

# 自己紹介

- 氏名：酒井 彰 (Sho Sakai)
- 卒業：鹿児島大学 数理情報科学科
- 所属：筑波大学大学院 数学学位プログラム 博士後期課程 1 年
- Podcast：Data Science LG: 学び合う統計とデータサイエンス  
ゲスト：Imperial、Oxford、大阪大学、立正大学、神戸大学、滋賀大学、総合研究大学院大学、東京大学、東京理科大学、筑波大学、早稲田大学、関西大学、日産自動車株式会社、dip株式会社、LINEヤフー株式会社、Sansan株式会社、みずほ第一フィナンシャルテクノロジー
- 勉強会の運営：ベイズ深層学習
- 骨髄バンクユースアンバサダー
  - ✓ 記事：骨髄ドナー（骨髄提供）体験レポートと関連情報
  - ✓ 語部公演：若い世代に骨髄移植を届ける：講演を通じて伝えた想い
- 献血
  - ✓ 70回目レポート：献血回数RTA 銀色有功章獲得！（70回到達）レポート + 献血のススメ

HP



in



# 高次元データにおける主成分回帰の推論理論

## 係数の仮説検定と予測誤差最小化の展望

酒井 彰 (Sho Sakai)<sup>1</sup>,  
矢田 和善 (Kazuyoshi Yata)<sup>1</sup>,  
青嶋 誠 (Makoto Aoshima)<sup>1</sup>  
<sup>1</sup> 筑波大学

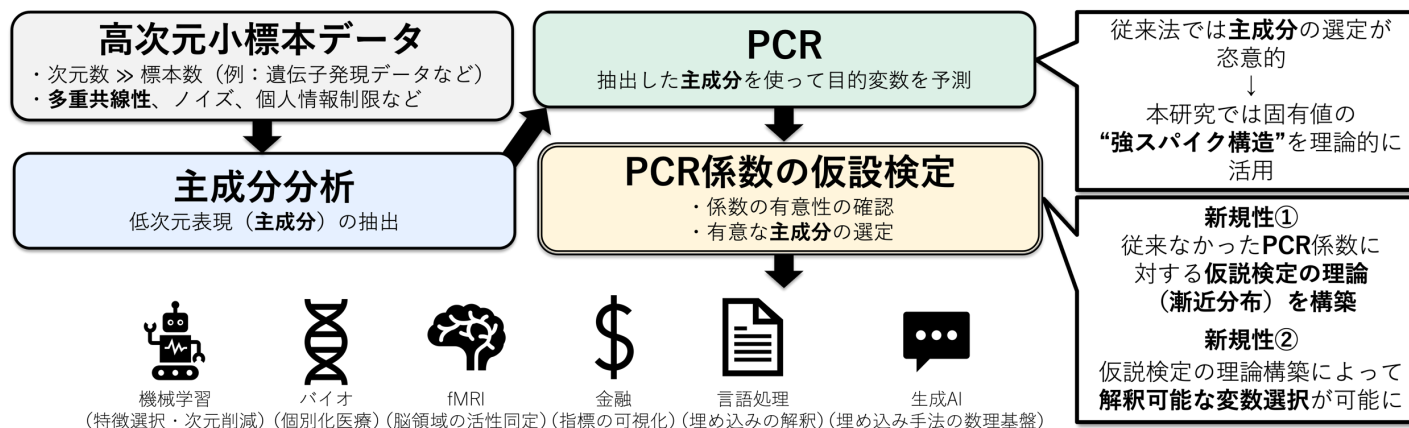
# 要点

## 提案：

- 強スパイク固有値 (**SSE**) モデルの下で、**主成分回帰 (PCR)** の各主成分係数に対する**仮説検定**を構築
- 回帰において有意な主成分が理論的にわかるだけでなく、射影 (次元削減) は主成分分析 (**PCA**) のままとし、評価と選択の段階で目的変数  $y$  を明示的に取り入れることで、従来の「**PCR は  $y$  を考慮していない**」という弱点も克服

