線形回帰の使用上の注意

正田 備也

masada@rikkyo.ac.jp

今日のお題

- 線形回帰における検定
  - ざっと流す程度。
- 正則化
  - Ridge回帰とLassoの話。重要。細かい理屈は抜きで。
- 主成分分析(PCA: principal component analysis)
  - 残り回数が以外に少ないことに気づき、あわてて追加。
  - 何週か後にもう一度話すかもしれません。

統計モデルの3つの使い途

https://www.slideshare.net/gshmueli/to-explain-to-predict-or-to-describe

Explanatory Model: ← e.g. 因果推論

test/quantify causal effect between constructs for “average” unit in population

Descriptive Model: ← e.g. 線形回帰における検定（今日の話題）

test/quantify distribution or correlation structure for measured “average” unit in population

Predictive Model: ← e.g. 深層学習

predict values for new/future individual units

統計モデルをdescriptive modelとして使う

線形回帰モデルの場合。

線形回帰における検定

- 説明変数の目的変数に対する影響が有意か調べたい
- 帰無仮説：特定のcoefficientまたはinterceptがゼロ
- 上記の帰無仮説が棄却できるか？という検定
  - 今日の参考資料：大阪大学「計量経済基礎」（谷崎先生）の講義資料
    - http://www2.econ.osaka-u.ac.jp/~tanizaki/class/2018/basic_econome/02.pdf

モデルの仮定（単回帰の場合）

$$y_i = b + ax_i + u_i$$

1. x_iは固定された値をとると仮定
2. すべてのiについて、誤差項u_iの期待値は0と仮定
3. すべてのiについて、誤差項u_iの分散はσ^2と仮定
4. すべてのi, jについて、誤差項u_iとu_jが無相関（$E[u_i u_j] = 0$）と仮定
5. すべてのiについて、誤差項u_iは平均0、分散σ^2の正規分布に従うと仮定
6. $N \to \infty$のとき、$\sum_{i=1}^{N}(x_i - \bar{x})^2 \to \infty$と仮定(ただし$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$)

多変量正規分布では、無相関⇔独立

- 例えば下記Webページを参照

  https://mathtrain.jp/uncorrelated

- 多変量正規分布では、無相関であることと、独立であることとは、同値
- よって、前のスライドの仮定から、誤差項 u_i は、すべて独立に、平均0、分散σ^2の正規分布に従うことが言える

単回帰の正規方程式

- 以下の連立方程式を解けば、傾きaの最小二乗推定量$\hat{a}$と、切片bの最小二乗推定量$\hat{b}$が求まる

$$\begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix} = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} {x_i}^2 \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix}$$

正規方程式を解くと…

$\hat{a}$も$\hat{b}$も、
正規分布に従う
独立な確率変数u_iの
線形和になっている

$$\hat{a} = a + \sum_{i=1}^{N} \omega_i u_i$$

$$\hat{b} = b - (\hat{a} - a)\bar{x} + \frac{1}{N}\sum_{i=1}^{N} u_i$$

ただし、 $\bar{x}$は$x_1, \ldots, x_N$の平均。 ω_iは下記のとおり。

$$\omega_i = \frac{(x_i - \bar{x})}{\sum_{j=1}^{N}(x_j - \bar{x})^2}$$

単回帰における検定(1/4)

- モデルの仮定より、以下のことが証明できる
  - 傾きaの最小二乗推定量$\hat{a}$は不偏推定量、つまり$E(\hat{a}) = a$
  - $\hat{a}$の分散は、$V(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
  - 切片bの最小二乗推定量$\hat{b}$は不偏推定量、つまり$E(\hat{b}) = b$
  - $\hat{b}$の分散は、$V(\hat{b}) = \frac{\sigma^2 \sum_{i=1}^{N} {x_i}^2}{N \sum_{i=1}^{N}(x_i - \bar{x})^2}$

単回帰における検定(2/4)

- $\hat{a}$も$\hat{b}$も、正規分布に従うことが示せる
  - 各u_iは独立に正規分布に従う
  - $\hat{a}$も$\hat{b}$も、 u_iの線形和で表される
  - 正規分布にしたがう確率変数の和は、正規分布に従う
    - http://joe.bayesnet.org/?p=4950
  - 以上のことから、 $\hat{a}$も$\hat{b}$も正規分布に従うことが言える

単回帰における検定(3/4)

- $\hat{a}$の分散の式も、$\hat{b}$の分散の式も、未知の値σを含んでいる
  - どうする？
- 以下のs^2が、誤差項の分散σ^2の不偏推定量となることが言える

$$s^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - \hat{b} - \hat{a}x_i\right)^2$$

- さらに$\frac{(N-2)s^2}{\sigma^2}$が自由度$N-2$のカイ自乗分布に従うことも言える

t検定のもとになっている命題(1/3)

命題1

母集団が正規分布$N(m, \sigma^2)$に従うとき、

- その標本$X_1, \ldots, X_N$から求めた標本平均$\bar{X}_N$と不偏標本分散$\bar{V}_N$は、独立である。
- $\bar{X}_N$は、正規分布$N(m, \frac{\sigma^2}{N})$に従う。　($\frac{\bar{X}_N - m}{\sigma}\sqrt{N}$は$N(0,1)$に従う。)
- $\frac{N-1}{\sigma^2}\bar{V}_N$は、自由度$N-1$のカイ二乗分布に従う。

t検定のもとになっている命題(2/3)

命題2

2つの確率変数Z, Yが独立で、Zが標準正規分布$N(0,1)$に従い、Yが自由度nのカイ二乗分布に従うならば、$T \equiv \frac{Z}{\sqrt{Y/n}}$は、自由度$n$の$t$分布に従う。

t検定のもとになっている命題(3/3)

