2020 年度修士論文

カルシウムイメージングデータのクラスタリング方法の一検討

タイトルは適当

5319E056-6　永山 瑞生

2020 年 12 月 14 日

早稲田大学 先進理工学研究科
電気・情報生命専攻
情報学習システム研究室

修 士 論 文 概 要 書

Summary of Master's Thesis

Date of submission: 02/29/2020

専攻名 (専門分野) Department	電気・情報生命	氏名 Name	夏目 漱石	指導教員 Advisor	村田 昇
研究指導名 Research guidance	情報学習システム	学籍番号 Student ID number	5320E123-4		
研究題目 Title	「我輩」の秘密に関する研究				

研究背景

問題設定

提案手法

応用例

まとめ

目次

1 序章 **1**
- 1.1 背景 . . . 1
- 1.2 カルシウムイメージングデータ . . . 1
- 1.3 関連研究 . . . 3
  - 1.3.1 脳データの解析 . . . 3
  - 1.3.2 カルシウムイメージングの解析例 . . . 3
  - 1.3.3 分解能の決定 . . . 4
- 1.4 目的 . . . 4

2 手法 **7**
- 2.1 類似度 A の推定 . . . 7
  - 2.1.1 数理モデル . . . 7
  - 2.1.2 NMF による A^* の推定 . . . 8
  - 2.1.3 NMF の一意性 . . . 9
  - 2.1.4 ブートストラップ法 . . . 9
  - 2.1.5 モデル平均 . . . 10
  - 2.1.6 NMF の基底数 . . . 11
- 2.2 A のクラスタリング . . . 11
  - 2.2.1 スペクトラルクラスタリングのクラスタ数の決め方 . . . 11
  - 2.2.2 評価 . . . 12

3 人工データ実験 **13**
- 3.1 シミュレーション . . . 13
  - 3.1.1 ネットワーク構造 . . . 13
  - 3.1.2 スパイクシミュレーション . . . 14
  - 3.1.3 カルシウムイメージングモデル . . . 15
  - 3.1.4 観測モデル . . . 15
  - 3.1.5 実験設定 . . . 15
- 3.2 A の推定に関する結果 . . . 16
  - 3.2.1 A の一意性 . . . 16
  - 3.2.2 手法の比較 . . . 17
  - 3.2.3 モデル平均の有用性 . . . 18
  - 3.2.4 NMF の基底数 . . . 18
- 3.3 A のクラスタリングに関する結果 . . . 20
  - 3.3.1 クラスタ数の決定 . . . 20

4 実データ解析 **21**
- 4.1 実データ . . . 21
- 4.2 結果 . . . 21

5 その他の検討事項 **23**
- 5.1 バイアス除去 . . . 23
- 5.2 重複除去 . . . 23

5.3 時間方向への制約 24
5.4 NMF のモデルエビデンス 24
5.4.1 ラプラス近似 25
5.4.2 ブートストラップによる近似 25

6 結論 **27**

第1章 序章

1.1 背景

睡眠は脳によって制御されており [1]，哺乳類にとって必要不可欠な生理現象である．その重要性にも関わらず，睡眠について解明されていないことが多い．その中でも，睡眠とはいかなる生物学的な状態か，という問いに対する明確な答えは未だない [2]．

哺乳類の睡眠状態は脳波によって定義される．しかし，哺乳類以外は脳波を計測することができないためふるまいでしか評価できない．そこで，ニューロンの活動から睡眠を新たに定義することができれば睡眠状態の解明に繋がると考えられる [3]．

脳内の情報伝達は複数個のニューロンによって行われている．また，睡眠時には多数のニューロンが活動してある現象が見られることが知られている．複数ニューロンの活動を解析することが重要である．

ニューロンの観察方法として，パッチクランプ法，細胞内記録法，細胞外記録法などの電気生理学的な手法が挙げられる．これらの手法は十分な時間分解能かつ細胞レベルでニューロンを観察することができる．しかし，電気生理学的な手法では観察できるニューロンの数は数十から多くても数百程度である．

より多くのニューロンを観察するために，蛍光イメージングの 1 つであるカルシウムイメージングという手法が用いられる．ニューロンで活動電位が発生（発火）すると細胞内の Ca^{2+} 濃度が上昇する．カルシウムイメージングでは，この Ca^{2+} 濃度上昇を蛍光で可視化する．具体的には，Ca^{2+} と結合すると蛍光強度が変化する蛍光分子を細胞内に発現させておき，Ca^{2+} 濃度を蛍光強度として蛍光顕微鏡で観察する．蛍光イメージングを用いる利点として，(1) 高い空間分解能，(2) 広い観察範囲，(3) 遺伝子工学と併用して興奮性／抑制性ニューロンの同定などをした上での観察ができることが挙げられる．一方，時間分解能が電気生理学的手法よりも低いことが蛍光イメージングの欠点である．Ca^{2+} 濃度の変化はニューロンの電気的変化よりも遅く，また，カルシウム感受性蛍光分子のキネティクスも影響する．さらに，カメラやレーザースキャンでのサンプリングレートは高くても 100Hz 程度であり，ニューロンの個々の発火を全て捉えるには不十分である．

本研究で扱うデータは，8Hz のサンプリングレートで観察された 100〜200 個のマウスのニューロンのカルシウムイメージングデータである．本研究では，低い時間分解能のカルシウムイメージングデータからニューロンをクラスタリングし，人工データ実験を通してどの程度の情報が抽出できるかを確認する．

1.2 カルシウムイメージングデータ

ニューロンは 1ms 単位で活動電位が発生する (発火)．活動電位は細胞内外のイオン濃度が局所的に変化することによって生じる．活動電位は細胞体から軸索を伝わり，シナプスを介して結合している別のニューロンに伝わる．哺乳類の皮質ニューロンにおいて，ニューロンからニューロンへシナプスを介して活動電位が伝わるには数十 ms かかる [4]．このように活動電位を伝えることによってニューロン間で情報がやり取りされる．

ニューロンが他のニューロンとコネクションを持つ状態のことをコネクティビティという．脳のコネクティビティには，synaptic connectivity と anatomical connectivity と functional connectivity の 3 種類がある．データによってどのコネクティビティの情報を取り出せるかは異なる．

1 個 1 個のニューロンを 1ms 単位で計測できればニューロン全ての活動を計測できるが，そのような技術は存在しない．ニューロンの計測方法には様々なものがあり，それぞれ計測可能な時間分解能と空間分解能が異なる．計測方法別の分解能については [5] の Fig 1 が分かりやすい．EEG，PET，fMRI は脳の一部のニューロンの活動によって生じた電位，血流，代謝量の変化を計測する．これらの手法ではニューロン単位の計測は行えないが，脳全体を計測することができる．電気生理学的な手法やカルシウムイメージングではニューロン単位で膜電位の変化やカルシウムイオン濃度の変化を計測する．これらの手法ではニューロン 1 個 1 個を計測できるが脳全体を計測することはできない．

カルシウムイメージングとはニューロン内のカルシウムイオン濃度を可視化することでニューロンの活動を計測する手法である．観察したい個体のニューロンにカルシウムイオンと結合する蛍光タンパク質を発現させると，細胞内のカルシウムイオン濃度に応じて蛍光強度が変化する．ニューロンが発火するとカルシウムイオンが流入するため蛍光強度が上昇し，その後徐々に蛍光強度は減少する．ニューロンを蛍光顕微鏡で観察することで図 1.1(B) のような蛍光強度を反映した画像が得られる．その画像から個々のニューロンの蛍光強度のデータを取得できる．

電気生理学的な手法と比べた時のカルシウムイメージングの利点として，ニューロンの位置情報が分かるため同じニューロンを複数回観察できることとより多くのニューロンの測定ができることが挙げられる．また，ニューロンには興奮性ニューロンと抑制性ニューロンの二種類があり，カルシウムイメージングではこの種類を見分けることができる．二種類の蛍光タンパク質を用いて，片方の蛍光タンパク質を抑制性ニューロンのみに発現させることで興奮性・抑制性が分かる．

カルシウムイメージングが電気生理学的な手法より劣る点として，時間分解能が挙げられる．カルシウムイメージングのサンプリングレートは測定機器に限界があり [6]，通常 1Hz-50Hz 程度である．ニューロンの発火は約 1ms で，シナプスを介して発火が伝わるのは 40ms 以内 [7] なので，ニューロンの発火を個別に観測することはできず，低いサンプリングレートでは発火の伝達も捉えることができない．また，カルシウムイオン濃度の変化は発火のタイムスケールより長い．発火後に蛍光強度が変化し始めるのは数ミリ秒後，蛍光強度がピークに達するのは数 100ms 後である．一方，蛍光強度が元の値に戻るまでには，数 100ms から数 1000ms かかる [8]．また，蛍光タンパク質の性能によっても時間遅れが生じる．

本研究で用いるデータは筑波大学の柳沢研究所で計測されたカルシウムイメージングデータである．本データは，2 光子多細胞カルシウムイメージングによって 1 匹のマウスの大脳皮質 1 次運動野第 2・3 層のニューロンのイメージング画像を得た後，人手で ROI がつけられ，ニューロンごとの数値データに直されたものである．1 時間おきに 15 分間のイメージングが計 6 回行われ，各イメージングのサンプリングレートは 8Hz(125ms ごと) である．このサンプリングレートは使用された機器の最大のものである．用いられた蛍光タンパク質は GCaMP6s である．実験系は図 1.1(A) の系で行われた [2]．観測されたニューロン数は 154 であった．時間方向には 4s ごとのマウスの状態 (wake, REM, NREM) のラベルが付いており，ニューロンごとに興奮性か抑制性のラベルがついている．ここのデータの説明は実データ実験で使うものによって変える

