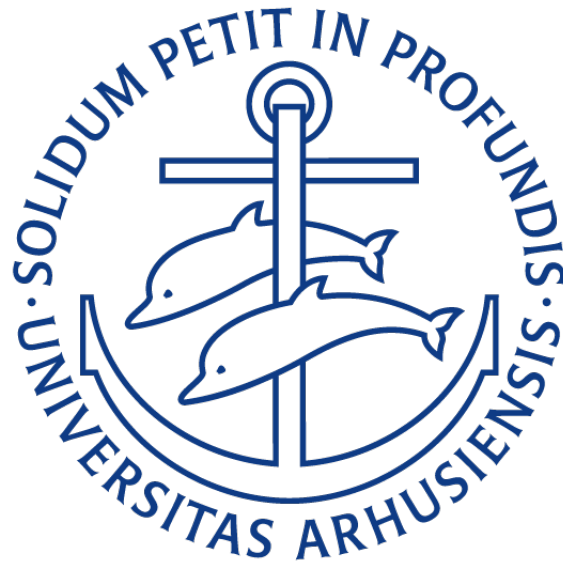


NON-NEGATIVE MATRIX FACTORIZATION
FOR CANCER GENOMICS

BY RAGNHILD LAURSEN



PHD DISSERTATION

DEPARTMENT OF MATHEMATICS
AARHUS UNIVERSITY

SEPTEMBER 2024

Non-negative matrix factorization for cancer genomics

PhD dissertation by
Ragnhild Laursen

Department of Mathematics, Aarhus University
Ny Munkegade 118, 8000 Aarhus C, Denmark

Supervised by
Professor Asger Hobolth

Submitted to Graduate School of Natural Sciences, Aarhus, September 17, 2024



Preface

This dissertation marks the end of my studies as a PhD student at the Department of Mathematics, Aarhus University. It was written under supervision from Professor Asger Hobolth in the period from January 2020 to September 2024.

The dissertation consist of an introductory review of the field and the following five self-contained papers:

- Paper A** A sampling algorithm to compute the set of feasible solutions for non-negative matrix factorization with an arbitrary rank
SIAM Journal on Matrix Analysis and Applications (Volume 43)
- Paper B** Model selection and robust inference of mutational signatures using Negative Binomial non-negative matrix factorization
BMC bioinformatics (Volume 24)
- Paper C** Flexible model-based non-negative matrix factorization with application to mutational signatures
Statistical Applications in Genetics and Molecular Biology (Volume 23)
- Paper D** Integration of opportunities and parametrized signatures to improve mutational signatures estimation
Paper draft
- Paper E** A simple extension of non-negative matrix factorization to find structures and spatially variable genes in multiple tissues
Paper draft

The first three papers are included in the published format of their respective journals, while the last two are paper drafts that are with-in the final stages before submission.

Paper A and C are joint work with Asger Hobolth, where I contributed extensively to the research, analysis and writing of these two papers. Paper B and D were written jointly with Marta Pelizzola and Asger Hobolth, where Marta Pelizzola and I contributed equally to the research, analysis and writing of these two papers. Additionally, Lasse Maretty contributed with biological insights and data curation for Paper B and D. In accordance with GSNS rules, most of Paper A and prelimi-

nary results of Paper B and C were used in my progress report for the qualifying examination in January 2022.

The last Paper E is a result of my research stay in autumn 2023, where I visited Professor Barbara Engelhardt at Gladstone Institutes in San Francisco, USA. It was written in collaboration with Barbara Engelhardt, where I undertook the majority of the research, analysis and writing.

The first chapter of the dissertation introduces the field of Non-negative Matrix Factorization (NMF) and its many statistical properties and challenges with a particular focus on applications to cancer genomics. It explains the current literature and how the five papers above contribute to the field. Paper A-D are all based on applications to mutational counts in cancer tumors, where Paper E uses applications to gene expression of cells in tissue slices.

Be aware that the notation of NMF in the introductory chapter is equivalent to the one found in Paper B, D and E, but Paper C deviates slightly and Paper A has its own separate notation. However, each paper explains its respective notation, leading to some repetitions between many of the papers, but also helps to avoid confusion for the papers where the notation deviates.

★ ★ ★

My PhD studies have been a long and rewarding journey, filled with both obstacles and triumphs along the way. It has been joyful thanks to the many fantastic people I have met along the way and the end has only been reached thanks to the invaluable support of many different people.

First of all I am grateful to my supervisor Asger Hobolth, who gave me the offer and opportunity to pursue a PhD degree. I appreciate our numerous discussions, your guidance and the endless positive support that you met me with in my many frustrations during my PhD studies. After our meetings I have always felt energized and filled with new ideas to move forward.

I also owe special thanks to Marta Pelizzola for choosing to come to Aarhus University from Vienna at a time when a new collaborator was essential in keeping me motivated and lifting my spirits. I highly appreciate all the great moments we have shared and our many technical discussion during my studies.

In the autumn of 2023 I visited Professor Barbara Engelhardt and her group, *the beehive*, in San Francisco, which is split between Stanford University and Gladstone Institutes. I am grateful to Barbara for allowing me to visit and for her support, both in research and financially, during my stay. I would especially like to thank Gladstone Institutes and her group there for being extremely welcoming and for involving me in their social and scientific activities.

All my colleagues at the department of mathematics also deserve a huge thanks for creating a great working environment. In particular I would like to thank Lota Copic, Jacob Thøstesen, Kenneth Borup, Helene Hauschultz, Rikke Eriksen and Anton Tjepner for cake breaks, friday beers and great discussion that gave me

renewed energy for my studies. This also includes Emil Dare, with whom I am happy to have shared all my years of study. I appreciate that you always checked in on me to ensure my spirits were high and that I never missed my daily dose of cake. I also thank Lars Madsen for his help in typesetting this thesis in L^AT_EX, which has greatly improved its appearance.

Finally, my family and friends deserve a huge thanks for their constant support and their questions and efforts in trying to understand my studies. Here, an extraordinary attention needs to be given to my husband Jeppe for being my biggest cheerleader and rock to get me through these studies. A significant part of my studies has included programming, and here he has been invaluable in various ways to improve my coding skills and prevented my computer from being thrown out the window during many frustrating moments. This PhD would surely not have been possible without your support.

Ragnhild Laursen
Aarhus, September 2024

Contents

Preface	i
Abstract	ix
Resumé	xi
Introduction	1
1 Non-negative Matrix Factorization (NMF)	4
2 Non-uniqueness	5
2.1 Requirements for a unique factorization	5
2.2 Finding the set of feasible solutions	6
Sampling algorithm (Paper A)	7
Perspectives	8
3 Distributional assumption	8
3.1 Poisson NMF	9
3.2 Negative Binomial NMF	10
Patient specific dispersion (Paper B)	10
Perspectives	11
4 Choosing the rank	12
4.1 Information Criteria	12
4.2 Cross-validation	13
SigMoS (Paper B)	14
4.3 Alternative methods	14
Perspectives	15
5 Regularization	15
5.1 Sparseness constraints of factors	15
5.2 Parametrizing entries in the factors	16
Flexible parametrization (Paper C)	17
Perspectives	18
6 Opportunities	18
6.1 Background signature	19
6.2 Probabilistic inclusion (Paper D)	19
Perspectives	20
7 Spatial data	21
7.1 Extension of NMF to model spatial data	21

	Neighborhood NMF (Paper E)	22
	Perspectives	23
8	Computational implementation	24
8.1	Floating point precision	24
8.2	Initialization	24
8.3	Stopping criteria	25
8.4	Boosting R Performance	25
	Public packages	26
	Public repositories	26
	References	27

**Paper A A sampling algorithm to compute the set of feasible solutions
for non-negative matrix factorization with an arbitrary rank 33**
by Ragnhild Laursen and Asger Hobolth

1	Introduction	35
2	The sampling algorithm	38
2.1	SFS for $N = 2$	38
2.2	Sampling the SFS for an arbitrary rank	39
2.3	Sampling a value λ in Λ_{ij}	39
2.4	Defining the size of the SFS and the stopping criteria	39
3	Applications and comparison with polygon inflation algorithm	40
3.1	Singular value decomposition to represent the SFS	40
3.2	Polygon inflation algorithm	43
4	Further analysis of the sampling algorithm for arbitrary rank	44
4.1	Influence of tuning parameter β and rank N	44
4.2	Running time	46
5	Taking advantage of random initialization	46
5.1	Finding the global minimum	47
5.2	Finding more than N subsets of the SFS	47
5.3	Noise in the data	47
6	Conclusion	49
	References	50

**Paper B Model selection and robust inference of mutational signatures
using Negative Binomial non-negative matrix factorization 53**
by Marta Pelizzola, Ragnhild Laursen and Asger Hobolth

1	Introduction	56
2	Results	58
2.1	Simulation Study	58
2.2	Breast cancer data	63
2.3	Prostate cancer data	65
3	Discussion	66

4	Methods	68
4.1	Negative Binomial model for mutational counts	68
4.2	Patient specific NB_N -NMF	69
4.3	Estimating the number of signatures	72
	References	76
	Supplementary material	79

Paper C Flexible model-based non-negative matrix factorization with application to mutational signatures **87**
by Ragnhild Laursen, Lasse Maretty and Asger Hobolth

1	Introduction	89
2	Determining the mutational signatures	91
2.1	Non-negative matrix factorization	91
2.2	Parametrization of a mutational signature	91
3	Results	94
3.1	Simulation study	94
3.2	Analysis of BRCA	95
3.3	Analysis of Liver data	100
3.4	Analysis of UCUT data	100
3.5	Choosing the number of signatures and parametrization	101
4	Methods	105
4.1	EM-algorithm for traditional non-negative matrix factorization	105
4.2	EM-algorithm for parametric non-negative matrix factorization	107
5	Discussion	109
	References	110

Paper D Integration of opportunities and parametrized signatures to improve mutational signatures estimation **111**
by Ragnhild Laursen, Marta Pelizzola, Lasse Maretty and Asger Hobolth

1	Introduction	113
2	Methods	116
2.1	Negative Binomial NMF	117
2.2	NB-NMF with opportunities	117
2.3	NB-NMF with opportunities and parametrized signatures	119
3	Results	120
3.1	Results on the breast cancer data set	120
3.2	Results on the liver cancer data set	124
4	Discussion	125
	Appendices	126
	References	129

Paper E	A simple extension of non-negative matrix factorization to find structures and spatially variable genes in multiple tissues	133
	<i>by Ragnhild Laursen and Barbara E Engelhardt</i>	
1	Introduction	135
1.1	Overview of neighborhood nonnegative matrix factorization (NNMF)	137
2	Results	138
2.1	10x Visium data from human brain	139
2.2	MERFISH data of mouse brain	142
2.3	MERFISH data of colon cancer	144
3	Method	144
3.1	Initialization	147
3.2	Length scale	147
3.3	Multiple slices and batching of large datasets	147
3.4	Computational efficiency	148
3.5	Weighting of genes	149
3.6	Avoiding celltype bias	149
4	Discussion	149
	References	150

Abstract

Non-negative Matrix Factorization (NMF) is one of the most popular methods used to analyze high-dimensional count data. It is used in various fields of research, and this dissertation serves as a guideline for analyzing count data with NMF, with a particular focus on its applications to mutational counts from tumors in cancer. The statistical properties and challenges of NMF are explained, along with proposed solutions.

In broad terms, NMF reduces high-dimensional count data into a factorization of two smaller non-negative matrices, with the goal of retaining the essential information in the data. To achieve this, several challenges of NMF need to be considered. These include the possible non-uniqueness of the factorization, the effects of the underlying distributional assumptions, how to choose the rank of the factorization, and how to regularize the results. These challenges and their interconnections are elaborated upon in the introduction of this dissertation, where Paper A-D explore solutions to these issues, with a particular focus on applications to mutational counts in cancer. The final paper discusses the application of NMF to spatial count data, where the locations of the observations are known. This is relevant to spatial transcriptomics data, where both the location and gene expression of single cells are known within tissue slices.

Although all the applications discussed focus on cancer genomics data, this dissertation should equip the reader to use NMF to analyze any type of count data and obtain an interpretable and robust factorization.

Resumé

Ikke-Negativ Matrix Faktorisering (NMF) er en af de mest populære metoder, der anvendes til at analysere høj-dimensionelle tælldata. Det bruges i forskellige forskningsfelter, og denne afhandling fungerer som en vejledning til analyse af tælldata med NMF, med særligt fokus på anvendelsen til mutationsdata fra kræft. De statistiske egenskaber og udfordringer ved NMF forklares, sammen med foreslåede løsninger.

I brede træk reducerer NMF høj-dimensionelle tælldata til en faktorisering af to mindre ikke-negative matricer med det formål at bevare den væsentlige information i dataene. For at opnå dette skal flere udfordringer ved NMF overvejes. Disse inkluderer den mulige ikke-entydighed af faktoriseringen, effekten af de underliggende fordelingsantagelser, hvordan man vælger rangen af faktoriseringen, samt hvordan man regulerer resultaterne. Disse udfordringer og deres sammenhænge uddybes i introduktionen af denne afhandling, hvor artiklerne A-D undersøger løsninger på disse problemer med særligt fokus på anvendelsen til mutationsdata fra kræft. Den sidste artikel diskuterer anvendelsen af NMF på rumlige tælldata, hvor observationernes placeringer er kendt. Dette er relevant for rumlig transkriptomisk data, hvor både placeringen og genudtrykket af enkelte celler er kendt i vævssnit.

Selvom alle de diskuterede anvendelser fokuserer på kræft genomisk data, bør denne afhandling give læseren de nødvendige værktøjer til at bruge NMF til at analysere enhver form for tælldata og opnå en fortolkelig og robust faktorisering.

Introduction

In recent years, the field of machine learning has become increasingly popular by enabling data-driven predictions and decision-making in a wide variety of fields. At its core, it relies on principles from statistical modeling, which has a long history in understanding relationships between variables and quantifying uncertainty. While statistical models focus on inference and understanding, machine learning emphasizes predictive accuracy and optimization.

However, in the rush to develop new machine learning algorithms, the statistical foundations that emphasize model validation, uncertainty quantification, and interpretability are often overlooked. This oversight can lead to overfitting, lack of model transparency, and unreliable predictions.

Recognizing the interplay between statistical modeling and machine learning is crucial. By combining machine learning techniques with solid statistical principles, we can develop models that are not only powerful but also interpretable, reliable, and generalizable. As the size of datasets are only getting larger and larger it is important to balance statistical insight with computational power to advance effective data analysis.

There exist many different machine learning methods to simplify large datasets, while retaining the essential information. One of the most popular methods is Principal Component Analysis (PCA), which arises from the principal components of a singular value decomposition. The method reduces a high-dimensional data matrix into the factorization of two smaller matrices, where the components are orthogonal to each other.

The method is powerful and efficient for many dimensionality reduction tasks, but it is hard to interpret on the recovered factorization because the entries can be both negative and positive. Another disadvantage of PCA is that it is minimizing the squared distance making it very sensitive to outliers if the data is not standardized. The issue arises because the underlying statistical model assumes the data follows a normal distribution, which is rarely validated.

Non-negative Matrix Factorization (NMF) is closely related to PCA, but is highly advantageous for handling count data. One of the main differences is that it imposes the constraint that all elements in the data and the extracted factor matrices must be non-negative. This non-negativity constraint ensures that the decomposed components are additive and non-subtractive, which aligns well with many real-world

Introduction

scenarios where negative values might not make sense.

Additionally the non-negative constraint leads to naturally more sparse components, which will give a more part based representation of the data (Lee and Seung, 2001). Another difference from PCA is that the components extracted in NMF are not required to be orthogonal. This allows the factorization to more freely capture the patterns and structures within the data. However, this flexibility also sometimes present a challenge, as it can lead to multiple solutions for the same dataset, making the factorization non-unique.

Another advantage of NMF is its ability to adapt to different types of distributions, allowing for more accurate modeling of noise in the data and eliminating the need for standardizing. A shared challenge of both PCA and NMF is the need to determine the number of principal components or lower dimensional rank to retrieve the smaller simplified factorization of the data. As this is often chosen based on the noise explained in the data it is especially advantageous that NMF can adjust to model the noise correctly. All these properties ensures the key advantage of NMF, which is interpretability.

The interpretability of NMF has made its use advantageous in many different fields. It has a long history in chemometrics, under the name *self modeling curve resolution* (Lawton and Sylvestre, 1971), and was later studied for environmental data, under the name *positive matrix factorization* (Paatero and Tapper, 1994). The method first became widely known under the name *non-negative matrix factorization* after Lee and Seung (1999) derived simple update rules to approximate the factorization. In Lee and Seung (2001) they also show its advantage in creating parts-based representations for images, which highly increased the interest in the method. In genomics, NMF was first applied to the gene expression of cells in Brunet et al. (2004) and later applied to mutational counts from cancer tumors in Alexandrov et al. (2013a).

In this thesis we will focus on the applications of NMF to mutational counts in cancer tumors (Figure 1a), where the mutational counts from a single patient is called a mutational catalogue. A mutational catalogue is constructed by comparing a tumor genome and a reference genome to count the mutations that has occurred. Each mutation type is defined by the nucleotide bases—Adenine, Thymine, Guanine, and Cytosine—that make up the genome, which are denoted by A , C , G and T . A specific mutation type refers to a base substitution, in which one of the four bases— A , C , G , or T —is replaced by one of the other three.

Naturally, there are $4 \cdot 3 = 12$ possible base substitutions, but the genome is double-stranded and symmetric. Since an A on one strand always pairs with a T on the opposite strand, and similarly, G pairs with C , a change from C to A on one strand is equivalent to a change from G to T on the opposite strand. As a result, it is common in the literature to focus only on the following six base substitutions: $C > A$, $C > G$, $C > T$, $T > A$, $T > C$ and $T > G$.