系

母集団が正規分布$N(m, \sigma^2)$に従うとき、その標本$X_1, \ldots, X_N$から求めた標本平均$\bar{X}_N$と不偏標本分散$\bar{V}_N$とについて、$(\bar{X}_N - m)\sqrt{\frac{N}{\bar{V}_N}}$は自由度$N-1$のt分布に従う。

- ポイント：標本から求めることのできる値（標本平均$\bar{X}_N$と不偏標本分散$\bar{V}_N$）だけを使って、母集団の平均mという未知の量に関する定量的な推定が可能になっている。

t検定の例

- 例：自由度20のt分布の両側5%点は2.0860

$$-2.0860 \leq (\bar{X}_N - m)\sqrt{\frac{N}{\bar{V}_N}} \leq 2.0860$$

  - 帰無仮説が$m = 0$のとき、 $\bar{X}_N\sqrt{\frac{N}{\bar{V}_N}} < -2.0860$または$\bar{X}_N\sqrt{\frac{N}{\bar{V}_N}} > 2.0860$ならば、有意水準5%において帰無仮説を棄却する。

単回帰における検定(4/4)

- 傾きaの最小二乗推定量$\hat{a}$については、$(\hat{a}-a)/s_{\hat{a}}$が自由度$N-2$のt分布に従うことが言える。ただし

$$s_{\hat{a}} = \frac{s}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}}$$

- 切片bの最小二乗推定量$\hat{b}$については、$(\hat{b}-b)/s_{\hat{b}}$が自由度$N-2$のt分布に従うことが言える。ただし

$$s_{\hat{b}} = s\sqrt{\frac{\sum_{i=1}^{N}x_i{}^2}{N\sum_{i=1}^{N}(x_i-\bar{x})^2}}$$

単回帰における検定をどう行うか

- 単回帰における最小二乗推定量$\hat{a}$と$\hat{b}$についても、以上の理屈により、t検定をおこなうことができる

- PythonならstatsmodelsのOLSを使えばよい
  - 実行結果に、ちゃんと検定統計量が表示される。

https://www.statsmodels.org/stable/regression.html

https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.fit.html

統計モデルをexplanatory modelとして使う

統計モデルを説明に使う（≒因果推論）

- （この話題は、この授業ではカバーしません）

「統計的因果効果推定の枠組みを通じて、得られたデータからまずは「ロバストな因果効果の推定」を行い、背後の共変量・交絡要因の影響を排除した「シンプルなメカニズムの理解」を行い”どのような介入が有効か”を示すことが解析実務では有用です。このような一連の解析はシステムに組み込まれていく汎用的機械学習モデルにはできない”データサイエンティストの腕の見せ所”になるのではないでしょうか？」

（星野崇宏「統計的因果効果の基礎」『調査観察データの統計科学-因果推論・選択バイアス・データ融合-』85-86頁）

統計モデルをpredictive modelとして使う

統計モデルを予測に使う（≒機械学習）

- 予測性能が良いのであれば、何が起きても構わない
  - 線形回帰の場合、いくつかの説明変数のp値が大きな値になろうが、予測性能が良いのであれば何の問題もない。
  - 多重共線性も問題にならない。
- モデルが正しいかどうかも、問題にならない
- 特に深層学習の世界では、予測性能の向上だけ考える
  - モデルの解釈性の無さも問題にならない。

正則化

Ridge回帰とLasso

[ESLII] Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. Chapter 3.

変数を選択することの問題点

- 説明変数が多いとき、例えばESLII, Sec. 3.2.1のExample: Prostate Cancerのように検定の結果を使ってnon-significantな変数を削ったりする
  - 同書3.3節には、もっと良い変数選択の方法が書かれてある。
- しかし、変数を選択するというのは、離散的な手続き
  - 予測対象となるデータ集合によって、性能に段差がつくことがある
- そこで、shrinkage methodsと呼ばれる連続的な手続きを採る

(Z score) = (Coefficient) / (Std. Error) による変数選択

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

[ESLII, p.50]

変数選択を連続的にする

- ある説明変数を使わない＝その説明変数の係数をゼロにする

ON/OFFではなく、連続的にすると・・・

- ある説明変数を使わない＝その説明変数の係数がゼロに近くなるようにする

Ridge回帰

- 通常の最小二乗法とは、最小化すべき関数が少し違う

$$l(\boldsymbol{a}) = \sum_{i=1}^{N}\left(y_i - a_0 - \sum_{j=1}^{d} a_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{d} {a_j}^2$$

- 説明変数の係数（切片は含まない）の2乗和も同時に最小化
  - 係数が全体的に$\mathbf{0}$のほうに近寄った値になる。
  - λでその強さをコントロールする。
  - λは交差検証などで決定する。（最小化の計算によっては決定できない。）