## 性能：

- シミュレーションでサイズ維持 + 検出力を確認
- 実データ (ゲノムクス) では**予測に本当に効く主成分の特定とMSE同等**を確認



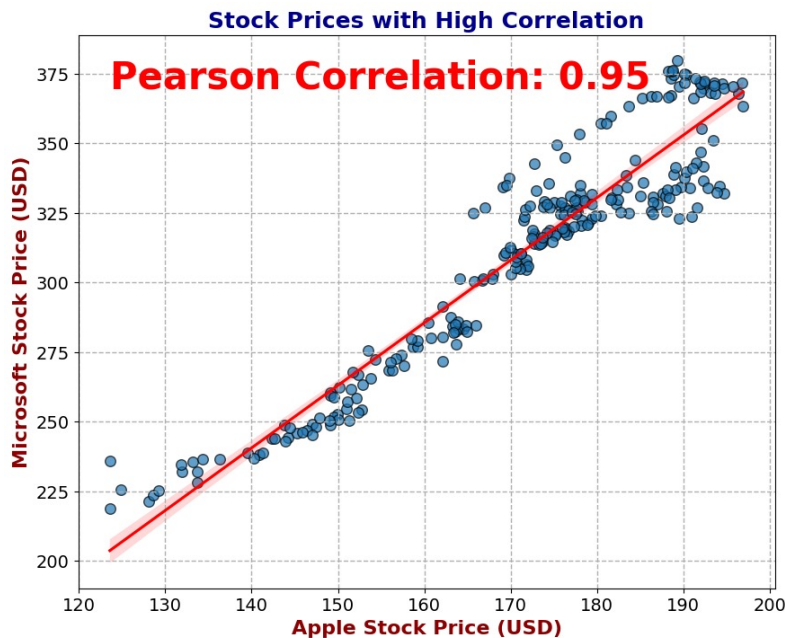
# 動機

線形回帰モデル： $y = X^T \beta + \varepsilon$

$$\hat{\beta}_{OLS} = (XX^T)^{-1}Xy$$

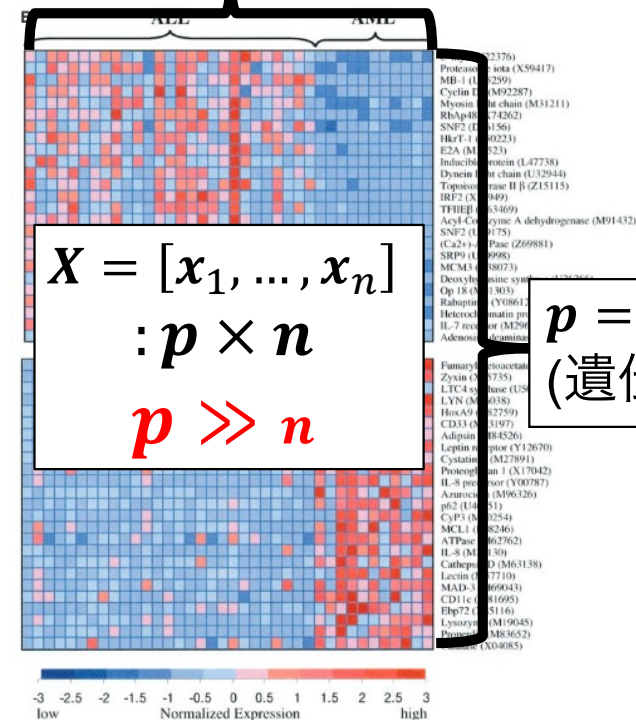
- 分散が大きくなる・解釈が困難
- $(XX^T)^{-1}$  は存在しない

## 多重共線性



## 高次元小標本データ

$n = 72$  (患者数)



遺伝子発現データ,  
Golub et al., 1999, Science

# 主成分回帰 (Principal Component Regression; PCR)

Original model:  $y = X^T \beta + \varepsilon$

## 1. PCAステップ: $Z = \Lambda^{-1/2} H^T (X - \mu)$

仮定:  $E(x_i) = \mu$  and  $\text{Cov}(x_i) = \sum_{j=1}^p \lambda_j \mathbf{h}_j \mathbf{h}_j^T$ ,  $\lambda_1 \geq \dots \geq \lambda_p \geq 0$

