるのか、幾つかの基本的事項を解明、理解する必要がある。本稿では「過適合と二重降下-Part I-」として典型的な例である線形回帰モデルを用いてこの統計的問題を検討する。この問題は実は線形回帰モデルという簡単な設定であっても、従来の統計分析では十分には考察していなかった問題であることをここで強調しておく。

さらに近年では例えばファイナンス (金融) という応用分野において Kelley et al. (2024) はデータ数より母数が多い線形回帰モデルにより金融リターンの予測の精度が良くなることを主張している。こうした主張がどういう意味で成り立つのか否かは統計分析の応用上では重要な意味があるだろう。

本稿ではシミュレーションや線形回帰モデルの数理的性質の検討事項を再吟味することにより、データ数より母数の数が多い場合に予測が良くなる現象とはどの様な状況であるのか、また説明変数の選択というよく行われている統計的分析はどの様な意味があるのか、と云う問題を吟味する。特に Hastie et al. (2022a) により得られた結果の一部を紹介するとともに二つのシミュレーション実験により二重降下現象が起きる状況の分析結果を報告する。「過適合と二重降下現象-Part II-」ではニューラ

[4]例えば日本語の文献としては佐和 (1979), 竹村 (1990) などが標準的である。

ルネットと深層学習モデルを扱う予定である。

以下では第 2 節で線形回帰問題における過適合と二重降下現象を考察する。次に第 3 節では Hastie et al. (2022a), Kelly et al. (2024) の説明に類似した二つのシミュレーション結果を報告する。第 4 節は幾つかの論点についての考察を述べ、第 5 節では本稿のまとめを述べる。補論 A:数理的補論として本稿で利用した 定理 1^* の導出、確率行列論 (RMT) への補足を述べ、付論 B にシミュレーション結果をまとめておく。

2 Double Decent Problem (二重降下問題)

観測可能な変数 Y を表現する確率変数 y_i $(i = 1, \cdots, n)$ のベクトル $\mathbf{y} = (y_i)$、p 個の説明変数からなる観測ベクトル $\mathbf{x}_i = (x_{ji})$ より $n \times p$ 行列 $\mathbf{X} = (x_{ij}), i = 1, \cdots, n; j = 1, \cdots, p$ とする。線形回帰モデルは

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{1}$$

で与えられる。ただし p 個の母数 β_j から母数ベクトル $\boldsymbol{\beta} = (\beta_j)$、誤差 $u_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ $(i = 1, \cdots, n)$ より誤差ベクトル $\mathbf{u} = (u_i)$, $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ とする。
線形回帰モデルにおける標準的仮定は (I) $\mathbf{E}(\mathbf{u}) = \mathbf{0}$, (II) $\mathbf{V}(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}_n$ ($\sigma^2 > 0$ は定数), (III) $\mathrm{rank}(\mathbf{X}) = p$ であり、最小二乗推定量は

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2}$$

で与えられる。
ここで条件 (III) については若干の注意が必要である。通常は $p < n$ が仮定されるが p が n に近くなると予測誤差は大きくなるが、$p = n$ の場合は与えられたデータに完全フィットすることが可能である。さらに $p \geq n$ の場合には「条件 (III)*:$\mathrm{rank}(\mathbf{X}) = n\ (\leq p)$」を仮定する。このとき $\mathrm{rank}(\mathbf{X}'\mathbf{X}) = n \leq p$ となる。ここで Moore-Penrose の一般化逆行列を用いて $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{y}$ であるが, $p \geq n$ の場合には $\hat{\boldsymbol{\beta}}_{OLS} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{y}$ と表現することも可能である。

誤差項の分散共分散行列が一般の場合 $\mathbf{V}(\mathbf{u}\mathbf{u}') = \boldsymbol{\Sigma}$ (> 0) には一般化最小二乗法 (GLS) は $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{+}\mathbf{X})^{+}\mathbf{X}'\boldsymbol{\Sigma}^{+}\mathbf{y}$、またリッジ回帰を λ を実数として $\hat{\boldsymbol{\beta}}_R = (\mathbf{X}'\mathbf{X} + n\lambda\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$ と表現すると、$\lim_{\lambda \to 0+} \hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_{OLS}$

である。

さて近年ではデータ分析の応用においてDeep Learning(深層学習) モデルの有効性が特に注目されている。従来から知られていたNeural Networksでの多層構造の層の数を大きくすると性能が飛躍的に上がるということが特に画像認識や大規模言語モデルなどの工学系の応用分野を注目にされているのである（最近の動向については例えばBishop and Bishop (2024)を参照されたい。Deep Learningモデルでは一般にデータから推定されるパラメター数が膨大になる。「パラメター数を膨大にとると予測精度が上がる」という主張はAICなど伝統的な統計学の伝承である「ケチの原理」(principle of parsimony) に矛盾しているように考えられる。統計学の伝承ではパラメター数を多くすると観測データ（training dataと呼ばれる）へのモデルのフィットは良くなるが、新たなデータ（test dataと呼ばれる）予測力は急速に減衰する、という考え方（ケチの原理) が統計学の基本として広く理解されている。この考え方の下でAICを始め、様々な統計的方法が開発されてきている。本稿は線形回帰モデルを用いて過適合を巡る統計的問題を解明する為の第一歩と位置づけられる。

ここで扱う線形回帰モデル $y_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + u_i$ $(i = 1, \cdots, n)$ ($\boldsymbol{\beta}$ は $p \times 1$ ベクトル) はHastie et al. (2022a) が検討した状況と同一であり、また統計的学習論の教科書Hastie et al. (2021)10.8節で二重降下現象の説明に用いられたシミュレーションと類似の実験を行い、その主張の再現を試みた。またKelly et al. (2024) で分析している統計モデルをシミュレーションにより再現を試みている。

ここでまず $(\mathbf{x}_i, u_i)(i = 1, \cdots, n)$ をi.i.d.系列（期待値と分散の存在は仮定)、あるいは $\mathbf{x}_i$ を所与とする u_i の条件付期待値・条件付分散の有限性を仮定する。さらに設定を簡単化して、$\mathbf{E}(\mathbf{x}_i) = \mathbf{0}$, $\mathbf{E}(\mathbf{x}_i\mathbf{x}_i^{'}) = \mathrm{I}_p$, $\mathbf{E}(u_i) = 0$, $\mathbf{E}(u_i^2) = \sigma^2$ (> 0)、p を n に依存させて $p(n)$、$p(n) \to \infty$ $(n \to \infty)$ とする。最小二乗推定量 (Hastie et al. (2022a) ではRidgeless least squares）は

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}\{\|\mathbf{b}\|_2 : \mathbf{b} \text{ minimizes} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2\} \tag{3}$$

で与えられるが、一般逆行列 (generalized inverse) を利用して $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{y}$ と表現する。このとき $p \geq n$ のときデータにたいして $y_i = \mathbf{x}_i^{'}\mathbf{b}$ $(i = 1, \cdots, n)$、すなわち n 組の観測データについては統計モデルは完全フィットするが、一般化逆行列は一意ではない。そこでMoore-Penroseの一般

化逆行列をとれば一意になる[5]。なお Hastie et al. (2022a) ではリッジ回帰、misspecified models、非線形回帰などの性質も詳しく議論しているが本稿では省略する。

ここでリスク関数はデータとは独立に得られる説明変数ベクトル $\mathbf{x}_0$ に対して予測二乗誤差

$$R_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) = \mathbf{E}[(\mathbf{x}_0^{'}\hat{\boldsymbol{\beta}} - \mathbf{x}_0^{'}\boldsymbol{\beta})^2|X]] \tag{4}$$

とする。この量は条件 $\mathbf{E}(\mathbf{x}_0\mathbf{x}_0^{'}) = \mathbf{I}_p$ を利用すると $R_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) = B_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) + V_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta})$,

$$B_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) = \mathbf{E}[|\mathbf{E}(\hat{\boldsymbol{\beta}}|X) - \boldsymbol{\beta}|^2\,, V_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta})] = \mathrm{Tr}[\mathbf{Cov}(\hat{\boldsymbol{\beta}}|X)] \tag{5}$$

となる。このとき

$$B_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) = \boldsymbol{\beta}^{'}\boldsymbol{\Pi}\boldsymbol{\beta}\,, V_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) = \frac{\sigma^2}{n}\mathrm{Tr}[\hat{\boldsymbol{\Sigma}}^{+}] \tag{6}$$