Lasso

- 通常の最小二乗法とは、最小化すべき関数が少し違う

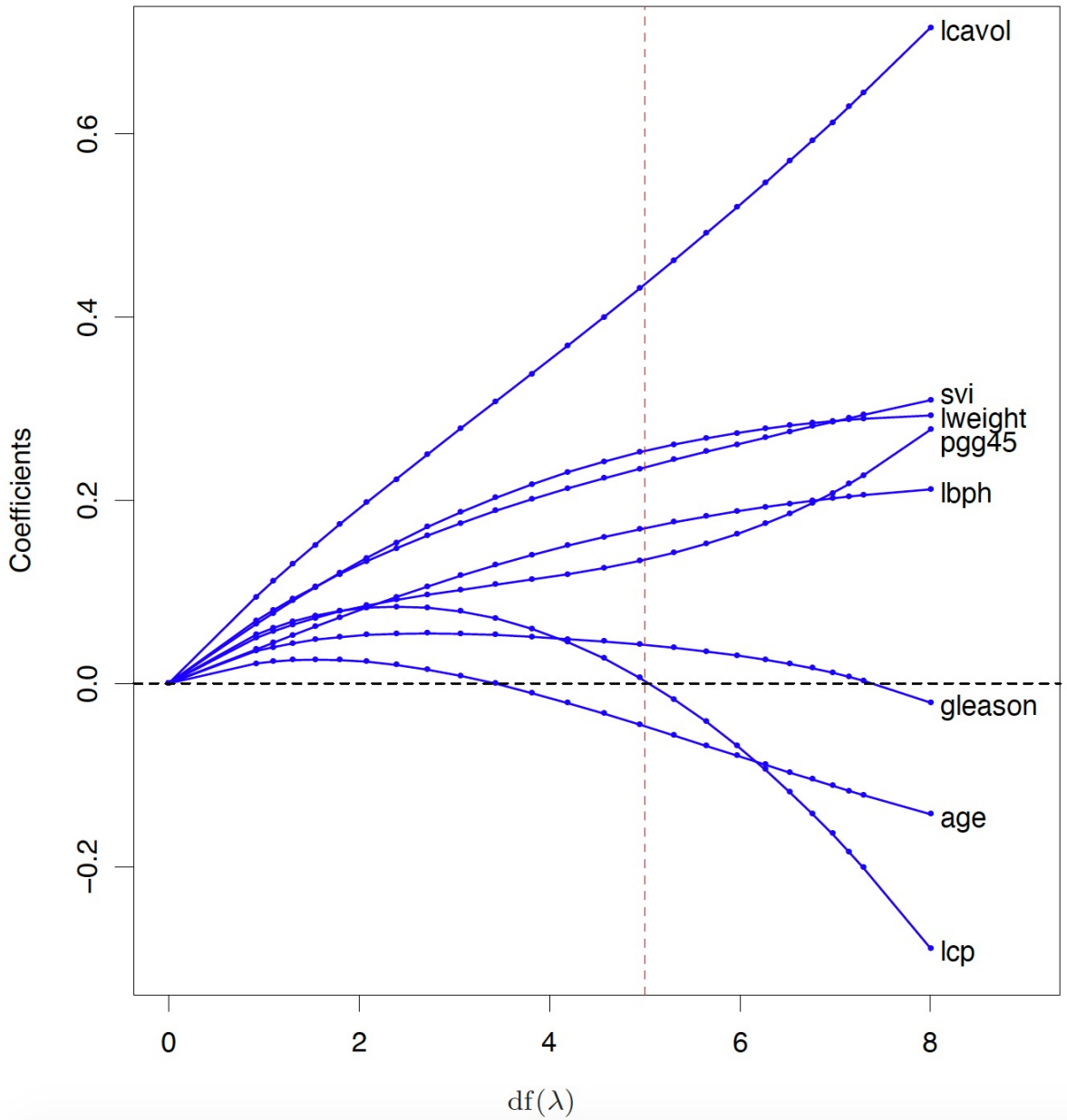
$$l(\boldsymbol{a}) = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - a_0 - \sum_{j=1}^{d} a_j x_{i,j}\right)^2 + \lambda \sum_{j=1}^{d} |a_j|$$

- 説明変数の係数（切片は含まない）の絶対値和も同時に最小化
  - 係数が全体的に$\mathbf{0}$のほうに近寄った値になる。
  - λでその強さをコントロールする。
  - λは交差検証などで決定する。（最小化の計算によっては決定できない。）

