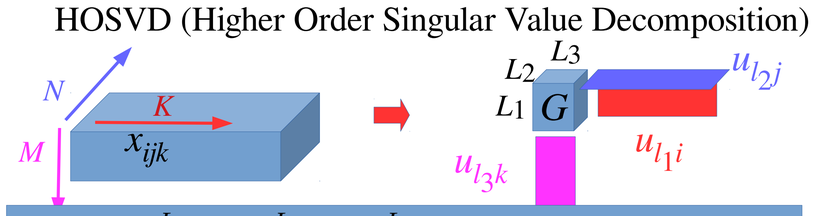
テンソル分解を用いた教師無し学習による変数選択法のバイオインフォマティクスへの応用

中央大学理工学部物理学科田口善弘

※テンソル分解とは？

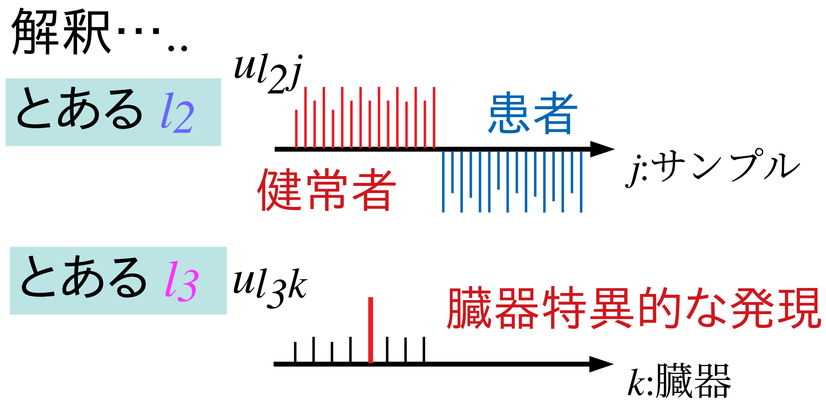
HOSVD (Higher Order Singular Value Decomposition)



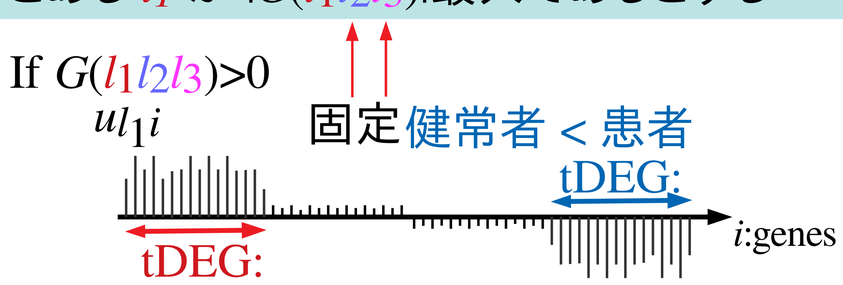
$$x_{ijk} \simeq \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \sum_{l_3=1}^{L_3} G(l_1 l_2 l_3) u_{l_1 i} u_{l_2 j} u_{l_3 k}$$

例　x_{ijk}: 遺伝子
N: 遺伝子数 (I),M: サンプル数 (j),
K: 臓器数 (k)

解釈‥..



とある l_1 が $|G(l_1 l_2 l_3)|$最大であるとする



tDEG: 健常者 > 患者
tissue specific Differentially Expressed Genes
臓器特異的に発現差がある遺伝子

遺伝子の選択基準:
$u_{l_1 i}$がガウス分布であると仮定（帰無仮説）
→ 累積χ二乗分布で遺伝子にP値を付与

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{l_{1i}}}{\sigma_{l_1}} \right)^2 \right] \qquad (1)$$