ここで  $\Lambda^{-1/2} = \text{diag}(\lambda_1^{-1/2}, \dots, \lambda_m^{-1/2})$ ,  $\lambda_1 \geq \dots \geq \lambda_m > 0$ ,

$H = [\mathbf{h}_1, \dots, \mathbf{h}_m]$ ,  $H^T H = I_m$

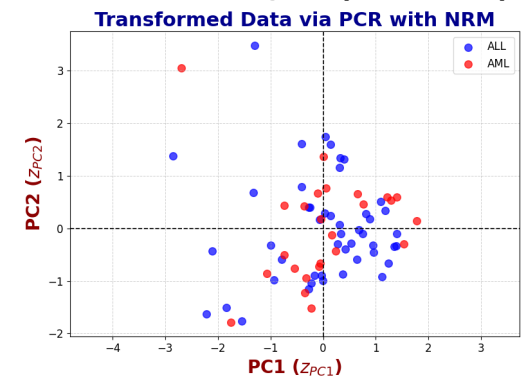
## 2. 回帰ステップ (PCR) : $y = Z^T \gamma + \varepsilon = \gamma_1 z_{PC1} + \dots + \gamma_m z_{PCm} + \varepsilon$

- 安定性:  $\text{Cov}(z_{i,PCj}, z_{i,PCj'}) = 0$ ,  $j \neq j'$
- 計算可能:  $m \leq \text{rank}(X)$

### 課題:

- どの PC 係数  $\gamma_k$  が  $y$  に有意に寄与するかを理論的に保証することができない
- モデル構築に目的変数  $y$  が考慮されていない

PCAの射影 ( $m = 2$ )



# 問題設定

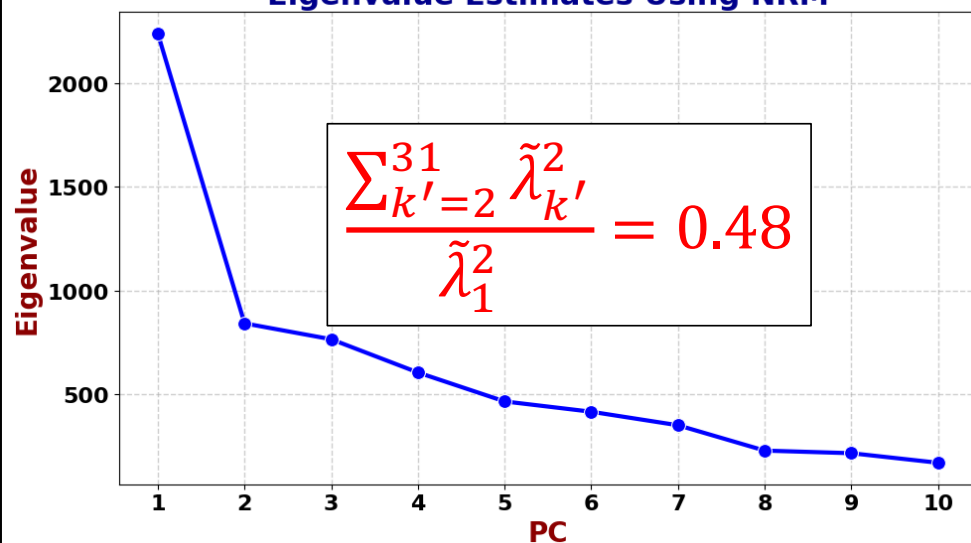
仮説 :  $H_0: \gamma_k = 0$  vs.  $H_1: \gamma_k \neq 0$

仮定 : **強スパイク固有値 (Strongly Spiked Eigenvalue; SSE) モデル**  
(Aoshima & Yata, 2018, *Statistica Sinica*) の一つ

$$\frac{\sum_{k'=m+1}^p \lambda_{k'}^2}{\lambda_m^2} = o(1) \text{ and } \frac{\lambda_{k'}}{\lambda_k} = o(1) \text{ for all } k < k' \leq m, \text{ as } p \rightarrow \infty.$$