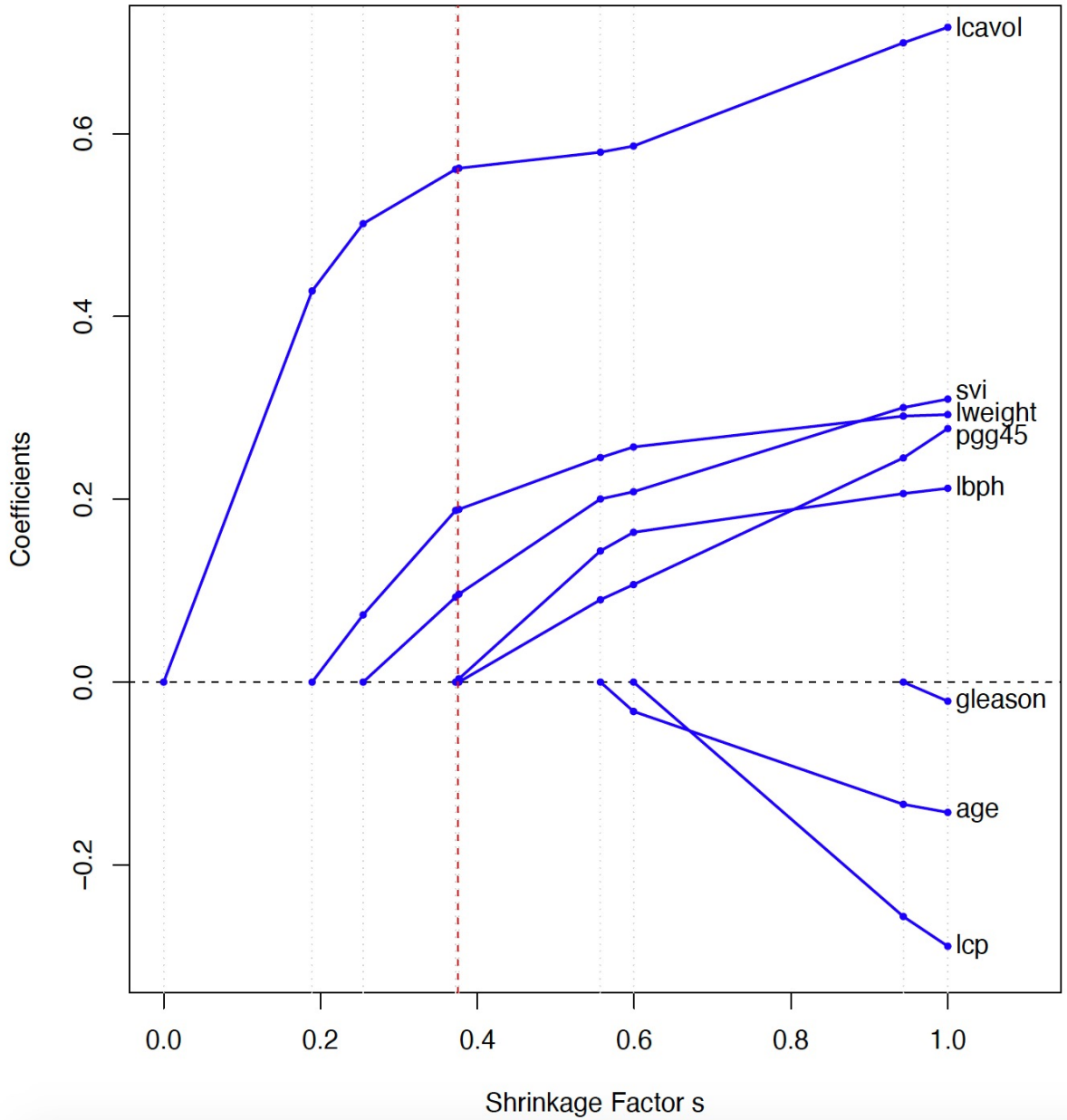
Ridge回帰とLassoの違い

- λを大きくして係数をゼロに近づける項の効きを強くすると…

- Ridge回帰では、すべての係数が全体的にゼロに近寄る
- Lassoでは、係数がひとつずつ、ほぼゼロの値になっていく



Ridge回帰 [ESLII, p.65]



Lasso [ESLII, p.70]

なぜ切片が正則化に含まれないのか(1/2)

- 例えばy_iに一斉に1を足して、推定をやり直したとすると…
  - 通常の最小二乗法：切片の推定値だけが変化する
  - Ridge回帰やLasso：切片を正則化に含ませると答え全体が変わる

- つまり、推定計算がy_iの原点をどこに採るかに依存してしまう
- よって、切片は、通常、正則化には含ませない

なぜ切片が正則化に含まれないのか(2/2)

- しかし、切片を含まない正則化を使った推定は、中心化されたデータを使うことで初めから切片を無視した正則化を使った推定と、全く同じ答えを与える
- また、前者の方法で得られる切片の推定値については、後者の方法で得られる他の係数の推定値を使って表現できる（下式）

$$\hat{a}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - \sum_{j=1}^{d} \hat{a}_j \left(\frac{1}{N}\sum_{i=1}^{N} x_{i,j}\right)$$

It has to be emphasized that in practice, the bias parameter θ_0 is left out from the norm in the regularization term; penalization of the bias would make the procedure dependent on the origin chosen for y. Indeed, it is easily checked that adding a constant term to each one of the output values, y_n, in the cost function would not result in just a shift of the predictions by the same constant, if the bias term is included in the norm. Hence, usually, ridge regression is formulated as

$$\text{minimize} \quad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \left(y_n - \theta_0 - \sum_{i=1}^{l} \theta_i x_{ni} \right)^2 + \lambda \sum_{i=1}^{l} |\theta_i|^2. \tag{3.43}$$

It turns out (Problem 3.11) that minimizing Eq. (3.43) with respect to θ_i, $i = 0, 1, \ldots, l$, is equivalent to minimizing Eq. (3.41) using *centered* data and neglecting the intercept. That is, one solves the task

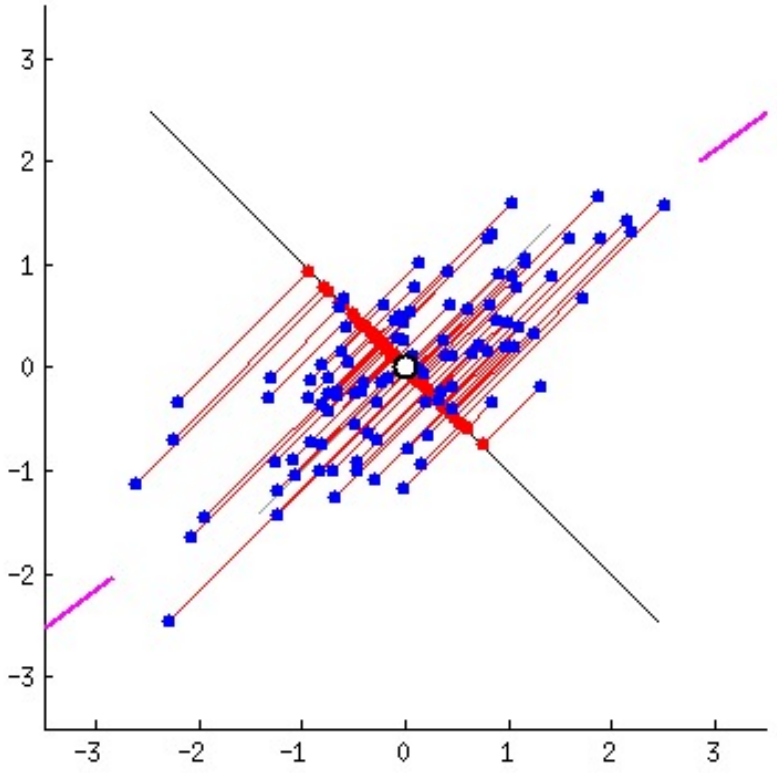
$$\text{minimize} \quad L(\boldsymbol{\theta}, \lambda) = \sum_{n=1}^{N} \left((y_n - \bar{y}) - \sum_{i=1}^{l} \theta_i (x_{ni} - \bar{x}_i) \right)^2 + \lambda \sum_{i=1}^{l} |\theta_i|^2, \tag{3.44}$$