→ P値を多重比較補正（Benjamini-Hochberg基準）
→ 補正P値がしきい値（通常0.01）以下の遺伝子を選択。

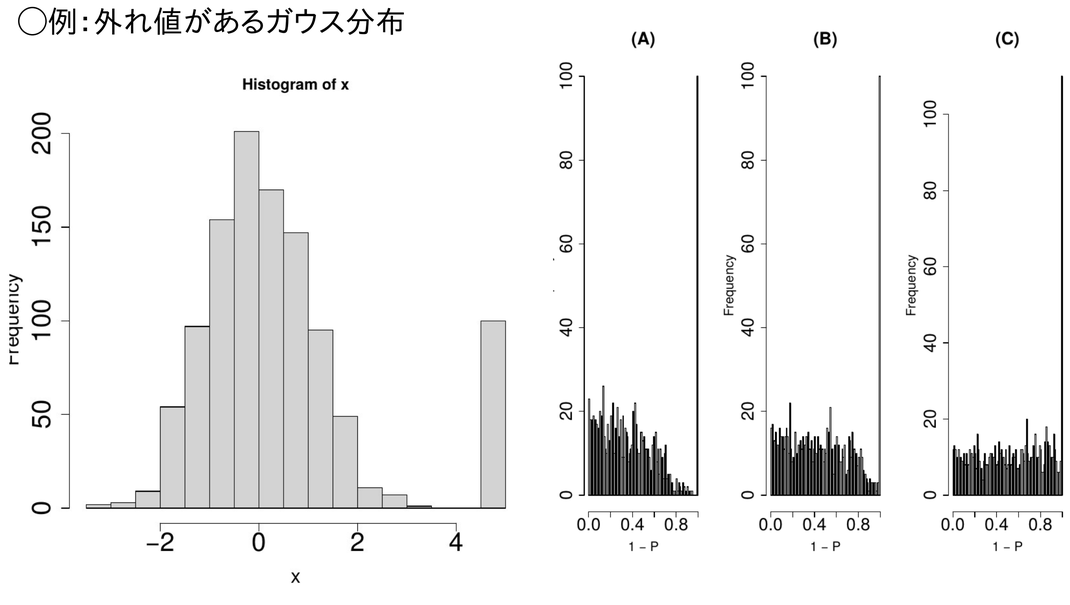
最近、σ_{l1i}の最適化をすると性能が大きく向上することを発見

1) Tensor decomposition- and principal component analysis-based unsupervised feature extraction to select more reasonable differentially expressed genes: Optimization of standard deviation versus state-of-art methods, Y-h. Taguchi, Turki Turki
bioRxiv 2022.02.18.481115; doi: https://doi.org/10.1101/2022.02.18.481115

2) Principal component analysis- and tensor decomposition-based unsupervised feature extraction to select more reasonable differentially methylated cytosines: Optimization of standard deviation versus state-of-the-art methods, Y-H. Taguchi, Turki Turki
bioRxiv 2022.04.02.486807; doi: https://doi.org/10.1101/2022.04.02.486807

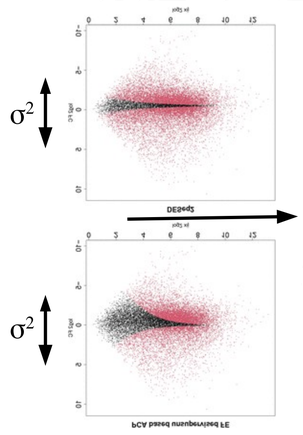
3) Tensor decomposition and principal component analysis-based unsupervised feature extraction outperforms state-of-the-art methods when applied to histone modification profiles
Sanjiban Sekhar Roy, Y-h. Taguchi
bioRxiv 2022.04.29.490081; doi: https://doi.org/10.1101/2022.04.29.490081

◯例：外れ値があるガウス分布



（A）　左の分布からσを計算してその値から（1）式でP値を出した場合の1−Pのヒストグラム
（B）　σを最適化した場合のP値を用いた場合の1−Pのヒストグラム
（C）　ガウス分布を生成したσを用いた場合のP値の1−Pのヒストグラム

応用例1　遺伝子発現プロファイルへの応用（文献1）
「発現量の多い遺伝子が選ばれやすいべき」という要請が「自動的に」満たされる。



DESeq2:経験則による分散関係を要請

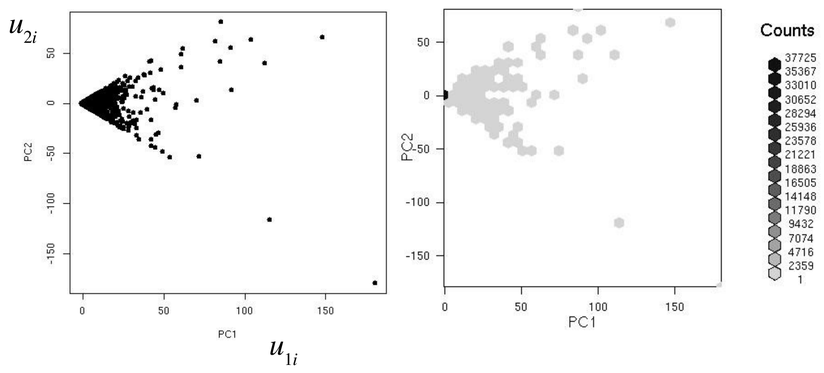
$$\frac{\sigma(\mu)^2}{\mu} = \sigma_0^2 + \frac{\sigma_1^2}{\mu}$$

主成分分析（PCA)を使うと「自動的に」満たされる

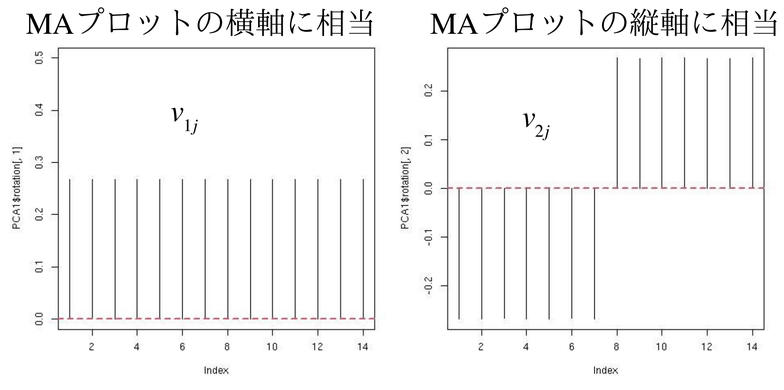
$$\Delta = \langle x_{ij} \rangle_{brain} - \langle x_{ij} \rangle_{cntl} \simeq u_{2j}$$

$$LFC = \log_2 \frac{\langle x_{ij} \rangle_{brain}}{\langle x_{ij} \rangle_{cntl}} = \log\left(1 + \frac{\Delta}{\langle x_{ij} \rangle_{cntl}} \right)$$