上位固有値のプロット

Eigenvalue Estimates Using NRM



ノイズ掃き出し法による推定量  $\tilde{\lambda}_j$  (Yata & Aoshima, 2012, *JMA*) を使用

T-cellデータ ( $n = 33$ ,  $p = 12,625$ ;  
Chiaretti et al., 2004, *Blood*)

# 主結果 1

**定理 1 (本提案)** : 強スパイク固有値モデルの下で適切な正則条件を満たすとき、  
 $k = 1, \dots, m$  について  $p, n \rightarrow \infty$  で

$$Z_k = \sqrt{\frac{n-1}{M_k}} \hat{\gamma}_k \Rightarrow N \left( \sqrt{\frac{n-1}{M_k}} \gamma_k, 1 \right)$$

ここで  $M_k = \frac{\gamma_k^2}{2} + \sum_{k' < k}^m \gamma_{k'}^2 + \sigma^2$ ,  $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$ .

$$\cdot \hat{\gamma}_k = \frac{1}{\sqrt{n-1}} \hat{\mathbf{u}}_k^T (\mathbf{y} - \bar{y} \mathbf{1}_n), \bar{y} = \frac{\mathbf{y}^T \mathbf{1}_n}{n}$$

$$\cdot \frac{\mathbf{x} - \bar{\mathbf{x}}}{\sqrt{n-1}} = \sum_{s=1}^{n-1} \hat{\lambda}_s^{-2} \hat{\mathbf{h}}_s \hat{\mathbf{u}}_s^T, \bar{\mathbf{X}} = [\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}] \text{ with } \bar{\mathbf{x}} = \frac{\sum_{i=1}^n \mathbf{x}_i}{n}$$

**補題 (本提案)** : 強スパイク固有値モデルの下、適切な正則条件のもとで直交  
 化主成分スコア  $\dot{\mathbf{u}}_{ok}$  に対して、 $\hat{\mathbf{u}}_k^T \dot{\mathbf{u}}_{ok} = 1 + o_P(n^{-1})$  が成り立つ。ここで  $p, n \rightarrow \infty$ 、  
 $k = 1, \dots, m$ 。

# 主結果2

**定理2 (本提案)** : 強スパイク固有値モデルの下で適切な正則条件を満たすとき、 $k = 1, \dots, m$  について  $p, n \rightarrow \infty$  で

$$P\left(\gamma_k \in \left[\hat{\gamma}_k - z^* \sqrt{\frac{\hat{M}_k}{n-1}}, \hat{\gamma}_k + z^* \sqrt{\frac{\hat{M}_k}{n-1}}\right]\right) = P\left(\sqrt{\frac{n-1}{\hat{M}_k}}(\hat{\gamma}_k - \gamma_k) \in [-z^*, z^*]\right) = 1 - \alpha + o(1)$$

ここで  $P(N(0, 1) \geq z^*) = \alpha/2$ ,  $\alpha \in (0, 1)$ .

$$\cdot \hat{M}_k = \frac{\hat{\gamma}_k^2}{2} + \sum_{k' < k}^m \hat{\gamma}_{k'}^2 + \hat{\sigma}^2, \hat{\sigma}^2 = \frac{\|\mathbf{y} - \hat{\gamma}_0 \mathbf{1}_n - \hat{\mathbf{Z}}^T \hat{\boldsymbol{\gamma}}\|^2}{n - m - 1}$$

$$\cdot \hat{\gamma}_0 = \bar{y}, \hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \dots, \hat{\gamma}_m)^T = (\hat{\mathbf{Z}} \hat{\mathbf{Z}}^T)^{-1} \hat{\mathbf{Z}} \mathbf{y} = \left( \frac{1}{\sqrt{n-1}} \hat{\mathbf{u}}_1^T (\mathbf{y} - \bar{y} \mathbf{1}_n), \dots, \frac{1}{\sqrt{n-1}} \hat{\mathbf{u}}_m^T (\mathbf{y} - \bar{y} \mathbf{1}_n) \right)^T$$

