「誤差は次元割る標本数」は遠い側の近似であり、標本数が次元に近づくと誤差は発散する

— ガウスの説明変数での最小二乗の余分な予測誤差は、よく使う d/n ではなく $d/(n-d-1)$ である

— 分離子は、標本数と次元の比が大きいか、1 に近いか — [論文743]

山岸 勇輝 (Yuuki YAMAGISHI)
Yamagishi Complexity Torus Lab, Japan｜山岸勇輝ラボ
2026年9月

— *From thinking to syncing.* —

要旨

回帰の余分な誤差は「次元割る標本数」でおおよそ見積もられる。それは標本数が次元よりずっと多いときの形である。厳密な形と並べて、近似が届く範囲と、外れ方を数える。新しい定理も法則も主張しない。 第一に、厳密な形は $d/(n-d-1)$ である——次元 10・標本 1000 で 0.010111、近似の 1.011122 倍（第2節）。 第二に、標本が減ると外れが広がる——標本 100 で 0.112360、20 で 1.111111、15 で 2.500000（第2節）。 第三に、これが本稿の芯である。標本数が次元に近づくと発散する——標本 12 で 10.000000、近似の 12.000000 倍（第2節）。 第四に、比を保てば次元に依らない——標本を次元の 10 倍にすると、次元 10 で 0.112360、次元 100 で 0.111235（第3節）。 第五に、シミュレーションが裏を取る——次元 5・標本 20 で 0.357921（厳密 0.357143）、標本 60 で 0.091716（0.092593）（第3節）。 第六に、分離子は、標本数と次元の比が大きいか、1 に近いかである——近似が効くのは前者だけである（第4節）。 キーワード：最小二乗、予測誤差、次元と標本数、近似の届く範囲、発散、標本数と次元の比。

本稿の射程（射程注記）：新しい定理も法則も主張しない——ガウスの説明変数での最小二乗の予測誤差 $\sigma^2 d/(n-d-1)$ は、既知である（シーバー 1977、ハスティー 2009）。説明変数を単純に置く——ガウスで互いに独立とし、相関や外れ値は扱わない。正則化を扱わない——係数を縮める手と変数を選ぶ手は扱わない。境目の手前だけを見る——標本数が次元＋1 以下では期待値が有限でないので扱わない。実在のデータを主張しない——個別のデータや学習器は扱わない。既刊との関係：論文733 は、誤差には最小が在るが、その最小の位置は手持ちのデータ数で動くことを示した——そこでは分散を $\sigma^2 p/n$ の形で置き、実在の学習器の振舞いは主張しないと明記した。ここではその形の厳密版を出し、標本数が次元に近いところで形そのものが外れることを示す。置いた形と、その形が届く範囲になる。論文182 も引く——「次元の呪い」が一つでないことの、標本数の側からの一例である。加えたのは、厳密な形と近似を同じ表に並べたこと、外れが 12.000000 倍まで広がることを示したこと、比を保てば次元に依らないことを示したこと、シミュレーションで裏を取ったこと、分離子を「標本数と次元の比」に置いたことである。

1. 記法と問い

設定：次元 d の説明変数をガウスで独立に置き、標本 n 個で最小二乗を行う。雑音の分散を σ^2 とすると、新しい点での余分な予測誤差の期待値は次のとおりである。

$$\mathrm{E}\big[\|\hat{\beta}-\beta\|^2\big] = \frac{\sigma^2 d}{n-d-1}, \qquad \text{近似 } \frac{\sigma^2 d}{n}$$

以下では $\sigma^2 = 1$ とする。

問い：「次元割る標本数」は、どこまで使えるのか。

2. 近似が外れ、発散する

計算 2.1：次元と標本数を変えて、厳密な形と近似とその比を出した。

次元	標本	厳密	近似	比
10	12	10.000000	0.833333	12.000000
10	15	2.500000	0.666667	3.750000
10	20	1.111111	0.500000	2.222222
10	50	0.256410	0.200000	1.282051
10	100	0.112360	0.100000	1.123596
10	1000	0.010111	0.010000	1.011122
100	1000	0.111235	0.100000	1.112347

観察 2.2

標本が多いと、近似はよく当たる。次元 10・標本 1000 で比 1.011122 である——「次元割る標本数」は、この側の形である。

観察 2.3

これが本稿の芯である。標本数が次元に近づくと、誤差は発散する。標本 12 で 10.000000、近似の 12.000000 倍になる——分母の $n - d - 1$ が 0 に向かうので、近似が小さく見積もる量は際限なく広がる。