8Hz というサンプリングレートはシナプス伝達一つを見るには不十分である．顕微鏡で観測すると活動電位が伝わる順番は分からなくなる．そのため，カルシウムイメージングデータからはニューロンの活動の相関の情報しか得ることができないと考えられる．また，脳の神経細胞はシナプスを 3 つか 4 つ介せば全て繋がると言われている．これより，このデータではニューロン間のシナプス伝達を推定するのではなく，機能的に同じニューロン，つまり同時に活動するニューロングループを推定する問題が適していると考えられる．機能的に同じニューロンは，3 つの脳のコネクティビティのうち functional connectivity で繋がっているニューロンを指す．

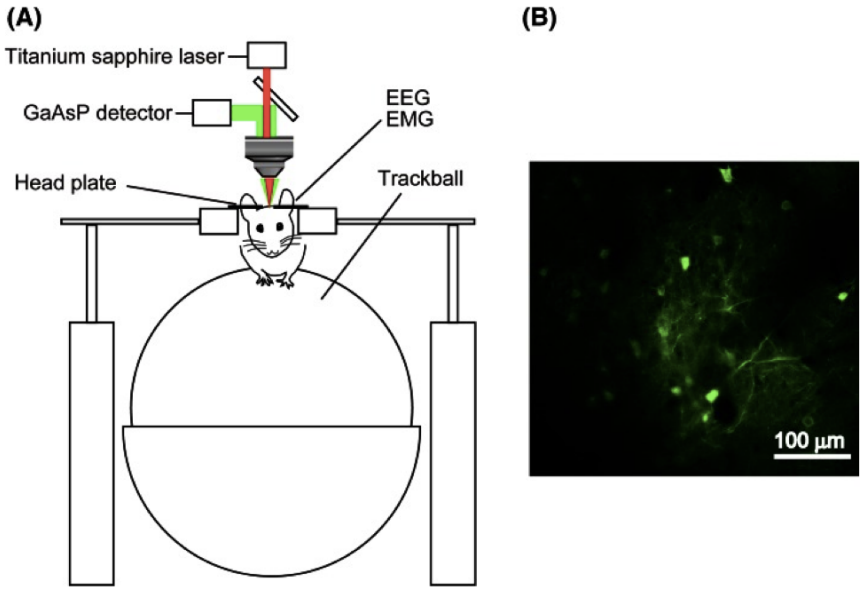


図 1.1: カルシウムイメージングの測定系

1.3 関連研究

1.3.1 脳データの解析

カルシウムイメージングは時間解像度が低く空間解像度が高いデータが取得できる手法である．同じ特徴である fMRI に対するデータの解析手法を紹介する．

fMRI データの解析は model based な手法と data-driven な手法がある [9]．Model based な手法の例は statistical parametric mapping (SPM) や cross-correlation analysis (CCA), coherence analysis (CA) などが上げられる．Data-driven な手法は更に decomposition と clustering に分けることができる. Decomposition には princinpal component analysis (PCA) や singular value decomposition (SVD)，independent component analysis (ICA) などが挙げられる．Clustering は fuzzy clustering analysis (FCA) や hierarchical clustering analysis (HCA) などがある．

1.3.2 カルシウムイメージングの解析例

カルシウムイメージングデータの解析には二種類考えられる．まずは，カルシウムイメージングデータからスパイクを推定し，そのスパイク列を解析する方法である．Vogelstein らは逐次モンテカルロ法を用いてカルシウムイメージングデータからスパイク推定を行なった [10]．しかし，カルシウムイメージングデータでも低いサンプリングレートで計測されたデータではこの方法は使えない．

もう一つは，データから直接ニューロンの活動を解析する方法である．Mishchencko はベイズ推定を用いたニューロンの結合推定を行なった [11]．しかし，カルシウムイメージングのサンプリングレートが 30Hz 以上でないと意味のある結合推定結果は得られないと報告している．また，Stetter らは Transfer Entropy を用いて培養された興奮性ニューロンの結合推定を行なった [12]．この手法では，モデルを仮定せずにデータからネットワーク推定を行なっている．また，ニューロンのネットワーク構造を仮定した人工のカルシウムイメージン

グデータを作成し，推定精度を議論している．Ikegaya らは蛍光強度データの一次微分から発火のタイミングの情報を取り出し，テンプレートマッチングによってリアクチベーション現象を解析した [13].

Molter らはカルシウムイメージングデータからニューロングループを抽出する方法を 8 つ人工データ実験と共に試した [14]．手法は大きく 2 つに分けられ，ニューロンペアの相関を見るものと，全てのニューロン活動の状態を見るものである．前者では固有値分解によって相関行列を作成した後，ICA や Promax rotation によってグルーピングを行う．後者では，ニューロンの活動から SVD，k-means ラスタリング，spectral クラスタリングなどを用いてグルーピングを行う．後者の方法では，各グループの時間方向の活動を平均をとるなどして，グループの活動としていた．人工データは，ニューロンをポアソン分布にしたがって発火させ，発火からカルシウムイメージングの観測データに変換していた．同じグループに所属するニューロンは発火確率を同じ時間帯にあげることで表現していた．ニューロンは 2 つのグループに所属する場合も考えられていた．最も良いと結論づけられていた手法は ICA-CS と SGC という手法だったが，安定して推定精度が高くなるのは観測時間が 1800s より長い場合だった．

Ghandor らは学習に関係するニューロン（engram cell）群について NMF を用いて解析を行った [15]．実験は複数セッションに分けて行われており，それぞれのセッションで NMF を用いてニューロングループとその活動に分解していた．基底数は AICc で決めていた．セッションごとに推定されたグループが近いかを cosine similarity で計った結果，engram cell では non-engram cell よりも繰り返し活動するグループが多かった．NMF によってニューロングループがどれほど抽出できるかは論じられていなかった．

1.3.3 分解能の決定

ニューロンの活動データの扱いには時間分解能と空間分解能の 2 つの側面から検討する必要がある．時間分解能については，蛍光強度データをそのまま用いる，時間窓に区切るなどが考えられる．空間分解能については，ニューロン 1 個を見る場合，2 個を見る場合，複数を見る場合が考えられる．手法によってどのレベルでデータを扱うかが異なる．表 1.1 にカルシウムイメージングデータを解析する際に使えそうな手法を載せる．

	生データ	時間窓で区切る
ペアで見る	時系列クラスタリング	glasso，類似度+クラスタリング
複数で見る	行列分解	ロジスティック回帰，時系列クラスタリング

表 1.1: カルシウムイメージングデータ解析に使えそうな手法

1.4 目的

カルシウムイメージングデータは上述のように多くのニューロンを観測できる利点がある一方，時間分解能が低いという欠点を持つ．本研究で扱うデータは低いサンプリングレートで観測されたデータである．ニューロンは複数で活動することで情報伝達を行っており，睡眠時に多数のニューロンが同時に活動する現象も確認されている．これより，本データでは同時に活動するニューロングループを推定して解析を行うことが適当である．

本研究の目的は，カルシウムイメージングデータから同時に活動するニューロングループを推定し，それらの睡眠・覚醒時の活動の違いを解析することである．ニューロングループの推定には，ニューロン同士の類似度行列を作成しそれをクラスタリングする方法をとる．また，人工データ実験によって，本手法によってカルシウムイメージングデータからニューロングループの情報がどの程度取り出せるかを確かめる．

本論文の構成は以下の通りである．第 2 章では数理モデルを元にした解析のアプローチを説明する．第 3 章では人工データの作成方法と人工データ実験の結果を述べる．第 4 章では実データ解析の結果を述べる．第 5 章では採用に至らなかった検討事項について述べ，第 6 章に結論を述べる．

第2章 手法

本論文の問題はニューロンのクラスタリングである．ニューロン間の類似度 $A \in [0,1]^{I \times I}$ をクラスタリングする問題を解く．ただし，I はニューロン数で a_{ij} はニューロン i と j の類似度である．本章では A の推定方法とクラスタリング手法の説明をする．

A は非負行列因子分解（NMF）で推定する．NMF を使う理由は数理モデルに基づいており，次節で説明する．クラスタリング手法にはスペクトラルクラスタリングを用いる．スペクトラルクラスタリングはグラフカットとして解釈でき，類似度行列 A をクラスタリングするのに適している．

2.1 類似度 A の推定

本節では NMF とブートストラップ法によって類似度 A を推定する方法を述べる．観測データ X は $X \in \mathbb{R}_+^{I \times J}$ とする．ただし，$\mathbb{R}_+$ を非負の実数の集合，J を観測時系列の長さとする．NMF で推定するのは $A^* \in \{0,1\}^{I \times I}$ であり，a^*_{ij} はニューロン i とニューロン j が同じグループに所属するか否かを表す．ブートストラップ法を用いて類似度 $A = E[A^*|X]$ を推定する．A の要素 a_{ij} はニューロン i とニューロン j が同じグループである確率 $Pr(a_{ij} = 1)$ を意味する．

2.1.1 数理モデル

カルシウムイメージングデータに対していくつかの仮定をおいた．

仮定 1
グループが K 個存在し，同じグループ内のニューロンは同時に活動する．ニューロンは複数のグループに所属することができる．観測時間内ではグループに属するニューロンは変化しない．

仮定 2
複数のグループが同時に活動する時，属するニューロンは被らない（ニューロンが属するグループは同時には活動しない）．

これらの仮説をもとに，数理モデルを構築する．ニューロン i の観測時系列は $x_{i:} \in \mathbb{R}_+^J$ である．

$c_{k:} \in \mathbb{R}_+^J$ $(k = 1, \ldots, K)$ をグループ k の活動の時系列とすると，仮定より $x_{i:}$ は $c_{i:}$ の重み付き和として表す：

$$x_{i:} = \sum_{k=1}^{K} d_{ik} c_{k:} + \eta_{i:}, \tag{2.1}$$

ただし，$d_{ik} \in \mathbb{R}^+$ で，$\eta_{i:} \in \mathbb{R}^J$ はガウスノイズの時系列である．カルシウムイメージングのノイズはポアソン分布に従う光子ノイズであるが，光子数が多い場合はガウス分布で近似できる [16]．