and the estimate of θ_0 in Eq. (3.43) is given in terms of the obtained estimates $\hat{\theta}_i$, i.e.,

$$\hat{\theta}_0 = \bar{y} - \sum_{i=1}^{l} \hat{\theta}_i \bar{x}_i,$$

主成分分析

PCAのイメージ



https://medium.com/@shiva1gandluri/principal-component-analysis-pca-in-machine-learning-c3f239249b73

PCAによる次元削減のイメージ

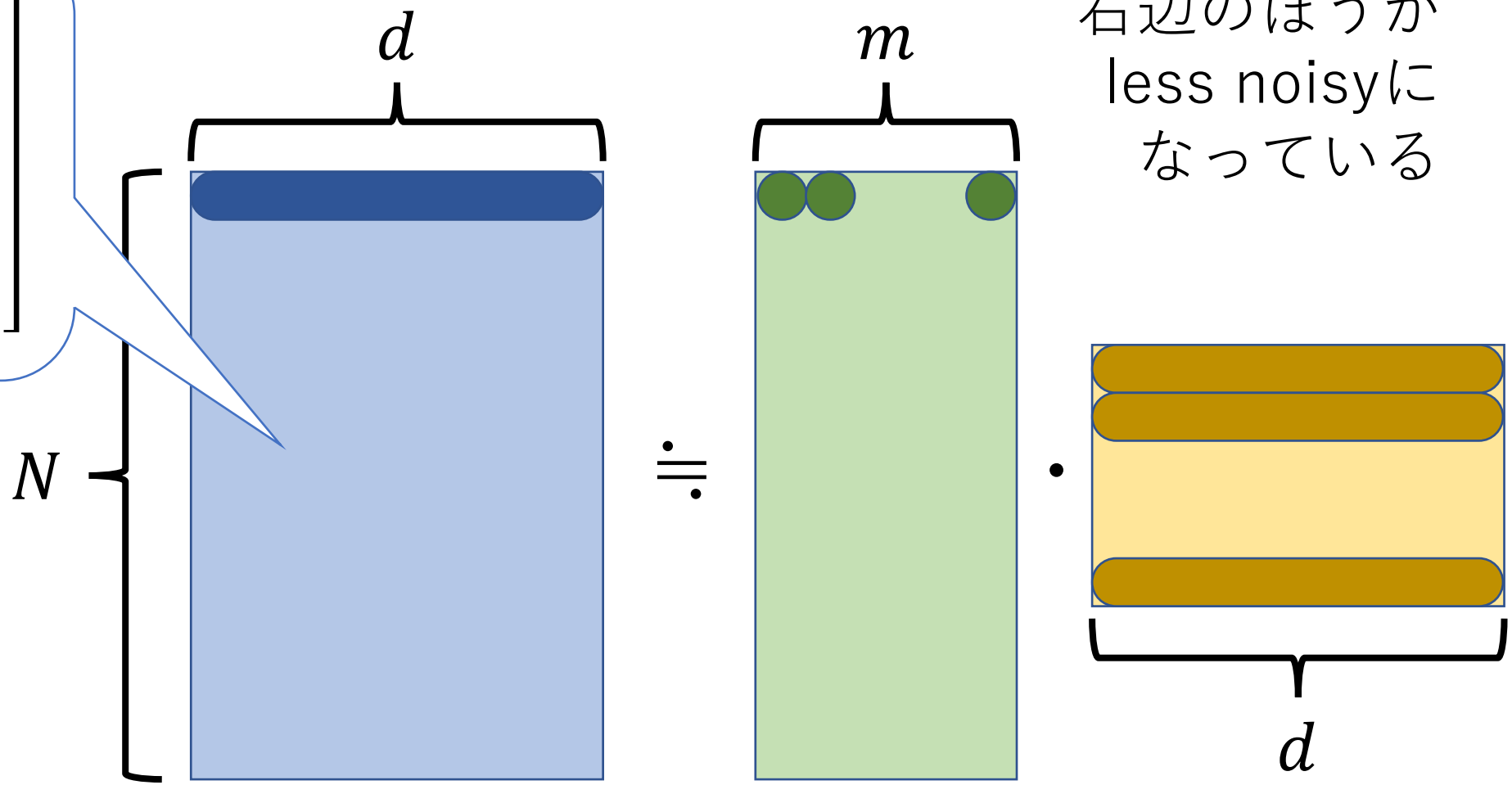
1. データをあらかじめ中心化しておく（平均を引いておく）
   - スケーリングもしておく（標準偏差で割っておく）
2. 原点を通る直線のうちデータに一番「近い」ものを見つける
3. その直線に垂直な平面へ、データを射影する
4. 2.に戻って、次元がひとつ落ちた空間で同じことを繰り返す

PCAの雰囲気

- データがどの方向に大きく散らばっているかを知ることは有益
- そこで・・・
- データがより大きく散らばっている向きを順番に見つけていく
  - 後から見つけた向きは、先に見つけた向きに直交しているようにする