A mutation type is denoted by the base substitution and the immediate flanking

nucleotide bases to the left and right (Figure 1b). An example of a mutation type is $A[C > A]T$, which indicates that a C changed to an A , where the immediate base flanking to the left was A and to the immediate right was a C . When more flanking bases are considered the context will simply be extended on both sides as $TA[C > A]TC$ for two immediate flanking bases.

This means the total number of mutation types is $M = 6 \cdot 4^{2f}$, where f denotes the number of flanking bases considered to each side of the base substitution. In the literature it is common only to consider one flanking nucleotide base to each side such that $M = 96$ (Alexandrov et al., 2013b, 2020), but there has evolved an interest in considering larger context as these can contain important contextual information (Shiraishi et al., 2015; Krawczak et al., 1998).

Mutational counts are of interest in the analysis of cancer, as it is commonly known that the main driver of cancer is due to somatic mutations in the genome. Mutational catalogues has therefore been collected by many different consortia including The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG) (2020) that collected 2,658 whole cancer genomes from 38 different cancer types.

To analyze the mutational counts it is assumed that each mutational catalogue arise from a mixture of a certain number of mutational processes. We assume that these mutational processes are distinct, such that each process leaves a different characteristic mark on the cancer genome.

These mutational processes are often represented as probability vectors over the different mutation types, where they are referred to as mutational signatures. This makes NMF superior to analyze this type of data, because it consists of counts i.e. non-negative and that these counts is assumed to arise from a positive exposure to

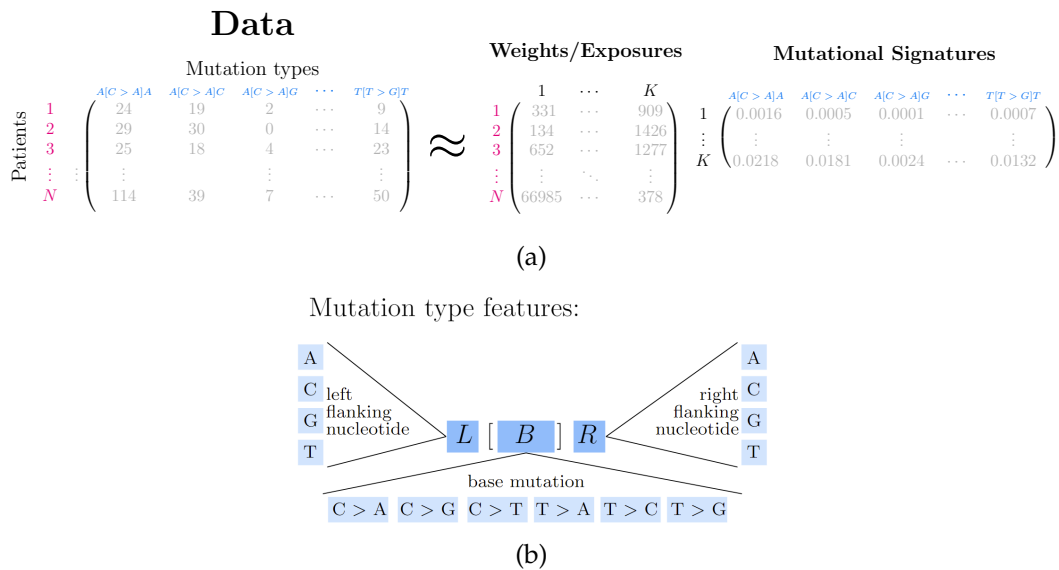


Figure 1: Overview of non-negative matrix factorization for mutational counts.

1. Non-negative Matrix Factorization (NMF)

different mutation signatures (Figure 1a).

The first analysis of mutational counts using NMF in [Alexandrov et al. \(2013a\)](#) laid the foundations for mutational signatures in the Catalogue Of Somatic Mutations In Cancer (COSMIC) ([Sondka et al., 2024](#)), which has been further expanded by many subsequent papers ([Alexandrov et al., 2015, 2016, 2020](#)).

COSMIC is a database of somatic variants in cancer, which include a large catalogue of around one hundred different mutational signatures, that has been recovered by analyzing mutational counts from different cancers using NMF. Some of these mutational signatures confirm already known biological aetiology such as UV-light([Shibai et al., 2017](#)) and tobacco([Alexandrov et al., 2016](#)), but for many of them the aetiology is still unknown.

The mutational signatures are used to get new insights into the drivers of cancer and the differences and similarities between different cancers. They also help decide in better treatment for patients by dividing them into different groups based on their exposure to certain mutational signatures ([Caruso et al., 2017](#); [Zhang et al., 2021](#)). It is therefore highly important that the retrieved mutational signatures are robust and reliable.

To ensure that NMF accurately reveals the true underlying structures of the data, several statistical properties must be addressed. In the following sections, these properties and extensions of standard NMF will be discussed, with a particular focus on my contributions to the field. The first section introduces the notation used in NMF and sections two through five will explore the following properties of NMF: Non-uniqueness, distributional assumption, choosing the rank and regularization. Section six and seven explain different extensions of NMF, where the model is adjusted for specific types of data. Section six explains the inclusion of mutational opportunities in the analyses of mutational counts and section seven explain the inclusion of a known location for the observations in the application of gene expression of cells. And the last section covers some of the general important points about the computational implementation of NMF.

1 Non-negative Matrix Factorization (NMF)

Let V be a non-negative data matrix of dimension $N \times M$, then NMF aims to approximate it by the product of two non-negative matrices in the following way

$$V \approx WH \tag{1}$$

where the matrices W and H only consist of non-negative entries and has dimensions $N \times K$ and $K \times M$, respectively. Usually, the rank K is chosen magnitudes smaller than N and M to construct a low dimensional representation of the data V .

For mutational counts each of the N rows of V represents a mutational catalogue from a patient over the counts of M different mutation types (Figure 1a). The rows of the matrix H consist of K mutational signatures, which consist of probability

vectors over the different mutation types. Each row in W represent the exposure or weight of the different mutational signatures for a certain patient.

2 Non-uniqueness

Apart from the non-negative constraint of the entries in the factorization, there is no further restrictions applied to NMF. This flexibility means that possibly several solutions can exist for the same dataset, making the factorization non-unique.

In fact there exist many invertible matrices $A \in \mathbb{R}^{K \times K}$, such that

$$\tilde{W} = WA^{-1} \geq 0 \quad \text{and} \quad \tilde{H} = AH \geq 0. \quad (2)$$

Here, \tilde{W} and \tilde{H} are a new solution with non-negative entries that give exactly the same approximation, $\tilde{W}\tilde{H} = WH$, of the data V .

Trivial solutions for A is a positive diagonal matrix that scales the entries in W and H or a permutation matrix that re-orders the columns in W and rows in H . The scaling can be prevented by the extra assumption that the columns in W or the rows in H need to sum to a specific value. For mutational counts it is usually assumed that the rows in H has to sum to one, as this makes the rows represent probability vectors over the different mutation types. The re-ordering is usually solved by matching the rows and columns of different solutions using any type of metric like cosine similarity.

2.1 Requirements for a unique factorization

It is common practice to say that an NMF solution is unique if the only possible solution for A is a scaled permutation matrix (Donoho and Stodden, 2003; Laurberg et al., 2008), which we will also use in the following. Though, in many cases other solutions for A exist than a scaled permutation matrix. The extent of this problem is highly influenced by the sparseness of both W and H . It has been proven by Donoho and Stodden (2003) that the factorization is indeed unique if both W and H includes a K -dimensional diagonal matrix after permutation of the rows in W and columns in H . This is however rarely fulfilled in reality as this requires there exist rows in the data that are only approximated by a single signature, but each row is often approximated by a combination of several signatures. Recently, in Gillis and Rajkó (2023) they constructed weaker conditions to identify partial uniqueness. They showed that certain columns of W and/or rows of H are unique if the matrices fulfill certain sparsity structures. There is still more to explore in this area as the literature often emphasize how these requirements are more strict than necessary for a factorization to be unique. One could also imagine that certain requirements of the matrices could give certain bounds to how much the factors W and H can change.

2. Non-uniqueness

2.2 Finding the set of feasible solutions

In the many cases where a factorization does not fulfill the established requirements for uniqueness, one need to uncover the extend of the problem by finding the possible Set of Feasible Solutions (SFS) for A in (2).

If $K = 1$ the factorization is simply a positive linear regression and A is only a scalar, which clearly makes the solution unique apart from scaling. In the case of $K = 2$ there exist simple analytical calculation to find the SFS for A as the matrix and its inverse has a simple form. Here, we can write A and A^{-1} in the following way

$$A = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \Rightarrow A^{-1} = \frac{1}{1-a-b} \begin{pmatrix} 1-b & -a \\ -b & 1-a \end{pmatrix}.$$

The rows of A has to sum to one to maintain the same scaling of the matrices, which only leaves two unknown parameters a and b . The bounds for a and b are found by rewriting the inequalities in (2). After assuming that $a + b < 1$ the bounds are given by the following lower bounds

$$\begin{aligned} a_{\min} &= \max_m \left\{ \frac{H_{1m}}{H_{1m} - H_{2m}} \middle| H_{2m} > H_{1m} \right\} \leq 0 \\ b_{\min} &= \max_m \left\{ \frac{H_{2m}}{H_{2m} - H_{1m}} \middle| H_{2m} < H_{1m} \right\} \leq 0 \end{aligned} \quad (3)$$

from the inequalities of \tilde{H} in (2) and following upper bounds

$$\begin{aligned} a_{\max} &= \min_n \left\{ \frac{W_{n1}}{W_{n1} + W_{n2}} \right\} \geq 0 \\ b_{\max} &= \min_n \left\{ \frac{W_{n2}}{W_{n2} + W_{n1}} \right\} \geq 0 \end{aligned} \quad (4)$$

from the the inequalities of \tilde{W} . This means the feasible values are $a \in [a_{\min}, a_{\max}]$ and $b \in [b_{\min}, b_{\max}]$. Notice, that the solution is unique when the matrices W and H include a diagonal matrix as this means $a_{\min} = a_{\max} = b_{\min} = b_{\max} = 0$. The assumption of $a + b < 1$ is fullfilled automatically, as the matrix A would otherwise permute the rows and columns of the factorization. The calculations of these bounds can both be found in [Moussaoui et al. \(2005\)](#) and [Laursen and Hobolth \(2022\)](#).

Clearly, the SFS for A is simple to uncover for $K = 2$, but already with $K = 3$ it becomes very computationally intensive ([Borgen and Kowalski, 1985](#)) and it has only been possible to solve analytically up to $K = 4$ ([Henry and Kim, 1990](#)). In higher dimension the number of hyper planes will increase extensively and there is no closed form solution for the bounds of the entries in A . This makes it computationally heavy to identify the bounds and others have therefore tried to approximate the SFS numerically ([Sawall et al., 2013](#); [Sawall and Neymeyr, 2017](#)). In [Sawall et al. \(2013\)](#) they introduce a polygon inflation algorithm to approximate the SFS for $K = 3$. In this case the SFS can be represented in a 2D plane and from an

original solution the whole SFS is identified by inflating polygons with an increasing number of vertices. For $K = 4$ another ray casting algorithm is proposed in [Sawall and Neymeyr \(2017\)](#). This method cast rays from origo and identify which rays that intersect with the SFS and in particular which part of the rays that intersect with the SFS.

Sampling algorithm (Paper A)

The current methods were very computational expensive and did not apply to K above four, but often K is chosen in the range between 5 to 20 for mutational counts. To uncover the extend of the non-uniqueness problem for mutational counts we therefore developed a new algorithm described in Paper A ([Laursen and Hobolth, 2022](#)) to find the SFS. This is a sampling algorithm that can find the SFS for an arbitrary K and much faster than previous methods. This algorithm is highly inspired by Gibbs sampling, as the SFS is iteratively found for each signature while the other signatures are fixed and then a new feasible value for the signature is sampled.

The algorithm wants to take advantage of the simple calculations for $K = 2$ and therefore restrict A in (2) to have the following form

$$(A_{ij}(\lambda))_{uv} = \begin{cases} 1 - \lambda & \text{if } u = v = i \\ 1 & \text{if } u = v \neq i \\ \lambda & \text{if } u = i, v = j \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

that depends on i, j and λ , where $i \neq j$. The transformation in (5) will change signature i to a convex combination of itself and signature j with the scale λ . Notice, that only two entries are different from the identity matrix and if $\lambda = 0$, then it is simply the identity matrix. The inverse matrix has a simple form as well, $A_{ij}^{-1}(\lambda) = A_{ij}(-\frac{\lambda}{1-\lambda})$, which makes it simple to find the SFS for λ . The range of feasible solutions for λ , such that both $WA_{ij}^{-1}(\lambda)$ and $A_{ij}(\lambda)H$ are non-negative, are similar to the range of a and b in (3) and (4). In this case, λ is bounded by

$$\begin{aligned} \lambda_{\min}(i, j) &= \max_m \left\{ \frac{H_{im}}{H_{im} - H_{jm}} \mid H_{jm} > H_{im} \right\} \leq 0 \\ \lambda_{\max}(i, j) &= \min_n \left\{ \frac{W_{ni}}{W_{ni} + W_{nj}} \right\} \geq 0 \end{aligned} \quad (6)$$

Given a solution $\hat{V} = WH$, where $W, H \geq 0$ and the rows of H sum to one, our sampling algorithm will find the SFS by repeatedly; sampling two random signatures i and j , find the range $[\lambda_{\min}, \lambda_{\max}]$, sample a random value λ in this range, change W to $WA_{ij}^{-1}(\lambda)$ and H to $A_{ij}(\lambda)H$ and save each of these new solutions. After repeating this multiple times it will leave a whole range of possible solutions

3. Distributional assumption

for each entry of W and H and the algorithm is stopped when this range does not expand any further. The SFS for each entry is then constructed as the interval between the minimum and maximum value recovered.

Results of the algorithm on mutational count data and comparisons to the polygon inflation algorithm are seen in Paper A. It also includes a more detailed explanation of the algorithm, running times and further details on how to recover the SFS.

I would like to note that the notation in Paper A is different to the notation presented here and in the rest of my papers. In Paper A we follow the notation of [Alexandrov et al. \(2013a\)](#) as it is the first and most popular paper to use non-negative matrix factorization for mutational counts and we want to make this field aware of the non-uniqueness problem. This notation is transposing the data matrix, such that the columns represent the patients and the rows represent different mutation types. Furthermore, the matrices and dimensions are denoted by different letters.

Perspectives

After applying our algorithm to various solutions from mutational count data, we concluded that the SFS is relatively small in this field, but still noteworthy in some cases. The SFS is often small for solutions from mutational counts because the data is relatively sparse and becomes even more so when the number of flanking nucleotides for the mutation types is increased.

We know that it is a more extensive issue for datasets in chemometrics and could potentially also be an issue with many other datasets. And here our sampling algorithm provide an easy and superior tool to check the non-uniqueness of a factorization with an arbitrary high rank.

Being aware of the potential issue is essential when recovering a factorization of a dataset with NMF. In particular we observed that the more similar or flat some of the signatures are across different features, the larger the SFS will be.

3 Distributional assumption

An often forgotten fact in machine learning is that behind the choice of the loss function, there is an underlying assumption about the distribution of the data. Checking whether the assumed distribution is a good fit for the data is unfortunately an often neglected task. For NMF this can result in wrong estimations of the factorization as the estimation of W and H is directly dependent on the chosen cost function.

As with many other methods, the most popular cost function for NMF is to minimize the squared Euclidean distance between the entries in the data matrix V and the approximation WH , also referred to as the Frobenius norm. This cost function assumes that the data follows a normal distribution. However, for NMF,

we know the data is non-negative and often contains only integer values, as seen for mutational counts. Therefore, another popular and more sensible assumption is the Poisson distribution.

3.1 Poisson NMF

For the Poisson distribution we assume the following:

$$V_{nm} \sim \text{Po}((WH)_{nm}) \quad n = 1, \dots, N \quad m = 1, \dots, M \quad (7)$$

which is equivalent to minimizing the Generalized Kullback-Leibler (GKL) divergence

$$d_{Po}(V||WH) = \sum_{n=1}^N \sum_{m=1}^M \{V_{nm} \log V_{nm} - V_{nm} \log((WH)_{nm}) - V_{nm} + (WH)_{nm}\}, \quad (8)$$

as it is equal to the negative log-likelihood of the Poisson distribution, up to an additive constant only dependent on V . Further derivations of this is seen in Paper A and C. The estimation of W and H under both the Frobenious norm and the GKL divergence was first introduced by [Lee and Seung \(1999, 2001\)](#), who was the first to construct multiplicative updates to estimate W and H .

An iterative procedure is necessary to find the estimate of W and H because a cost function like in (8) only is convex if either W or H is fixed. In [Lee and Seung \(2001\)](#) they introduce the following updates to minimize the GKL in (8)

$$W \leftarrow W \otimes \left(\frac{V}{WH} H^T \right) \quad H \leftarrow H \otimes \left(W^T \frac{V}{WH} \right) \quad H \leftarrow \frac{H}{H \mathbb{1}_{M \times M}}, \quad (9)$$

where $\mathbb{1}_{M \times M}$ is an $M \times M$ matrix only containing ones, and \otimes and $-$ is entry wise multiplication and division, respectively. Notice, that the last update of H simply normalizes the rows to sum to one.

After initializing W and H by random non-negative entries, the updates are iterated until convergence of the cost function in (8). These updates ensures that the entries in W and H remain positive and the cost function is decreasing with each iteration, if all the initialized values in W and H are positive. In [Lee and Seung \(1999\)](#) they recover the updates from an MM-algorithm, but equivalent updates are obtained with an EM-algorithm, which is shown in the methods of Paper C.

The non-convexity of the divergence in (8) means that the updates in (9) only ensure a local minimum is found. To increase the chance of a global minimum, it is common to initialize the algorithm multiple times. Initialization, convergence and other details on the implementation of the algorithm is further elaborated in Section 8.