$$\cdot \hat{\mathbf{Z}} = [\hat{\mathbf{z}}_{\text{PC}1}, \dots, \hat{\mathbf{z}}_{\text{PC}m}]^T, \hat{\mathbf{z}}_{\text{PC}i} = \hat{\lambda}_i^{-1/2} (\mathbf{X} - \bar{\mathbf{X}})^T \hat{\mathbf{h}}_i$$

$$\mathbf{H}_0: \gamma_k = 0 \text{ を棄却} \Leftrightarrow \sqrt{\frac{n-1}{\hat{M}_k}} \hat{\gamma}_k \notin [-z^*, z^*]$$

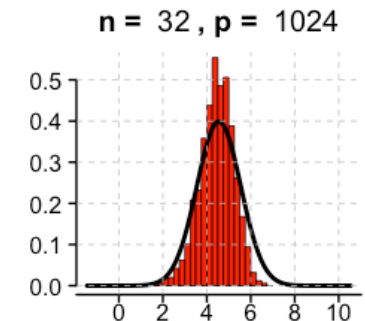
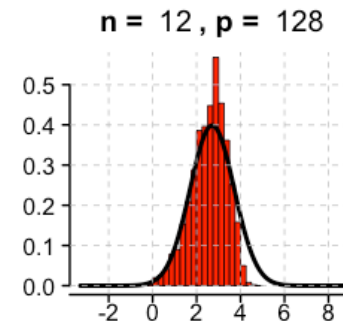
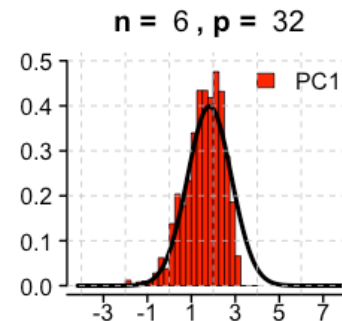
# シミュレーション1

設定：

- $E(\mathbf{x}_i) = \mathbf{0}$ ,  $\text{Cov}(\mathbf{x}_i) = \text{diag}(\mathbf{p}, \mathbf{p}^{3/5}, 1, \dots, 1)$
- $n = \lfloor \mathbf{p}^{1/2} \rfloor$ ,  $\mathbf{p} = 2^s$ ,  $s = 5, \dots, 10$ .
- $\mathbf{y} = \mathbf{2} \times \mathbf{z}_1 + \mathbf{0} \times \mathbf{z}_2 + \boldsymbol{\varepsilon}$ ,  $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, 2I)$

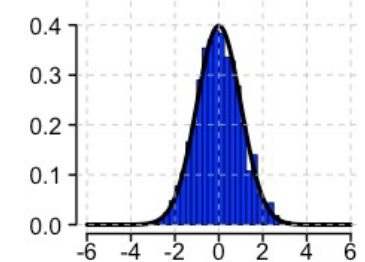
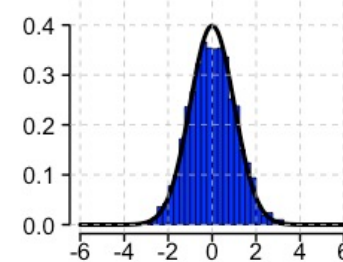
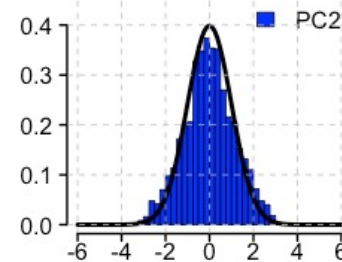
**PC1** ( $H_1: \gamma_1 \neq 0$ ):