仮定 2 を置くことによって，式 (2.1) の線形モデルを考えることができる．仮定 2 がなかった場合，あるニューロンの蛍光強度は複数グループからの影響によって上限なく上昇できてしまう．実際は，ニューロンの蛍光強度の最大値は蛍光タンパク質の量で決まるため，蛍光

強度は最大値で飽和する．その場合式 (2.1) には飽和を組み込まなければいけない．本論文では，複数グループが同時に活動する場合はそれらも同じグループとみなし，単純な線形モデルで表す．

式 (2.1) は行列形式で以下のように表現できる：

$$\begin{aligned} Y &= DC, \\ X &= Y + H. \end{aligned} \tag{2.2}$$

ただし，$D \in \mathbb{R}_+^{I\times K}$, $C \in \mathbb{R}_+^{K\times J}$, $H \in \mathbb{R}^{I\times J}$ である．また，D の要素 (i,k) は d_{ik}，C の i 行は $c_{i:}$，H の i 行は $\eta_{i:}$ である．

2.1.2 NMF による A^* の推定

非負行列因子分解 (nonnegative matrix factorization; NMF）[17] は行列分解の手法の一つである．NMF は以下の最適問題を解く：

$$\underset{D\geq 0, C\geq 0}{\arg\min}\ ||X - Y||_F^2.$$

基底数 k の NMF のモデルを $\mathcal{M}_k$ とおく．ノイズ行列 H の各要素が正規分布 $\mathcal{N}(0,\sigma^2)$ に従う $\mathcal{M}_k$ の尤度は以下である．

$$p(X|Y_k,\mathcal{M}_k) = \prod_{i,j}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{([Y_k]_{ij}-x_{ij})^2}{2\sigma^2}),$$

ただし，Y_k はモデル $\mathcal{M}_k$ における推定量である．対数尤度は以下のようになる．

$$\log p(X|Y_k,\mathcal{M}_k) = -\frac{IJ}{2}(\log 2\pi + 2\log\sigma) - \frac{1}{2\sigma^2}\sum_{ij}([Y_k]_{ij}-x_{ij})^2.$$

NMF の寄与率行列 $P \in \mathbb{R}^{I\times K}$ の要素を次のように定義する：

$$\begin{aligned} p_{ik} &= \frac{||d_{ik}c_{k:}||_1}{\sum_{l=1}^K ||d_{il}c_{l:}||_1} \\ &= \frac{d_{ik}||c_{k:}||_1}{\sum_{l=1}^K d_{il}||c_{l:}||_1}. \end{aligned}$$

要素 p_{ik} はニューロン i に対する基底 k の寄与率という意味である．

A^* は寄与率行列 P から作成する．まず，P と同じサイズの行列 $G \in \{0,1\}^{I\times K}$ を作る．G は，P の各行について最大値のみを 1，それ以外を 0 とした行列である．

$$g_{ij} = \begin{cases} 1 & (j = index(max(p_{i:}))) \\ 0 & (otherwise) \end{cases}$$

推定量 A^* を

$$A^* = GG^\top,$$

と定義する．

この推定量は cluster ensemble でも用いられている pairwise similarity[18] と似たものになっている．

2.1.3 NMF の一意性

NMF の推定には一意性がなく，ある正則行列 Q を考えた時，

$$\begin{aligned} X &= DC \\ &= DQRC \\ &= D'C', \\ R &= Q^{-1}, \\ D' &= DQ, \\ C' &= RC, \end{aligned} \tag{2.3}$$

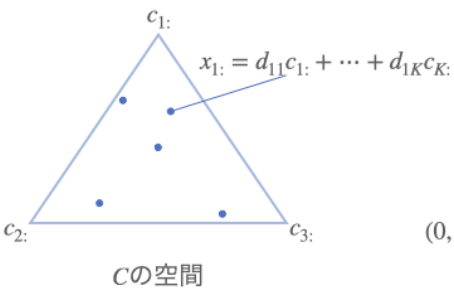
のように別の D' と C' が推定される可能性がある．

NMF に一意性がある条件は [19] などでまとめられている．しかし，一意性を持たせるにはかなり条件が狭められる．

NMF と同じように非負行列を分解する手法に nonnegative rank factorization (NRF) [20] がある．NRF ではデータ行列 X を $X = DC$ に分解できる最小の基底数を nonnegative rank $\mathrm{rank}_+(X)$ と定義し，$\mathrm{rank}(X) = \mathrm{rank}(X)_+$ となる行列 X を対象とする．全ての非負行列は NMF できるが，NRF ができるとは限らない．NRF の計算は NP 困難であるが，[20] では NRF の計算と，NRF を持たない行列に対して MNRS という分解の計算方法を提示している．しかし，X にノイズが乗っている場合に NRF が存在する条件は調査されていない．

この場合の D' と C' を用いて作られる寄与率行列 P' と元の寄与率行列 P の関係を考える．簡単のため，X に行和 1 の正規化を加え C に行和 1 の制約を加える．この時 D の行和も 1 となり，$P = D$ となる．式 (2.3) より，$P' = PQ$ となる．この時，G の作り方には一意性がないため推定量 A にも一意性がない．

$K = 3$ の場合の図を用いて説明をする．図 2.1 に C の空間内の $x_{i:}$ と D の空間内の $d_{i:}$ を表す．$d_{i:}$ の和は 1 であるが，軸は図 2.3 のように動くことができる．簡単のため，G を作る際に累積寄与率の閾値を低くしてハードクラスタリングを行うと，図 2.3 のようにニューロンが所属する基底が変化することがある．その結果，推定量 A も変化するため一意性はない．



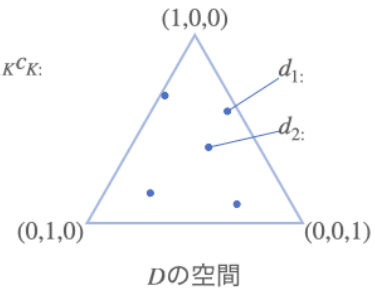


図 2.1: C の空間と D の空間．

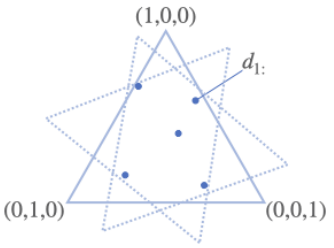


図 2.2: D の空間の軸は回転・移動する．

2.1.4 ブートストラップ法

ブートストラップ法 [21] とは，推定量の分布を近似する方法である．データ X が分布 F に従うとき，確率変数 $R(X, F)$ を推定する問題を考える．ブートストラップサンプル X^* を作成して，$R^* = R(X^*, \hat{F})$ を推定すると，R^* の分布は R の分布を近似する．

今回の問題ではブートストラップサンプル X^* から A^* を推定する．類似度 A を A^* の期

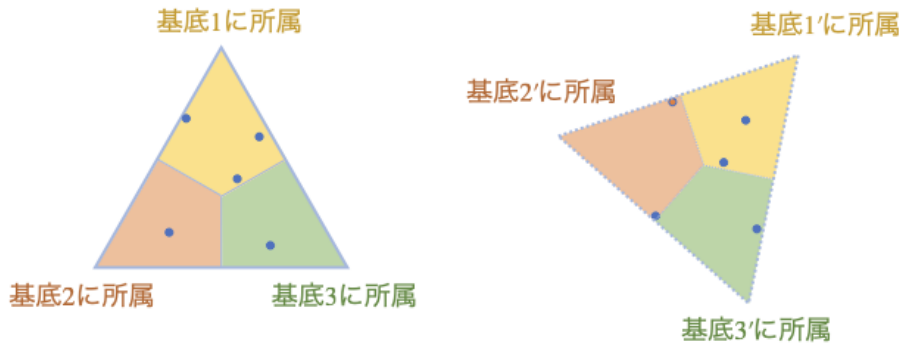


図 2.3: D の空間の軸の変化によってニューロンが所属する基底が変化する．

待値 $E[A^*|X]$ として推定する：

$$\begin{aligned} A &= E[A^*|X] \\ &= \frac{1}{B}\sum_{b=1}^{B} A^{*b}, \end{aligned}$$

ただし，B はブートストラップサンプル数，A^{*b} は X^{*b} から推定された推定量である．

NMF のブートストラップ方法にはいくつかの方法が考えられる．簡単な 3 種類の方法について説明する．列のサンプリングによるブートストラップは，データ行列 X の各列がデータサンプルとしてみなせるので，列をサンプリングして X^* を作る方法である．ブロックブートストラップでは，複数列を塊としてサンプリングする．この方法は時系列データでのブートストラップで用いられており，時間方向に制約の入った NMF などでは有効である．残差型ブートストラップは一回 NMF のモデルを推定し，推定後の残差をサンプリングして推定した DC に足す方法である．

$$X^* = \hat{D}\hat{C} + H^*,$$

ただし，$\hat{D}$ と $\hat{C}$ は最優推定量で，H^* は $X - \hat{D}\hat{C}$ をリサンプリングした行列である．ニューロンごとにノイズの大きさが異なることが想定される場合は，行ごとにリサンプリングを行うのが適切だと考えられる．NMF のモデル式 (2.2) では H は i.i.d. なノイズという仮定を置いているので，残差型ブートストラップが一番モデルに沿ったブートストラップ方法と言える．