The updates in (9) have been broadly used and the [Lee and Seung \(2001\)](#) paper, that includes the multiplicative updates to minimize both the GKL in (8) and the Frobenius norm, has been cited over 16.000 times.

3. Distributional assumption

A more recent discovery is that the Poisson distribution is over-dispersed for many datasets. It is because the variance is fixed to be equal to the mean, which makes it sensitive to high dispersion or outliers in the data. A natural way to extend the Poisson distribution to include more variability is to use the Poisson-gamma distribution, which is equivalent to the Negative Binomial distribution shown in [Pelizzola et al. \(2023\)](#).

3.2 Negative Binomial NMF

For the Negative Binomial distribution the mean and variance has the following form

$$\mathbb{E}[V_{nm}] = (WH)_{nm} \quad \text{and} \quad \text{Var}(V_{nm}) = (WH)_{nm} \left(1 + \frac{(WH)_{nm}}{\alpha} \right). \quad (10)$$

The extra parameter α will control the level of dispersion, where low values correspond to a high dispersion and $\alpha \rightarrow \infty$ will converge to the Poisson distribution, where the variance is equal to the mean. The negative log-likelihood of the Negative Binomial distribution is equal to the following divergence

$$d_{NB}(V||WH) = \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log \left(\frac{V_{nm}}{(WH)_{nm}} \right) - (\alpha + V_{nm}) \log \left(\frac{\alpha + V_{nm}}{\alpha + (WH)_{nm}} \right) \right\}, \quad (11)$$

up to an additive constant that only depends on V and α . The derivation are further described in the methods of Paper B. As expected the limit case for $\alpha \rightarrow \infty$ is the GKL in (8). The multiplicative updates to minimize (11) is given by

$$W \leftarrow W \otimes \frac{\frac{V}{WH} H^T}{\frac{\alpha+V}{\alpha+WH} H^T} \quad H \leftarrow H \otimes \frac{W^T \frac{V}{WH}}{W^T \frac{\alpha+V}{\alpha+WH}} \quad H \leftarrow \frac{H}{H \mathbb{1}_{M \times M}}, \quad (12)$$

Notice, that both in (9) and (12) the updates of W and H are controlled by the fraction $\frac{V}{WH}$ that tells whether the approximated entries are too high or too low. The difference in (12) is the additional division by a term that depends on α . If there is an entry V_{nm} that is much smaller than α , this division has little effect, as $\left(\frac{\alpha+V}{\alpha+WH} \right)_{nm} \approx 1$. On the other hand, if V_{nm} is much larger than both α and $(WH)_{nm}$, then $\left(\frac{\alpha+V}{\alpha+WH} \right)_{nm} > 1$, which scales down the update, preventing excessive adjustment to this potential outlier.

These updates in (12) were first introduced and derived by [Gouvert et al. \(2020\)](#) and have later been applied to mutational counts in [Lyu et al. \(2020\)](#) and [Pelizzola et al. \(2023\)](#).

Patient specific dispersion (Paper B)

In Paper B ([Pelizzola et al. \(2023\)](#)) we make a comparison, on mutational count data, between assuming the Negative Binomial and the Poisson distribution. We argue

why the Negative Binomial model makes sense for mutational counts biologically and supports the hypothesis with residual plots of different datasets that clearly show overdispersion in the Poisson model.

In [Gouvert et al. \(2020\)](#) they choose the dispersion, α , from a grid search of the three values 1, 10, and 100, but we suggest to estimate α as the Maximum Likelihood Estimate (MLE) of the Negative Binomial, where WH is fixed to the estimates from Poisson NMF (full details in methods of Paper B). This makes it possible for a much more precise estimate of the dispersion. Another main contribution is that we observe a high difference in dispersion between patients and therefore introduce a Negative Binomial NMF, where there is a patient specific dispersion. In this model the variance is given by

$$\text{Var}(V_{nm}) = (WH)_{nm} \left(1 + \frac{(WH)_{nm}}{\alpha_n} \right), \quad (13)$$

such that there is a dispersion parameter for each patient, $\alpha_1, \dots, \alpha_N$, that needs to be estimated. It is possible to have this very fine grained distinction of the dispersion between patients now that we have developed a precise estimation procedure of the dispersion. Under the model with patient specific dispersion we have derived new updates, which are very similar to the updates in (12). These are given by

$$W_{nk} \leftarrow W_{nk} \frac{\sum_{m=1}^M \frac{V_{nm}}{(WH)_{nm}} H_{km}}{\sum_{m=1}^M \frac{V_{nm} + \alpha_n}{(WH)_{nm} + \alpha_n} H_{km}} \quad H_{km} \leftarrow H_{km} \frac{\sum_{n=1}^N \frac{V_{nm}}{(WH)_{nm}} W_{nk}}{\sum_{n=1}^N \frac{V_{nm} + \alpha_n}{(WH)_{nm} + \alpha_n} W_{nk}}, \quad (14)$$

which simply adds different values of α for each patient compared to the updates in (12).

Perspectives

It is hard to construct a quantile plot for the NMF models because as seen in (7) and (10), there is only one observation for each mean. We therefore validated the different models from the residual plots. The method that we used in Paper B to argue for the Negative Binomial distribution is to plot the residuals against the mean and show the 95% confidence region of the residuals under the different distributions. Here, if the residuals increase with the mean the normal distribution is rejected and if large parts of the data also lie outside of the confidence region for the Poisson distribution one should use the Negative Binomial distribution.

One of the really crucial issues with assuming a wrong distribution for the data is the estimation of the rank K . If the chosen model is overdispersed there will be an overestimation in the rank K as extra signatures are needed to explain the additional dispersion in the model. This was also one of the main points in Paper B, where we also developed a new computational method called SigMoS. This method is robust to estimating the true rank even when the underlying model is misspecified. The estimation of the rank K and our method SigMoS is further elaborated in the next section.

4. Choosing the rank

4 Choosing the rank

In NMF, the rank K is a critical hyperparameter that determines the quality and interpretability of the factorization. A too high rank will lead to overfitting, which will lose the interpretability of the factors, but a too low rank can give a poor accuracy of the data and the factors can lack important information. Choosing the rank for any dataset is a challenging and recurrent problem. The noise in a dataset makes it difficult to find the true lower dimensional rank.

4.1 Information Criteria

Some of the classical statistical methods to find the correct model for a dataset are Akaike Information Criterion (AIC) or Bayesian Information Criteria (BIC). These two information criteria are based on the log-likelihood and are highly related. They both penalize the log-likelihood by the number of parameters in the model, but the penalty term is larger for BIC than for AIC. The BIC is given by

$$\text{BIC} = -2\ell + \ln(n_{\text{obs}})n_{\text{prm}} \quad (15)$$

where ℓ is the estimated log-likelihood value, n_{obs} is the number of observations and n_{prm} the number of parameters to be estimated. For AIC the term $\ln(n_{\text{obs}})$ is simply replaced by the scalar 2, which means the penalty term is higher for the BIC already with more than 7 observations.

Some of the methods used to estimate the rank of NMF for mutational counts are also based on BIC (Fischer et al., 2013; Rosales et al., 2017). In Fischer et al. (2013) they use the BIC directly, where the number of observations is the number of patients N and the number of parameters is $K \cdot M$ (i.e. the number of entries in the signature matrix H). The optimal rank is then chosen as the K that gave the smallest BIC. In Rosales et al. (2017) they are using a more bayesian approach to estimate W and H , which means they have a sequence of estimates of W and H for a specific rank K . They therefore calculate a median BIC for each possible rank K and choose the minimum median BIC as the optimal K . To calculate the BIC they choose the number of observations as the number of patients N as in Fischer et al. (2013), but the number of parameters n_{prm} is defined as $K \cdot (N + M)$ i.e. the number of entries in both W and H . Together with the method presented in Rosales et al. (2017) they also made an R package available called SigneR to estimate the rank and calculate the NMF factorization.

In Paper B and C we set n_{prm} similar to Rosales et al. (2017) as the entries in both W and H , but the number of observations is chosen as the number of positive entries in the data matrix. We found that only using N as the number of observations was not representative of the large data matrix of observations and found the number of positive entries more representative. It is therefore not consistent how the BIC is calculated, which is an issue for consistency and comparison.

One of the other issues with using the AIC or BIC is that it depends a lot on the assumed distribution through the likelihood. If a model is misspecified and overdispersed these information criteria will select additional signatures to compensate for this overdispersion in the model. The AIC and BIC are therefore heavily dependent on a right distributional assumption of the data. This is illustrated in Paper B (Pelizzola et al., 2023), where the rank K is heavily overestimated for both AIC and BIC, when the true distribution is Negative Binomial but the Poisson distribution is assumed.

4.2 Cross-validation

Cross-validation is another model selection procedure that is widely applied in supervised settings, such as regression and classification. It is often shown to have a better performance compared to other model selection procedures including AIC and BIC (Gelman et al., 2014; Bro et al., 2008).

Though, implementing the setup into an unsupervised setting, such as PCA or NMF, is a more challenging task. The simple idea of cross-validation is to split the data into multiple subsets and train multiple models on the data, where each of the subsets are left out, and then validate the trained model on the held out subset. The data used for training is referred to as the training set and the held out subset is called the test set.

A first thought would be to hold out rows or columns of the data V to train a model, but that would remove corresponding rows of W or columns of H making it hard to validate. Because the trained model cannot validate the held out rows or columns alone.

Instead many methods are holding out a random set of entries in the data matrix, such that none of the rows or column in V are completely excluded. This idea was introduced by Wold (1978) to identify the dimension of PCA. One of the issues with completely excluding entries of V is that the multiplicative updates will be more complex and therefore more time consuming. A critical review of different cross-validation methods for PCA are found in Bro et al. (2008). They highlight two methods as being best for estimating the rank of component analysis. The first is an improved version of Wold (1978), where they exclude random elements of the data. The other method comes from a package called "Eigenvector" Wise et al. (2007), where they exclude rows of the data and then each variable in the left out samples is predicted based on the model and the samples excluding that one variable.

For NMF analysis on mutational counts Lal et al. (2021) proposed a cross-validation method, where they do not have missing data, but instead replace certain entries of the data by zero. These entries are then simultaneously updated by the estimate of the factorization while the factors are estimated. The error of the estimates of these "zero replaced" entries are then evaluated and used to determine the correct rank. The method in Lal et al. (2021) is implemented as a package

4. Choosing the rank

called SparseSignatures, which both can estimate the rank and create sparse NMF solutions.

SigMoS (Paper B)

In Paper B ([Pelizzola et al., 2023](#)) we developed a new method, SigMoS, to estimate the underlying rank of a dataset with inspiration from cross-validation to utilize its efficiency. Our method is also implemented as an R packages called SigMoS.

Our method first estimate the factors W and H from the full data and then the number of patients N is split into ten random subsets. Afterwards factors W_{train} and H_{train} are estimated for the training datasets, where each subset of patients are left out. The left out patients are then estimated based on the matrix product between W from the full data and H_{train} from the training data. The minimum mean error of the ten left out subsets then determine the optimal rank K .

The idea behind this method is that when the optimal rank is chosen, then the estimated signatures from the full data and the training data will be similar. A matrix product between the weights of the full model W and the signatures of the training data H_{train} should therefore give a good estimate of the left out patients when the optimal rank is chosen. The method is described in further detail in Paper B.

4.3 Alternative methods

Another approach to find the optimal rank, that has also been applied in cancer genomics is stability analysis. In [Alexandrov et al. \(2013a\)](#) they investigate how stable the solution of the signatures in H is to different bootstrap samples of the original data. The reproducibility of the signatures was evaluated by the silhouette width of all the solutions of H from the different bootstrap samples. This reproducibility is getting worse as K increases, while the cost function will decrease. The optimal K is chosen at a point, where the reproducibility is high and the cost function is low. This method is implemented in a package called `SignatureAnalyzer`, that can also recover the mutational signatures from mutational counts using NMF.

Sometimes the optimal K is only determined from the cost function, at the place where the cost function plateaus. This is also referred to as the elbow method and is implemented as the model selection procedure in a package called `sigfit` ([Gori and Baez-Ortega, 2018](#)), that again also can analyze mutational counts with NMF to find mutational signatures.

In ([Shiraishi et al., 2015](#)) they use another method based on splitting. This means the K is chosen such that no two signatures (i.e. rows) in H have a strong correlation. The idea behind this method is that when K is too large then a signature would start to split into two, creating almost identical signatures.

Perspectives

One of the main advantages of our SigMoS package is that it performs well under model misspecifications compared to AIC, BIC and other packages including SigneR, SparseSignatures, SignatureAnalyzer and sigfit. In Paper B we show that our method performed significantly better than other packages when data was simulated under the Negative Binomial model, which is because the other packages in the literature assume the underlying distribution is Poisson, which is not always the case.

Even though cross-validation and stability methods is significantly more accurate in estimating the rank of the factorization they are more computationally heavy. This is because they estimate a model for the data several times for a fixed K , where different parts of the data are left out or bootstrap samples are created. On the other hand for AIC or BIC only one model needs to be estimated for each K . In Paper B we see that the information criteria, in particular BIC, performs well when the right distribution is assumed. One could therefore argue that our SigMoS method is good when the true underlying distribution is unknown, but when the true distribution is known it can be computationally advantageous to use BIC. If a dataset is very computationally heavy and the distribution of the noise is unknown, it is also sometimes necessary to use the more simple model selection procedures such as the elbow method.

5 Regularization

All the entries in the factors W and H are free parameters that need to be estimated, which is causing the non-uniqueness problem elaborated on in Section 2. The large amount of free parameters also make the factorization less stable. A natural thought is therefore to regularize or impose extra constraints on the factors such that the parameters are less free.

5.1 Sparseness constraints of factors

Even though the non-negative constraint of NMF naturally imposes more sparse factors, there has been a high interest in adding additional sparseness constraints to the factors (Hoyer, 2004; Pascual-Montano et al., 2006; Pauca et al., 2006). In particular it has been implemented in cancer genomics for both gene expression data (Kim and Park, 2007) and mutational counts (Lal et al., 2021). The most common way to enforce sparseness is to use LASSO, where an extra regularization term is added to the cost function (Lal et al., 2021; Pauca et al., 2006).

For mutational counts, it is particularly beneficial to enforce additional sparsity on the weight matrix W to categorize patients into distinct groups. When analyzing several cancer types at once some signatures should only be present for certain cancers (Alexandrov et al., 2020). Or when analyzing patients with the same cancer

5. Regularization

there is a high interest in splitting the patients into different treatment groups. Examples of this is seen in [Zhang et al. \(2021\)](#), where they want to split lung cancer patients into different subgroups, based on the exposures of a specific signature, to enhance the development of personalized medicine.

Another approach to enforcing sparser signatures is to use convex NMF, which adds the constraint that the signatures must be convex combinations of the data matrix V . An interesting observation in [Ding et al. \(2008\)](#) is that k-means can be viewed as a special case of convex NMF, highlighting its suitability for clustering tasks.

5.2 Parametrizing entries in the factors

In the analysis of mutational counts there is another way to restrict the entries in the signature matrix H by the natural connection between the features.

Recall from Figure 1b, that the features for mutational count data are given by different mutation types. These include information on the base substitution which can take six different values and its flanking nucleotides, which can be one of the four A, C, G and T . The number of flanking nucleotides included in the mutation types differ, but the most common is only to include the immediate flanking nucleotides, which leaves $4 \cdot 6 \cdot 4 = 96$ different features of mutation types.

Notice, that in this case there will be 16 different mutation types with the same base substitution. It would be natural to think that the mutation rate is connected for mutation types with the same base substitution, but possibly also for mutation types where the same nucleotide is flanking to the left or right. An idea proposed in [Shiraishi et al. \(2015\)](#) is therefore to view the base substitution and each of the flanking nucleotides as separate factor variables, that has 6 or 4 possible levels. We denote the base substitution as the factor variable B with six levels and the first left and right flanking nucleotide as the factor variable L_1 and R_1 with four levels as shown in Figure 1b. The model proposed in [Shiraishi et al. \(2015\)](#) then restrict each signature row in H with the following log additive factor model

$$\log(h_{\ell br}) = \beta_b^B + \beta_\ell^{L_1} + \beta_r^{R_1} - C \quad (16)$$

where $h_{\ell br}$ represents an entry of a row in H , where $\ell, r = 1, \dots, 4$ and $b = 1, \dots, 6$, and the constant C assures that the rows of H sum to one. Here, the constant is given by

$$C = \log \left(\sum_{\ell br} \exp \left(\beta_b^B + \beta_\ell^{L_1} + \beta_r^{R_1} \right) \right).$$

The model in (16) is a simple model with only 12 parameters compared to the standard, where there is a parameter for each of the 96 mutation types. The standard model is equivalent to including all interaction terms between these different factors. This simple model in (16) includes 12 free parameters because there is $6 + 4 + 4 = 14$ different variables, but two of these will be confounding in the factor model.

In [Shiraishi et al. \(2015\)](#) they showed that the more simple parametrization was especially beneficial when more flanking nucleotides was considered. In the paper they considered a dataset with two flanking nucleotides, which has $M = 6 \cdot 4^4 = 1536$ mutation types. Using a simple additive model similar to (16) they could reduce the number of parameters to $6 + 4 \cdot 4 - 4 = 18$, as there is four confounding variables in this case. They showed that this large reduction in the number of parameters gave more robust signatures, that had a high reproducibility when the data was down-sampled. A disadvantage of the simple model from [Shiraishi et al. \(2015\)](#) is that it gives a much worse fit to the data compared to the standard model, which made us propose a middle ground between the standard and simple model in [Laursen et al. \(2024\)](#).

Flexible parametrization (Paper C)

In Paper C ([Laursen et al., 2024](#)) we sought to find the middle ground between the very simple model constructed in [Shiraishi et al. \(2015\)](#) and the standard NMF model for mutational counts in [Alexandrov et al. \(2013a\)](#). We also wanted to construct a more flexible model where one could specify the appropriate parametrization of the features and where each signature could have its own parametrization.