である。ただし $\boldsymbol{\Pi} = \mathbf{I}_p - \hat{\boldsymbol{\Sigma}}^{+}\hat{\boldsymbol{\Sigma}}$. $\hat{\boldsymbol{\Sigma}} = \mathbf{X}'\mathbf{X}/n$ とする。
ここで線型回帰モデル $y_i = \mathbf{x}^{'}\boldsymbol{\beta} + u_i$ $(i = 1\cdots, n)$ が $p(n) \to \infty (n \to \infty)$ のとき意味を持つには $\|\boldsymbol{\beta}\|$ が有界となる必要がある。このとき次の結果を報告している。

定理 1 (Hastie et al. (2022a)) : 誤差項は i.i.d. 系列, 条件 (I),(II) を仮定, $\mathbf{x}_i$ $(i = 1, \cdots, n$ は i.i.d. 系列, $\mathbf{E}[x_{ij}] = 0$, $\mathbf{E}[x_{ij}x_{ik}] = \delta_{jk}$ $(\delta_{jj} = 1.\delta_{ik} = 0$ if $j \neq k))$, x_{ij} の $4+\delta$ $(\delta > 0$ 次積率の存在を仮定する. 係数ベクトル $\boldsymbol{\beta}$ $(p \times 1)$ を基準化して $r^2 = \|\boldsymbol{\beta}\|_2^2$, $p/n \to \gamma$ $(\gamma \geq 0)$ ($n \to \infty$) とする。
(i) $\gamma < 1$ のとき

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

(ii) $\gamma > 1$ のとき

$$R_X(\hat{\boldsymbol{\beta}};\boldsymbol{\beta}) \to r^2(1-\frac{1}{\gamma}) + \sigma^2\frac{1}{\gamma-1}\ \ (a.s.)\,. \tag{8}$$

---

[5] $m \times n$ 行列 $\mathbf{A}$ に対して Moore-Penrose 逆行列 $n \times m$ 行列 $\mathbf{A}^{+}$ とする。このとき $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$, $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$ が成り立つ。一般に $\mathbf{A}\mathbf{A}^{+}$ と $\mathbf{A}^{+}\mathbf{A}$ は射影行列 (projection matrices) になるなど Moore-Penrose の一般化逆行列の説明については例えば竹内 (1974) 第 3 章を参照されたい。ここでは説明変数行列 $\mathbf{X}$ について $\mathbf{A} = \mathbf{X}'\mathbf{X}$, $\mathrm{ranl}(\mathbf{X}) = n$ のときは $\mathbf{A}^{+} = \mathbf{X}^{'}(\mathbf{X}\mathbf{X}^{'})^{-1}(\mathbf{X}\mathbf{X}^{'})^{-1}\mathbf{X}$ となることを利用する。($\mathbf{A}^{+}$ は Moore-Penrose の一般化逆行列の条件を満たしていることが分る。)

(**注意 1**) $\gamma < 1$ ならバイアスは存在せず γ が大きくなると 1 で分散は発散する。$\gamma > 1$ のときにはシグナル・ノイズ比 $c = r^2/\sigma^2$ に依存して結果が変わる。もし $c \leq 1$ なら $\gamma > 1$ が増大するにつれてリスクは単調に減衰する。しかし $c > 1$ のときにはバイアスと分散のトレードオフが登場、γ が増大するとき局所最小解 $(\gamma^* = \sqrt{c}/(\sqrt{c}-1))$ が存在する。

ここで直観的には $\mathbf{X}'\mathbf{X}/n \xrightarrow{p} \mathbf{M}$ $(p \times p)$ なら $\mathrm{tr}[(\mathbf{X}'\mathbf{X}/n)^{-1}] \xrightarrow{p} \mathrm{tr}[\mathbf{M}^{-1}] \sim p/(n-p)$ であるから、適当な条件の下で $\mathrm{tr}[(\mathbf{X}'\mathbf{X}/n)^{-1}] \sim \mathrm{tr}[\mathbf{M}^{-1}]$ となりそうであるが、次元 $p(n)$ が n とともに増大して $\mathbf{x}_i$ $(i = 1, \cdots, n)$ が確率的に変動する場合には以下で示すように正確に評価する必要がある。

(**注意 2**) 定理 1 は興味深い結果であるが、定理 1 とその拡張の Hastie et al.(2022a) による証明では Random Matrix Theory(確率的行列論) の一般的結果および Hastie et al.(2022b), Knowles and Yin (2017) による最新の結果などが必要となる。

また数理的に一般化にする（例えば定理 1 の一般化）には Hastie et al. (2022b) で詳しく説明しているように RMT(Random Matrix Theory, 確率的行列論) の長い議論が必要となるので定理 2 の重要な論点のみを補論で言及する。

(**注意 3**) 簡単な場合として、$\gamma < 1$ のとき説明変数ベクトル $\mathbf{x}_i (i = 1, \cdots, n)$ が多次元正規分布 $N_p(\mathbf{0}, \mathbf{\Omega})$ にしたがうと仮定すれば, 逆ウイッシャート分布の性質 (例えば Anderson (2003) p.273, Lemma 7.7.1) より

$$\mathbf{E}[(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{'})^{-1}] = \mathbf{E}[(\mathbf{X}'\mathbf{X})^{-1}] = \frac{1}{n-p-1}\mathbf{\Omega}^{-1} \tag{9}$$

を利用して類似の結果が証明できる。(ここで $\mathbf{\Omega} > 0$ を仮定する)。ここで $\mathbf{\Omega} = \mathbf{I}_p$ とおき、行列の trace を求めると $p/(n-p-1) \sim \gamma/(1-\gamma)$ が導けることが鍵となる。

このことから正規性を仮定すると次のようなに確率収束の意味ではあるが、統計的多変量解析による初等的に証明することが可能である。問題の理解に有用と考えられるので本稿の数理的補論にその証明を与えておく。また補論の議論から明らかなように、説明変数についての正規性の仮定はさらに緩めることが可能である。

定理 $\mathbf{1}^*$ ： 誤差項は i.i.d. 系列, 条件 (I),(II) を仮定, $\mathbf{x}_i$ $(i = 1, \cdots, n$ は i.i.d. 系列 x_{ij} は互いに独立な N(0,1) にしたがうことを仮定する. $p/n \to$

γ ($\gamma \geq 0$) ($n \to \infty$) とする。
(i) $\gamma < 1$ のとき

$$R_X(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}) \xrightarrow{p} \sigma^2 \frac{\gamma}{1-\gamma} \ , \tag{10}$$

(ii) $\gamma > 1$ のとき

$$R_X(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}) \xrightarrow{p} r^2(1 - \frac{1}{\gamma}) + \sigma^2 \frac{1}{\gamma - 1} \ . \tag{11}$$

(**注意 4**) 定理 1 の範囲では $c \leq 1$ であればリスクは σ^2 より大きくなる。Hastie et al. (2022a) はリッジ回帰 (Ridge Regression) 推定、$\mathbf{x}_i$ に共分散構造がある場合、線形モデルが misspecified である場合、非線形性などの場合、などをかなり包括的に検討している。設定されている線形回帰モデルが誤っている (misspecified) 場合には, ある意味では当然の結果とはいえるが、γ が大きくなる時にノルムを最小化する推定に基づく予測のリスクが単調に減衰する場合もあり得ることが示されている。ただし高次元モデルの場合、予測誤差の挙動を系統的に分析するには RMT が利用する必要があるのでかなり困難な問題となる。数理的付論では一つの結果について言及しておくが、本稿の主な目的は「過適合と二重降下現象」であるので詳細な議論は省略する。

(**注意 5**) Hastie et al. (2022a) では誤差項の分散 σ^2 (> 0) が一定, 説明変数ベクトル $\mathbf{x}_i$ ($i = 1, \cdots, n$ の分散共分散が均一の場合 (isotropic features), 不均一かつ相関構造のある場合 (correlated features) の場合を固有値の挙動を利用して詳しく分析している。必ずしも誤差項が i.id. にしたがうとは限らない場合の議論は今後の課題と思われる。

3　シミュレーション実験

3.1　シミュレーション 1

Hastie et al. (2021, ISLR (2021) とも呼ぶ)10 章では真のモデルが $Y = \sin(X) + \epsilon$, $X \sim U[-5, +5]$, $\epsilon \sim N(0, 0.3^3)$ として $n = 20$ として 3 次スプライン関数をフィットすると二重降下現象が確認できることを説明して

いる。我々は少し拡張して統計モデル

$$Y = a\sin(X + c) + \epsilon\ ,\ X \sim U[-h, +h]\ ,\ \epsilon \sim N(0, \sigma^2) \qquad (12)$$

を利用する。ここで a, c を様々な値に設定することで Hastie et al. (2021) の主張を確認する。ISLR (2021) では $a = 1, c = 0$, $n = 20, h = 5, \sigma = 0.3$ の場合の結果を Interpolation における double decent 現象の典型例として説明している。ここで右辺第 1 項は $\sin(x + c) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}(x + c)^{2n+1}$ であるから、真の回帰関数は無限次の多項式（滑らかな関数）である場合にスプライン関数による回帰を利用したデータ分析を行う状況に対応する。