機械学習の分野でブートストラップ法が多く使われる場面はバギング [22] である．バギングとは，ブートストラップ法によって学習器を増やしその出力の平均をとる学習方法である．今回のブートストラップ法の使い方もバギングとして見ることができる．

2.1.5 モデル平均

モデル平均とは異なるモデルの推定量の平均をとって精度向上を測る方法である．本論文で用いる推定量 A^* は基底数によって行列サイズが変化しないので，異なる基底数の推定結果の平均をとることができる．その場合の推定量は：

$$A = \frac{1}{K_{max} - K_{min} + 1}\sum_{k=K_{min}}^{K_{max}} A_k,$$

ただし，平均する最小の基底数を K_{min}，最大の基底数を K_{max} とする．

次節でも触れるが，NMF の真の基底数を求めるのは難しい問題である．モデル平均によって，異なる基底数の推定結果を平均して精度をあまり落とさないようにする．今回扱うデータから推定される A は基底数が違う時に大きくは変化しない．真の基底数周りで平均した A の方が真の基底数でないモデルから推定された A_k よりも精度が落ちないと考えられる．

アンサンブル学習では，それぞれのモデルにある程度の推定精度があり [23]，答えに多様性がある方が精度が上がる [24] と言われている．本論文でのモデル平均の使い方は，真の基底数が分からない時に推定結果をなまして推定を間違うリスクを下げるという使い方をしている．

2.1.6 NMF の基底数

NMF の基底数の決め方にはいくつかのアプローチがある．まずは，専門家や解析者の知識に基づいて決めることである．この方法はデータに対して十分な知識がない時には使えない．

次に BIC [25] や AIC [26] を用いる方法である．これらは漸近理論に基づいた近似を行った情報量基準である．NMF はデータが増えるほどパラメータ数 $(I + J) * K$ が増えるという特徴があり，これらを用いるのは本来不適切である．

次に R の NMF パッケージにも組み込まれている Brunet ら [27] の方法を紹介する．彼らは NMF の推定結果からノード同士が同じ基底に所属するかしないかを表す connectivity matrix $A \in \{0, 1\}^{I \times I}$ を作成する．本論文の推定量 A と同じ意味の行列である．初期値を 20-100 回変化させて C の平均 $\bar{A} \in [0, 1]^{I \times I}$ を計算する．彼らは真の基底数ではこの推定量が 0 か 1 に寄るようになると仮定して，最も $\bar{A}$ が安定する基底数を求める．安定度は $1 - \bar{A}$ とその cophenetic correlation coefficient の Pearson correlation から計算する．

Ubaru ら [28] はブートストラップを用いて NMF を行っており，D がそれほど変わらない基底数を採用している．推定した D_b 同士の相互相関行列について dissimilarity [29] を測りその平均が最小となる基底数を用いる．

Hutchins らはテストデータに対する RSS（residual sum of squares）が真の基底数以降になるとあまり下がらなくなると論じている [30]．

Bayesian NMF ではギブスサンプリングなどを用いてモデルエビデンスを計算している [31]．

上記で述べた方法の他にも様々な方法が考案されているが，全てのデータに当てはめられるような枠組みは存在しない．

2.2 A のクラスタリング

A のクラスタリングにはスペクトラルクラスタリングを用いる．スペクトラルクラスタリングは，グラフをクラスタリングする．隣接行列からグラフラプラシアンを求め，その固有ベクトルを k-means などでクラスタリングする．他のクラスタリング手法も使えるが，A をグラフとして見ることができるため，スペクトラルクラスタリングを用いる．

2.2.1 スペクトラルクラスタリングのクラスタ数の決め方

クラスタリングにおけるクラスタ数の決め方には様々な方法がある．その中でもクラスタリング手法によらない方法は，安定性を見る方法 [32] や Gap 統計量 [33] がある．スペクトラルクラスタリングに特化した方法としては，固有値ギャップを見る方法 [34] がある．

固有値ギャップは固有値を小さい順にプロットした際に，一気に固有値が大きくなる箇所である．そこをクラスタ数とする．これは目で見て判断しなければならないが，今回は固有値の差分を取って大津の二値化にかけ，差分が大きいグループの最大の固有値の箇所を固有値ギャップとして計算した．

クラスタが明瞭な場合に固有値ギャップを見る方法が最も安定してクラスタ数を決定できたため，実験では固有値ギャップを見る方法を採用する．

2.2.2 評価

比較手法には [14] で性能の良かった ICA-CS を用いる．また，相関行列とも比較を行う．

クラスタリングの評価方法には同論文で用いられていた Best Match score を用いる．二つのクラスタリング結果 $\mathcal{A} = \{A_1, \ldots, A_{|\mathcal{A}|}\}$ と $\mathcal{A}' = \{A'_1, \ldots, A'_{|\mathcal{A}'|}\}$ があった場合を考える．$\mathcal{A}$ と $\mathcal{A}'$ の Best Match distance [35] を

$$BestMatch_d = \sum_{A \in \mathcal{A}} min_{A' \in \mathcal{A}'} d(A, A') + \sum_{A' \in \mathcal{A}'} \min_{A \in \mathcal{A}} d(A', A),$$

とする．ただし，

$$d(A, A') = 1 - \frac{|A \cap A'|}{|A \cup A'|},$$

である．これは Jaccard 係数を 1 から引いた量である．Best Match score は

$$BestMatchscore = 1 - \frac{1}{|\mathcal{A}| + |\mathcal{A}'|} BestMatch_d(\mathcal{A}, \mathcal{A}'),$$

である．

第3章 人工データ実験

3.1 シミュレーション

ニューロン集団のカルシウムイメージングデータをシミュレーションによって作り，解析手法を評価する．シミュレーションでは 1）ニューロンのネットワーク構造を作成し，2）スパイクのシミュレーションを行い，3）蛍光強度の観測データに変換する．

3.1.1 ネットワーク構造

シミュレーションに用いるニューロンの個数を N として，ニューロンのネットワーク構造を $S \in 0, 1^{N \times N}$ とする．s_{ij} はニューロン i からニューロン j へ活動電位が伝わるかを表している．本節では S の作り方を説明する．

ニューロンのネットワーク構造には small world network[36] を用いる．Small world network はノード数，張り替え確率，初期次数を決めることによってネットワークを作成するアルゴリズムである．初期次数は，ニューロンが平均何個のニューロンとシナプス結合を持つかという変数である．張り替え確率は，初期次数によって作成された規則的なグラフのエッジをランダムに張り替える確率である．そのため，エッジのうち何割が遠くのニューロンとつながっているかを表す変数である．

実際のニューロンを small world network によって表すために，初期次数と張り替え確率を実データから決める．今回はこの値はニューロンのコネクションの割合と相互のコネクションの割合から決める．興奮性ニューロン同士の 6.7%であり，そのうち双方向のコネクションの割合は 24%である [37]．発達中マウスの興奮性ニューロンから抑制性ニューロンへのコネクティビティと抑制性ニューロンから興奮性ニューロンへのコネクティビティはどちらも 78%であった [38]．成熟したマウスではより少ないと思われるが，データが見つからなかったため，40%とした．相互のコネクションの割合がランダムにエッジを作るよりも高いのは，近いニューロンにコネクションが作られやすいからだと考えられる．これらのデータを実現するように初期次数と張り替え確率を調整した．用いたパラメータを表 3.1 に示す．抑制性ニューロン同士のコネクティビティは分からないため，興奮性ニューロンと同じにしている．

結合の種類	初期次数	張り替え確率
同種類のニューロン間	$0.0335N$	0.3
興奮性ニューロンと抑制性ニューロン間	$0.2N$	0.3

表 3.1: ネットワーク構造のパラメータ

実際のネットワーク構造の作り方を説明する．ネットワーク構造は興奮性ニューロン同士の結合，抑制性ニューロン同士の結合，興奮性ニューロンと抑制性ニューロン間の結合の 3 つに分けて作成する．まず，全ニューロンのうち抑制性ニューロンと興奮性ニューロンのインデックスを決めておく．全てのニューロンについて表 3.1 に従ってネットワークを作成し，それぞれに対応する隣接行列の上三角または下三角行列を取り出して結合する．作成したいのは向きのある有向グラフなので，上三角行列と下三角行列を分けて作成する．

3.1.2 スパイクシミュレーション

スパイクのシミュレーションに Izhikevich モデル [39] を用いる．このモデルは Hodgikin-Huxley モデルをもとにしており，計算コストが低い．Izhikevich モデルでは，あるニューロンの膜電位が閾値を超えると発火したとみなし，あらかじめ定義したニューロンのネットワーク構造に従って結合を持つニューロンの膜電位を上昇させる．このシミュレーションで設定しなければいけないのは，個々のニューロンの特徴パラメータ，重み付きのネットワーク構造，外部からのランダムな入力である．

まず，個々のニューロンの特徴パラメータについて説明する．このモデルではニューロンごとに 4 つのパラメータを設定する必要があり，そのパラメータでニューロンを特徴づける．本論文では興奮性ニューロンには regular spiking neurons，抑制性ニューロンには fast spiking neurons を用いる．それらのパラメータを 表 3.5 に示す．ただし，r_e と r_i は 0 から 1 の一様分布に従う確率変数である．

ニューロンの種類	a	b	c	d
興奮性ニューロン	0.02	0.2	$-65+15r_e^2$	$8-6r_e^2$
抑制性ニューロン	$0.02+0.08r_i$	$0.25-0.05r_i$	-65	2