Recognizing that the model in [Shiraishi et al. \(2015\)](#) is simply multinomial logistic regression made it possible to construct a more general setup where each row in H was modeled in the following way:

$$\log(H_k) = X_k \beta_k - C. \quad (17)$$

where H_k represents the k^{th} row of H and the constant C assures that the rows in H sum to one and is given by $\log\left(\sum_{m=1}^M \exp(X_k \beta_k)_m\right)$. Here, the parametrization can be arbitrarily chosen through the design matrix X . This makes it possible to have any parametrization of the signatures including the simple model from [Shiraishi et al. \(2015\)](#) and the standard model in [Alexandrov et al. \(2013a\)](#). In particular we introduced the following parametrization:

$$\log(h_{\ell br}) = \beta_b^B + \beta_{\ell b}^{L \times B} + \beta_{br}^{B \times R} - C, \quad (18)$$

where the interaction between the neighboring features is included. We found this parametrization more biologically plausible and showed that it both had the robustness of the simple parametrized signatures in [Shiraishi et al. \(2015\)](#), while still maintaining the good fit of the data as the standard model.

The parametrization of the mutational signatures in H are imposed by adding an extra update step to (9), that ensure the rows follow the predefined parametrization. Given a list of design matrices X_1, \dots, X_K for each signature, the update of H in (9) is replaced by updating each row H_k in the following way:

6. Opportunities

1. Calculate the original update of H_k from (9) as

$$y_k = H_k \otimes \left(W_k^T \frac{V}{WH} \right)$$

2. Estimate the Poisson regression

$$\log(y_k) = X_k \beta_k \tag{19}$$

by finding the parameter vector $\hat{\beta}_k$.

3. Update the row H_k following this parametrization

$$H_k \leftarrow \frac{\exp(X_k \hat{\beta}_k)}{\mathbf{1}' \exp(X_k \hat{\beta}_k)}$$

Notice, that Poisson regression is preformed instead of multinomial logistic regression and then the estimates are simply scaled to sum to one afterwards. This is known as the 'Poisson trick' [Lee et al. \(2017\)](#), which is further described in Paper C.

Perspectives

In Paper C we compare different parametrizations on three different datasets, where two of them have one flanking nucleotide with $M = 96$ and a last dataset that is equivalent to the one in [Shiraishi et al. \(2015\)](#) with two flanking nucleotides at each side i.e. $M = 1536$. We show that our more flexible model is a good middle ground that both gives robust signatures, while maintaining a good fit to data. To support the extension of different parametrizations for each signatures, we recovered the top factor variables that best explain the different mutational signatures. Here, we saw that some include three way interaction, but others mainly include two-way interactions and a few of the mutational signatures can be explained by the simple additive model, which supports that different parametrizations is beneficial.

In Paper C we introduced our parametrized NMF model for poisson NMF, but it can be easily extended to any NMF updates by changing step 1 in (19). For example in Paper D the procedure in (19) is implemented for the Negative Binomial NMF updates.

6 Opportunities

In the analysis of data it is important to consider if certain unwanted bias are included in data, which could affect the analysis and thereby conclusions. These bias are sometimes caused by the absence of important covariates, which is the case for mutational counts.

Recall, that mutational counts are the number of times a certain mutation occurred in a genome. The mutational catalogue for a genome could for example

include 5 of type $A[T > A]G$ and 50 of type $A[T > A]T$, which would indicate that there is a much higher probability of a mutation when a T is flanking to the right compared to a G . But if there is 10 times as many occurrences of the triplet ATT compared to ATG in the genome, then that could be the sole explanation of this difference.

If the four nucleotides in the genome were placed at random with equal probability, then this would not be an issue, but this is not the case. Indeed there is an intrinsic system of how the nucleotides are placed in the genome and therefore a variation in the occurrences of different nucleotide contexts. The occurrence of a certain nucleotide context in the genome is denoted as the opportunity of the corresponding mutation type. The name opportunity arises because it reflects at how many cites in the genome the mutation type had an opportunity to occur. Failing to account for these differences in the opportunity can lead to an undesirable bias towards certain mutation types in the signatures. In the literature there has been implemented different extensions to the analysis of mutational counts to account for the difference in the opportunity of the mutation types.

6.1 Background signature

In [Shiraishi et al. \(2015\)](#) and [Lal et al. \(2021\)](#) they include a background signature in the analysis of mutational counts to account for the intrinsic composition of the genome. Compared to the factorization in (1) they include an extra term to the factorization in the following way:

$$V \approx w_0 h_0 + WH \quad (20)$$

where $w_0 \in \mathbb{R}_+^N$ is a column vector of weights for the background signature and $h_0 \in \mathbb{R}_+^M$ is a row vector including the background signature. Each entry in the background signature is set as the opportunity for the corresponding mutation type and then divided by the total to obtain a probability vector. After the background signature is fixed they estimate the corresponding weights in w_0 for each patient. The matrix product $w_0 h_0$ is then subtracted from the data V before the estimation of W and H . The idea is that the background signature can subtract the potential biases in the data, which are caused by the intrinsic composition of the genome.

6.2 Probabilistic inclusion (Paper D)

Another incorporation of the mutational opportunity is shown in [Fischer et al. \(2013\)](#) and Paper D. Here, the opportunity is incorporated into the probabilistic model in the following way:

$$V_{nm} \sim \text{Pois}((WH)_{nm} O_m), \quad (21)$$

where O_m denotes the opportunity for mutation m . The estimated mutational signatures in (21) will be adjusted for the opportunity, such that the probabilities will reflect the actual relative differences in the mutation types.

6. Opportunities

In Paper D we further explain the inclusion of opportunities as in (21) and extend it to the Negative Binomial model, where $(WH)_{nm}$ is again replaced by $(WH)_{nm}O_m$ in the model i.e.

$$\mathbb{E}[V_{nm}] = (WH)_{nm}O_m \quad \text{and} \quad \text{Var}(V_{nm}) = (WH)_{nm}O_m \left(1 + \frac{(WH)_{nm}O_m}{\alpha}\right). \quad (22)$$

The multiplicative updates under the Negative Binomial model in (22) are derived in Paper D and given by

$$W_{nk} \leftarrow W_{nk} \frac{\sum_{m=1}^M \frac{V_{nm}}{(WH)_{nm}O_m} H_{km}}{\sum_{m=1}^M \frac{V_{nm} + \alpha_n}{(WH)_{nm}O_m + \alpha_n} H_{km}} \quad H_{km} \leftarrow H_{km} \frac{\sum_{n=1}^N \frac{V_{nm}}{(WH)_{nm}O_m} W_{nk}}{\sum_{n=1}^N \frac{V_{nm} + \alpha_n}{(WH)_{nm}O_m + \alpha_n} W_{nk}} \quad (23)$$

for each entry $n = 1, \dots, N$, $k = 1, \dots, K$ and $m = 1, \dots, M$. Notice, that the difference from the updates in (12) is again $(WH)_{nm}$ which is replaced by $(WH)_{nm}O_m$.

Actually, we can note that including opportunities into the model is only a matter of multiplying a constant. The original factorization in (1) can simply be rewritten as

$$V \approx WH = W \underbrace{H \begin{pmatrix} \frac{1}{O_1} & 0 & \dots & 0 \\ 0 & \frac{1}{O_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & & \frac{1}{O_M} \end{pmatrix}}_{\tilde{H} \text{ signature matrix for the model with opportunities}} \begin{pmatrix} O_1 & 0 & \dots & 0 \\ 0 & O_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & & O_M \end{pmatrix} = W\tilde{H}\mathbf{O}$$

where \mathbf{O} is the diagonal matrix with entries O_1, \dots, O_M .

This means the optimal estimate of H in the model without opportunities in (1) simply need to be scaled by a diagonal matrix to include the opportunities. Applying the updates in (23) is thereby equivalent to applying the original Negative Binomial updates in (12) and afterwards multiplying H by \mathbf{O}^{-1} . The estimation of W and H does therefore not necessarily have to change under the standard model, as the opportunities can be incorporated afterwards by scaling. Though, this is not the case when the signatures are parametrized as explained in (19).

Perspectives

In the parametrized models it will have a high influence if the opportunities is included in the multiplicative updates, which is shown in Paper D. The opportunities will change the landscape of the signatures, which means different and often simpler models would be able to parametrize the signatures. In Paper D, we demonstrate that a simple parametrized model provides a better fit for the signatures when opportunities are included, compared to when they are not. This shows that much of the variation in the data can be explained by the opportunities, which is also supported by showing that the opportunities and the mutational counts are positively

correlated. We show that this correlation increases, when larger contexts, with more flanking nucleotides are considered.

We also show that for all models including the standard model, the inclusion of opportunities improves prediction of mutations in new regions, where the opportunity is known. In particular it improves the prediction of mutation types with a very large or small opportunity, which is also shown in the paper.

7 Spatial data

Another possible extension of NMF is the incorporation of spatial information, which leverages a known location of each observation in the data matrix. This extension is of particular interest in genomics, where a new technology from 2015 (Chen et al., 2015), called MERFISH, can now map the location of cells together with their gene expression in a slice of tissue. This type of data makes it possible to get a better understanding of the structures in tissue (Velten and Stegle, 2023). In cancer genomics it is highly beneficial to further understand the biological processes in and around a tumor. Here, these new technologies make it possible to identify inflammation and anti-tumor regions of cancer tumors Pelka et al. (2021).

In this context, the data matrix V contains gene expression counts, where the N rows correspond to different cells and the M columns represent different genes. The rows of H represent distinct gene signatures, that often consist of gene markers for a certain cell type. The rows of W are the weight or activity of the different gene signatures for each cell, which are often sparse because many of the gene signature are cell type specific.

Gene expression data on its own has been available for decades and was already analyzed with NMF in Brunet et al. (2004), but the recent addition to this data is the known location of each cell in the tissue. This contributes an additional data matrix $X \in \mathbb{R}^{N \times 2}$ that tells the location of each cell in a tissue slice. Sometimes several slices are created from the same tissue making the location available in 3D and not only 2D.

The goal is to incorporate the location into the model to recover spatial gene signatures, structures and interesting multicellular neighborhoods in tissue. Within the last few years there has been developed a large range of different methods to model spatial genomics data (Li and Zhou, 2022; Zhao et al., 2021; Hu et al., 2021; Chen et al., 2020; Yuan et al., 2022; Yuan, 2024; Dong and Zhang, 2022), including some that builds on extensions of NMF (Ma and Zhou, 2022; Townes and Engelhardt, 2023; Chidester et al., 2023).

7.1 Extension of NMF to model spatial data

The extension of NMF to model spatial data adds an additional constraint to the columns of W , as they represent the activity landscape of gene signatures over the

7. Spatial data

cells in the tissue. The shared goal is to make weights of neighboring cells more correlated.

The methods in [Ma and Zhou \(2022\)](#) and [Chidester et al. \(2023\)](#) essentially add a Markov Random Field (MRF) assumption to the columns of W . The difference is that in [Ma and Zhou \(2022\)](#) they use a gaussian MRF, but in [Chidester et al. \(2023\)](#) it is a simple MRF as each cell only depends on the nearest neighbor and not on a neighborhood determined from a Gaussian kernel. In [Townes and Engelhardt \(2023\)](#) they add a Gaussian prior to the columns of W , where the correlation matrix is defined from a Gaussian kernel as in [Ma and Zhou \(2022\)](#).

For the MRF it is assumed that each entry in W can be determined by its neighbors, which means each entry in W is adjusted as an average of its neighbors. For the Gaussian prior the entries in W are sampled from a multivariate normal with the specified correlation structure.

In [Ma and Zhou \(2022\)](#) they have derived update rules, but only in for updating W , as they assume the gene signatures in H are known. In both [Chidester et al. \(2023\)](#) and [Ma and Zhou \(2022\)](#) they use gradient descent to optimize a posterior likelihood. Gradient descent is really time consuming compared to deriving multiplicative updates and especially in [Townes and Engelhardt \(2023\)](#) that has the bottleneck of having to inverse the correlation matrix of dimension N .

The number of cells N are between thousands to a million in one slice and M can be any subset of the 30,000 known genes. The data is therefore significantly larger than the mutational counts data, where N is usually in the hundreds and M is typically 96. This means the computational aspect needs to be more highly considered for this type of data in the development of new methods. It is important to balance the statistical insights with computational power to make sure that the model and analysis can be obtained for the data in reasonable time.

The other methods that model spatial genomics data ([Li and Zhou, 2022](#); [Zhao et al., 2021](#); [Hu et al., 2021](#); [Chen et al., 2020](#); [Yuan et al., 2022](#); [Yuan, 2024](#); [Dong and Zhang, 2022](#)) are based on extensions of PCA or graph neural networks, which are often faster than NMF but most of them only cluster the tissue, where the underlying interpretation of the clusters are unknown and need to be found from different post processing steps. The unique ability of NMF to produce directly interpretable results makes it especially compelling to explore methods that extend NMF to model high-dimensional spatial data, which was the motivation for Paper E.

Neighborhood NMF (Paper E)

In Paper E we have developed a new method that is computationally efficient while maintaining the interpretability of the results. This is possible because we are using the fast multiplicative updates instead gradient descent.

We are also incorporating the spatial information in a very simple, but intuitive way that adds an additional fast multiplicative update. This makes our method

feasible to run on million of cells, while preserving the gene expression of each cell and not bagging cells together or analyzing a reduced set of *metagenes* constructed from PCA like many other methods in the literature (Li and Zhou, 2022; Chen et al., 2020; Yuan, 2024).

Our algorithm combines the multiplicative updates from Lee and Seung (2001) together with a Gaussian smoothing step, such that the updates are as follows:

$$H \leftarrow \text{rnorm} \left(H \otimes \left(W^T \frac{V}{WH} \right) \right) \quad (24)$$

$$W \leftarrow W \otimes \left(\frac{V}{WH} H^T \right) \quad (25)$$

$$W \leftarrow \text{rnorm}(S_\phi(X))W \quad (26)$$

where $S_\phi(X)$ is the Gaussian kernel given by

$$(S_\phi(X))_{ij} = \exp \left(-\frac{\|x_i - x_j\|_2^2}{\phi^2} \right).$$

Here, ϕ determine the length scale of how large a neighborhood to correlate to and x_i is the two or three dimensional coordinates of cell i . The function $\text{rnorm}(\cdot)$ defines a matrix transformation that row normalizes. This means $\text{rnorm}(S_\phi(X))$ is a normalized version of $S_\phi(X)$, where the rows sum to one to ensure a neighborhood average and H are normalized as seen in the usual Poisson updates in (9).

The additional update in (26) is not derived to minimize a specific cost function. Instead it is inspired by the statistical rule, that given a vector of independent standard normal distributions $Z \in \mathbb{R}^N$, then the vector $S_\phi(X)Z$ will have covariance $S_\phi(X)S_\phi(X)^T = S_\phi(X)^2$. And because $(S_\phi(X))_{ij}$ is positive for neighboring cells i and j , we also know that the correlation is increased between neighbors. The normalization of the rows of $S_\phi(X)$ is necessary to assure that the mean value of the entries in W is not changed.

All in all the updates (24) and (25) will minimize the GKL, while the last update (26) will increase the correlation in W between neighboring cells.

Perspectives

The huge advantage of the updates for NMF in Paper E is that they are simple, intuitive and performs well on real data compared to other methods in the literature. The other methods that are extending NMF use the very time consuming gradient decent, as the models are too complex to derive multiplicative updates. This means they do not apply to the increasing amount larger datasets with up to a million cells or multiple tissues.

In Paper E, NMF is applied to three different datasets, where one of these include a dataset of two slices with almost two million cells in total. This shows its ability to apply to very high-dimensional data. It can also be seen from the

8. Computational implementation

results that the method gives a more detailed distinction of smaller structures in the tissue compared to other methods that are bagging cells and not preserving the information of each cell. Additionally, the method is compared to two of the best performing methods in the literature (Li and Zhou, 2022; Yuan, 2024). Here, NNMF runs several times faster than the method that produce comparable results and the other method is faster than NNMF, but performs significantly worse in the results. Results and further details on the method can be found in Paper E.

8 Computational implementation

During the implementation of different algorithms for NMF I have faced a lot of non-statistical challenges. Some of the most important and crucial ones to be aware of is mentioned below.

8.1 Floating point precision

In some cases the entries of W and H reach zero in the estimation, even though the theory states that this is impossible if they are initialized by positive entries (Lee and Seung, 1999). This is caused by a common issue in programming, which is floating point precision. A floating number has a limit in the number of decimals saved, which means very small decimal numbers in W or H are rounded to zero. This can result in division by zero in the multiplicative updates, which causes an error in the estimation. In our implementations we have solved this by adding a lower limit of $1e^{-10}$ to the entries of W and H during estimation.

8.2 Initialization

As mentioned in section 3 the updates for W and H from one initialization only assures a local minimum. It is therefore important to initialize the updates multiple times to increase the chance of finding the global minimum. The algorithms for NMF are only suboptimal, as there is no guarantee of recovering the global minimum. However, we assume the global minimum is reached when the solution is consistent across runs.

Sometimes a very large number of initializations need to be given to retrieve the global minimum. This is for example the case for the flexible parametric model in Paper C, which often required at least 100 random initialization to make sure that it recovered the global minimum. It requires an extensive amount of computational power to run the algorithm to convergence 100 times, so often other approaches are used to circumvent this.

In Biernacki et al. (2003) they test a lot of different ways of initializing an EM-algorithm to recover the highest likelihood. They show that the most efficient way is to initialize a very large number of times for a few iterations, and then only run the best one until convergence.

This procedure was implemented for the parametric NMF algorithm in Paper C, where we initialized the algorithm 100 times. For each initialization we ran 50 iterations, and then we selected the best performing initialization to run until convergence.