$$\sqrt{\frac{n-1}{\hat{M}_1}} \hat{\gamma}_1 \Rightarrow N(\sqrt{n-1}, 1)$$



**PC2** ( $H_0: \gamma_2 = 0$ ):

$$\sqrt{\frac{n-1}{\hat{M}_2}} \hat{\gamma}_2 \Rightarrow N(0, 1)$$



# シミュレーション2

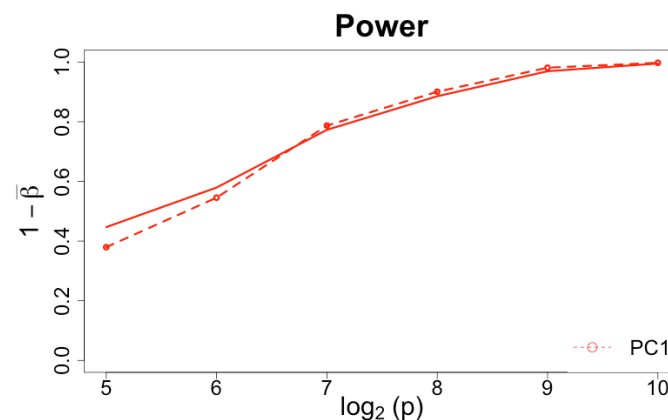
- ・ 有意水準 :  $\alpha = 0.05$
- ・  $y = 2 \times z_1 + 0 \times z_2 + \varepsilon$

## PC1 ( $H_1: \gamma_1 \neq 0$ )

$$P_r = \begin{cases} 1 & \Leftrightarrow H_0: \gamma_1 = 0 \text{ が棄却された} \\ 0 & \Leftrightarrow H_0: \gamma_1 = 0 \text{ が棄却されなかった} \end{cases}$$

$$\text{Power: } 1 - \bar{\beta} = 1 - \frac{1}{2000} \sum_{r=1}^{2000} P_r \approx 1$$

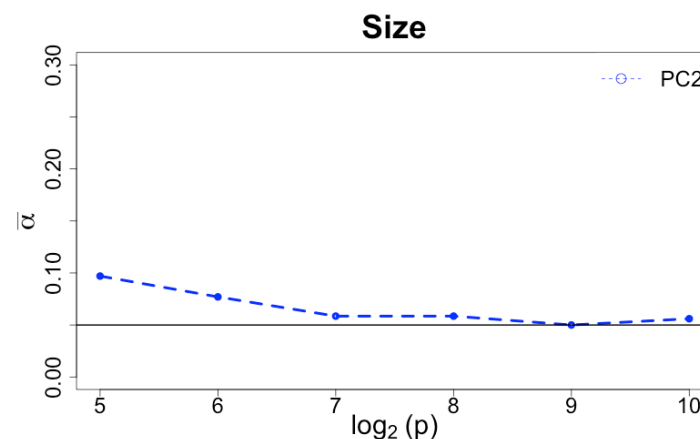
Solid line: Asymptotic Power



## PC2 ( $H_0: \gamma_2 = 0$ )

$$P_r = \begin{cases} 1 & \Leftrightarrow H_0: \gamma_2 = 0 \text{ が棄却された} \\ 0 & \Leftrightarrow H_0: \gamma_2 = 0 \text{ が棄却されなかった} \end{cases}$$

$$\text{Size: } \bar{\alpha} = \frac{1}{2000} \sum_{r=1}^{2000} P_r \approx 0.05$$



# 実データ解析 1

データセット：Genomics of Drug Sensitivity in Cancer (GDSC), Garnett et al., 2012, *Nature*

・ サンプル：がん細胞

$n = 712$

・ 説明変数：遺伝子発現・DNAメチル化・変異・CNV 等

$p = 17,849$

・ 目的変数：薬剤 BMS-536924 の IC50

① データ分割：80% 訓練データ, 20% テストデータ (20回)

② PCA & 仮設検定 ( $\alpha = 0.05$ ):