まず Hastie et al. (2021) による $n = 20$ として X の自然スプライン関数 (natural spline) をフィットすると二重降下現象が確認できるという主張の再現を試みた。データ数 n=20, スプライン関数は 3 次関数 (パラメター 4) であり端点での制約条件 (パラメター数 4) を説明変数の knot 結節点数を K とすると定数項 1 より $y_i = \beta_0 + \sum_{j=1}^{K+3} \beta_j b(x_i) + u_i\ (i = 1, \cdots, n)$ と云う形式の回帰モデルになる。3 次スプラインの場合は $1, x, x^2, (x - \xi)^3$ (ξ は結節点) であるが端点に natural spline の制約がある。

ここでデータ数 n, パラメター数は $d = K + 4 - (4 - 1) = K + 1$、自由度 (degrees of freedom) は $n - d$ となる。（ISLR の Page 299 参照。) $d < n$ の場だけでなく $d = n$,$d > n$ の場合も検討する。$d < n$ の場合は $p \to n$ とすると予測誤差が急速に増大、$d > n$ のときには予測誤差が減衰したのちまた増大、最小二乗推定によるフィットの図と予測誤差の幾つかの図を付論 B に示しておく。

付論 B に掲載した図より次のような事項が観察される。

(i) $X \sim U[-h, +h]$, $h = 5$ として d を大きくとるとしばしば推定スプライン関数が途中で発散的な挙動を示すことが生じる。これは離散的なノードがたまたまま隣接して観察されたとき、多項式でフィットしようとするために生じると考えられる。そこでフィットするモデルを固定してデータ数を増加させてみたのが図 1 と図 2 である。予想通り、データ数を増加すると予測値は安定し、データ数を増加すると予測誤差が減少することが観察される。

(ii) データ数を固定してパラメター数 d を増大するシミュレーションを行って見た結果が図 3 である。$d > n$ の場合には観測データは完全フィットとなる。この時に予測誤差をプロットしたのが図 4 である。$d \to n$ のとき予測誤差は発散するが、その後に降下する現象が始まり、図 4 のように二重降下の現象が観察されている。

(iii) 図 5 では二重降下現象を漸近的に説明した Hastie et al.(2022) の理論値（定理 1）を (黒) 曲線で同時に示した。青線（右上の 0）は複数回のシミュレーションを行った結果の median の値を示している。なお、ここでのシミュレーションの設定は真のモデルは無限次の多項式で表現される非線形モデルという misspesified の場合なので定理 1(Hastie et al. (2022)) の理論的検討と必ずしも整合的とは言えない。しかし類似の二重降下の状況は確認できる。
(vi) シミュレーションにおける $d < n$ の場合の予測誤差の最小化点と $d > n$ の場合の予測誤差の最小化点の比較は興味深い。前者が後者より小さい場合は古典的な場合（パラメター数がデータ数より小さい場合）と思われる。例えば図 5 からはモデルの設定が正しくない場合に回帰分析を行うと予測誤差が漸近的評価より小さくなり得ないことを示しているが、一般的に成立するか更なる検討が必要である。後者が前者より小さくなる場合はどのように特徴づけられるか、また d を大きくすると限りなく予測誤差が単調に小さくなる場合が存在するのかは興味深い論点である。

3.2 シミュレーション 2

ファイナンス分野では例えば Kelly et al. (2024) の研究が論争的であるのでその主張がどこまで正しいか否かを実験で確かめることは興味深い。Kelly et al. (2024) では線形回帰モデル $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ を仮定している。ただし彼らの仮定 2・3 では独立な説明変数は分散共分散 $\boldsymbol{\Psi}$, $\mathrm{tr}(\boldsymbol{\Psi})/p$ の有界性、仮定 4 では $\boldsymbol{\beta} = (\beta_j)$ はランダム $\mathbf{E}(b_i) = 0$, $\mathbf{E}[\boldsymbol{\beta}\boldsymbol{\beta}^{'}] = b_{*,p}/p\mathbf{I}_p$、$\mathrm{tr}(\mathbf{E}(\boldsymbol{\beta}\boldsymbol{\beta}^{'}))/p$ の有界性などを仮定している。この場合、double decent な現象が起きるが、さらに反転することがないことを主張しているように解釈できる。したがって、Kelly et al. (2024) の議論は Hastie et al. (2022) の結果とは異なるが、想定している統計モデルの設定は類似しているが、異なる仮定を含んでいる。

シミュレーション 2 として定数項なしの線形回帰モデル $y_i^{(p)} = \sum_{j=1}^{p} x_{i,j}^{(p)} \beta_j^{(p)} + \epsilon_i^{(p)}$ $(i = 1, \cdots, n)$ を利用する。ここで誤差項は $\mathbf{E}[\epsilon^{(p)}] = 0, \mathbf{E}[[\epsilon_i^{(p)}[^2 = \sigma^2$ であるが簡単化の為 $N(0, \sigma^2)(\sigma = 1)$ とする。係数の数と観測数は $p/n = c, n \to \infty, p(n) \to \infty, n = 10, 100, 500, \cdots$, $x_{ij} \sim N(0, 1)$（$i = 1, \cdots, n; j = 1, \cdots, p(n)$）として
(Case 1) $\beta_j^{(p)} = r/\sqrt{p}$ $(j = 1, \cdots, p)$, $i.e. r^2 = \|\boldsymbol{\beta}\|_2^2$,
(Case 2) $\beta_j \sim N(0, r^2/p)$ 互いに独立,

の場合を設定した。シミュレーションにより発生させたtraning-dataに対して最小二乗フィットを行い、さらにtest-dataに対して予測誤差を評価した。このシミュレーションではCase 1がHastie et al. (2022a)のIsotropic featureの標準的な場合、Case 2がKelly et al. (2024)の最も単純な場合に対応する。

付論Bの図6・図7はCase 1とCase 2の実験結果を示している。なお、pが大きい時説明変数x_{ij}のオーダーに注意する必要がある。統計モデルの分散は

$$\mathbf{E}(y_i^{(p)2}) = \mathbf{E}\left[\sum_{j=1}^{p} \beta_j^{(p)} X_{ij}\right]^2 + \sigma^2 \tag{13}$$

が対応するので、シミュレーション2は$p(n), n$が大きい時には$\mathbf{Var}[y] = r^2 + \sigma^2$が対応する。ここで$r^2$はシステマチック部分、$\sigma^2$はノイズ部分の貢献である。

さらに$X_{ij} \sim N(0, 1/p)$として実験した結果を図8・図9に示しておく。$p(n), n$が共に大きい時には$r \to 0$となりシステマティック部分の変動が無視されて、ノイズの貢献が回帰モデルを支配することになる。二つの論文の一見すると矛盾するような主張の背後の議論に関係する事項であり、Kelly et al. (2024)の主張に対応するのではないかと解釈できるだろう。

4　若干の考察

本稿で議論している過適合と二重降下現象を巡る問題から幾つかの論点を議論しよう。

線形回帰モデルにおける過適合と二重降下現象についての研究の経緯、本稿と異なる高次元確率論を用いたリスク評価についてはBartlett et al. (2020), Tsigler and Bartlett (2023)が詳しい。高次元の場合のリスクの上限と下限を導き、説明変数の個数pと観測数nが必ずしも比例的でない場合にも二重降下現象は起きる条件を示しているという意味ではより一般的ではあるが、高次元確率論における重要な仮定、確率分布がsub-exponentialクラスに限定されるという意味では制約的である。

ここで前節で利用した定数項なしの線形回帰モデル$y_i^{(p)} = \sum_{j=1}^{p} x_{i,j}^{(p)} \beta_j^{(p)} + \epsilon_i^{(p)}$ $(i = 1, \cdots, n)$を利用して、離散モデルと連続モデルの関係を考察しよう。離散メッシュ$i(n) = 1, \cdots, n$に対し$i(n)/n \to t$, $j(p) = 1, \cdots, p$に対し$j(p)/p \to s$をとり$n, p(n) \to \infty$となる状況を考える。ここで変数