Another way to get consistent results is to be more accurate in the initialization. In Paper E we initialize standard NMF multiple times and iterate each 50 times. Neighborhood NMF is then initialized by the best factorization from the standard model. This initialization procedure is often referred to as a *warm start*, which is popular in machine learning. The idea is to create a more sensible initialization of an algorithm to make convergence faster.

8.3 Stopping criteria

The stopping criteria of the iterative updates are most commonly chosen based on the convergence of the cost function. Such that the algorithm is assumed to have converged and stopped, when the cost function changes less than ϵ .

As the algorithms for NMF often has a slow convergence in the end it is essential to chose ϵ such that an optimal solution is recovered, but also avoid that it runs forever with minimal change in the cost. The value of the cost function can deviate extensively between different datasets, which can make it hard to recover a shared optimal threshold. In our implementations we therefore divide the difference in the cost function by the current value to look at the relative decrease in the cost function. Then we instead stop the algorithm, when this relative difference changes less than ϵ . This makes it easier to construct a common value for ϵ , which we often set between $1e^{-5}$ to $1e^{-10}$.

In all our implementations we have also added an additional stopping criteria of a maximum number of iteration to make sure they do not continue forever. This is set to 10.000 for most of the papers, except for paper E where it is set to 500 iterations.

8.4 Boosting R Performance

All the algorithms and packages from the included papers were constructed in R, which is one of the most popular language for statistical computing. Even though it is very intuitive to use it has a disadvantage in its computational speed. One way to minimize this disadvantage of R is to utilize its Foreign Function Interface with C++ using the package Rcpp. This enables the user to easily and with minimal overhead make function calls to compiled C++ code. Writing code in C++ enables manual memory management and allows the user to write more performant code. For all our implementations we have moved the core iterations into C++ to greatly speed up the estimation. Below, the publicly available packages and repositories from the included papers are listed. Further details and documentations can be found on the respective links.

8. Computational implementation

Public packages

SFS : www.github.com/ragnhildlaursen/SFS (Paper A)
SigMoS : www.github.com/MartaPelizzola/SigMoS (Paper B)
NNMF : www.github.com/ragnhildlaursen/NNMF (Paper E)

Public repositories

Paper A : www.github.com/ragnhildlaursen/sampleSFS_paper
Paper C : www.github.com/ragnhildlaursen/paramNMF_ms

References

- L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259, 2013a.
- L. B. Alexandrov, S. Nik-Zainal, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013b.
- L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton. Clock-like mutational processes in human somatic cells. *Nature genetics*, 47(12):1402–1407, 2015.
- L. B. Alexandrov, Y. S. Ju, K. Haase, P. Van Loo, I. Martincorena, S. Nik-Zainal, Y. Totoki, A. Fujimoto, H. Nakagawa, T. Shibata, P. J. Campbell, P. Vineis, D. H. Phillips, and M. R. Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, nov 2016. doi: 10.1126/SCIENCE.AAG0299.
- L. B. Alexandrov et al. The repertoire of mutational signatures in human cancer. *Nature* 2020 578:7793, 578(7793):94–101, feb 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-1943-3.
- C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.
- O. S. Borgen and B. R. Kowalski. An extension of the multivariate component-resolution method to three components. *Analytica Chimica Acta*, 174:1–26, 1985.
- R. Bro, K. Kjeldahl, A. K. Smilde, and H. Kiers. Cross-validation of component models: a critical look at current methods. *Analytical and bioanalytical chemistry*, 390:1241–1251, 2008.
- 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*, 101(12):4164–4169, 2004.
- D. Caruso, A. Papa, S. Tomao, P. Vici, P. B. Panici, and F. Tomao. Niraparib in ovarian cancer: results to date and clinical potential. *Therapeutic advances in medical oncology*, 9(9):579–588, 2017.
- K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.

References

- Z. Chen, I. Soifer, H. Hilton, L. Keren, and V. Jojic. Modeling multiplexed images with spatial-lda reveals novel tissue microenvironments. *Journal of Computational Biology*, 27(8): 1204–1218, 2020.
- B. Chidester, T. Zhou, S. Alam, and J. Ma. Spicemix enables integrative single-cell spatial modeling of cell identity. *Nature genetics*, 55(1):78–88, 2023.
- C. H. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.
- K. Dong and S. Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.
- D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16:1141–1148, 2003.
- A. Fischer, C. J. Illingworth, P. J. Campbell, and V. Mustonen. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, 14(4):1–10, 2013.
- A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24:997–1016, 2014.
- N. Gillis and R. Rajkó. Partial identifiability for nonnegative matrix factorization. *SIAM Journal on Matrix Analysis and Applications*, 44(1):27–52, 2023.
- K. Gori and A. Baez-Ortega. sigfit: flexible bayesian inference of mutational signatures. *bioRxiv*, page 372896, 2018.
- O. Gouvert, T. Oberlin, and C. Fevotte. Negative Binomial Matrix Factorization. *IEEE Signal Processing Letters*, 27:815–819, 2020. ISSN 15582361. doi: 10.1109/LSP.2020.2991613.
- R. C. Henry and B. M. Kim. Extension of self-modeling curve resolution to mixtures of more than three components: Part 1. Finding the basic feasible region. *Chemometrics and Intelligent Laboratory Systems*, 8(2):205–216, 1990.
- P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9), 2004.
- J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.
- H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.
- M. Krawczak, E. V. Ball, and D. N. Cooper. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *The American Journal of Human Genetics*, 63(2):474–488, 1998.

- A. Lal, K. Liu, R. Tibshirani, A. Sidow, and D. Ramazzotti. De novo mutational signature discovery in tumor genomes using SparseSignatures. *PLOS Computational Biology*, 17(6): e1009119, jun 2021. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1009119.
- H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen. Theorems on positive data: On the uniqueness of NMF. *Computational intelligence and neuroscience*, 2008.
- R. Laursen and A. Hobolth. A sampling algorithm to compute the set of feasible solutions for nonnegative matrix factorization with an arbitrary rank. *SIAM Journal on Matrix Analysis and Applications*, 43(1):257–273, 2022.
- R. Laursen, L. Maretty, and A. Hobolth. Flexible model-based non-negative matrix factorization with application to mutational signatures. *Statistical Applications in Genetics and Molecular Biology*, 23(1):20230034, 2024.
- W. H. Lawton and E. A. Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3): 617–633, 1971.
- D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, oct 1999. ISSN 00280836. doi: 10.1038/44565.
- D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- J. Y. Lee, P. J. Green, and L. M. Ryan. On the “Poisson Trick” and its extensions for fitting multinomial regression models. *arXiv:1707.08538*, 2017.
- Z. Li and X. Zhou. Bass: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome biology*, 23(1):168, 2022.
- X. Lyu, J. Garret, G. Rätsch, and K. V. Lehmann. Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics*, 36 (Supplement_1):I154–I160, jul 2020. ISSN 14602059. doi: 10.1093/BIOINFORMATICS/BTAA473.
- Y. Ma and X. Zhou. Spatially informed cell-type deconvolution for spatial transcriptomics. *Nature biotechnology*, 40(9):1349–1359, 2022.
- S. Moussaoui, D. Brie, and J. Idier. Non-negative source separation: range of admissible solutions and conditions for the uniqueness of the solution. In *Proceedings.(ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–289. IEEE, 2005.
- P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.
- A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). *IEEE transactions on pattern analysis and machine intelligence*, 28(3):403–415, 2006.

References

- V. P. Pauca, J. Piper, and R. J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear algebra and its applications*, 416(1):29–47, 2006.
- M. Pelizzola, R. Laursen, and A. Hobolth. Model selection and robust inference of mutational signatures using negative binomial non-negative matrix factorization. *BMC bioinformatics*, 24(1):187, 2023.
- K. Pelka, M. Hofree, J. H. Chen, S. Sarkizova, J. D. Pirl, V. Jorgji, A. Bejnood, D. Dionne, H. G. William, K. H. Xu, et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18):4734–4752, 2021.
- R. A. Rosales, R. D. Drummond, R. Valieris, E. Dias-Neto, and I. T. Da Silva. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics*, 33(1):8–16, 2017.
- M. Sawall and K. Neymeyr. A ray casting method for the computation of the area of feasible solutions for multicomponent systems: Theory, applications and facpack-implementation. *Analytica chimica acta*, 960:40–52, 2017.
- M. Sawall, C. Kubis, D. Selent, A. Boerner, and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: concepts and applications. *Journal of Chemometrics*, 27(5):106–116, 2013.
- A. Shibai, Y. Takahashi, Y. Ishizawa, D. Motooka, S. Nakamura, B.-W. Ying, and S. Tsuru. Mutation accumulation under UV radiation in *Escherichia coli*. *Scientific Reports*, 7(1):1–12, nov 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-15008-1.
- Y. Shiraishi, G. Tremmel, S. Miyano, and M. Stephens. A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS genetics*, 11(12):e1005657, 2015.
- Z. Sondka, N. B. Dhir, D. Carvalho-Silva, S. Jupe, Madhumita, K. McLaren, M. Starkey, S. Ward, J. Wilding, M. Ahmed, et al. Cosmic: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*, 52(D1):D1210–D1217, 2024.
- The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020.
- F. W. Townes and B. E. Engelhardt. Nonnegative spatial factorization applied to spatial genomics. *Nature methods*, 20(2):229–238, 2023.
- B. Velten and O. Stegle. Principles and challenges of modeling temporal and spatial omics data. *Nature Methods*, 20(10):1462–1474, 2023.
- B. M. Wise, N. B. Gallagher, R. Bro, J. Shaver, W. Windig, and R. S. Koch. Pls toolbox 4.0, 2007.
- S. Wold. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20(4):397–405, 1978.
- Z. Yuan. Mender: fast and scalable tissue structure identification in spatial omics data. *Nature Communications*, 15(1):207, 2024.

References

- Z. Yuan, Y. Li, M. Shi, F. Yang, J. Gao, J. Yao, and M. Q. Zhang. Sotip is a versatile method for microenvironment modeling with spatial omics data. *Nature Communications*, 13(1): 7330, 2022.
- T. Zhang, P. Joubert, N. Ansari-Pour, W. Zhao, P. H. Hoang, R. Lokanga, A. L. Moye, J. Rosenbaum, A. Gonzalez-Perez, F. Martinez-Jimenez, et al. Genomic and evolutionary classification of lung cancer in never smokers. *Nature genetics*, 53(9):1348–1359, 2021.
- E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uyttingco, S. E. Taylor, P. Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.

Paper

A large, bold, white capital letter 'A' is centered within a solid black rectangular box.

**A sampling algorithm to compute the set of feasible solutions
for non-negative matrix factorization with an arbitrary rank**

by Ragnhild Laursen and Asger Hobolth

Published in SIAM Journal on Matrix Analysis and Applications

A SAMPLING ALGORITHM TO COMPUTE THE SET OF FEASIBLE SOLUTIONS FOR NONNEGATIVE MATRIX FACTORIZATION WITH AN ARBITRARY RANK*

RAGNHILD LAURSEN[†] AND ASGER HOBOLTH[†]

Abstract. Nonnegative matrix factorization (NMF) is a useful method to extract features from multivariate data, but an important and sometimes neglected concern is that NMF can result in nonunique solutions. Often, there exist a set of feasible solutions (SFS), which makes it more difficult to interpret the factorization. This problem is especially ignored in cancer genomics, where NMF is used to infer information about the mutational processes present in the evolution of cancer. In this paper the extent of nonuniqueness is investigated for two mutational counts data, and a new sampling algorithm that can find the SFS is introduced. Our sampling algorithm is easy to implement and applies to an arbitrary rank of NMF. This is in contrast to state of the art, where the NMF rank must be smaller than or equal to four. For lower ranks we show that our algorithm performs similar to the polygon inflation algorithm that is developed in relation to chemometrics. Furthermore, we show how the size of the SFS can have a high influence on the appearing variability of a solution. Our sampling algorithm is implemented in the R package SFS (<https://github.com/ragnhildlaursen/SFS>).

Key words. identifiability, mutational processes, nonnegative matrix factorization (NMF), sampling, uniqueness

AMS subject classifications. 15A23, 62-04, 62P10

DOI. 10.1137/20M1378971

1. Introduction. The applications of nonnegative matrix factorization (NMF) are many, and one of them is in pure component analysis for analytical chemistry. In this field, it is a big obstacle that the solution from NMF is nonunique, such that there exist a set of feasible solutions (SFS) and not only one. As a consequence, there exists a vast literature in chemometrics on finding the SFS, both for general applications of NMF and more specific applications to chemical data [3, 7, 14, 21]. Having a nonunique solution to NMF makes it problematic to interpret the factorization. NMF is an unsupervised learning method that factorizes a nonnegative data matrix $M \in \mathbb{R}_+^{K \times G}$ into two nonnegative matrices $P \in \mathbb{R}_+^{K \times N}$ and $E \in \mathbb{R}_+^{N \times G}$ [15]. The rank N is usually chosen much smaller than K and G . This means the factorization is an approximation

$$M \approx PE.$$

In the remainder of this paper we assume that such an approximation is found; we refer to section 5 for a discussion of this issue. All the entries in P and E are free parameters that need to be estimated. The problem with the factorization is the large amount of free parameters such that other solutions can be constructed by finding invertible matrices $A \in \mathbb{R}^{N \times N}$ that fulfill

$$(1.1) \quad \tilde{P} = PA \geq 0 \quad \tilde{E} = A^{-1}E \geq 0.$$

Then the product of \tilde{P} and \tilde{E} gives the exact same approximation of M as the product of P and E . Trivial ambiguities exist when A is either a diagonal matrix

*Received by the editors November 6, 2020; accepted for publication (in revised form) by P. Drineas October 28, 2021; published electronically February 22, 2022.

<https://doi.org/10.1137/20M1378971>

[†]Department of Mathematics, Aarhus University, Aarhus C, 8000, Denmark (ragnhild@math.au.dk, asger@math.au.dk).

1. Introduction

or a permutation matrix, which scales or reorders the columns in P and rows in E . These ambiguities are always possible, so it is standard to define P and E as a unique solution to NMF if the only possible ambiguities are scaling and/or reordering [4, 11, 13, 17]. Here, the SFS for P and E are defined as

$$(1.2) \\ \mathcal{M}(P) = \left\{ \tilde{P} \in \mathbb{R}_+^{K \times N} \mid \exists \text{ invertible } A \in \mathbb{R}^{N \times N} : \tilde{P} = PA \geq 0 \text{ and } \tilde{E} = A^{-1}E \geq 0 \right\}, \\ \mathcal{M}(E) = \left\{ \tilde{E} \in \mathbb{R}_+^{N \times G} \mid \exists \text{ invertible } A \in \mathbb{R}^{N \times N} : \tilde{P} = PA \geq 0 \text{ and } \tilde{E} = A^{-1}E \geq 0 \right\},$$

where A is normalized such that the columns of $\tilde{P} = PA$ sum to one, as this removes the scaling ambiguity. This can be shown as follows:

$$\sum_{j=1}^K \tilde{P}_{jn} = \sum_{j=1}^K \left\{ \sum_{i=1}^N P_{ji} A_{in} \right\} = \sum_{i=1}^N A_{in} \underbrace{\sum_{j=1}^K P_{ji}}_{=1} = \sum_{i=1}^N A_{in}$$

for each $n = 1, \dots, N$ and therefore constructing A such that the columns sum to one would automatically ensure the columns of \tilde{P} to sum to one, such that the whole SFS, (\tilde{P}, \tilde{E}) , have the same scaling. The problem of reordering of the entries can be solved by ordering by cosine similarity.

In chemometrics, the SFS for a given NMF solution has been investigated since 1971, where Lawton and Sylvester [14] first introduced the analytical calculation for finding the SFS for rank two, i.e., $N = 2$. A vast literature has later evolved on finding the SFS for a higher rank, $N \geq 3$, both by analytical calculations and with numerical algorithms [3, 7, 10, 18]. The analytical calculations of the SFS have been extended to ranks of three and four [3, 10] but require a significant amount of calculations when the size of K and G are large, as it increases the number of inequalities in (1.1). For a higher rank, the polygon inflation algorithm [21] and the ray casting algorithm [22] are numerical methods for approximating the SFS. We refer to [9] for a review of methods for determining the SFS. Here, we will focus on the recent polygon inflation algorithm [21] to compare with our new sampling algorithm. Our sampling algorithm has the advantage of being able to compute the SFS for an arbitrary rank of the factorization, which is needed for many NMF applications. In particular we will focus on applications in cancer genomics, where NMF is used to infer information about the mutational processes in cancer evolution.