表 3.2: Izhikevich モデルのパラメータ

次に，重み付きのネットワーク構造 $W \in \mathbb{R}^{I\times I}$ について説明する．ニューロン i から j へ結合があった場合，w_{ij} はニューロン i が発火した時にニューロン j の膜電位をどれだけ上昇させるかという数値である．W は，前節で作成した S の非ゼロ要素を数値で置き換えることで作成する．興奮性ニューロンからの結合は一様分布 $U(0,0.5)$ からサンプルし，抑制性ニューロンからの結合は一様分布 $U(-2,0)$ からサンプルする．

$$w_{ij} = \begin{cases} U(0,0.5) & (s_{ij}=1 \ and \ i \in \ excitatory \ neuron) \\ U(-2,0) & (s_{ij}=1 \ and \ i \in \ inhibitory \ neuron) \\ 0 & (s_{ij}=0) \end{cases} \tag{3.1}$$

最後に外部からのランダムな入力について説明する．ニューロンには観測範囲外からの入力がある（以降，外部入力とする）．そのため，シミュレーション中も外部からの電位を乱数としてニューロンの電位に足す．本論文では，ニューロンの活動も外部入力の大きさで表現する．活動していない興奮性ニューロンと抑制性ニューロンにはそれぞれ，$\mathcal{N}(0,5)$ と $\mathcal{N}(0,2)$ に従う乱数を足す．活動している興奮性ニューロンと抑制性ニューロンにはそれぞれ，$\mathcal{N}(1,5)$ と $\mathcal{N}(0.4,2)$ に従う乱数を足す．これらを表 3.3 に示す．活動していないニューロンへの外部入力は [39] で用いられていたものを採用した．ただし，興奮性ニューロンの活動時の外部入力は変化させた実験もある．

ニューロンの種類	活動時の外部入力	活動していない時の外部入力
興奮性ニューロン	$\mathcal{N}(0.8,3)$	$\mathcal{N}(0,5)$
抑制性ニューロン	$\mathcal{N}(0.4,0.1)$	$\mathcal{N}(0,2)$

表 3.3: シミュレーションに用いる外部入力の値

本論文では同時に活動するニューロンを推定するのが目的の 1 つである．ある時間帯にあるニューロングループが活動する時，そのニューロングループには平均値を上げた外部入力を足し，それ以外のニューロンには平均 0 の外部入力を足す．こうすることで，ニューロングループの活動のみ上がる（つまり蛍光強度が上がる）．実際の脳でもこのように外部からの入力によってニューロンの活動を制御していると考えられる．あるニューロングループを活動させるには，そのグループのハブとなるニューロンにのみ強い外部入力を与える方法も

考えられるが今回は採用しない．なぜなら，ネットワーク構造をかなり工夫しないと実現できないためである．

実際にマウスのニューロンの発火頻度がどれくらいなのか [40] を元に表 3.4 に示す．

ニューロンの種類	覚醒時（Hz）	ノンレム睡眠時（Hz）	レム睡眠時（Hz）
興奮性ニューロン	0.76 ± 1.53	0.69 ± 0.86	0.88 ± 1.33
抑制性ニューロン	5.59 ± 7.25	4.69 ± 5.62	4.25 ± 9.43

表 3.4: ニューロンごとの発火頻度の中央値

3.1.3 カルシウムイメージングモデル

スパイクデータからカルシウムイオン濃度を計算する [10] のモデルを用いる:

$$[Ca^{2+}]_{i,t} - [Ca^{2+}]_{i,t-1} = -\frac{\Delta}{\tau}([Ca^{2+}]_{i,t-1} - [Ca^{2+}]_b) + An_{i,t} + \sigma_c\sqrt{\Delta}\epsilon_{i,t},$$

ただし，$[Ca^{2+}]_{i,t}$ をニューロン i の時刻 t でのカルシウムイオン濃度，$[Ca^{2+}]_b$ をカルシウムイオン濃度のベースライン，Δ を時間幅，τ は時定数，A は 1 つのスパイクでのカルシウムイオン濃度の上がり幅，$n_{i,t} \in \{0, 1\}$ はニューロン i の時刻 t でのスパイク，σ_c はノイズの分散，$\epsilon_{i,t}$ は標準正規分布に従う確率変数である．この人工データでは saturation は考えないこととする．

次に，同論文のモデルを使ってカルシウムイオン濃度 $[Ca^{2+}]_{i,t}$ をカルシウムイメージングで計測される蛍光強度 $F_{i,t}$ に変換する：

$$F_{i,t} = \alpha[Ca^{2+}]_{i,t} + \beta + \sigma_F\epsilon_{i,t},$$

α は強度，β はバイアス，σ_F はノイズの分散である．

表 3.5 に使用したパラメータを示す．

$[Ca^{2+}]_b$	Δ	τ	A	σ_c	α	β	σ_F
0.1	0.001	0.5	5.0	1.0	1.0	0	1.0

表 3.5: カルシウムイメージングモデルでのパラメータ

3.1.4 観測モデル

実データは 8 Hz でサンプリングされたデータなので，シミュレーションした蛍光強度を 8 Hz で足し合わせる：

$$x_{i,t'} = \sum_{t=1}^{125} F_{i,t},$$

ここで，t' はサンプリング後の時刻を表す．

上記の方法で作成した人工データ時系列と観測時系列をそれぞれ図 3.1 と図 3.2 に示す．

3.1.5 実験設定

800 個の興奮性ニューロンと 200 個の抑制性ニューロンについてネットワーク構造 S を作成し，式 (3.1) に従って重み付きネットワーク構造 W を作成した．W は全実験を通して固定である．

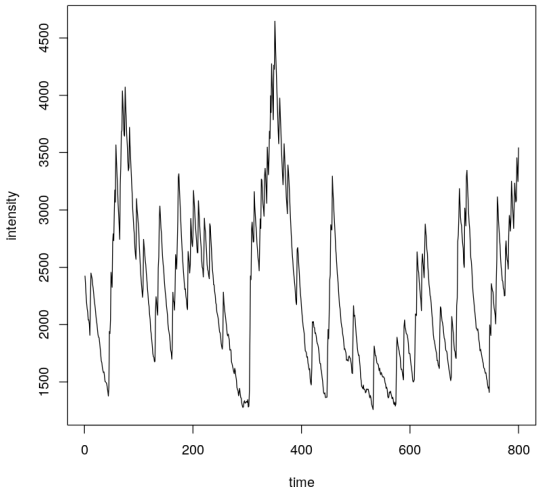


図 3.1: 1 つのニューロンの人工時系列.

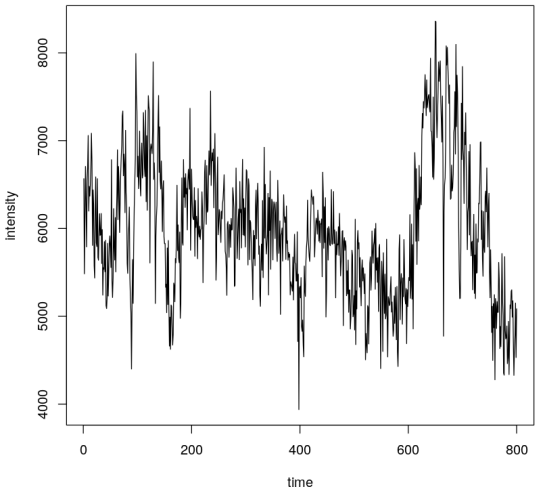


図 3.2: 1 つのニューロンの観測時系列.

W を元にカルシウムイメージングのシミュレーションを行った．1 つのグループに所属するニューロン数は 50〜200 個とした．グループが活動する時間は 5s ごとに変えた．全 1000 個のニューロンのうち，解析に用いるのは固定された 100 個の興奮性ニューロンのみとする．

1. ニューロンは 1 つのグループに必ず所属する
2. ニューロンは 1 つのグループに所属するかグループに所属しない
3. ニューロンは 1 つか 2 つのグループに必ず所属し，同じニューロンが所属しているグループ同士の活動は被らない

3.2 A の推定に関する結果

本節では A の推定に関する結果について述べる．実験は，5s ごとに 1 つか 2 つのグループが活動するデータを 105s シミュレーションさせた結果を用いた．ただし，最初の 5s はシミュレーション数値の安定性のため解析から除外した．

3.2.1 A の一意性

前章で述べた通り，NMF には一意性がないため A にも一意性がない．1 つの人工データについて初期値を変化させて NMF を行うと推定される A は同じではない．A の上三角の要素について 1 と推定された頻度を調べる．収束性と前章で述べたように寄与率が D の空間で取りうる値の範囲を狭めるため，C の行和を 1 とする制約を入れている．X に正規化を加えなかった結果を図 3.3 に，X に行和 1 の正規化を加えた結果を図 3.4 に示す．X に行和 1 の正規化を加えると，D の自由度は下がる．

A に一意性がある場合は，ある A の要素が 1 と推定される回数は 0 か 1000 になる．結果より，そうなっていないので A に一意性がないのがわかる．また，図 3.4 より X を正規化した方がよりばらつきが大きくなっているのがわかる．これは，正規化をしない方が X の中の大きな値に推定が引っ張られ，同じ局所解に落ちやすくなっているからだと考えられる．また，この正規化の仕方は物理的には，ある期間にニューロンの活動の総和が一定であるという意味をもつ．実際はそうではないので，今回の正規化の仕方はこの実験のみに留める．

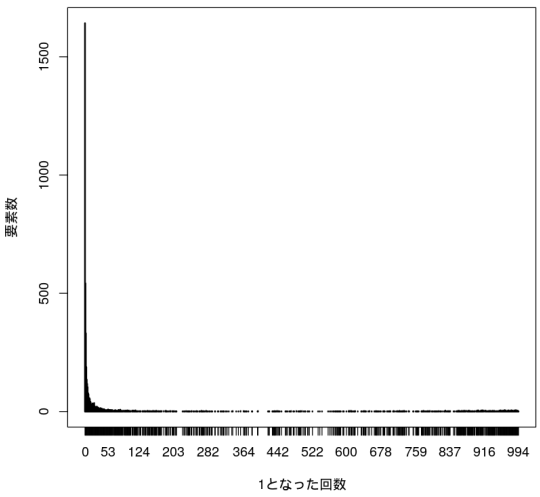


図 3.3: X を正規化せずに初期値を 1000 回変化させて NMF から A を推定し, 各要素について 1 と推定された頻度をプロットした.