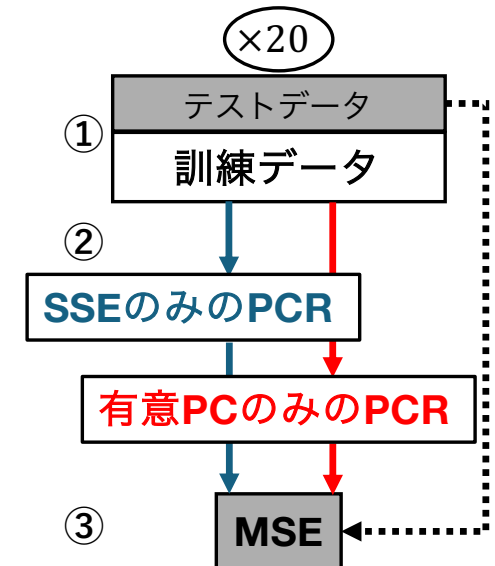
1. 双対共分散  $S_D$  ( $n \times n$ ) を推定し、PCスコア  $\hat{u}_k$  と係数推定

$$\hat{\gamma}_k = \frac{1}{\sqrt{n-1}} \hat{u}_k^T (\mathbf{y} - \bar{y} \mathbf{1}_n) \text{ を得る}$$

2. スパイク数  $m$  を Aoshima & Yata, 2018, *Statistica Sinica* より推定する

3. 統計量  $Z_k$  を作り、標準正規の臨界値 ( $\alpha = 0.05$ ) と比較。有意PCの選択後、PCRモデルを再構築

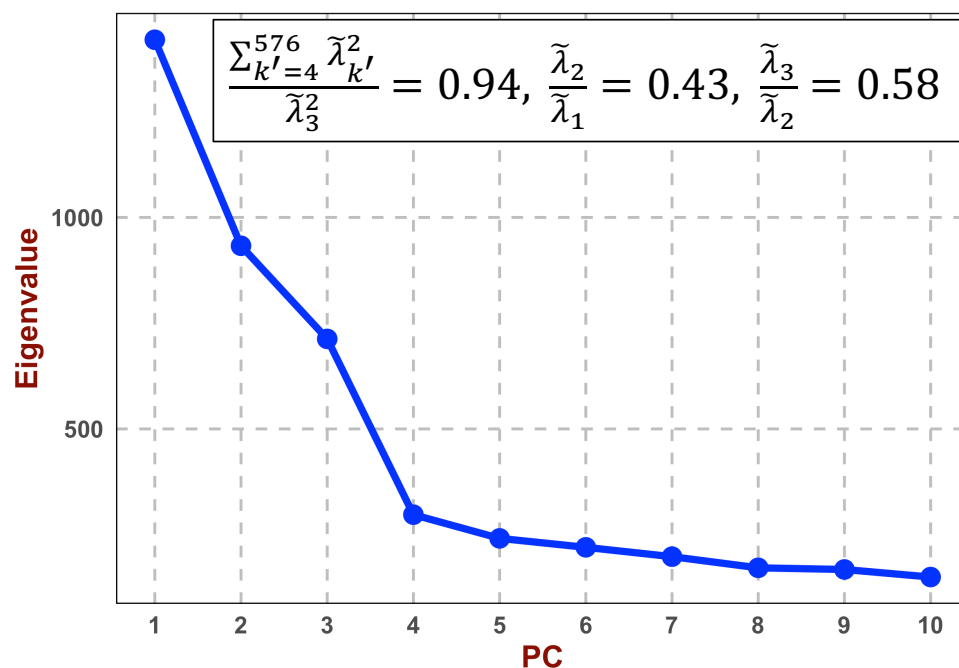
③ 比較：SSEのみのPCRと有意PCのみのPCRでMSEを比較



# 実データ解析 2

例 (ほとんどの分割において同様) :  $m = 3$

Eigenvalue Estimates (Training Data)