In the field of cancer genomics, the assumption is that all the mutations in a cancer genome come from a spectrum of N mutational process that each have their own effect on the cancer genome. The overall goal is to identify these mutational processes and use it to enhance the understanding of cancer evolution [19]. A popular method to infer information on the mutational processes from a matrix of mutational counts is NMF. Here, M is a $K \times G$ matrix of mutational counts. The G columns of dimension $K \times 1$ represent the mutational catalog for each genome, which are found from sequencing a cancer genome. The number of different mutation types, K , in the data sets is 96 as they include the 6 base substitutions C>A, C>G, C>T, T>A, T>C, T>G, and the immediate 3' and 5' neighboring bases, i.e., $6 \cdot 4 \cdot 4 = 96$ different mutation types. To recover the mutational processes from the mutational catalogs, we assume that each mutational catalog is a positive linear combination of a certain number of mutational processes, N . This means approximating M by the factorization of two nonnegative matrices P and E of dimension $K \times N$ and $N \times G$,

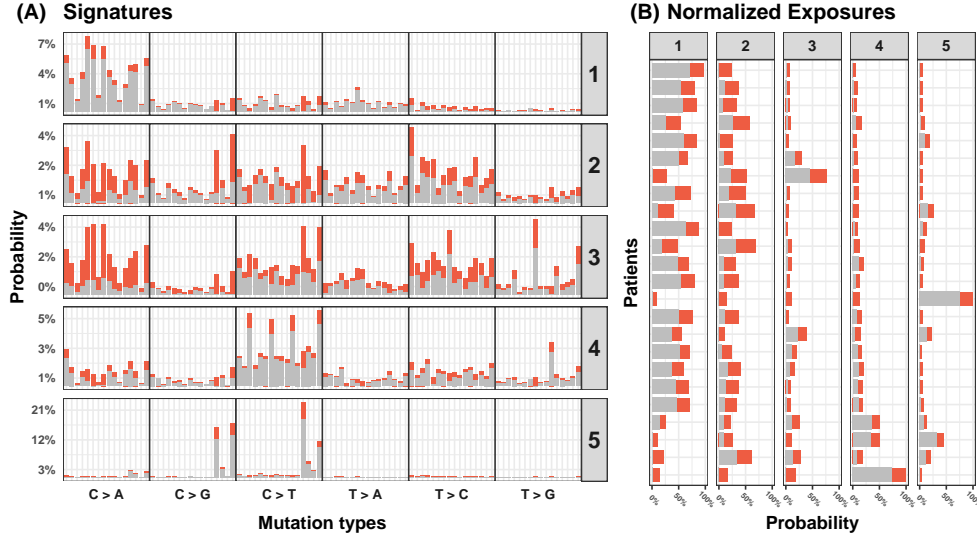


FIG. 1. The signatures and exposures for the Lung A. cancer data, when assuming five mutational processes. In (A) the signatures are depicted, with the SFS marked in red from the minimum feasible value to the maximum feasible value. In (B) the normalized exposures are depicted, where the SFS is again marked in red. The signatures are arranged according to their total exposure, i.e., $\sum_{g=1}^G E_{ng}$ in decreasing order.

respectively, such that $M \approx PE$. The matrix P represents the N signatures for the mutational processes and the matrix E represents their exposures for each mutational catalog. The size of N is usually chosen magnitudes smaller than K and G , which makes P and E a low-dimensional representation of M describing its main features.

Through this paper we refer to two different data sets of mutations in whole genomes. One of them is the 21 Breast cancer genomes, which is a commonly known and analyzed data set in relation to mutational processes in cancer genomics [2, 5, 20], which we refer to as the Breast cancer data. The data is a matrix of dimension 96×21 consisting of the mutational counts for the 96 different mutation types in 21 different breast cancer genomes. The other data set consists of the mutational counts from 24 patients having Lung Adenocarcinoma cancer taken from [1]. The data set is referred to as the Lung A. cancer data and is a 96×24 matrix of mutational counts. The Lung A. cancer data is chosen because it has a large SFS, even for large sizes of N . The SFS found with our sampling algorithm for the Lung A. cancer when assuming five mutational processes is depicted in Figure 1. The red bars illustrate the SFS from the minimum feasible value to the maximum feasible value for each of the different entries. In Figure 1 we observe that the range of the SFS varies over the different mutation types and patients, as the changes are very dependent on the structure of one another. The code and data for this paper is available at https://github.com/ragnhildlaursen/sampleSFS_paper.

The paper is structured as follows. In section 2 we introduce and explain our sampling algorithm. In section 3 we apply our algorithm on the Breast and Lung A. cancer data and compare it to the polygon inflation algorithm. Section 4 is a further analysis of the optimal choice of parameters as well as running time of our sampling algorithm. In section 5, we at last cover the problem of identifiability of P and E in relation to initialization for the updates made by Lee and Seung [15] and noise in the data. These identifiability problems are compared to the influence of the SFS. The

2. The sampling algorithm

paper ends with concluding remarks regarding our sampling algorithm, choice of rank N , and the lack of uniqueness of the matrix factorization.

2. The sampling algorithm. Recall that our starting point is two nonnegative matrices $P \in \mathbb{R}_+^{K \times N}$ and $E \in \mathbb{R}_+^{N \times G}$ that approximate our data $M \in \mathbb{R}_+^{K \times G}$ and that we want to find the SFS in (1.2) for both P and E . The general idea of our algorithm is to use the simple analytical calculation for rank two and adapt it to higher dimensions through sampling. First, the SFS is described in the simple setting of $N = 2$ to ease the understanding of the general setting.

2.1. SFS for $N = 2$. For the case of a rank two factorization, the SFS for P and E can be found in a closed form. The calculations we make here are similar to the ones found in [17], but here we find the SFS for one column of P at a time. Assume we would like to find the SFS for the first column of P . Then we set

$$(2.1) \quad A_{12}(\lambda) = \begin{pmatrix} 1 - \lambda & 0 \\ \lambda & 1 \end{pmatrix}.$$

The inverse $A_{12}^{-1}(\lambda)$ is simple as well and can be directly expressed in terms of the original matrix as

$$(2.2) \quad A_{12}^{-1}(\lambda) = \frac{1}{1 - \lambda} \begin{pmatrix} 1 & 0 \\ -\lambda & 1 - \lambda \end{pmatrix} = \begin{pmatrix} 1 + \frac{\lambda}{1 - \lambda} & 0 \\ -\frac{\lambda}{1 - \lambda} & 1 \end{pmatrix} = A_{12} \begin{pmatrix} -\frac{\lambda}{1 - \lambda} \\ 1 \end{pmatrix}.$$

The simple inverse eases both the calculations and computation time for the algorithm. The feasible values of λ must fulfill that all entries of $\tilde{P} = PA_{12}(\lambda)$ and $\tilde{E} = A_{12}^{-1}(\lambda)E$ remain nonnegative, which can be formulated as

$$(2.3) \quad 0 \leq PA_{12}(\lambda) = \begin{pmatrix} P_{11}(1 - \lambda) + P_{12}\lambda & P_{12} \\ \vdots & \vdots \\ P_{K1}(1 - \lambda) + P_{K2}\lambda & P_{K2} \end{pmatrix} = \begin{pmatrix} P_{11} - \lambda(P_{11} - P_{12}) & P_{12} \\ \vdots & \vdots \\ P_{K1} - \lambda(P_{K1} - P_{K2}) & P_{K2} \end{pmatrix}$$

and

$$(2.4) \quad \begin{aligned} 0 \leq A_{12}^{-1}(\lambda)E &= \frac{1}{1 - \lambda} \begin{pmatrix} E_{11} & \dots & E_{1G} \\ -E_{11}\lambda + E_{21}(1 - \lambda) & \dots & -E_{1G}\lambda + E_{2G}(1 - \lambda) \end{pmatrix} \\ &= \frac{1}{1 - \lambda} \begin{pmatrix} E_{11} & \dots & E_{1G} \\ E_{21} - \lambda(E_{11} + E_{21}) & \dots & E_{2G} - \lambda(E_{1G} + E_{2G}) \end{pmatrix}. \end{aligned}$$

A general requirement is $\lambda < 1$ to assure the entries in (2.4) remain nonnegative. The entries in (2.4) result in an additional upper bound for λ given by

$$(2.5) \quad \bar{\Lambda}_{12} = \min_{g=1, \dots, G} \left\{ \frac{E_{2g}}{E_{1g} + E_{2g}} \mid E_{1g} + E_{2g} > 0 \right\} \geq 0.$$

In (2.3), the requirement of $\lambda < 1$ assures $P_{k1} - \lambda(P_{k1} - P_{k2}) \geq 0$ for all k where $P_{k1} \geq P_{k2}$. For the other case where $P_{k2} > P_{k1}$ we get the following lower bound for λ

$$(2.6) \quad \underline{\Lambda}_{12} = \max_{k=1, \dots, K} \left\{ \frac{P_{k1}}{P_{k1} - P_{k2}} \mid P_{k2} > P_{k1} \right\} \leq 0.$$

This means that all $\lambda \in [\underline{\Lambda}_{12}, \bar{\Lambda}_{12}]$ give feasible solutions for the first column of P , while the second column is fixed. The bounds of λ for the second column of P will

be similar, but where 1 and 2 are simply switched in (2.5) and (2.6). In this simple setting, the individual feasible intervals of $[\underline{\Lambda}_{12}, \bar{\Lambda}_{12}]$ and $[\underline{\Lambda}_{21}, \bar{\Lambda}_{21}]$ give the whole SFS for both P and E . In the case of a rank higher than two the analytical calculations for the SFS are substantially more complicated. Calculations for N equal to 3 can be seen in [3, 10].

2.2. Sampling the SFS for an arbitrary rank. Our algorithm uses the analytical calculations for $N = 2$ above and is inspired by Gibbs sampling [6]. The idea is to change each column of P with an affine combination of another column chosen by random sampling. This is done sequentially for each column of P while updating E correspondingly such that the matrix product remains the same. For the approach we define a general transformation matrix $A_{ij}(\lambda)$, of dimension $N \times N$, similar to the one in (2.1):

$$(2.7) \quad (A_{ij}(\lambda))_{uv} = \begin{cases} 1 - \lambda & \text{if } u = v = i, \\ 1 & \text{if } u = v \neq i, \\ \lambda & \text{if } u = j, v = i, \\ 0, & \text{otherwise.} \end{cases}$$

The transformation $A_{ij}(\lambda)$ changes column P_i to an affine combination of column P_i and P_j , where $j \neq i$. Let Λ_{ij} denote the feasible interval for λ given a solution P and E , such that $PA_{ij}(\lambda) \geq 0$ and $A_{ij}^{-1}(\lambda)E \geq 0$. The endpoints of this interval are similar to the ones in (2.5) and (2.6), i.e., a function of column i and j in P and row i and j in E

$$(2.8) \quad \begin{aligned} \underline{\Lambda}_{ij} &= \underline{f}(P, E, i, j) := \max_k \left\{ \frac{P_{ki}}{P_{ki} - P_{kj}} \mid P_{kj} > P_{ki} \right\} \leq 0, \\ \bar{\Lambda}_{ij} &= \bar{f}(P, E, i, j) := \min_g \left\{ \frac{E_{jg}}{E_{ig} + E_{jg}} \mid E_{ig} + E_{jg} > 0 \right\} \geq 0 \end{aligned}$$

such that $\Lambda_{ij} = f(P, E, i, j) = [\underline{\Lambda}_{ij}, \bar{\Lambda}_{ij}]$.

2.3. Sampling a value λ in Λ_{ij} . The most simple way to choose λ is uniformly at random in the interval Λ_{ij} , but this choice often gives a slow convergence. The endpoints of the interval often lead to larger changes in P and E and are therefore more favorable. The choice of $\lambda \in \Lambda_{ij}$ is chosen as a shifted symmetric beta distribution with equal shape parameters denoted by β . Setting $\beta = 1$ results in a uniform choice, and setting $\beta < 1$ gives a higher probability at the endpoints. The different choices of β are further described in subsection 4.1 but for now we fix $\beta = 0.5$. Sampling $\lambda \in \Lambda_{ij}$ is done by sampling x from a regular beta distribution with the shape parameters equals to β and then setting $\lambda = x \cdot \bar{\Lambda}_{ij} + (1 - x) \cdot \underline{\Lambda}_{ij}$. After sampling a $\lambda \in \Lambda_{ij}$, the matrices P and E are updated to $P_{new} = PA_{ij}(\lambda)$ and $E_{new} = A_{ij}^{-1}(\lambda)E$. In one iteration, this is done sequentially for $i = 1, \dots, N$, where there is chosen a random $j \neq i$ to mix with for each i .

2.4. Defining the size of the SFS and the stopping criteria. After each iteration the values of P and E are saved, so after \mathcal{S} iterations we have the following samples from the SFS for both P and E :

$$\begin{aligned} \mathbf{P}^{\mathcal{S}} &= \{P^0, P^1, \dots, P^{\mathcal{S}}\}, \\ \mathbf{E}^{\mathcal{S}} &= \{E^0, E^1, \dots, E^{\mathcal{S}}\}. \end{aligned}$$

3. Applications and comparison with polygon inflation algorithm

This means every entry in P and E have S samples besides the initial solution (P^0, E^0) . We often observe that for some entries all samples are equivalent and for others they spread across an interval. Each new sample is created as an affine combination of the previous sample including itself, which means all the values between two samples in the interval will be feasible as well. The SFS for each entry is therefore defined as the interval between the minimum and the maximum value of these S samples. Similar to the polygon inflation algorithm [21], our algorithm also assumes that the SFS consists of connected sets as the samples are affine combinations of one another. This assumption has not appeared to restrict the SFS for any of the data sets we have used, but in theory the solution could have a larger SFS than what is found by our algorithm. We define the size of \mathbf{P}^S as the average change of each entry across the full sample:

$$(2.9) \quad \text{avg}\langle \mathbf{P}^S \rangle = \frac{1}{K \cdot N} \sum_{n=1}^N \sum_{k=1}^K \left\{ \max_{s=0,1,\dots,S} P_{kn}^s - \min_{s=0,1,\dots,S} P_{kn}^s \right\}.$$

To assure that the algorithm stops at the right time, the size of the SFS is calculated after each \mathcal{T} iterations, which we have chosen to fix at $\mathcal{T} = 1000$. The algorithm is stopped when the size of the SFS has changed less than ϵ within these \mathcal{T} iterations

$$\text{avg}\langle \mathbf{P}^S \rangle - \text{avg}\langle \mathbf{P}^{S-\mathcal{T}} \rangle < \epsilon,$$

where we set $\epsilon = 10^{-10}$. This assures that our algorithm continues until convergence and stops when no more changes are available. One could add a similar stopping criteria for \mathbf{E}^S , but in our experiments we found the proposed criteria satisfactory to find the whole SFS for both P and E . The sampling algorithm is summarized in Algorithm 3.1.

3. Applications and comparison with polygon inflation algorithm. In this section results from the sampling algorithm are compared with the polygon inflation algorithm. The comparison is made for both the Breast and Lung A. cancer data, when assuming three mutational processes. The results from our sampling algorithm are illustrated in Figure 2 with the representation defined in (1.2). Before making the comparison we have to define another representation of the SFS that is more commonly used in chemometrics [3, 7, 10, 18, 21] and in particular for the polygon inflation algorithm. After introducing this representation, the assumption of three mutational processes for the comparison is more easily justified.

3.1. Singular value decomposition to represent the SFS. In chemometrics it is common to define the SFS relative to the singular value decomposition (SVD). Note, any matrix $PE \in \mathbb{R}_+^{K \times G}$ of rank N can be decomposed into

$$PE = U \Sigma V',$$

where $U \in \mathbb{R}^{K \times N}$ and $V \in \mathbb{R}^{G \times N}$ are the N eigenvectors for $PE(PE)'$ and $(PE)'PE$, respectively, and $\Sigma \in \mathbb{R}^{N \times N}$ is the diagonal matrix of singular values. The factorization into U and $\Sigma V'$ consist of both positive and negative entries but can be transformed by a matrix $T \in \mathbb{R}^{N \times N}$ such that $UT \geq 0$ and $T^{-1}\Sigma V' \geq 0$ [16]. To remove the scaling ambiguities of the factorization, the columns of T are normalized and defined by

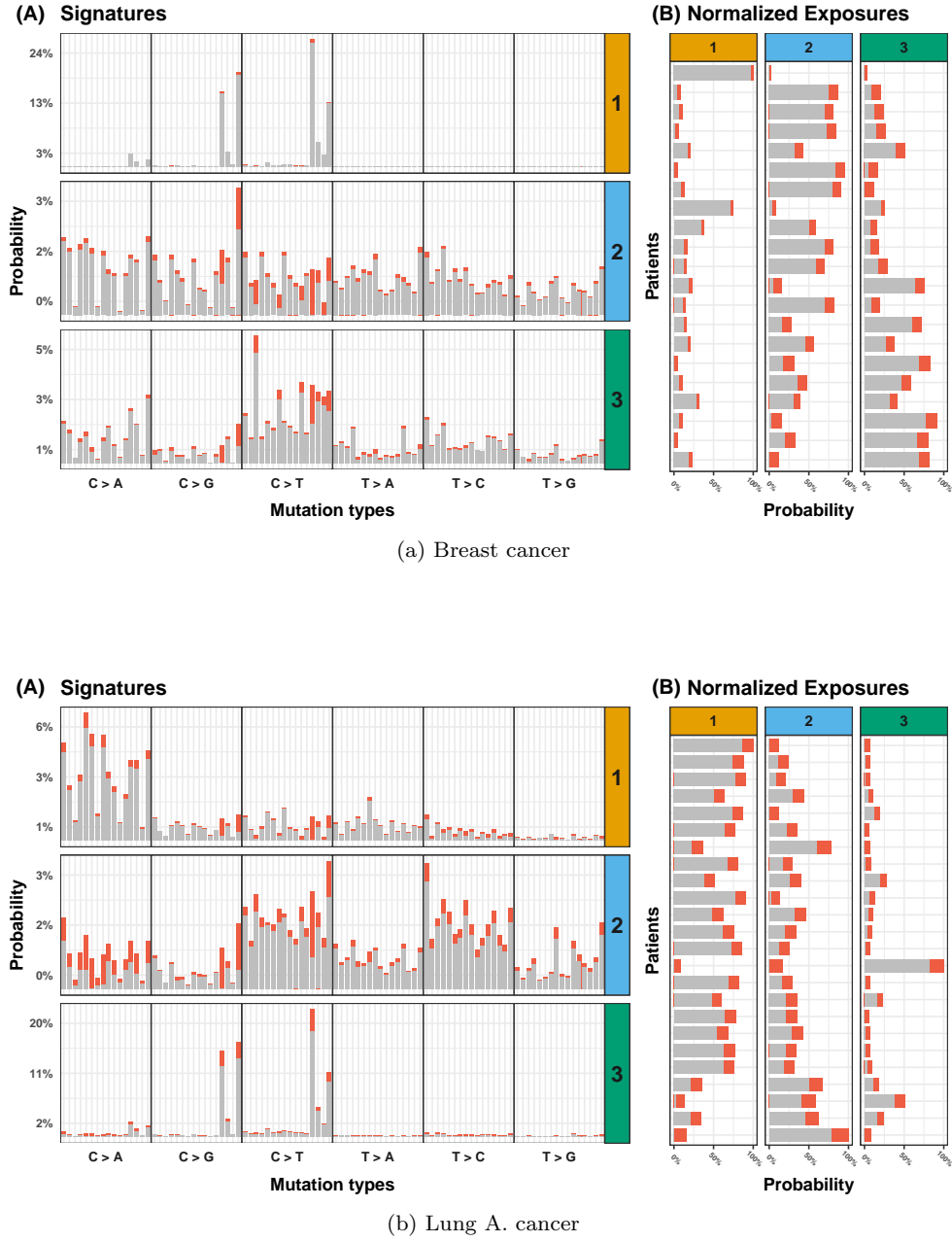


FIG. 2. The signatures and exposures for the Breast and Lung A. cancer data, when assuming three mutational processes. In (A) the signatures are depicted, with the SFS marked in red from the minimum feasible value to the maximum feasible value. In (B) the normalized exposures are depicted, and the SFS is again marked in red. The signatures are arranged according to their total exposure, i.e., $\sum_{g=1}^G E_{ng}$ in decreasing order.