を基準化して $y^{(p)}_{i(n)}/\sqrt{n} = \Delta y_{i(n)/n}$, $x^{(p)}_{i(n),j(p)}/\sqrt{n} \sim \sigma_x W(\Delta i(n)/n, j(p)/p)$, $\beta^{(p)}_j = \beta_{j(p)/p}$, $\sigma^2/n \sim \mathbf{Var}(d\epsilon_{i(n)/n})$, $(1/p) \sim \Delta s$ とすると、離散近似 $(1/\sqrt{n})y^{(p)}_i = \sum_{j=1}^{p}(1/\sqrt{n})x^{(p)}_{i,j}(1/\sqrt{p})\beta^{(p)}_j + (1/\sqrt{n})\epsilon^{(p)}_i$ により $i(n)/n \to t$, $j(p)/p \to s$ のとき弱収束 (weak convergence) の意味で連続モデル表現

$$dy_t = \int_0^1 \beta_s \sigma_x \dot{\mathbf{W}}(t,s)ds + \sigma_\epsilon dB_t \ \ (0 \le t, s \le 1) \tag{14}$$

が得られルだろう[6]。ここで詳細な議論は省略するが, B_t $(0 \le t \le 1)$ はブラウン運動、$\mathbf{W}(t,s)$ $(0 \le t, s \le 1)$ は任意の時刻 t に対する（$B(t)$ とは独立な）柱状ブラウン運動である。こうした表現はあまり見かけないが, $y_t = \int_0^t \int_0^1 \beta_s \sigma_x \dot{\mathbf{W}}(t,s)ds + \sigma_\epsilon B_t$ と表現すると連続時間では $\mathbf{E}[\dot{\mathbf{W}}(t,s)\dot{\mathbf{W}}(t',s)] = dt$ より $\mathbf{Var}[y_t] = \sigma_x^2 \int_0^1 \beta_s^2 ds + \sigma_t^2$, $\sigma_t^2 = t\sigma_\epsilon^2$ $(0 \le t \le 1)$ が得られる。

5 おわりに

伝統的な回帰分析で想定している条件 $p << n$ から一旦離れて、$p > n$ の場合 (過学習, overparametrization の場合と呼ばれている) を考察すると、過学習における二重降下 (double decent) 現象が観察される。このとき回帰モデルにおける説明変数の数 $p(n)$ の選択問題なども再び浮上してくる。例えば予測誤差をデータから推定した基準を最小化することがどの様な状況で有用なのか、などを理解する必要があるだろう。しかしながら $p > n$ の場合には $p < n$ の場合と異なる解決すべき幾つかの統計的問題も浮かび上がってくる。画像解析や大規模言語モデルなど原理的に n が非常に大きくとれる工学的応用に限らず、例えば近年のミクロ経済分析などでは多数のダミー変数を利用することなどが行われているので本稿の議論は様々な応用に置いてあながち意味のないこととは云えないだろう。

過適合と二重降下現象の解明は統計的モデリング, 例えばDeep Learning モデルにおける入力変数の選択、Layer 数の決定などに役立つ可能性がある。また統計的問題としては説明変数の係数 $\boldsymbol{\beta}$（$p \times 1$ ベクトル）の次元

[6]舟木・乙部・謝 (2019) を参考としたが、$\mathbf{W}(t,x)$ $(0 \le t, x \le 1)$ は Hilbert 空間値をとる確率変数（柱状ブラウン運動, cylindrial Brownian Motion）, $\dot{\mathbf{W}}(t,x)$ $(0 \le t, x \le 1$) は任意の t におけるホワイトノイズを意味する。連続値をとる任意の x をとめると t についてブラウン運動であるが、その存在は数学的には自明ではない。

が大きく、無限に操作母数 (incidental parameters) の数が大きくなるという統計学における古典的問題に関連している。

本稿では特に線形回帰モデルに絞って二重降下現象にを検討した。近年に注目されている深層学習は画像認識や大規模言語モデルなど幾つかの工学分野での応用上での有効性が既に確認されているが, 多くの問題において多数の母数を含む統計モデルがどこまで有用であるかはなお未知な部分が少なくない。標本数 n のとき母数 $p(n)$ をどの様にとり選択すると適合度 (goodness of fitting) ではなく予測精度 (prediction precision) が向上するか、統計科学では AIC(赤池情報量規準) を始め長い間議論されているがいまだ決定的な結果は得られていない。ここで観察データへの完全フィッテング, 過適合な統計モデルの利用可能性という新たな問題が実務的な観点からも浮上している。有限標本理論と共に本稿のような漸近的な評価, シミュレーション実験などのアプローチからの検討が今後の議論の一つのきっかけになることを期待したい[7]。

参考文献

[1] 現代数理統計学 , 竹村彰通, 2020, 学術図書出版.

[2] 回帰分析, 佐和隆光, 1979, 朝倉書店。

[3] 竹内啓, 線形数学, 1974, 培風館。

[4] 国友直人・湯浅良太編 (2025),「R と Python による統計的学習入門: ISLR・ISLP 実習 (日本語版)」, 統計エキスパート DP), 統計数理研究所.

[5] 舟木直久・乙部厳己・謝賓 (2019) 「確率偏微分方程式」, (岩波書店)

[6] Anderson, T.W. (2003), *An Introduction to Statistical Multivariate Analysis*, 3rd Edition, Wiley.

[7] Bai, Z. and J. Silverstein (2009), *Spectral Analysis of Large Dimensional Random Matrices*, 2nd Edition, Springer.

[7]統計的学習についての R と Python の日本語プログラム (ISLR 及び ISLP) は例えば国友編 (2025) にある。

[8] Bartlett, P., Long,P.,Lugosi, G. and Tsigler, A. (2020), “Benign overfitting in linear regression, “ *PNAS*, 48-117, 30063-30070.

[9] Bishop, C.M. and Bishop, H. (2024), *Deep Learning*, Springer.

[10] Hastie,T., Montanari, A., S. Rosset and R.J. Tibshirani (2022a),”Surprises in High-Dimensional Ridgeless Least Squares Interpolation,” *Annals of Statistics*, 50-2, 949-986.

[11] Hastie,T., Montanari, A., S. Rosset and R.J. Tibshirani (2022b),”Supplement to Surprises in High-Dimensional Ridgeless Least Squares Interpolation,” arXiv:1903:08560.

[12] James, G., D.Witten, Hast, T, and Tibshirani, R.. (2021), *An Introduction to Statistical Learning with Applications in R*, Springer.

[13] Kelly, B, S. Malamud, and K. Zhou (2024), ” The Virtue of Complexity in Return Prediction”, *Journal of Finance*, LXXIX-1, 459-503.

[14] Knowles, Antti and Jun Yin (2017), ”Anisotoropic local laws for random matrices,” *Probability Thepry and Related Fields*, 169(1-2), 257-352.

[15] Tsigler, A. and Bartlett, P. (2023), “Benign overfitting in ridge regression, “ *Journal of Machine Learning Research*, 1-74.

[16] Yao, J., Zheng, S. and Z. Bai (2015), *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge U.P.

付論 A:数理的補論

この補論では前節まで述べた幾つかの定理や命題について数理的導出および追加事項を補論として述べておく。

A.1 定理 1^* **の証明** :
(I) $p \times 1$ ベクトル列 $\mathbf{x}_i$ $(i = 1, \cdots, n)$ が互いに独立に $N(\mathbf{0}, \mathbf{I}_p)$ にしたがうとする。$\mathbf{S}_n = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{'}$ は Wishart$(n, p, \mathbf{I}_p)$ にしたがう。$1\mathbf{S}_n^{-1}$ の対角要素、非対角要素の分布は対称性から同一であるから期待値は c_1, c_2 がと

れて $\mathbf{E}[\mathbf{S}_n^{-1}] = c_1\mathbf{I}_p + c_2\mathbf{1}_p\mathbf{1}_p'$ となる。(ここで $\mathbf{1}_p' = (1, \cdots, 1)$ とした。) 任意の直交行列 $\mathbf{Q}$ にたいして $\mathbf{Q}\mathbf{E}[\mathbf{S}_n^{-1}]\mathbf{Q}' = \mathbf{E}[\mathbf{S}_n^{-1}]$ であるから $c_2 = 0$ となる。次に Anderson (2003) の Theorem 5.2.2(Hotelling の T 統計量の導出) の証明より例えば $[1 + (p-1)] \times [1 + (p-1)]$ に分割して

$$\mathbf{S}_n^{-1} = \begin{bmatrix} s_{11n} & \mathbf{s}_{1n}' \\ \mathbf{s}_{1n} & \mathbf{S}_{22n} \end{bmatrix}^{-1}$$