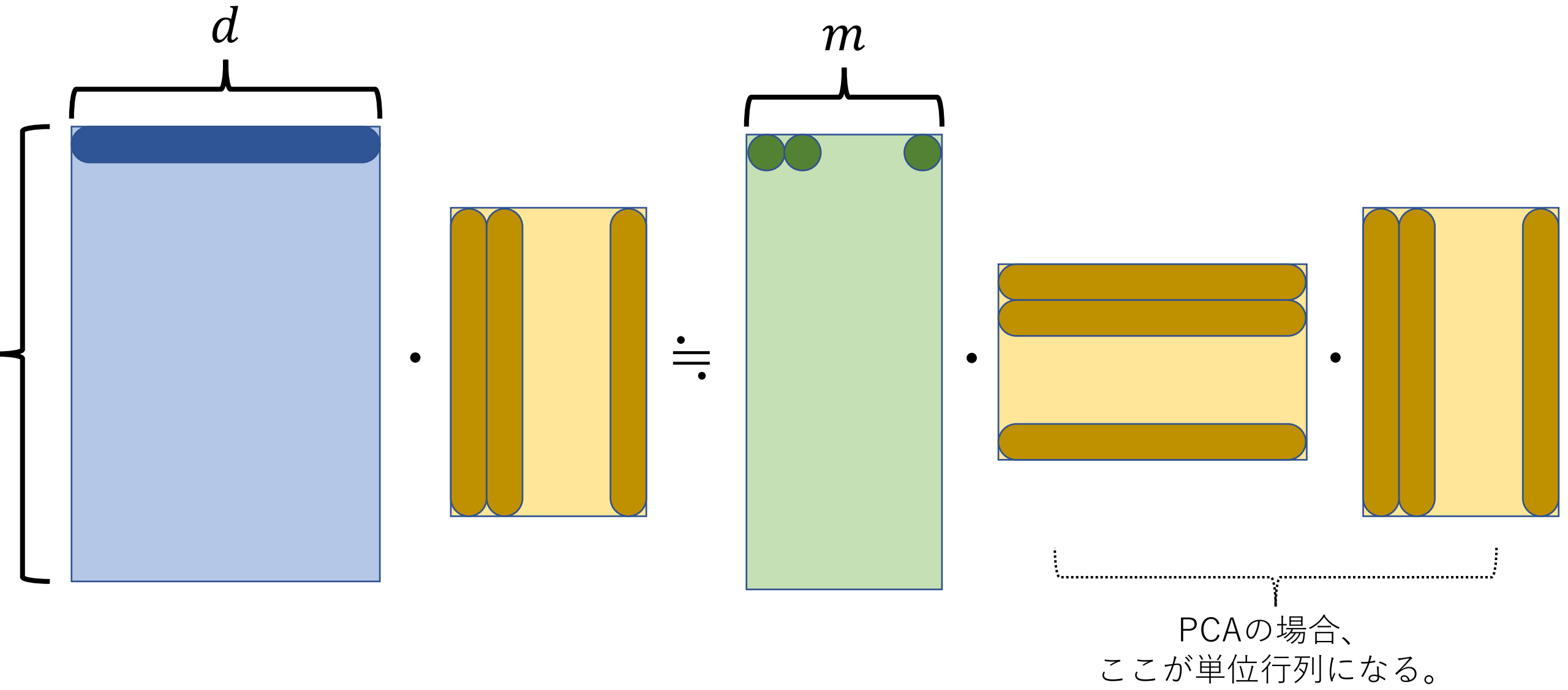
PCAで計画行列をless noisyにする

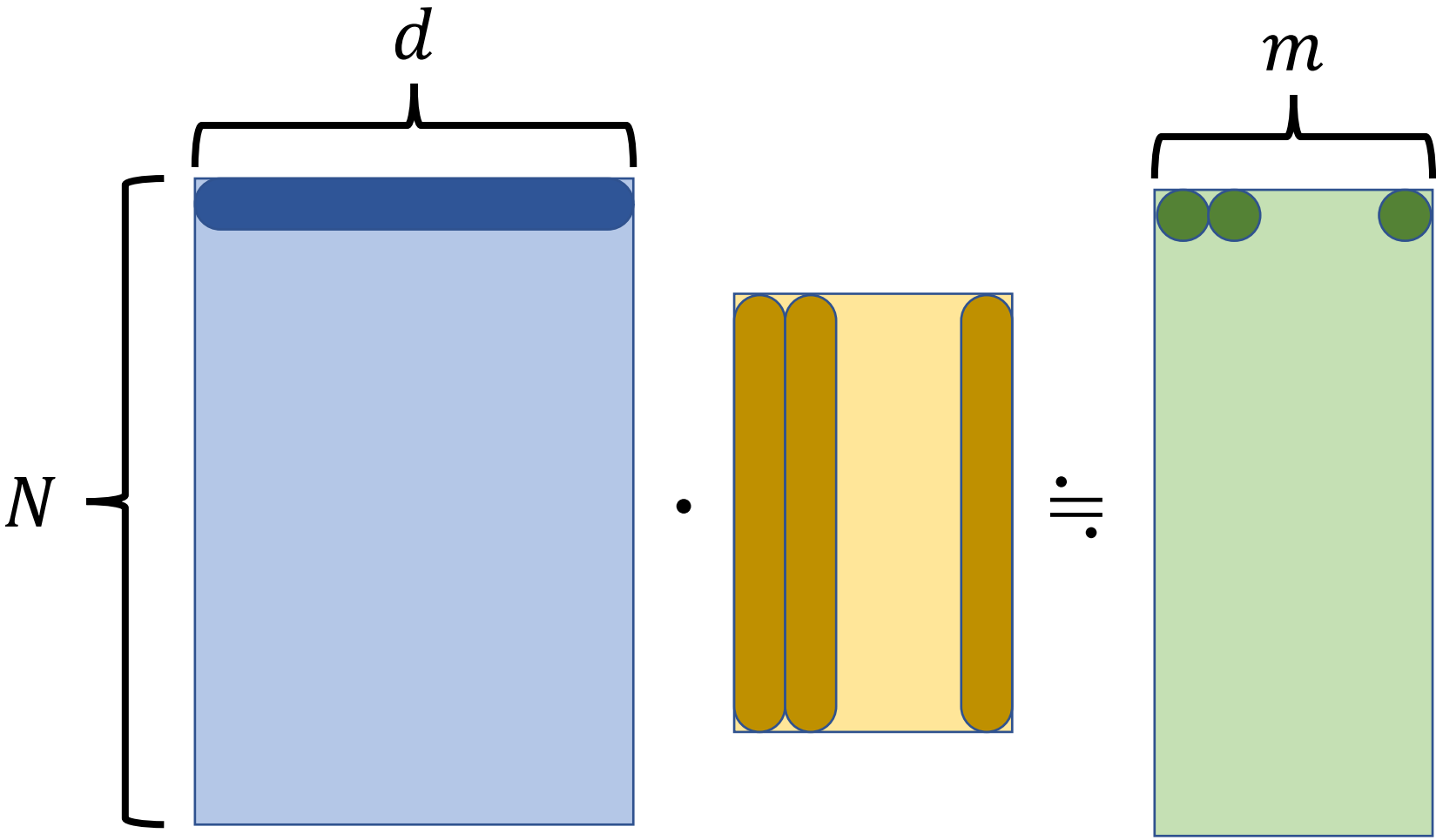
$$\begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,d} \end{bmatrix}$$