$$(3.1) \quad T = \left(\begin{array}{c|ccc} 1 & 1 & \cdots & 1 \\ \alpha_1 & w_{1,1} & \cdots & w_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N-1} & w_{N-1,1} & \cdots & w_{N-1,N-1} \end{array} \right) = \begin{pmatrix} 1 & e' \\ \alpha & W \end{pmatrix}$$

3. Applications and comparison with polygon inflation algorithm

Algorithm 3.1 Finding the SFS given an initial solution (\hat{P}, \hat{E}) .

```

1: Initialize  $P^0 = \hat{P}$  and  $E^0 = \hat{E}$ 
2: for  $s = 1, 2, \dots$  do
3:    $P_1 \leftarrow P^{s-1}$ 
4:    $E_1 \leftarrow E^{s-1}$ 
5:   for  $i = 1, \dots, N$  do
6:      $j \leftarrow \text{random element of } \{1, \dots, N\} \setminus \{i\}$ 
7:      $\Lambda_{ij} = f(P_i, E_i, i, j)$ 
8:      $x \leftarrow \text{Beta}(\beta, \beta)$ 
9:      $\lambda = x \cdot \Lambda_{ij} + (1 - x) \cdot \underline{\Lambda}_{ij}$ 
10:     $P_{i+1} = P_i A_{ij}(\lambda)$ 
11:     $E_{i+1} = A_{ij}(-\frac{\lambda}{1-\lambda}) E_i$ 
12:   end for
13:    $P^s = P_{N+1}$ 
14:    $E^s = E_{N+1}$ 
15: end for if  $\text{avg}\langle \mathbf{P}^s \rangle - \text{avg}\langle \mathbf{P}^{s-T} \rangle < \epsilon$ 

```

such that $\alpha = (\alpha_1, \dots, \alpha_{N-1})' \in \mathbb{R}^{(N-1)}$, $W \in \mathbb{R}^{(N-1) \times (N-1)}$, and $e' = (1, \dots, 1) \in \mathbb{R}^{(N-1)}$. The scaling of the columns are made such that the first entry of each column is 1. Here, the SFS is defined by

$$(3.2) \quad \begin{aligned} \mathcal{M}_{SVD}(P) &= \left\{ \alpha \in \mathbb{R}^{(N-1)} \mid \exists W \in \mathbb{R}^{(N-1) \times (N-1)} : UT \geq 0 \text{ and } T^{-1}\Sigma V' \geq 0 \right\}, \\ \mathcal{M}_{SVD}(E) &= \left\{ \alpha \in \mathbb{R}^{(N-1)} \mid \exists W \in \mathbb{R}^{(N-1) \times (N-1)} : U(T')^{-1} \geq 0 \text{ and } T'\Sigma V' \geq 0 \right\}, \end{aligned}$$

which is similar to the definition in [18, 21], where a detailed description of this construction can be seen. As the set in (3.2) is defined in terms of $\alpha \in \mathbb{R}^{(N-1)}$, we choose to assume $N = 3$ such that the SFS can be visualized in two dimensions. A big advantage of \mathcal{M}_{SVD} is that it removes the problem of reordering columns and rows as it only focuses on possible transformations of the SVD of PE and not the actual values of P and E . The connection between the definition in (3.2) and the one from (1.2) is mostly different normalizations of P and E . Constructing an invertible T with columns consisting of different $\alpha \in \mathcal{M}_{SVD}(P)$ gives

$$PE = U\Sigma V' = \underbrace{UTD_1^{-1}}_{\tilde{P}} \underbrace{D_1 T^{-1}\Sigma V'}_{\tilde{E}},$$

where $\tilde{P} \in \mathcal{M}(P)$, $\tilde{E} \in \mathcal{M}(E)$, and $D_1 = \text{diag}(e'UT) \in \mathbb{R}^{N \times N}$. The diagonal matrix D_1 scales the matrices such that \tilde{P} has columns summing to one. Given $\tilde{P} \in \mathcal{M}(P)$, the values of α can be found in the columns of

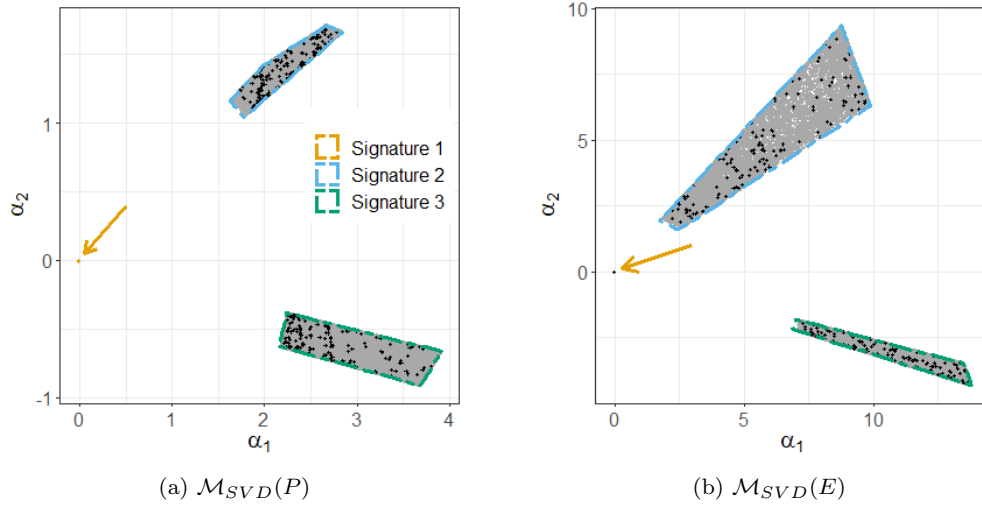
$$T = U' \tilde{P} D_2^{-1}$$

as the eigenvectors in U are orthonormal and $D_2 = \text{diag}(e_1 U' \tilde{P}) \in \mathbb{R}^{N \times N}$, where $e_1 = (1, 0, \dots, 0)$. The diagonal matrix D_2 assures that the first row of T consists of ones as in (3.1).

The set in (1.2) is easier to interpret as this visually shows the direct possible changes on P and E , but the one in (3.2) gives a more simple visualization of the SFS and is better for comparison to the polygon inflation algorithm.

3.2. Polygon inflation algorithm. The polygon inflation algorithm is an algorithm used to approximate the SFS. The algorithm is introduced in [21] and a further analysis and proofs are made in [18]. The algorithm assumes the set \mathcal{M}_{SVD} for both P and E , at most, consists of N separated whole free subsets or one connected set. In the examples in Figure 3 they all have three separated whole free subsets, such that all the points within the polygons are feasible solutions. Having, at most, N whole free subset or one connected set is also an assumption made for the sampling algorithm to find the whole SFS.

Breast Cancer



Lung A. Cancer

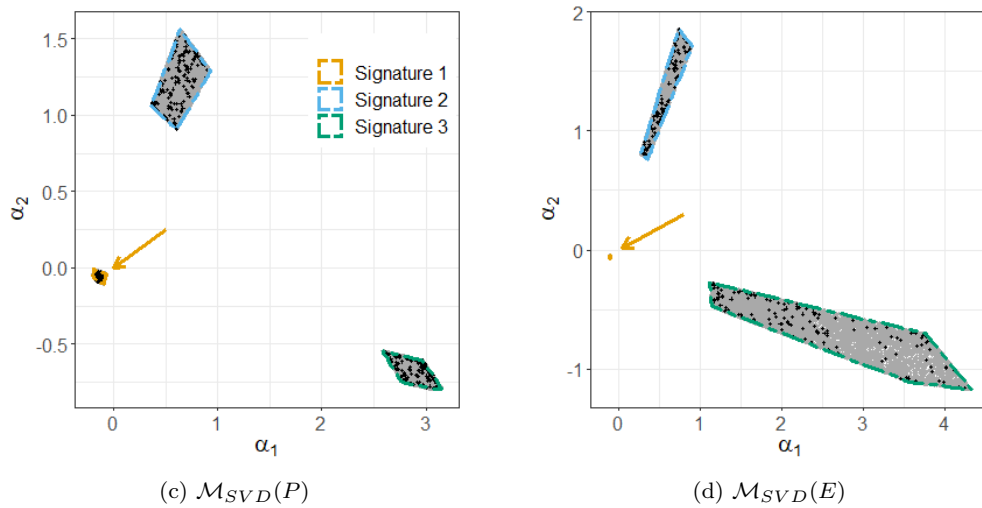


FIG. 3. The SFS for the Breast and Lung A. cancer data in the \mathcal{M}_{SVD} representation, where the gray dots show the results of all the iterations from the sampling algorithm and the black dots highlight the results from the last 500 samples. The colored polygons are the SFS found from the polygon inflation algorithm. Coloring of the three signatures is in accordance with Figure 2.

4. Further analysis of the sampling algorithm for arbitrary rank

The idea of the polygon inflation algorithm is to approximate the subsets of the SFS in terms of the representation in (3.2) by inflating an initial solution in \mathcal{M}_{SVD} to the boundaries which creates a polygon. The algorithm continues to inflate the edges of the polygon with new vertices on the boundary of the SFS until the area stops increasing.

In the package FAC-PACK (<http://www.math.uni-rostock.de/facpack/>) they have applied the algorithm for rank three and four. The results from the polygon inflation algorithm and sampling algorithm with rank three are seen in Figure 3, where the gray and black dots illustrate the samples from the sampling algorithm, and the colored polygons are the SFS found from the polygon inflation algorithm. Note the colored lines of the different areas in Figure 3 correspond to the signature with the same strip colors in Figure 2. The sampling algorithm proposed here clearly gives similar results as the polygon inflation algorithm, which has been the case for all examples we have tried. The SFS for E is also completely found, even though the stopping criteria and sampling is made in relation to the signature matrix P . The matrix used as a reference is therefore unimportant in terms of finding the SFS for both P and E .

Curious observation is that the size of the areas in Figure 3 are hard to directly transfer to the size of the changes in different entries in Figure 2, although the representation in Figure 3 gives a more simple visualization of the SFS. The advantages of our sampling algorithm compared to the polygon inflation algorithm are that it is easier to implement and can scale to an arbitrary dimension of N , as shown in Figure 5. This is especially important for cancer genomics where the number of mutational processes varies a lot depending on, e.g., the number of samples and number of cancer types. Potentially, this could also be advantageous in other fields, where NMF is applied with higher rank of N .

4. Further analysis of the sampling algorithm for arbitrary rank. Here we will discuss the choice of the parameter β and the running time of the sampling algorithm. The sampling algorithm is applied to an assumed rank between 2 and 10, which makes it possible also to comment on how the size of the SFS and computation time is influenced by an increase in the rank.

4.1. Influence of tuning parameter β and rank N . The influence of the parameter β is found by running the sampling algorithm multiple times for $\beta \in \{0.1, 0.5, 1\}$ on the Breast and Lung A. cancer data. We test for $\beta = 1$ which is equivalent to choosing λ uniformly at random and also for $\beta = 0.1$ that mainly samples at the endpoints. At last we also include $\beta = 0.5$ that is something in-between the two cases. An illustration of 1000 samples from the algorithm with $\beta \in \{0.1, 0.5, 1\}$ is seen in Figure 4. Visually we observe from Figure 4 that $\beta = 0.5$ strikes a good balance between sampling many points close to the edges and at the same time covering the full region of the SFS.

In Figure 5, the size of the SFS, $\text{avg}(\mathbf{P}^S)$, is depicted as a function of N , i.e., the number of assumed mutational processes. The size of the SFS for the Breast cancer data is close to zero for five mutational processes and above. Even below five mutational processes the average variations of the entries seem to be fairly small.

On the contrary the size of the SFS for the Lung A. cancer is increasing with the number of signatures. The increase of the SFS when N is equal to five instead of three can be seen by comparing Figure 1 and Figure 2(b), where there are more red areas with N equal to five than with three. Here, it is clear that an increase in the number of mutational processes does not necessarily decrease the size of the SFS. The

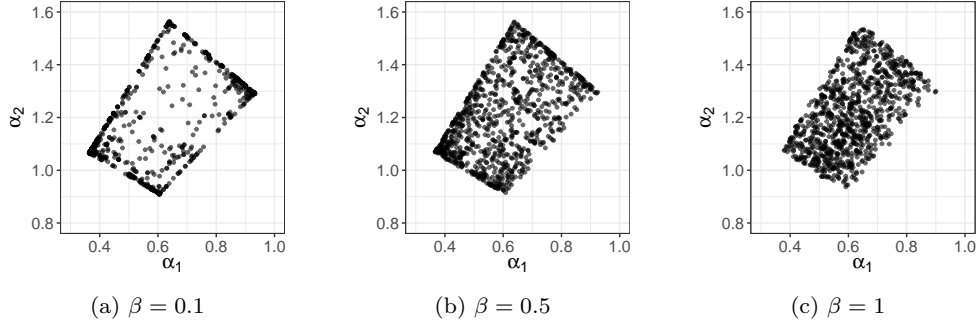


FIG. 4. Sampling 1000 points in the SFS $\mathcal{M}_{SVD}(P)$ for different choices of β . Here, the results are illustrated on the second signature of Lung A. cancer in Figure 3.

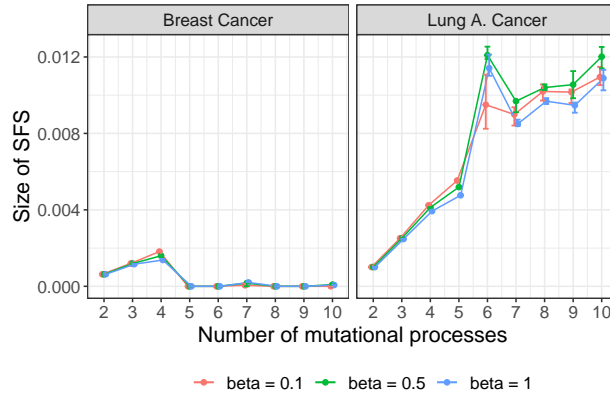


FIG. 5. The size of the SFS for different number of assumed mutational processes, i.e., N . Each point is based on the results of running the sampling algorithm 50 times for each choice of β and N . The error bars show the percentile 0.25 and 0.75 of the 50 different runs.

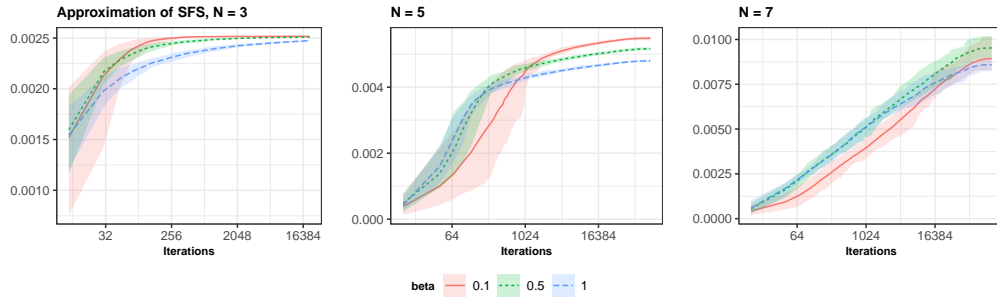


FIG. 6. Approximation for the size of the SFS on the Lung cancer data. The approximation is shown on a logarithmic scale for different numbers of assumed mutational processes, N , and different β .

values $\text{avg}\langle \mathbf{P}^S \rangle$ on the y-axis can appear small, but remember from the definition in (2.9) that we divide our sum by $K \cdot N$.

The choice of β and N seem to have an influence on both the stability and final size of $\text{avg}\langle \mathbf{P}^S \rangle$, which can be seen in Figure 6. For five mutational processes and below, the smallest $\beta = 0.1$ performs best, and for more than five mutational processes $\beta = 0.5$ is preferable. Sampling $\lambda \in \Lambda_{ij}$ uniformly, i.e., $\beta = 1$ separates the

5. Taking advantage of random initialization

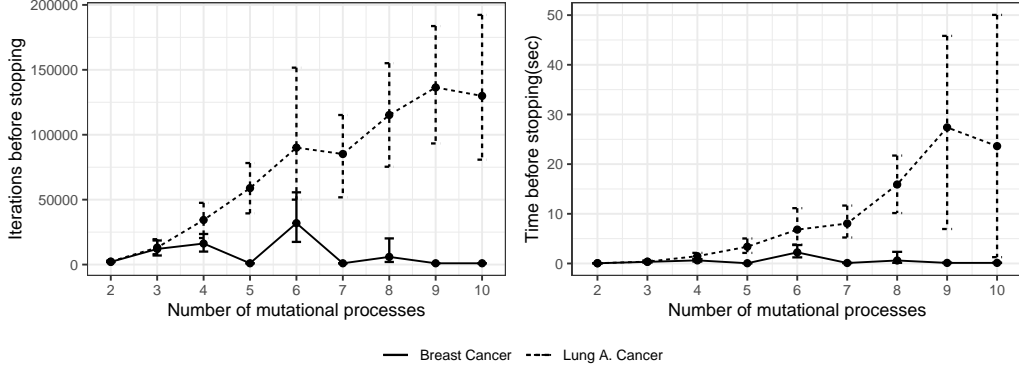


FIG. 7. *The number of iterations and time needed to compute the SFS for different number of mutational processes, where the algorithm was repeated 50 times. In the figures above, the average is reported together with the quantiles 0.05 and 0.95.*

sample equally across the area, which gives a very low probability of sampling in the endpoints and thereby gives a smaller size of the SFS, which is clearly seen in both Figure 5 and 6. On the contrary, sampling $\lambda \in \Lambda_{ij}$ with a lot of weight on the endpoints, i.e., $\beta = 0.1$ excludes sampling of points in the center. This works well for a small number of mutational processes, but when the number of mutational processes increases and they can be mixed more across each other it can be hard to reach certain areas of the SFS that requires sampling from the center at some point. This means $\beta = 0.1$ gives a large volatility of the size of the SFS for a larger number of mutational processes. One could also argue to sample β from a predefined distribution when N exceeds five.