とおくと対角成分 (1,1) 要素は $V = s_{11n} - \mathbf{s}_{1n}'\mathbf{S}_{22n}^{-1}\mathbf{s}_{1n}$ の逆数の分布に一致する。$V \sim \chi^2(n-(p-1))$ より Gamma 分布の性質を利用すると $\int_0^\infty v^{-1}c(m)v^{(m-2)/2}e^{-v/2}dv = \int_0^\infty c(m)v^{(m-2-2)/2}e^{-v}dv = c(m)/c(m-2) = 1/(m-2)$ （ただし $m = n-(p-1)$, $(c(m) = 1/2^{m/2}\Gamma(m/2))$) となる。同様に $\int_0^\infty v^{-2}c(m)v^{(m-2)/2}e^{-v/2}dv = \int_0^\infty c(m)v^{(m-2-4)/2}e^{-v}dv = c(m)/c(m-4) = 1/[(m-2)(m-4)]$ であるから, 自由度 $m = n-(p-1)$ より

$$\mathbf{E}[V^{-1}] = \frac{1}{n-p-1}\ ,\ \mathbf{E}[V^{-2}] = \frac{1}{(n-p-1)(n-p-3)}$$

となるので

$$\mathbf{Var}[V^{-1}] = \frac{2}{(n-p-1)^2(n-p-3)} < \frac{2}{(n-p-3)^3}$$

が得られる。
次に $p \times p$ 行列 $\mathbf{A}_n = (1/n)\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' = (a_{ij})$, $\mathbf{A}_n^{-1} = (a^{ij})$ とする。ここで

$$\frac{1}{n}\mathrm{tr}(\mathbf{A}_n^{-1}) - \frac{p}{n}\frac{1}{1-\gamma} = \frac{p}{n}(\frac{1}{p})\sum_{i=1}^p[(a^{ii} - \mathbf{E}(a^{ii})) + (\mathbf{E}(a^{ii}) - \frac{1}{1-\gamma})]$$

を評価する。
(i)$n \to \infty$ のとき $\mathbf{E}[a^{ii}] = n/(n-p-1) \to 1/(1-\gamma)$,
(ii)

$$\begin{aligned}\mathbf{Var}[\frac{1}{p}\sum_{i=1}^p(a^{ii} - \mathbf{E}(a^{ii})] &= (\frac{1}{p})^2\sum_{i,j=1}^p \mathbf{E}[(a^{ii} - \mathbf{E}(a^{ii}))(a^{jj} - \mathbf{E}(a^{jj})] \\ &\leq (\frac{1}{p})^2\sum_{i,j=1}^p[\mathbf{Var}[(a^{ii})\mathbf{Var}(a^{jj})]^{1/2}\end{aligned}$$

である。
$\mathbf{A}_n$ はi.i.d. の仮定の下で対称であるから $n \to \infty$ $(p \to \infty)$ のとき $\frac{1}{n}\mathrm{tr}(\mathbf{A}_n^{-1}) - \frac{p}{n}\frac{1}{1-\gamma}$ は 0 に確率収束する。

(II) $0 < \gamma < 1$ のときには既に求まったので、$1 < \gamma$ のときを扱う。

補題 A.1 ： 固有値 $s_i = \lambda_i(\mathbf{X}'\mathbf{X})$ $(i = 1, \cdots, n)$, $t_i = \lambda_i(\mathbf{X}\mathbf{X}')$ $(i = 1, \cdots, n)$ とすると、ゼロでない固有値について $\sum_{i=1}^n (1/s_i) = \sum_{i=1}^n (1/t_i)$ となる。

(**補題 A.1 の証明**) (i) $n \times p$ $(n \geq p)$ 行列 $\mathbf{A}$ とする。$(n+p) \times (n+p)$ 分割行列 (例えば Anderson (2003) の Theorem A.3.2) について関係

$$
\begin{aligned}
\left| \begin{array}{cc} \lambda \mathbf{I}_n & \mathbf{A} \\ \mathbf{A}' & \mathbf{I}_p \end{array} \right| &= |\lambda \mathbf{I}_n - \mathbf{A}\mathbf{A}'||\mathbf{I}_p| \\
&= |\lambda \mathbf{I}_n||\mathbf{I}_p - \mathbf{A}'(\lambda \mathbf{I}_n)^{-1}\mathbf{A}| = \lambda^{n-p}|\lambda \mathbf{I}_p - \mathbf{A}'\mathbf{A}|
\end{aligned}
$$

より二つの行列のゼロでない固有値が一致することが分かる。
$p \geq n$ のときは n と p を交換すれば (i) に帰着できる。
（**Q.E.D.**）

行列 $\mathbf{X}'\mathbf{X}$ の固有値が行列 $\mathbf{X}\mathbf{X}'$ に一致する (Singlar Value Decompotion) ので $\mathrm{tr}(\mathbf{E}[(\mathbf{X}\mathbf{X}')^{-1}]) = [1/(p-n-1)]\mathrm{tr}(\mathbf{I}_n)$ より $n/(p-n-1) \sim 1/(\gamma - 1)$ が求まる。すなわち, 固有値 $s_i = \lambda_i(\mathbf{X}'\mathbf{X})$ $(i = 1, \cdots, n)$, $t_i = \lambda_i(\mathbf{X}\mathbf{X}')$ $(i = 1, \cdots, n)$ とすると、ゼロでない固有値は $\sum_{i=1}^n (1/s_i) = \sum_{i=1}^n (1/t_i)$ より

$$
\begin{aligned}
\sigma^2 \mathrm{Tr}[(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i')^+] &= \sigma^2 \sum_{i=1}^n \frac{1}{s_i} \\
&= \sigma^2 \sum_{i=1}^n \frac{1}{t_i} \\
&\sim \sigma^2 \frac{n}{p} / [1 - \frac{n}{p}] = \frac{\sigma^2}{\gamma - 1}
\end{aligned}
$$

となる (補題 A.1 を参照)。バイアス部分は

$$
B_X(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \boldsymbol{\beta}'[\mathbf{I}_p - (\mathbf{X}'\mathbf{X})^+(\mathbf{X}'\mathbf{X})]\boldsymbol{\beta} \tag{A.15}
$$

より $\mathbf{E}[B_x(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})] = r^2(1 - n/p)$ を次のようにすると導ける。任意の直交行列 $\mathbf{U}$ と単位ベクトルを用いては $\mathbf{U}\boldsymbol{\beta} = r\mathbf{e}_i$ $(i = 1, \cdots, p)$, $r^2 = \boldsymbol{\beta}'\boldsymbol{\beta}$ と

する。このとき $\mathbf{E}[\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})^{+}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}] = r^2\mathbf{E}[\mathbf{e}_i'\mathbf{U}'(\mathbf{X}'\mathbf{X})^{+}(\mathbf{X}'\mathbf{X})\mathbf{U}\mathbf{e}_i]$
より $(1/p)\sum_{i=1}^{p}[\cdot\cdot]$ をとると $(1/p)\mathbf{E}[\mathrm{tr}(\mathbf{X}'\mathbf{X})^{+}(\mathbf{X}'\mathbf{X})] = n/p$ となるからである。
(**Q.E.D.**)

A.2. 定理 1^* についての考察 :
定理 1^* における主張は説明変数が確率的な場合、必ずしも多次元正規分布とは限らなくても成立する。

例えば, 説明変数の積率条件 $\mathbf{E}[x_{ij}^4] < \infty$ が一様に成立することを仮定しよう。任意の i $(i = 1, \cdots, n$ について $K_n = p - 1$ として $(1 + K_n)$ ベクトル $\mathbf{x}_i = (y_i, \mathbf{z}_i')'$ に分割、$\mathbf{y} = (y_i)$, $\mathbf{Z} = (\mathbf{z}_i')$, $\mathbf{P}_Z = \mathbf{Z}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$ $(\mathrm{rank}(\mathbf{P}) = K_n)$ とする。このとき $\mathbf{P}_Z$ は射影行列であるから $\mathbf{E}[\mathbf{y}'\mathbf{P}_Z\mathbf{y}] = K_n$, $\mathbf{E}[(\mathbf{y}'\mathbf{P}_Z\mathbf{y})^2] = \kappa_4\sum_{i=1}^{n}\mathbf{E}[p_{ii}^2] + K_n^2 + 2K_n$ より
$\mathbf{Var}[\mathbf{y}'\mathbf{P}_Z\mathbf{y}] = \kappa_4\sum_{i=1}^{n}\mathbf{E}[p_{ii}^2] + 2K_n$ となる。(ここで $\kappa_4 = \mathbf{E}[x_{ji}^4] - 3$ とする。) この項は $O(K_n)$ である。同様に $\mathbf{Var}[\mathbf{y}'(\mathbf{I}_p - \mathbf{P}_Z)\mathbf{y}] = \kappa_4\sum_{i=1}^{n}\mathbf{E}[(1 - p_{ii})^2] + 2(n - K_n)$ より $O(n - K_n)$ となる。したがって $(1/n)\mathbf{y}'(\mathbf{I}_p - \mathbf{P}_Z)\mathbf{y} - (1 - \gamma) \to 0$ $(n \to \infty)$ となる。$(\frac{1}{n})\mathbf{y}'(\mathbf{I}_p - \mathbf{P}_Z)\mathbf{y} \to 1 - \gamma$ (> 0) $(n \to \infty)$ であるから $n \to \infty$ のとき

$$[(\frac{1}{n})\mathbf{y}'(\mathbf{I}_p - \mathbf{P}_Z)\mathbf{y}]^{-1} - 1/(1 - \gamma) = \frac{-(\frac{1}{n})\mathbf{y}'(\mathbf{I}_p - \mathbf{P}_Z)\mathbf{y} + (1 - \gamma)}{(1 - \gamma)(\frac{1}{n})\mathbf{y}'(\mathbf{I}_p - \mathbf{P}_Z)\mathbf{y}} \xrightarrow{p} 0$$