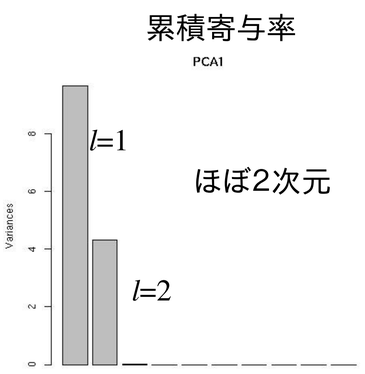
PCAによる遺伝子埋め込み　　密度分布



サンプル付与主成分
MAプロットの横軸に相当　　MAプロットの縦軸に相当

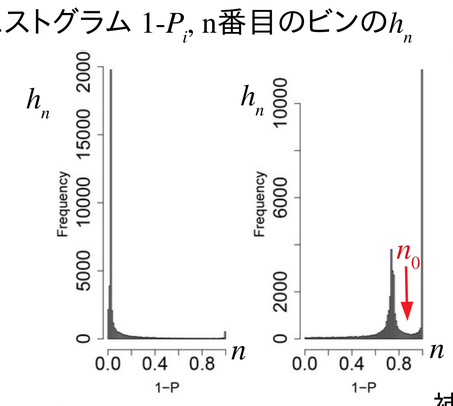


累積寄与率



ほぼ2次元

ヒストグラム 1-P_i, n番目のビンのh_n



補正 $P(n_0)$=0.1

帰無仮説
u_{2i} はGaussian分布 → 累積 χ^2 分布

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{2i}}{\sigma} \right)^2 \right]$$

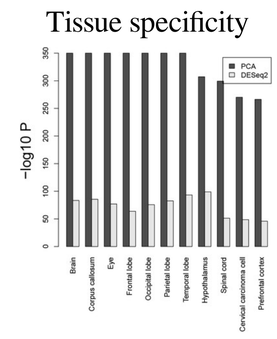
左:　$\sigma = \sqrt{\frac{\sum_i (u_{2i} - \langle u_{2i} \rangle)^2}{N}}$

右: 最適 σ は σ_hを最小化

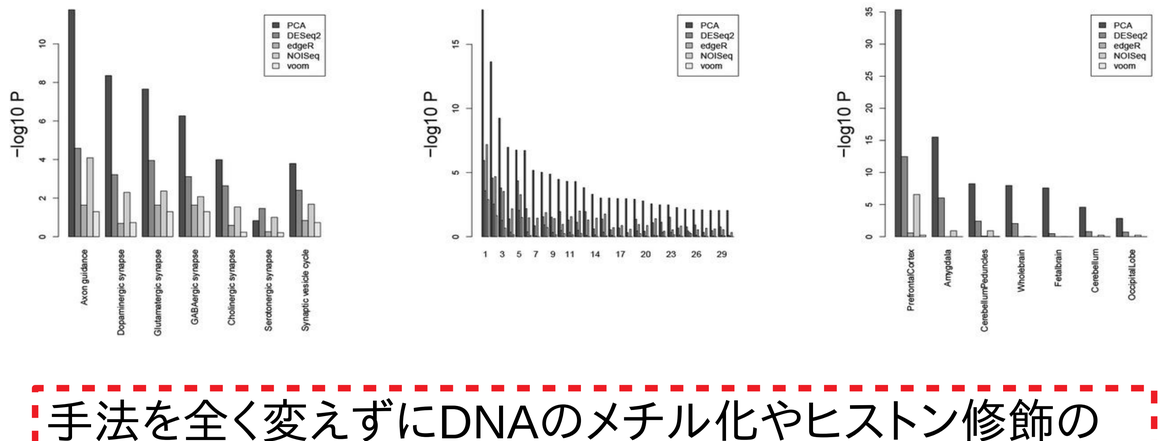
$$\sigma_h = \sqrt{\frac{\sum_{n<n_0} (h_n - \langle h_n \rangle)^2}{N_n(n<n_0)}}$$

補正P値 P_i <0.1の遺伝子を選択

生物学的な評価：PCA対DESeq2



(A) KEGG (B) GO BP (C) Human gene atlas
PCA vs DESeq2 vs edgeR vs NOISeq vs voom



手法を全く変えずにDNAのメチル化やヒストン修飾の差分解析にも使えることが解った（文献2, 3）.
DEGやメチル化、ヒストン修飾解析のデファクトスタンダードになるように努力したい。

演題発表内容に関連し、発表者に開示すべきCOI関係にある企業などはありません。