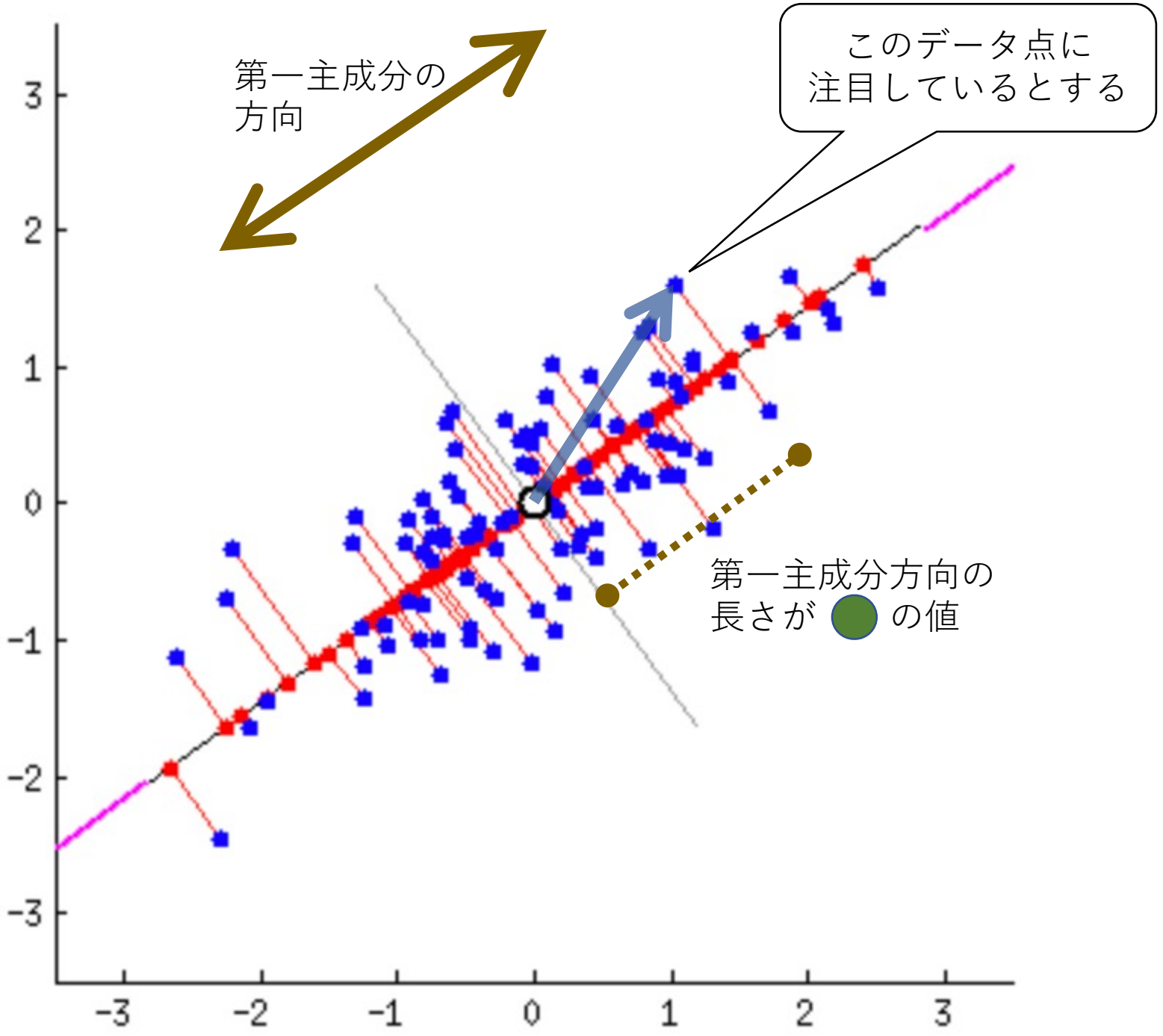
右辺のほうが
less noisyに
なっている

左辺は
元々の
計画行列


d
m
PCAの場合、
ここが単位行列になる。



- 元のd次元空間のなかに、m本の直交する座標軸を取り…
- その軸を使って、元のデータ点の座標値を決め直す


第一主成分の
方向
このデータ点に
注目しているとする
第一主成分方向の
長さが
の値
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