となる。これより $\frac{1}{n}\mathrm{tr}(\mathbf{A}_n^{-1}) - \frac{p}{n}\frac{1}{1-\gamma}$ は 0 に確率収束することが期待できる。
なおここでの直観的な議論は数理的には p, n は同時に大きくなる状況での追加の議論が必要である。

また現時点では多くの応用統計家にとり確率行列 (RMT) の理論は自明とは云えない。説明変数の分散分散行列 $\boldsymbol{\Sigma}$ が単位行列に比例するとは限らない場合は分析が複雑になるが初等的な分析も望まれる。4 節で言及した連続過程での近似も課題である。

A.3. RMT と Hastie et al.(2022a) の定理 2 について :
ランダムな非負定符号行列 $\mathbf{A}$ $(p \times p)$ の固有値 λ_j $(j = 1, \cdots, p)$ とすると固有値の経験分布 ESD(empirical specral distribution) は $F^A = (1/p)\sum_{j=1}^{p}\delta_{\lambda_j}$ で与えられる。(ここで δ_λ はデラック記号を意味する。) 有限測度 μ の Stietjes Transform (or Cauchy Transform) は $s_\mu(z) = \int \frac{1}{x-z}\mu(dx)$

($z \in \mathbf{C}\backslash\mathbf{\Gamma}_\mu$) ($\mathbf{\Gamma}_\mu$ は μ のサポート) で定義され、

$$s_A(z) = \int \frac{1}{x-z} F^A(dx) = \frac{1}{p}\mathrm{Tr}(\mathbf{A} - z\mathbf{I})^{-1} \tag{A.16}$$

で与えられる。ここで次元 p が n に比例する高次元の場合には多変量解析の通常の議論を修正する必要があるが、理論的分析に有用なスティルチェス変換 (Stietljes Transform) と逆変換については Yao, Zheng, and Bai (2015) の 2.2 節 (Theorem 2.7 など) が分かり易い。実軸上の有限測度 μ に対して Stieltjes 変換は $S_\mu(z) = \int \frac{1}{x-z}\mu(dx)$ ($z \in \mathbf{C}\backslash\mathbf{\Gamma}_\mu$) で定義されるが, Hermite 行列 $\mathbf{A}$ に対して (A.14) で与えられる。また、確率行列論 (random matrix theory, RMT) における Marchenko-Pastur 分布など基本的な内容は例えば Bai and Silverstein (2009) が解説している。

互いに独立な確率変数ベクトルから作られる $p \times p$ ランダム行列 $\mathbf{S}_n = (1/n)\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^{'}$ に対してスティルチェス変換 (Stietljes Transform) を $s_n(z) = (1/p)\mathrm{tr}[\mathbf{S}_n - z\mathbf{I}_p]^{-1}$ とすると、$\mathbf{E}[s_n(z)]$ が満たす方程式は $p/n \to y > 0$ のとき $s_(z) = 1/[1 - z - (y + yzs(z))]$ となる。特に $z = 0$ とすると $s(0) = 1/(1-y)$ が得られるので $p/n \sim \gamma$ より極限は y を γ に変更して $(p/n)(1-y) \sim \gamma/(1-\gamma)$ となる。

ここで説明変数ベクトル $\mathbf{x}(p \times 1)$ の分散共分散行列 $\mathbf{\Omega}$, $\mathbf{x} = \mathbf{\Omega}^{1/2}\mathbf{z}$ となる場合を考察しよう[8]。$\mathbf{\Omega}$ は非負定符号行列であるから $\mathbf{\Omega} = \sum_{i=1}^p s_i\mathbf{v}_i\mathbf{v}_i^{'}$ ($s_1 \geq s_2 \geq \cdots s_p \geq 0$) と固有値分解する。ここで二つの確率分布を

$$\hat{H}_n(s) = \frac{1}{p}\sum_{i=1}^p 1_{\{s \geq s_i\}}\ ,\ \hat{G}_n(s) = \frac{1}{\|\boldsymbol{\beta}\|_2^2}\sum_{i=1}^p < \boldsymbol{\beta}, \mathbf{v}_i >^2 1_{\{s \geq s_i\}} \tag{A.17}$$

で定める。
さらに $c_0 = c_0(\gamma, \hat{H}_n)$ を方程式

$$1 - \frac{1}{\gamma} = \int \frac{1}{1 + c_0\gamma s} d\hat{H}_n(s) \tag{A.18}$$

を満たす正定数とする。Hastie et al. (2022a) は定理 2(Theorem 2) として n と p が共に大きく比例的であり、固有値の有界性などを仮定すると、

[8] ここで確率変数ベクトル $\mathbf{z}$ は $\mathbf{E}[\mathbf{z}] = \mathbf{0}$, $\mathbf{E}[\mathbf{z}\mathbf{z}^{'}] = \mathbf{I}_r$ ($0 < r \leq p$ を満たすとする。なお Hastie et al. (2022a, b) では $\mathbf{x}_i$ の分散共分散行列に対して記号 $\mathbf{\Sigma}$ を利用しているが, 混乱を避ける意味で (9) を含め記号 $\mathbf{\Omega}$ を利用した。

予測誤差のバイアスと分散はそれぞれ漸近的に

$$\mathcal{B} = \|\boldsymbol{\beta}\|_2^2[1+\gamma c_0 \frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}] \int \frac{s}{(1+c_0\gamma s)^2} d\hat{G}_n(s) \qquad \text{(A.19)}$$

および

$$\mathcal{V} = \sigma^2\gamma c_0[\frac{\int \frac{s^2}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}{\int \frac{s}{(1+c_0\gamma s)^2} d\hat{H}_n(s)}] \qquad \text{(A.20)}$$

で与えられることを示している。導出は確率行列論（RMT) における標本分散共分散行列の極限に関するかなり込入った議論（証明の詳細は Hastie et al. (2022b) に与えられているが, 鍵となるのは Knowles and Yin (2017) による Anisotropic local law （異方的局所法則) と呼ばれる数理的結果である）が必要となる。ただし定理 1 のように結果を分布を解析的に表現することは困難であり, 上のような積分表現で与えられる.

ここで説明変数ベクトル $\mathbf{x}$ $(p \times 1)$ の分散・共分散行列が $\boldsymbol{\Omega} = \mathbf{I}_p$ のとき、$\hat{H}_n(s) = \hat{G}_n(s) = 1$ $(s = 1)$ より $1 - 1/\gamma = 1/[1 + c_0\gamma]$ であるから $1 + c_0\gamma - 1/\gamma - c_0 = 1$, $c_0 = 1/[\gamma(\gamma - 1)]$ となるので $\gamma c_0 = 1/(\gamma - 1)$, $(1+\gamma c_0)/(1+\gamma c_0)^2 = (\gamma-1)/\gamma$ となる（ここで $c_0 > 0$ は $\gamma > 1$ を意味する)。したがって $\mathcal{B} = \|\boldsymbol{\beta}\|_2^2(1-1/\gamma)$ および $\mathcal{V} = \sigma^2/(\gamma-1)$ より $\gamma > 1$ の時の 定理 1 の結果に一致することが確認できる。

なお $0 < \gamma < 1$ の場合には分散共分散行列 $\boldsymbol{\Omega} = \mathbf{I}_p$ であるので $c_0 < 0$ (Hastie et al. (2022a) の (12) 式) となる。したがって, この場合には Hastie et al. (2022a) の定理 2 の公式は成立しないことから、定理 2 などの結果については主張は若干の修正 (例えば論文の中で利用している記号 c_0 の定義など) が必要と思われる。