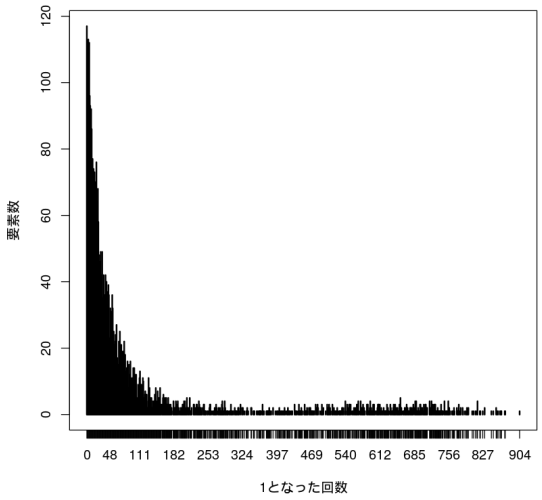


図 3.4: X を行和 1 で正規化して初期値を 1000 回変化させて NMF から A を推定し, 各要素について 1 と推定された頻度をプロットした.

3.2.2 手法の比較

NMF，PCA，ICA，logistic regression，glasso の性能の比較を行った．Logistic regression と glasso については筆者の卒論を参照されたい．どちらも時間窓をスライドさせてネットワークを行う．今回の実験では時間窓を 40，スライド幅を 20 とした．Glasso のハイパーパラメータを $\rho = 0.3$ とした．一回でもエッジが張られたニューロン同士は同じグループとして推定量 A と同じ行列を作成した．

PCA と ICA は NMF と同じく一般化線形成分分析の手法 [41] である．PCA と ICA では D に相当する行列で NMF と同じく推定量 A を作成する．

2 のタイプについて 100 種類のデータを生成し，A を閾値 0.5 で切って $0, 1^{I \times I}$ の行列にした時の F1 score を比較した．図 3.5 より，NMF の精度が最も高いことが分かる．NMF の

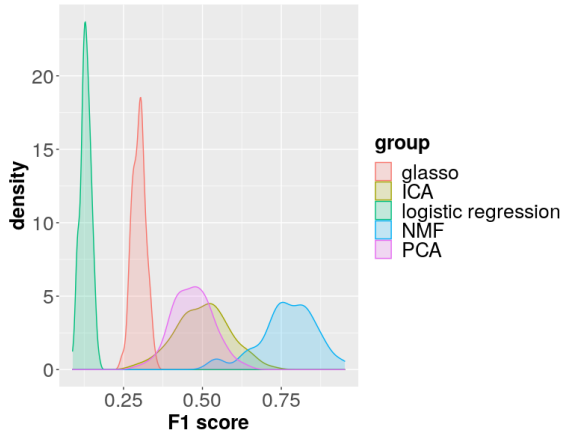


図 3.5: NMF，PCA，ICA，logistic regression，glasso の F1 score の密度分布

非負制約がデータに合っているためだと思われる．Glasso と logistic regression についてはニューロングループを推定するという実験設定はやや不利で合った．各窓ごとのニューロンネットワークの活動を反映している可能性があるので悪い手法とはいえない．

3.2.3 モデル平均の有用性

バギングの有用性を確認するために各人工データについて，1 回 NMF を行った結果，30 回初期値を変えた結果，30 回ブートストラップした結果の F1 score を 図 3.6 に示す．なお，NMF は 20 回初期値を変化させて再構成ごさが最小となる結果を 1 回の結果として用いた．

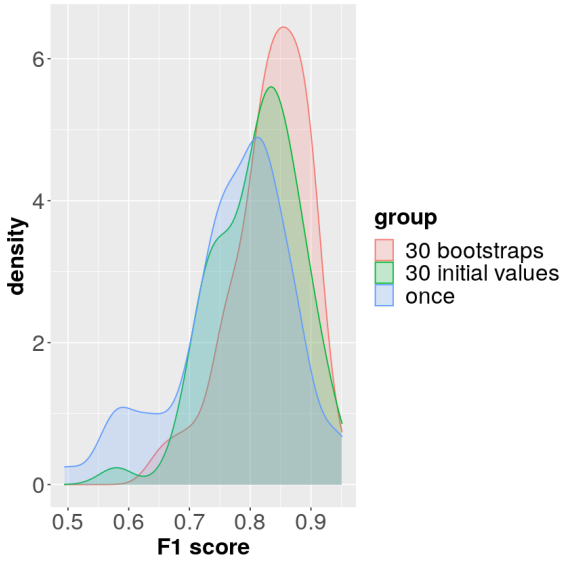


図 3.6: NMF を 1 回行った時の A，30 回初期値を変えた A の平均，30 回ブートストラップを行った A の平均それぞれの F1 score の分布．

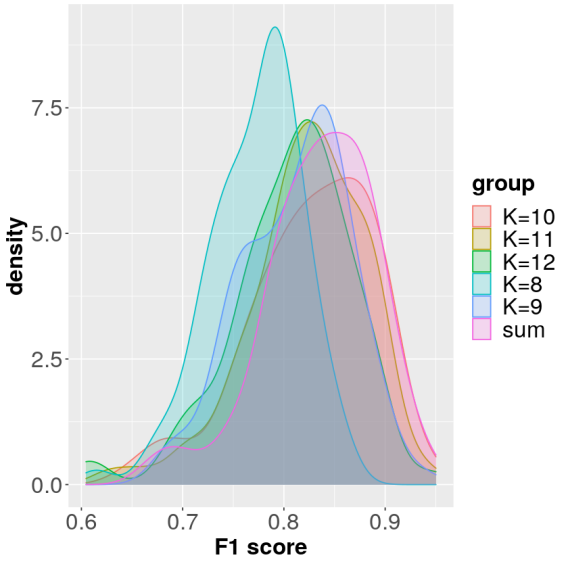


図 3.7: 基底数ごとにブートストラップを行った時の A の F1 score と全ての A の平均をとった時の F1 score の分布．

図 3.6 より，ブートストラップを行った方が精度がよくなることがわかる．

基底数別に推定された A の平均をとった時の F1 score を 図 3.7 に示す．これより，真の基底数周りの A の平均をとることである程度の精度は保たれることがわかる．

3.2.4 NMF の基底数

NMF の基底数を決める方法をいくつか試した．人工データ 86 個について Brunet らと Ubrau らの方法で基底数を決めた時に各基底数が何回選ばれるかを 図 3.8．図 3.9 に示す．Brunet らの方法では真の基底数 10 に近い基底数が選ばれているが，Ubaru らの方法では小さい基底数が選ばれる傾向にあった．

また，1 つの人工データについて AIC と AICc を計算した結果を 図 3.10，図 3.11 に示す．どちらも基底数が大きくなるごとに減少する傾向があった．他の人工データについても同様の結果であった．

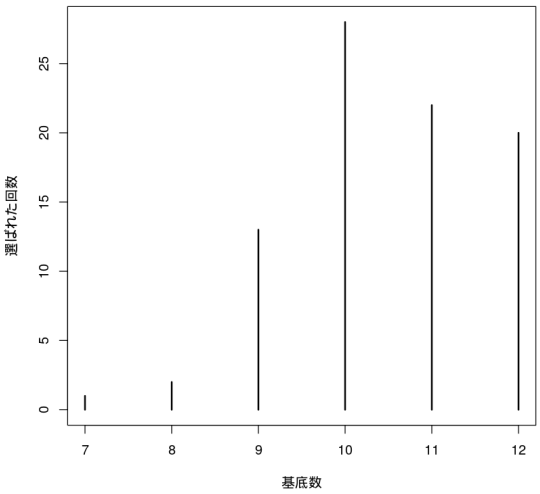


図 3.8: Brunet らの方法で基底数を決めた時に各基底数が選ばれた回数（真の基底数は 10）.

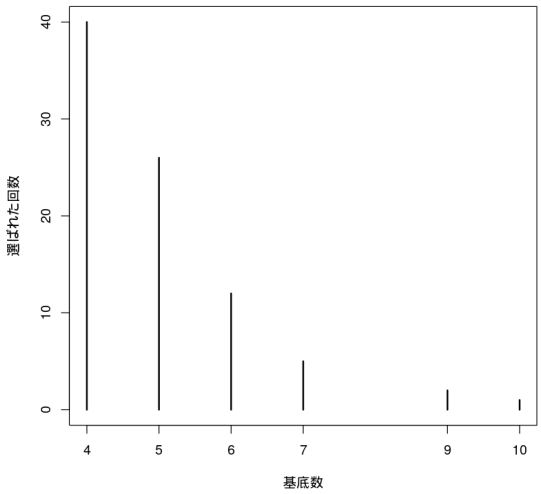


図 3.9: Ubaru らの方法で基底数を決めた時に各基底数が選ばれた回数（真の基底数は 10）.