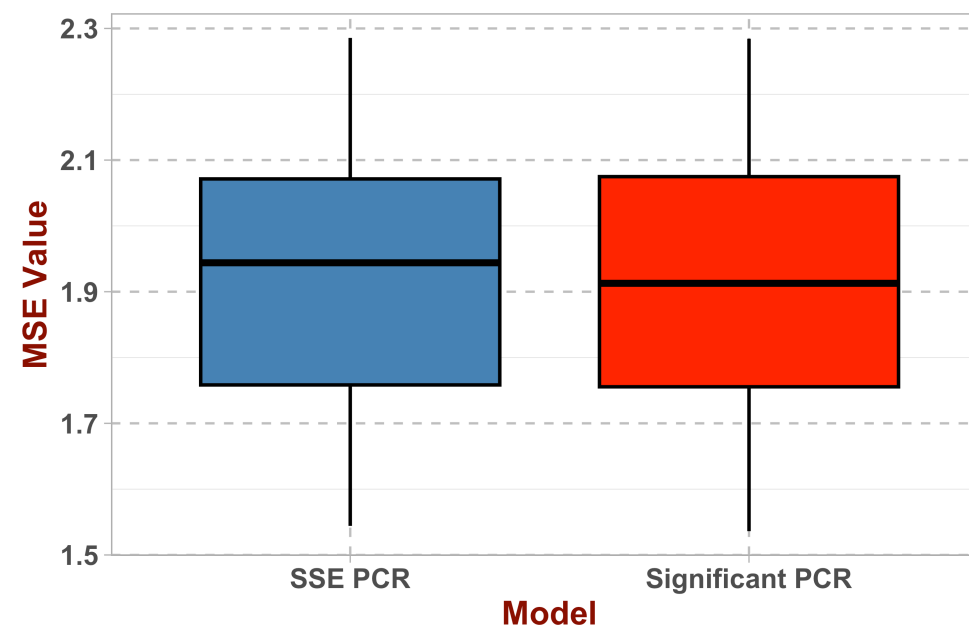


PC	1	2	3
P-values	0.00	0.63	0.30

結果 : **PC1**が予測に有効

	MSE (95% CI)
SSEのみのPCR	(1.80, 2.01)
有意PCのみのPCR	(1.80, 2.00)

MSE Distribution by Model



# まとめ

## 提案：

- 強スパイク固有値（**SSE**）モデルの下で、**主成分回帰（PCR）**の各主成分係数に対する**仮説検定を構築**
- 射影（次元削減）は主成分分析（**PCA**）のままとし、評価と選択の段階で目的変数  $y$  を明示的に取り入れることで、従来の「**PCR は  $y$  を考慮していない**」という弱点を克服

## 性能：

- シミュレーションでサイズ維持＋検出力を確認
- 実データ（ゲノミクス）では**予測に本当に効く主成分の特定とMSE同等を確認**

## 次のステップ：

- リスクベースのモデル構築：強スパイク固有値モデルと二重下降（**Double Descent**）を踏まえた解析的な予測リスク式を導出し、最適なPC数  $m$  を選ぶ（先行研究：Green et al., 2025, AoS）。仮説検定と組み合わせることができれば、**モデル構築・変数選択の両面で  $y$  を考慮**できる。
- 統一的枠組み：このリスク基準を本検定と組み合わせ、部分最小二乗法（**PLS**）やマッチング相関分析（**MCA**）、正準相関分析（**CCA**）へ一般化。
- ロードマップ：独立成分分析（**ICA**）、因果探索構造探索（**LiNGAM**）へ拡張し、高次元モデル選択のツールキットを目指す。

# 参考文献

1. Aoshima, M., & Yata, K. (2018). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, **28**, 43–62.
2. Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., & Foa, R. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.
3. Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R. J., Bignell, G. R., Tam, A. T., Davies, H., Stevenson, J. A., ... & Benes, C. H. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.
4. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
5. Green, A., & Romanov, E. (2025). The high-dimensional asymptotics of principal component regression. *Annals of Statistics*, **53(4)**, 1697–1727.
6. Yahoo Finance. (2025). *Apple Inc. (AAPL) stock prices*. Retrieved from <https://finance.yahoo.com/quote/AAPL/>
7. Yahoo Finance. (2025). *Microsoft Corporation (MSFT) stock prices*. Retrieved from <https://finance.yahoo.com/quote/MSFT/>
8. Yata, K., & Aoshima, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis*, **105**, 193–215.