課題

Solubility（溶解性）データセット(1/2)

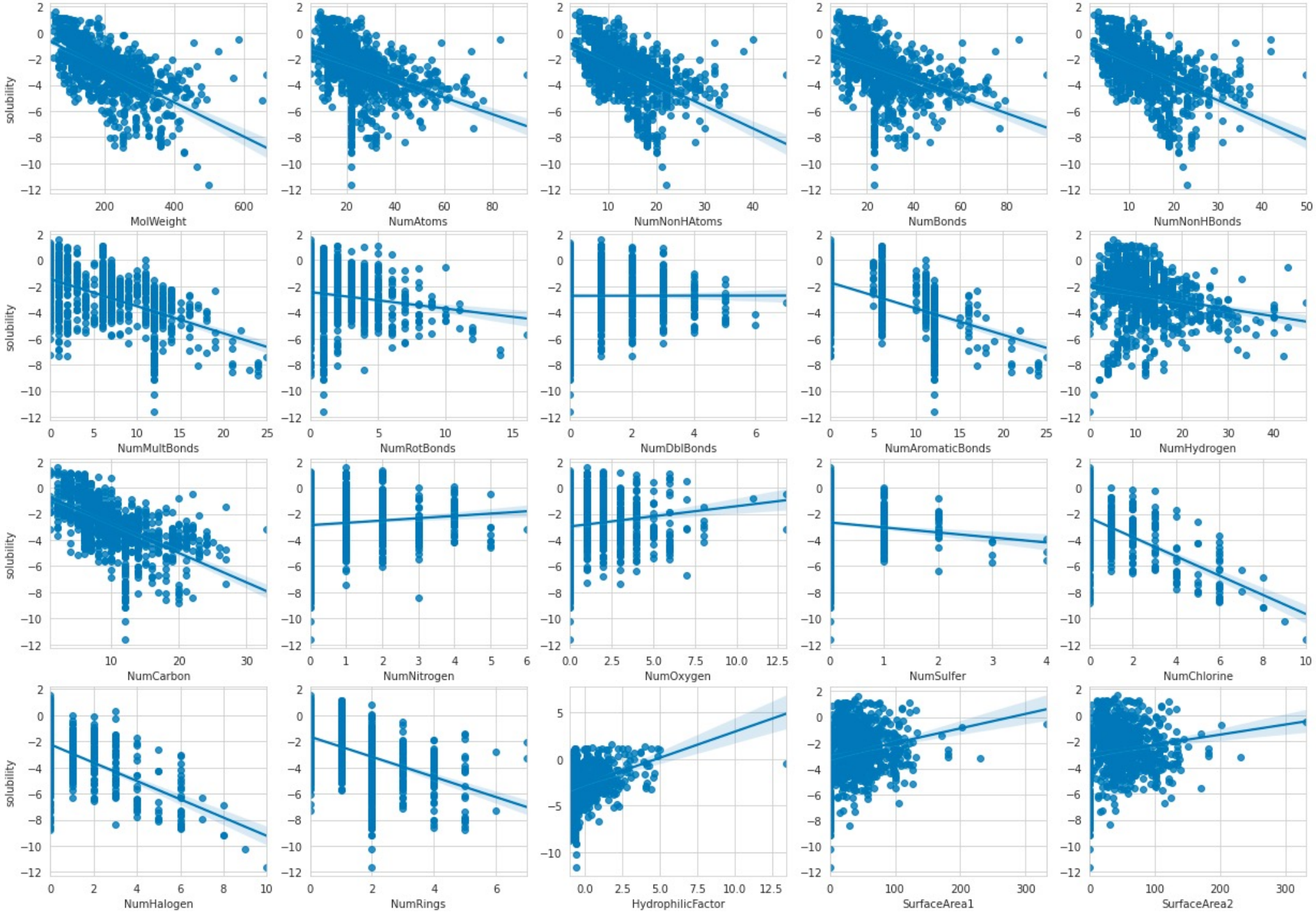
- Max Kuhn and Kjell Johnson. Applied Predictive Modeling. Springer, 2013. に出てくるデータセット (Section 6.1)
- 説明変数は下記の228個
  - Two hundred and eight binary “fingerprints” that indicate the presence or absence of a particular chemical substructure.
  - Sixteen count descriptors, such as the number of bonds or the number of bromine atoms.
  - Four continuous descriptors, such as molecular weight or surface area.
- 目的変数はlog solubility
  - 範囲は－11.6 to 1.6、平均は－2.7

Solubility（溶解性）データセット(2/2)

- 訓練データ951件
  - これをさらに訓練データと検証データに分割してモデル選択。
  - 交差検証の方法は自由。
- テストデータ316件
  - 最後に一回、手法の最終的な評価に使うだけ。

課題8

- solubilityデータセットのテストデータに対して、できるだけ予測性能の良いモデルを見つけよう
  - Ridge回帰やLassoを使ってもいいです。
  - 特徴量はどのように加工してもいいです。


solubility
MolWeight
NumAtoms
NumNonHAtoms
NumBonds
NumNonHBonds
NumMultBonds
NumRotBonds
NumDblBonds
NumAromaticBonds
NumHydrogen
NumCarbon
NumNitrogen
NumOxygen
NumSulfer
NumChlorine
NumHalogen
NumRings
HydrophilicFactor
SurfaceArea1
SurfaceArea2