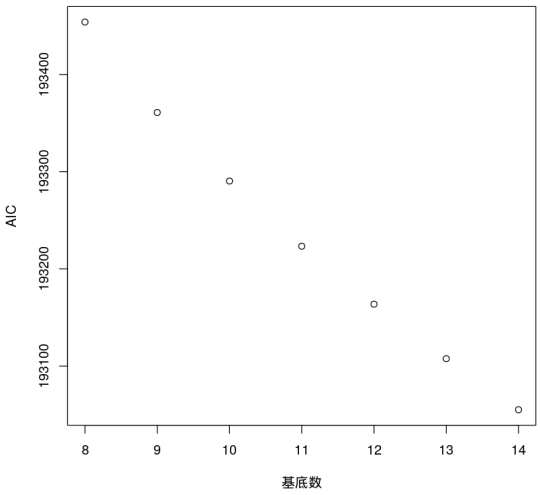


図 3.10: あるデータについて AIC を計算した時の結果.

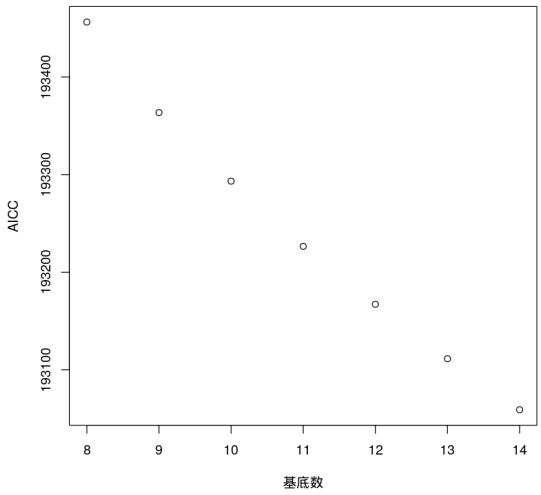


図 3.11: あるデータについて AICc[42] を計算した時の結果.

3.3 A のクラスタリングに関する結果

3.3.1 クラスタ数の決定

スペクトラルクラスタリングでクラスタ数を決定する方法はいくつかあるが，その中でも小さい固有値を見る方法と，Gap 統計量を用いた方法を試す．A と比較する類似度行列として，データの相互相関行列と分散共分散行列を用いる．

第4章 実データ解析

4.1 実データ

4.2 結果

第5章 その他の検討事項

本節では，データの特徴を考慮して考案したNMFの制約とモデルエビデンスの計算方法を紹介する．

5.1 バイアス除去

ニューロンごとに発現している蛍光タンパク質の量や細胞の大きさが異なる．そのため，ニューロンごとのバイアスが観測データに載っていると考え，バイアスを除去する方法を試した．この時の数理モデルは以下のようになる：

$$X = Y + H + B,$$

ただし，$B \in \mathbb{R}_+^{I\times J}$ は行ごとに同じ数値が入ったバイアス行列である．

バイアスの推定方法は，D に1列を足し，C に $\mathbf{1}$ の1行を足してNMFを更新する．D の列にバイアスが推定されることを期待した．

簡単な人工データ実験を行った結果，足したバイアスよりも大きいバイアスが推定されてしまうことがわかった．また，そもそもの数理モデルが異なると考え直した．

蛍光タンパク質の量や細胞の大きさが異なるというモデルは以下のように表される：

$$X = A(Y + H),$$

ただし，$A \in \mathbb{R}_+^{I\times I}$ は対角行列である．

5.2 重複除去

置いた仮定では，あるニューロンが複数のグループに所属する時，グループの活動は被らないとしている．しかし，NMFの推定時にそのような制約は入れていないので，NMFで推定した結果この仮定が破られているようであれば制約は入れなければならない．

以下の目的関数を考えた：

$$\underset{D\geq 0, C\geq 0}{\arg\min}\ ||X - DC||_F^2 - \lambda \sum_{k=1}^{K}\sum_{l\neq k}^{K} \left(||d_{:l} - d_{:k}||_1 ||c_{l:} - c_{k:}||_1\right). \tag{5.1}$$

更新則を導出する．参考にしたのは [43] である．式 (5.1) のLagrange関数 L は，

$$L = \mathrm{Tr}(X^T X) - 2\mathrm{Tr}(X^T DC) + \mathrm{Tr}(C^T D^T DC) - \mathrm{Tr}(\Phi_C C^T) - \mathrm{Tr}(\Phi_D D^T) - \lambda \mathrm{Tr}(F^T C H^T S^T DF),$$

であり，KKT条件は，

$$\begin{aligned}
\frac{\partial L}{\partial C} = \frac{\partial L}{\partial D} &= 0 \\
D &\geq 0 \\
C &\geq 0 \\
\Phi_C &\geq 0 \\
\Phi_D &\geq 0 \\
\Phi_C C = \Phi_D D &= 0
\end{aligned}$$

である．ただし，Φ_D と Φ_C はそれぞれ $D \geq 0$，$C \geq 0$ に対する Lagrange 乗数で，$F \in [0,1]^{K\times(K-1)!}$ は 2 つの時間の組み合わせを表現した以下のような行列である：

$$F = \begin{pmatrix} 1 & 1 & \dots & 0 & \dots & 0 \\ -1 & 0 & \dots & 1 & \dots & 0 \\ 0 & -1 & \dots & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & \dots & -1 \end{pmatrix}.$$

また，$S = sign(DF)$，$H = sign(F^T C)$ とおく．

D を求めるには以下の式を解く：

$$\frac{\partial L}{\partial D} = -2XC^T + 2DCC^T - \Phi_D - \lambda SHC^T FF^T = 0.$$

これを求めると D の要素の更新は以下である：

$$d_{ik} \leftarrow d_{ik}\frac{2[XC^T]_{ik} + \lambda[SHC^T FF^T]^+_{ik}}{2[DCC^T]_{ik} - \lambda[SHC^T FF^T]^-_{ik}},$$

ただし，$[\cdot]^+$ は行列の中の正の要素，$[]^-$ は負の要素である．

同様に，C の要素の更新は以下の通りである：

$$c_{kj} \leftarrow c_{kj}\frac{2[D^T X]_{kj} + \lambda[FF^T D^T SH]^+_{kj}}{2[D^T DC]_{kj} - \lambda[FF^T D^T SH]^-_{kj}}.$$

5.3 時間方向への制約

カルシウムイメージングデータはスパイク情報を反映するのが遅く，一度上がった蛍光強度は緩やかに下がっていく．そのため，NMF で分解を行う際も，行列 C の時間方向に前時刻の値と近くなるような制約を入れることでより正確なニューロングループの抽出が行えると考えられる．C の偶数列を前後の列の平均とする NMF も提案されている [44]．しかし，これはかなりスムースになる制約だと考えられる．そこで，以下のような制約を加えた目的関数が考えられる：

$$\underset{D\geq 0, C\geq 0}{\arg\min}\ ||X - DC||_F^2 + \lambda \sum_t ||c_{:t} - c_{:t-1}||_1.$$

これは fused lasso [45] と同じような制約である．

更新則は，D はユークリッド型 NMF と同じだが C は異なる．更新則は以下である：

$$c_{kj} \leftarrow c_{kj}\frac{2[D^T X]_{kj} - s_{kj}}{2[D^T DC]_{kj}},$$

ただし，$s_{:j} = sign(c_{:j} - c_{:j-1})$ であり，1 列目のみ $s_{:1} = \mathbf{0}$ である．

5.4 NMF のモデルエビデンス

NMF の基底数を決める際にブートストラップから計算されたモデルエビデンスを用いることを考える．BIC は最尤推定量の尤度からモデルエビデンスを近似して扱っている．ブートストラップによってモデルエビデンスを近似計算できると考えられる．本節ではその説明をする．

今回の問題を Bayesian model averaging の枠組みに当てはめると，

$$p(A|X) = \sum_k p(A|\mathcal{M}_k, X)p(\mathcal{M}_k|X)$$

で $p(A|X)$ を求めることになる．今回は異なる基底数の NMF から求まった A をモデルの事後確率で重み付けて足し合わせることになる．

モデルの事後確率は

$$p(\mathcal{M}_k|X) = \frac{m_k p(\mathcal{M}_k)}{\sum_l m_l p(M_l)}$$

で表される．ただし，

$$m_k = \int p(X|Y_k, \mathcal{M}_k)p(Y_k|\mathcal{M}_k)dY_k$$

である．これはモデルエビデンスや marginal likelihood と呼ばれる．また，$p(\mathcal{M}_k)$ はモデルが正しい確率である．

ある条件下で，母数の事後確率の密度関数はブートストラップによる最尤推定量の分布と同じになる．そのため，ブートストラップサンプルから推定した A の分布をもとに事後確率を計算すれば良い．

現在の設定では $p(\mathcal{M}_k)$ は一様なため，エビデンスを見る必要がある．

5.4.1 ラプラス近似

エビデンスの計算には事前分布を決めなければならず，積分も計算しなければいけない．そこでラプラス近似により m_k は以下のような $\hat{m}_k$ で近似できる．

$$\log \hat{m}_k = \log p(X|\hat{Y}_k, \mathcal{M}_k) - \frac{d_k}{2}\log n$$

ここで，d_k は母数の数（$I \times K + K \times J$），n は観測データ数（J）である．この近似を使ってログ Bayes 因子を計算したのが BIC である．

$\log p(X|\hat{Y}_k, \mathcal{M}_k)$ は初期値を変えて NMF を行い，尤度が最も大きくなった対数尤度を用いる．また，$\sigma^2 = Var(X - Y_k)$ として計算する．

5.4.2 ブートストラップによる近似

ブートストラップの推定量の分布は最尤推定量の分布を近似する（ブートストラップサンプルが生成されたパラメータ分バイアスは乗る）ので，m_k は以下のように近似できる．

$$\begin{aligned} m_k &= \int p(X|Y_k, \mathcal{M}_k)p(Y_k|\mathcal{M}_k)dY_k \\ &\sim \frac{1}{B}\sum_b \prod_{i,j} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{([Y_k^b]_{ij} - X_{ij}^b)^2}{2\sigma^2}\right) \end{aligned}$$