付論 B：関連する図

この付論 B ではシミュレーションの結果の図を掲載する。プログラムは Python により新たに開発した。

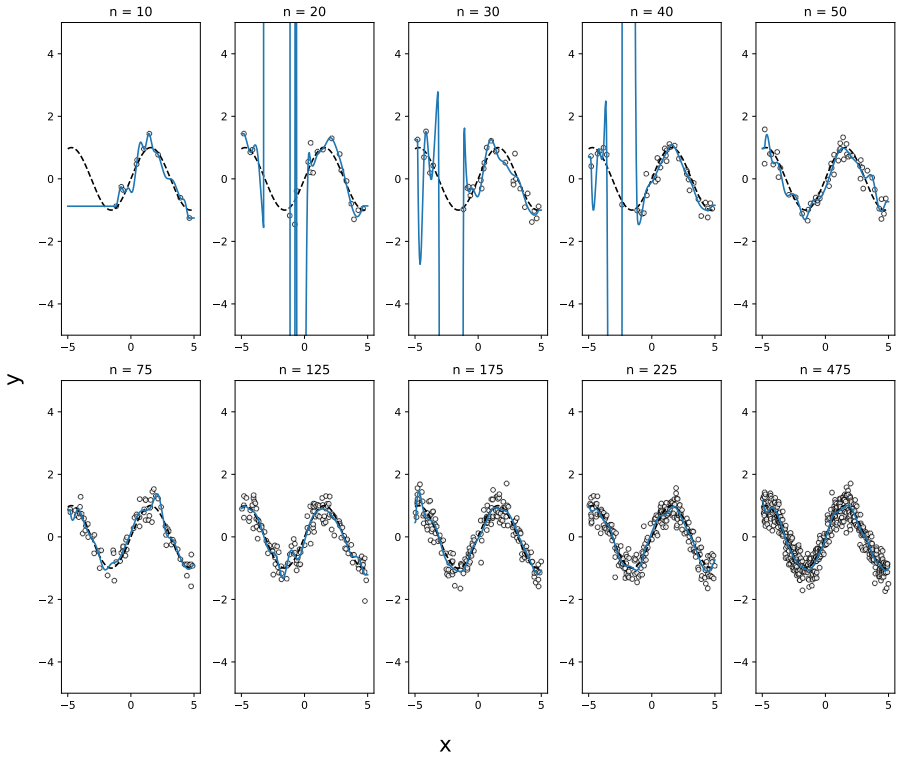

図 1: (最小二乗フィット 1) パラメータ数を $d = 20$ に固定した 3 次スプライン回帰モデルに対し、学習サンプル数 n を変化させたときの推定関数の比較。黒破線は真の関数 $f(x) = \sin(x)$、点はノイズ付き学習データを表し、各パネルは異なる n に対応している。

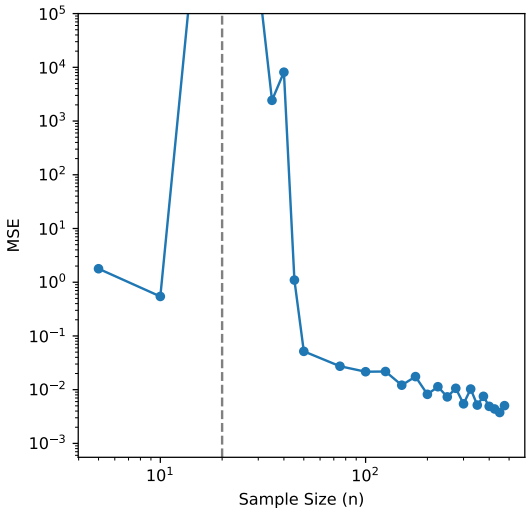

図 2: (最小二乗フィット 2) 3 次スプライン回帰モデルにおける学習サンプル数 n に対するテスト平均二乗誤差（MSE）の依存性。テスト誤差は真の関数 $f(x) = \sin(x)$ に対して評価した。縦軸および横軸は対数スケールで表示している。破線は学習サンプル数がパラメータ数が等しくなる点（$n = d = 20$）。

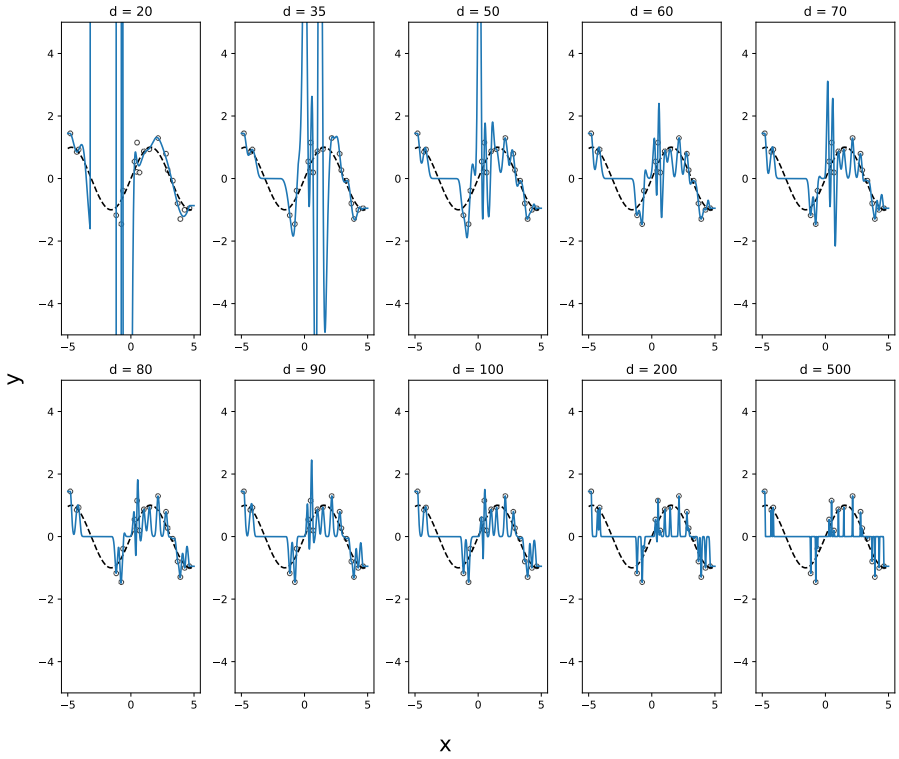

図 3: (最小二乗フィット 3) ノイズを含む学習データ（観測数 $n = 20$）に対して、パラメータ数 d の異なるキュービックスプライン回帰モデルを適用した際の推定関数の比較。黒破線は真の関数 $f(x) = \sin(x)$、点はノイズ付き学習データを表し、各パネルは異なるパラメータ数 d に対応している。

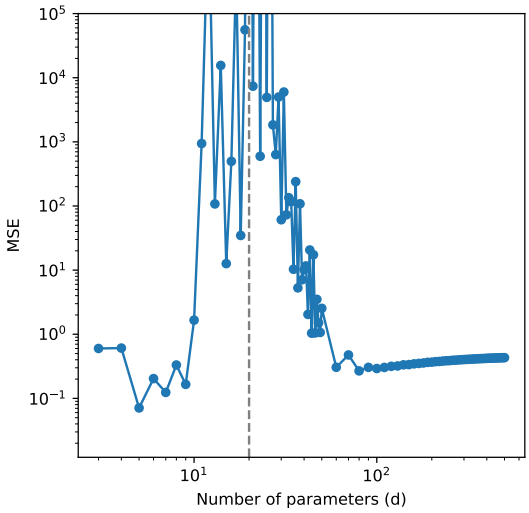

図 4: (二重降下現象 1) 3 次スプライン回帰モデルにおけるパラメータ数 d に対するテスト平均二乗誤差（MSE）の依存性。テスト誤差は真の関数 $f(x) = \sin(x)$ に対して評価した。縦軸および横軸は対数スケールで表示している。破線は学習サンプル数がパラメータ数と等しくなる点（$d = n = 20$）