観察 2.4

外れは一様ではない。比は標本 50 で 1.282051、20 で 2.222222、15 で 3.750000 と、境目に近いほど急に増える——近似の誤差を「一割くらい」と見るのは、遠い側だけで正しい。

3. 比を保てば、次元に依らない

> **観察 3.1**
>
> 標本を次元の 10 倍に保つと、次元を変えてもほぼ同じになる。次元 10・標本 100 で 0.112360、次元 100・標本 1000 で 0.111235 である——効いているのは次元や標本数そのものではなく、その比である。

計算 3.2：次元 5 で、係数・説明変数・雑音をすべて乱数で作り、最小二乗を解いて余分な誤差を平均した。

次元・標本（回数）	シミュレーション	厳密	相対差
5・20（6000）	0.357921	0.357143	0.002179
5・60（4000）	0.091716	0.092593	0.009468

> **観察 3.3**
>
> シミュレーションは厳密な形に乗る。標本 20 でも相対差は 0.002179 である——近似 $5/20 = 0.25$ ではなく、0.357143 のほうが実際の値である。

> **観察 3.4**
>
> 次元を増やすなら、標本も同じ割合で増やさねばならない。比が保たれていれば誤差はほぼ変わらず、比が 1 に近づけば発散する——次元が多いこと自体より、比の悪化が効く。

4. 分離子

> **観察 4.1**
>
> 分離子は、標本数と次元の比が大きいか、1 に近いかである。近似が効くのは前者だけで、後者では誤差の見積もりそのものが崩れる——「次元割る標本数」を使うときは、比がどちら側かを先に言わねばならない。

> **観察 4.2**
>
> 論文733 と並べる。733 は、誤差の最小の位置が手持ちのデータ数で動くことを、分散を $\sigma^2 p/n$ と置いて示した——ここではその置いた形の厳密版を出し、形が届く範囲を切る。論文182 の言う「呪いは一つではない」の、標本数の側の一例でもある。

> **観察 4.3**
>
> 言わないこと。説明変数の相関も外れ値も正則化も、境目の向こうも扱わなかった——示したのは、近似の届く範囲と、その外での発散である。

5. 点呼と結び

1. 標本 1000 で **0.010111**（計算 2.1）：近似が当たる。
2. 比 **1.011122**（計算 2.1）：遠い側。
3. 標本 100 で **0.112360**（計算 2.1）：外れ始める。
4. 標本 20 で **1.111111**（計算 2.1）：比 **2.222222**。
5. 標本 12 で **10.000000**（観察 2.3）：本稿の芯。
6. 近似の **12.000000** 倍（観察 2.3）：発散する。
7. 次元 100 で **0.111235**（観察 3.1）：比を保てば同じ。
8. シミュレーション **0.357921**（計算 3.2）：厳密 **0.357143**。
9. 標本 60 で **0.091716**（計算 3.2）：厳密 **0.092593**。
10. 比が大きいか、1 に近いか（観察 4.1）：分離子・論文733・182。

結び：「誤差は次元割る標本数」は遠い側の近似であり、標本数が次元に近づくと誤差は発散する。ガウスの説明変数での最小二乗の余分な予測誤差は $d/(n-d-1)$ で、次元 10 なら標本 1000 で 0.010111（近似の 1.011122 倍）とよく当たるが、標本 100 で 0.112360、20 で 1.111111、15 で 2.500000 と外れが広がる——そして標本 12 では 10.000000、近似の 12.000000 倍になる。一方、標本を次元の 10 倍に保つと、次元 10 で 0.112360、次元 100 で 0.111235 とほぼ同じで、効いているのは比である——次元を増やすなら標本も同じ割合で増やさねばならず、比が 1 に近づけば見積もりそのものが崩れる。シミュレーションは次元 5・標本 20 で 0.357921 と厳密な 0.357143 に乗り、近似の 0.25 ではなかった。分離子は、標本数と次元の比が大きいか、1 に近いかである。既刊との位置——論文733 は分散の形を置いて最小の動きを読んだ。ここはその形が届く範囲を切る。正直に言えば——説明変数の相関も外れ値も正則化も扱わず、境目の向こうは扱わなかった。

参考文献

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
2. Seber, G. A. F. (1977). *Linear Regression Analysis*. Wiley.
3. 山岸勇輝 (2026). 『「次元の呪い」は一つの呪いではない』[論文182]. Zenodo. 10.5281/zenodo.22104170
4. 山岸勇輝 (2026). 『誤差には最小が在るが、その最小の位置は模型の性質ではなく、手持ちのデータ数で動く』[論文733]. Zenodo. 10.5281/zenodo.23015055

— From thinking to syncing. —

山岸 勇輝 (Yuuki YAMAGISHI)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)