ここで，Y_k^b はブートストラップサンプル b から計算された Y_k で，B はブートストラップサンプル数である．また，$\sigma^2 = Var(X^b - Y_k^b)$ として計算する．

ラプラス近似は本来 NMF には不適切なので，ブートストラップによる近似でモデルエビデンスを計算することができれば基底数を推定できるかもしれない．

第6章 結論

考察でいいのか？

本論文では，数理モデルと人工データ実験に基づいて NMF がニューロンのグループ推定に有効であることを示した．NMF では基底数に寄らない推定量を用いて，バギングによって基底数を決めなければいけない問題を緩和した．人工データ実験によってどの程度の情報が取れるかも示した．

謝辞

本研究を進めるにあたり指導教員の村田昇先生には多くのご指導をいただきました．また，今回用いた技術に関する問題の面白さを知ることができました．深く感謝いたします．赤穂先生，日野先生，有竹さんにも貴重なお時間を割いて助言をいただき，質問に答えていただきました．深く感謝いたします．また，貴重なデータを提供してくださり，カルシウムイメージングやニューロンに関する知見について教えていただいた筑波大学の柳沢研究室の上田助教に感謝いたします．

研究室の先輩，同期，後輩にはゼミなどを通じて多くのアドバイスをいただきました．特に同期の皆様には細かい相談や精神面でも多く支えていただきました．ありがとうございました．

参考文献

[1] J. A. Hobson, “Sleep is of the brain, by the brain and for the brain”, *Nature*, vol. 437, no. 7063, pp. 1254–1256, Oct. 2005.

[2] T. Kanda, N. Tsujino, E. Kuramoto, Y. Koyama, E. A. Susaki, S. Chikahisa, and H. Funato, “Sleep as a biological problem: an overview of frontiers in sleep research”, *The Journal of Physiological Sciences*, vol. 66, no. 1, pp. 1–13, Jan. 2016.

[3] T. Kanda, T. Miyazaki, and M. Yanagisawa, “Imaging Sleep and Wakefulness”, in *Make Life Visible*, Springer Singapore, 2020, pp. 169–178.

[4] E. M. Izhikevich, J. A. Gally, and G. M. Edelman, “Cerebral Cortex V 14 N 8 Spike-timing Dynamics of Neuronal Groups”, *Cortex August*, vol. 14, pp. 933–944, 2004.

[5] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon, “Putting big data to good use in neuroscience.”, *Nature neuroscience*, vol. 17, no. 11, pp. 1440–1, Nov. 2014.

[6] 中村 健, “神経細胞内局所的カルシウム濃度変化のリアルタイムイメージング法”, *Folia Pharmacol. Jpn*, vol. 121, no. 5, pp. 357–364, 2003.

[7] G. Q. Bi and M. M. Poo, “Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type.”, *The Journal of neuroscience: the official journal of the Society for Neuroscience*, vol. 18, no. 24, pp. 10 464–10 472, 1998.

[8] 平 理一郎, “脳神経計算原理の解明を目指した 2 光子多細胞イメージングの情報技術展開［II・完］”, *Animal Genetics*, vol. 101, no. 9, pp. 926–931, 2018.

[9] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, “Review of methods for functional brain connectivity detection using fMRI”, *Computerized Medical Imaging and Graphics*, vol. 33, no. 2, pp. 131–139, Mar. 2009.

[10] J. T. Vogelstein, B. O. Watson, A. M. Packer, R. Yuste, B. Jedynak, and L. Paninski, “Spike Inference from Calcium Imaging Using Sequential Monte Carlo Methods”, *Biophysical Journal*, vol. 97, no. 2, pp. 636–655, Jul. 2009.

[11] Y. Mishchencko, J. T. Vogelstein, and L. Paninski, “A Bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data”, *The Annals of Applied Statistics*, vol. 5, no. 2B, pp. 1229–1261, 2011.

[12] O. Stetter, D. Battaglia, J. Soriano, and T. Geisel, “Model-Free Reconstruction of Excitatory Neuronal Connectivity from Calcium Imaging Signals”, *PLoS Comput Biol*, vol. 8, no. 8, p. 1 002 653, 2012.

[13] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, and R. Yuste, “Synfire chains and cortical songs: temporal modules of cortical activity.”, *Science (New York, N.Y.)*, vol. 304, no. 5670, pp. 559–64, Apr. 2004.

[14] J. Mölter, L. Avitan, and G. J. Goodhill, “Detecting neural assemblies in calcium imaging data”, *BMC Biology*, vol. 16, no. 1, p. 143, Nov. 2018.

[15] K. Ghandour, N. Ohkawa, C. C. A. Fung, H. Asai, Y. Saitoh, T. Takekawa, R. Okubo-Suzuki, S. Soya, H. Nishizono, M. Matsuo, M. Osanai, M. Sato, M. Ohkura, J. Nakai, Y. Hayashi, T. Sakurai, T. Kitamura, T. Fukai, and K. Inokuchi, "Orchestrated ensemble activities constitute a hippocampal memory engram", *Nature Communications*, vol. 10, no. 1, pp. 1–14, Dec. 2019.

[16] L. Sjulson and G. Miesenböck, "Optical Recording of Action Potentials and Other Discrete Physiological Events: A Perspective from Signal Detection Theory", *Physiology*, vol. 22, no. 1, pp. 47–55, Feb. 2007.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[18] T. Boongoen and N. Iam-On, *Cluster ensembles: A survey of approaches with recent extensions and applications*, May 2018.

[19] X. Fu, K. Huang, N. D. Sidiropoulos, and W. K. Ma, "Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications", *IEEE Signal Processing Magazine*, vol. 36, no. 2, pp. 59–80, 2019.

[20] B. Dong, M. M. Lin, and M. T. Chu, "Nonnegative rank factorization—a heuristic approach via rank reduction", *Numerical Algorithms*, vol. 65, no. 2, pp. 251–274, Feb. 2014.

[21] B. Efron, "Bootstrap Methods: Another Look at the Jackknife", *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, Jan. 1979.

[22] L. Breiman, "Bagging Predictors", Tech. Rep., 1996, pp. 123–140.

[23] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[24] L. I. Kuncheva and D. P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1798–1808, 2006.

[25] l. wasserman larry, "bayesian model selection and model averaging", *journal of mathematical psychology*, vol. 44, no. 1, pp. 92–107, Mar. 2000.

[26] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[27] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 12, pp. 4164–4169, Mar. 2004.

[28] S. Ubaru, K. Wu, and K. E. Bouchard, "UoI-NMF cluster: A robust nonnegative matrix factorization algorithm for improved parts-based decomposition and reconstruction of noisy data", in *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, vol. 2017-December, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 241–248.

[29] S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise, "Stability-driven nonnegative matrix factorization to interpret Spatial gene expression and build local gene networks", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 16, pp. 4290–4295, 2016.

[30] L. N. Hutchins, S. M. Murphy, P. Singh, and J. H. Graber, “Position-dependent motif characterization using non-negative matrix factorization”, *BIOINFORMATICS ORIGINAL PAPER*, vol. 24, no. 23, pp. 2684–2690, 2008.

[31] A. T. Cemgil, “Bayesian inference for nonnegative matrix factorisation models”, *Computational Intelligence and Neuroscience*, vol. 2009, 2009.

[32] A. Ben-Hur, A. Elisseeff, and I. Guyon, “A stability based method for discovering structure in clustered data”, Tech. Rep.

[33] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic”, Tech. Rep.

[34] U. Von Luxburg, “A Tutorial on Spectral Clustering”,

[35] M. K. Goldberg, M. Hayvanovych, and M. Magdon-Ismail, “Measuring similarity between sets of overlapping clusters”, in *Proceedings - SocialCom 2010: 2nd IEEE International Conference on Social Computing, PASSAT 2010: 2nd IEEE International Conference on Privacy, Security, Risk and Trust*, 2010, pp. 303–308.

[36] D. J. Watts and S. H. Strogatz, “Collective dynamics of ’small-world’ networks”, *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[37] J.-S. Jouhanneau, J. Kremkow, A. L. Dorrn, and J. F. Poulet, “In Vivo Monosynaptic Excitatory Transmission between Layer 2 Cortical Pyramidal Neurons”, *Cell Reports*, vol. 13, no. 10, pp. 2098–2106, Dec. 2015.

[38] C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter, *Pyramidal cell communication within local networks in layer 2/3 of rat neocortex*, Aug. 2003.

[39] E. M. Izhikevich, “Simple Model of Spiking Neurons”, *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 14, no. 6, 2003.

[40] B. O. Watson, D. Levenstein, J. P. Greene, J. N. Gelinas, and G. Buzsáki, “Network Homeostasis and State Dynamics of Neocortical Sleep”, *Neuron*, vol. 90, no. 4, pp. 839–852, May 2016.

[41] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Nonnegative Matrix and Tensor Factorizations*. Chichester: Wiley Publishing, 2009.

[42] M. R. Symonds and A. Moussalli, *A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike’s information criterion*, Aug. 2011.

[43] M. Babaee, S. Tsoukalas, M. Babaee, G. Rigoll, and M. Datcu, “Discriminative Nonnegative Matrix Factorization for dimensionality reduction”, *Neurocomputing*, vol. 173, pp. 212–223, Jan. 2016.

[44] V. C. Cheung, K. Devarajan, G. Severini, A. Turolla, and P. Bonato, “Decomposing time series data by a non-negative matrix factorization algorithm with temporally constrained coefficients”, in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-Novem, Institute of Electrical and Electronics Engineers Inc., Nov. 2015, pp. 3496–3499.

[45] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, “Sparsity and smoothness via the fused lasso”, Tech. Rep., 2005, pp. 91–108.