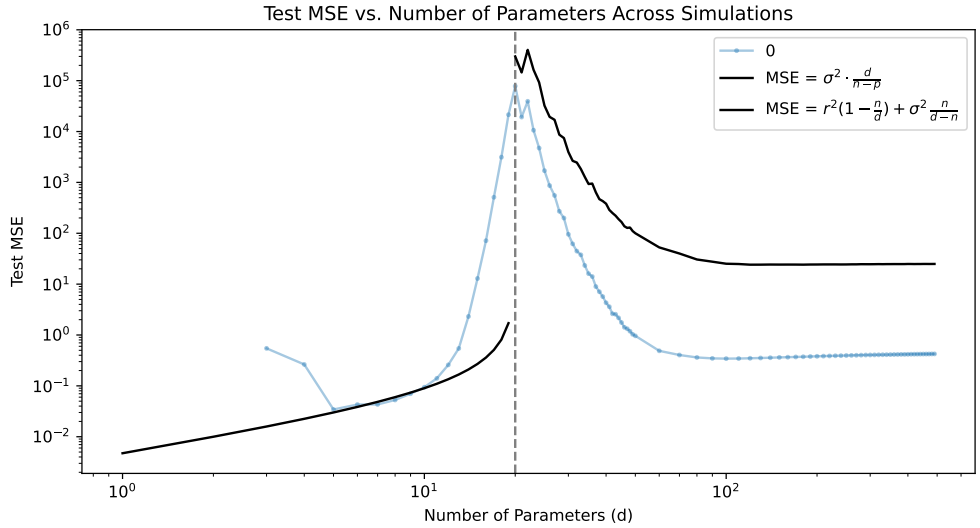

図 5: (二重降下現象 2) 3 次スプライン回帰モデルにおけるパラメータ数 d に対するテスト平均二乗誤差（MSE）の挙動。各 d に対するテスト誤差は、学習データを独立に再サンプリングした 1000 回の反復計算の中央値を用いて評価した。破線は学習サンプル数とパラメータ数が等しくなる点（$d = n$）。黒実線は、$d < n$ の領域における MSE の理論値 (Hastie et al. (2022a))。

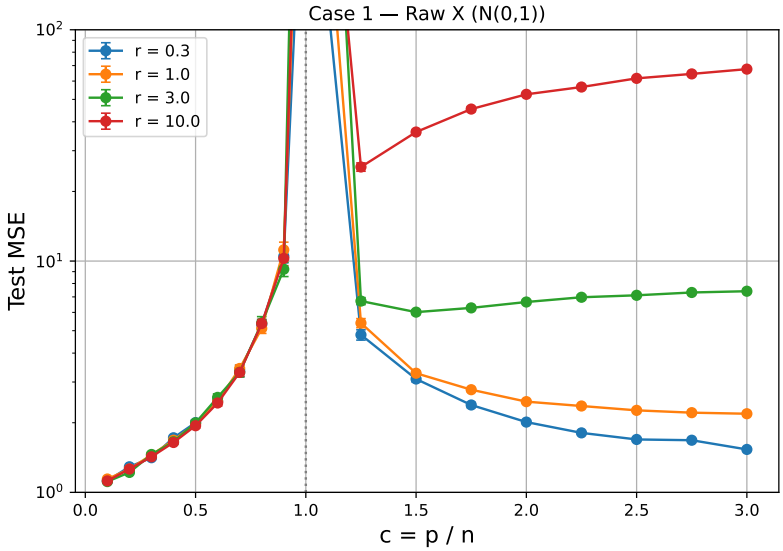

図 6: (二重降下現象 3) 線形回帰モデル $y = X\beta + \varepsilon$ における説明変数次元とサンプルサイズの比 $c = p/n$ とテスト平均二乗誤差（MSE）の関係。回帰係数 β は固定，ユークリッドノルムが $\|\beta\| = r$ となるよう正規化している。説明変数行列 X の各成分は独立に $X_{ij} \sim \mathcal{N}(0, 1)$ に従い，誤差項は $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ とした。各曲線は異なる信号強度パラメータ r に対応し，点は複数回の独立試行に基づくテスト MSE の標本平均，エラーバーはその標準誤差を表す。縦の点線は説明変数次元とサンプルサイズが等しくなる点（$p = n$）。縦軸は対数スケールで表示している。

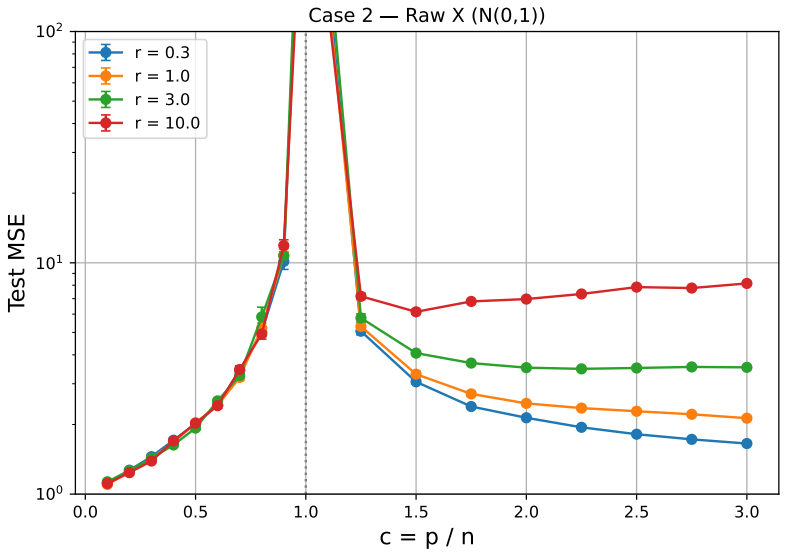

図 7: (二重降下現象 4) 説明変数次元とサンプルサイズの比 $c = p/n$ とテスト平均二乗誤差（MSE）の関係。図 6 と同様の線形回帰モデルを，用い説明変数行列 $X_{ij} \sim \mathcal{N}(0, 1)$ も同様に生成するが，各試行ごとに回帰係数 β をランダムに生成し，そのユークリッドノルムが $\|\beta\| = r$ となるよう正規化している。その他の表示方法，評価手法，および記号の定義は図 6 と同一。

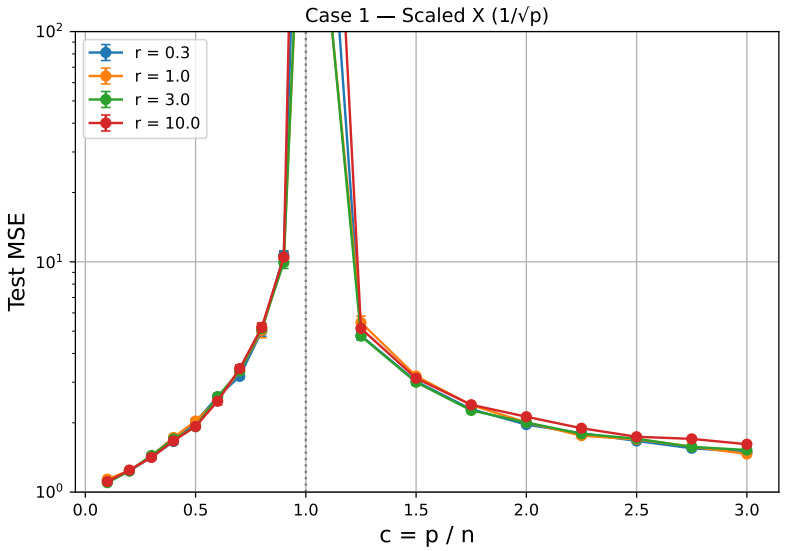

図 8: (二重降下現象 5) 説明変数次元とサンプルサイズの比 $c = p/n$ とテスト平均二乗誤差（MSE）の関係。図 6 と同様であるが，本図では説明変数行列の各成分を $1/\sqrt{p}$ で正規化している。

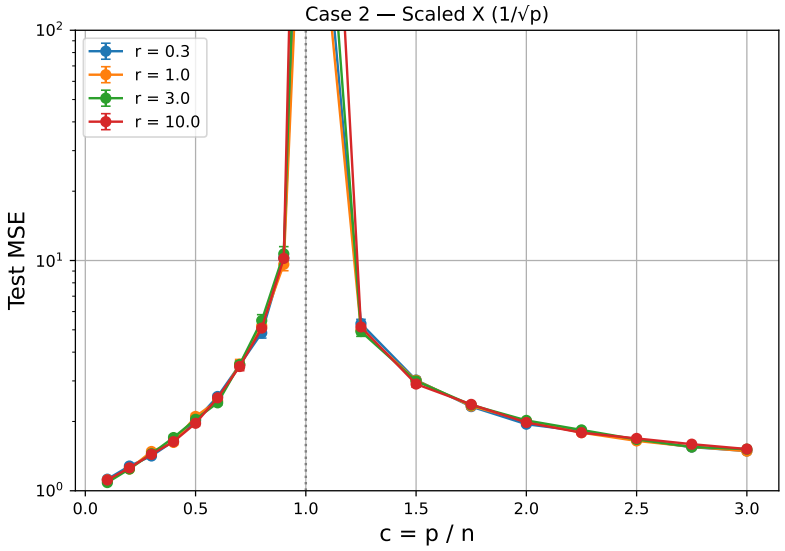

図 9: (二重降下現象 6) 説明変数次元とサンプルサイズの比 $c = p/n$ とテスト平均二乗誤差（MSE）の関係。図 7 と同様であるが，本図では説明変数行列の各成分を $1/\sqrt{p}$ で正規化している。