4.2. Running time. Our sampling algorithm is not only simple to implement, but is also very fast even for higher ranks of the factorization. In Figure 7, the average number of iterations and computation time (using a regular laptop with Intel Core i7-8565U CPU @ 1.80 GHz and 16 GB RAM) is shown on the two cancer data sets for different sizes of N , where the algorithm was repeated 50 times for each choice of N . In general the computation time is highly influenced by the size of the SFS and the rank N .

5. Taking advantage of random initialization. The identification of P and E is not only influenced by the size of SFS but will also be influenced by noise in data and the initialization of the algorithm by Lee and Seung [15]. Here, we will cover the influence of these aspects and compare it to the SFS. We investigate how random initialization can be used to determine the global minimum and to explore the size of the SFS.

In cancer genomics the most natural assumption is that each mutational count is drawn from a Poisson distribution

$$(5.1) \quad M_{kg} \sim \text{Pois}((PE)_{kg}).$$

Under this assumption, a natural way to estimate P and E is to maximize the log-likelihood function

$$(5.2) \quad \ell(P, E; M) = \sum_{k=1}^K \sum_{g=1}^G M_{kg} \log((PE)_{kg}) - (PE)_{kg} - \log(M_{kg}!)$$

$$(5.3) \quad = -D(M|PE) + C,$$

where

$$(5.4) \quad D(M|PE) = \sum_{k=1}^K \sum_{g=1}^G M_{kg} \log \left(\frac{M_{kg}}{(PE)_{kg}} \right) - M_{kg} + (PE)_{kg}$$

is the generalized Kullback–Leibler divergence and C is a constant only dependent on M . Hence, maximizing the likelihood function is equivalent to minimizing the generalized Kullback–Leibler divergence $D(M|PE)$. This is the divergence appearing in Lee and Seung [15], which is decreased through each update. The objective function is convex in P or E but not in both variables together. As mentioned in Lee and Seung [15], the update rules can therefore not ensure a global minimum, but only a local minimum.

5.1. Finding the global minimum. In [24] they prove that it is NP-hard to solve NMF to optimality. The Lee–Seung updates are an example of the majorize-minimize (MM) algorithm, and [12] provides a simple proof of the update rules. We suggest the standard procedure of running the MM algorithm multiple times with different initializations in order to find the global minimum.

In this paper we have performed at least five random initializations when finding an NMF solution. The influence of an increase in the number of initializations can be seen in Figure 8(a), which shows how just a few initializations will stabilize the reached minimum to a large extent. As the minimum stops changing we assume to have found the global minimum. Even though the algorithm reaches the same minimum we will still observe variability in the estimates of P and E , when they have a variability in the SFS. Different random initializations with the same minimum of $D(M|PE)$ will still give different results in the SFS, which is illustrated in Figure 8(b). In Figure 8(b) it is clear how the initial solutions are more keen to reaching certain parts of the SFS apposed to others.

5.2. Finding more than N subsets of the SFS. Although we have seen no practical applications with more than N subsets, it is possible. In this rare case our sampling algorithm would fail to find the whole SFS. Here we have proposed a useful supplementary method to help find the additional subsets that are seen in the constructed example from [13]. This method is to use random initializations, as described above. In Example 3 of [13] they show that determining the matrices P and E in the following way:

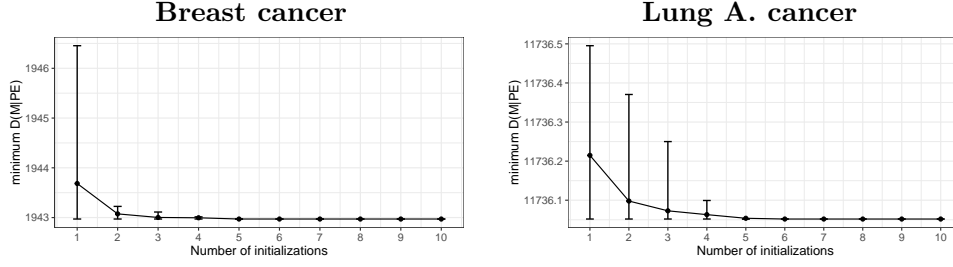
$$(5.5) \quad E = \begin{pmatrix} 0.7 & 1 & 1 & 0.7 & 0 & 0 \\ 1 & 0.7 & 0 & 0 & 0.7 & 1 \\ 0 & 0 & 0.7 & 1 & 1 & 0.7 \end{pmatrix},$$

$$(5.6) \quad P = E^T,$$

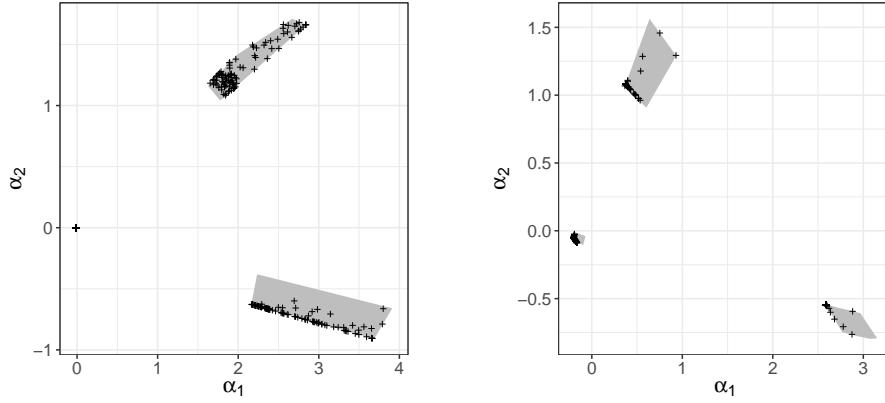
$$(5.7) \quad M = PE,$$

where $N = 3$ gives six subsets for the SFS of M . Here, our algorithm would fail to recover the whole SFS and only find three of the six subsets. This is because the algorithm starts from a single initial solution and can only recover the subsets of the SFS that includes this initial solution. However, different random initializations with the same minimum may end up in different parts of the SFS and thereby also different subsets. Running the sampling algorithm from these different subsets would help recover all parts of the SFS in this special case. In Figure 9 the gray area shows the SFS, and the black crosses show the reached solutions from 20 random

5. Taking advantage of random initialization



(a) Effect of more initializations



(b) Final result from random initializations in $\mathcal{M}_{SVD}(P)$

FIG. 8. Results from different initializations. In (a), the average of the minimum $D(M|PE)$ for a certain number of initializations is depicted together with error bars that show the quantiles 0.05 and 0.95. The average and quantiles are based on 100 runs. The reached minimum of $D(M|PE)$ is 1943 and 11736 for Breast and Lung cancer, respectively. In (b), the results from the 100 different runs with ten initializations are shown in $\mathcal{M}_{SVD}(P)$ relative to the solution found previously in Figure 3(a).

initializations. The solutions from random initialization reach all six subsets just after these 20 random initialization, but we cannot be sure of the reliability of this method. Incorporating more initial solutions does not guarantee that we have found the whole SFS but could possibly help to recover more of the SFS.

5.3. Noise in the data. Noise in the data obviously has an influence on the identification of P and E and their true rank. Sometimes the rank is determined based on the variance in P and E , but this can be a very unreliable method if the factorization has a large SFS. A large SFS can make variance of P and E appear much larger than it actually is. We will illustrate this on the Breast cancer data assuming three mutational processes. Assume we have found a global minimum factorization $\hat{P}\hat{E} \in \mathbb{R}_+^{96 \times 21}$ that approximates the Breast cancer data $M \in \mathbb{N}_0^{96 \times 21}$. The influence of variance is now illustrated through parametric bootstrapping using the assumption in (5.1). Specifically, we create 100 bootstrap samples

$$M^{*b} \sim \text{Pois}(\hat{P}\hat{E})$$

for $b = 1, \dots, 100$. The NMF algorithm from Lee and Seung [15] is then run on $(M^{*1}, \dots, M^{*100})$ for both 10 different random initializations that will reach different parts of the SFS and the same 10 initializations that should reach the same part of the

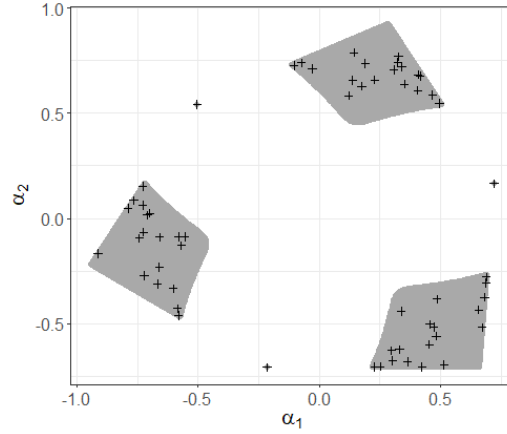


FIG. 9. Example 3 from [13], where there are six subsets. The black crosses show the results from 20 random initializations of the NMF algorithm, and the gray area shows the SFS found from the sampling algorithm by starting in the 20 different initializations. Notice, all the initializations obtain $D(V|WH) = 0$.

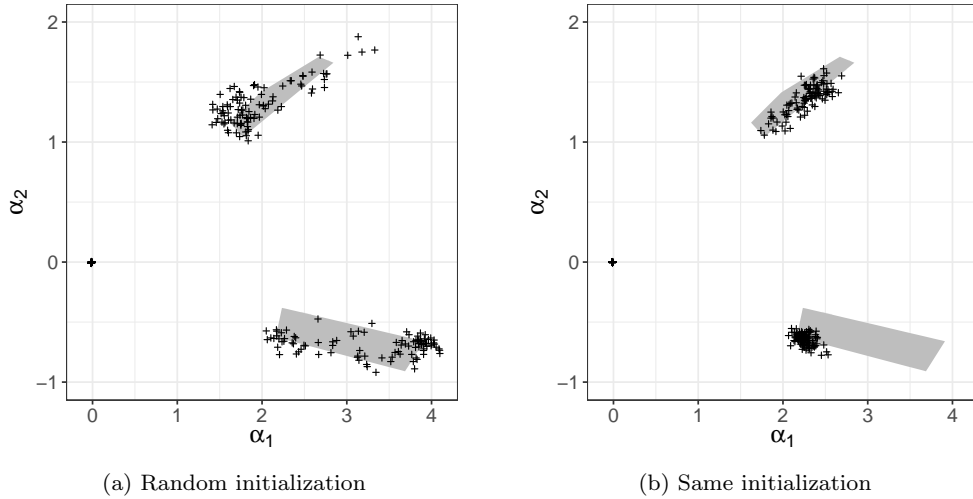


FIG. 10. Visualization of the identification problem due to variance, where (a) shows the results from random initialization and (b) shows the results using the same ten initialization for each sample. Both show the global minimum from 100 bootstrap samples.

SFS, which is illustrated in Figure 10. We clearly see how the appearing variability of the results is influenced by the SFS through random initialization.

A large variance could therefore appear as a low reproducibility of the signatures but is actually just a result of the SFS. Remember, having a large SFS does not make the assumed number of mutational processes less correct but is instead a result of the structure of the signatures and exposures.

6. Conclusion. We have introduced a novel sampling algorithm that can find the SFS for an arbitrary rank of NMF. The algorithm is based on the solution for two dimensions together with random sampling to explore the possible changes of the initial solution. The algorithm gives good approximation of the SFS for rank three

6. References

compared to current algorithms, most notably the polygon inflation algorithm. For higher ranks we do not know how well the sampling algorithm actually approximates the SFS, as we have no reference for ranks above four.

The problem of nonuniqueness for NMF in relation to mutational counts have also been highlighted here. The size of the SFS is shown to depend strongly on the specific data set at hand and the assumed number of mutational processes. Furthermore, the importance of several initializations is emphasized together with the important fact of how a large variance in a signature could potentially be caused by the SFS.

An important aspect of NMF is to determine the true rank of the factorization, and despite a detailed geometric understanding [8] it remains a challenging problem. In cancer genomics there have both been introduced methods based on the Bayesian information criteria (BIC) [5, 20] and another method that both includes the variability of the signatures and how well P and E approximates M [2]. The latter procedure could be problematic, as there could appear large variations due to the size of the SFS. One method to remove the problem of nonuniqueness of the solution is to decrease the variability of the signatures through fewer parameters. This was implemented in a simple form in [23].

Our sampling algorithm is very simple to implement and fast, which makes it easy to check whether an NMF solution is unique or not. This article has enlightened the problem of nonuniqueness of the NMF and at the same time showed a simple way to check this uncovered problem. It is important to remember that the size of the SFS is not a tool to determine the correct choice of N but instead a way to find all the solutions that give the same global minimum approximation of the data. Variability in the SFS is not the result of a poor factorization but instead a lack of uniqueness of the matrix factorization.

Our sampling algorithm can approximate the SFS, and for the NMF rank 3 problem we have empirical evidence that we find a similar SFS as for the polygon inflation algorithm. It could be important to show rigorous mathematical results about the ability of our sampling algorithm to recover the full SFS, but that is a topic for future research.

Acknowledgments. We thank the three anonymous reviewers, Dan Ariel Søndergaard, and Kenneth Borup for their constructive suggestions and comments on earlier versions of this manuscript.

REFERENCES

- [1] L. B. ALEXANDROV, S. NIK-ZAINAL, D. C. WEDGE, S. A. J. APARICIO, S. BEHJATI, A. V. BIANKIN, G. R. BIGNELL, N. BOLLI, A. BORG, A.-L. BØRRESEN-DALE, S. BOYALT, B. BURKHARDT, A. P. BUTLER, C. CALDAS, H. R. DAVIES, C. DESMEDT, R. EILS, J. E. EYFJÖRD, J. A. FOEKENS, M. GREAVES, F. HOSODA, B. HUTTER, T. ILICIC, S. IMBEAUD, M. IMIELINSKI, N. JÄGER, D. T. W. JONES D. JONES, S. KNAPPSKOG, M. KOOL, S. R. LAKHANI, C. LÓPEZ-OTÍN, S. MARTIN, N. C. MUNSHI, H. NAKAMURA, P. A. NORTHCOTT, M. PAJIC, E. PAPAEMMANUIL, A. PARADISO, J. V. PEARSON, X. S. PUENTE, K. RAINE, M. RAMAKRISHNA, A. L. RICHARDSON, J. RICHTER, P. ROSENSTIEL, M. SCHLESNER, T. N. SCHUMACHER, P. N. SPAN, J. W. TEAGUE, Y. TOTOKI, A. N. J. TUTT, R. VALDÉS-MAS, M. M. VAN BUUREN, L. VAN'T VEER, A. VINCENT-SALOMON, N. WADDELL, L. R. YATES, AUSTRALIAN PANCREATIC CANCER GENOME INITIATIVE, ICGC BREAST CANCER CONSORTIUM, ICGC MMML-SEQ CONSORTIUM, ICGC PEDBRAIN. J. ZUCMAN-ROSSI, P. A. FUTREAL, U. McDERMOTT, P. LICHTER, M. MAYERSON, S. M. GRIMMOND, R. SIEBERT, E. CAMPO, T. SHIBATA, S. M. PFISTER, P. J. CAMPBELL, AND M. R. STRATTON, *Signatures of mutational processes in human cancer*, Nature, 500 (2013), pp. 415–421.

- [2] L. B. ALEXANDROV, S. NIK-ZAINAL, D. C. WEDGE, P. J. CAMPBELL, AND M. R. STRATTON, *Deciphering signatures of mutational processes operative in human cancer*, Cell Rep., 3 (2013), pp. 246–259.
- [3] O. S. BORGES AND B. R. KOWALSKI, *An extension of the multivariate component-resolution method to three components*, Anal. Chim. Acta, 174 (1985), pp. 1–26.
- [4] D. DONOHO AND V. STODDEN, *When does non-negative matrix factorization give a correct decomposition into parts?*, Adv. Neural Inf. Process. Syst., 16 (2003), pp. 1141–1148.
- [5] A. FISCHER, C. J. ILLINGWORTH, P. J. CAMPBELL, AND V. MUSTONEN, *EMU: Probabilistic inference of mutational processes and their localization in the cancer genome*, Genome Biol., 14 (2013), pp. 1–10.
- [6] A. E. GELFAND AND A. F. SMITH, *Sampling-based approaches to calculating marginal densities*, J. Am. Stat. Assoc., 85 (1990), pp. 398–409.
- [7] P. J. GEMPERLINE, *Computation of the range of feasible solutions in self-modeling curve resolution algorithms*, Anal. Chem., 71 (1999), pp. 5398–5404.
- [8] N. GILLIS AND F. GLINEUR, *On the geometric interpretation of the nonnegative rank*, Linear Algebra Appl., 437 (2012), pp. 2685–2712.
- [9] A. GOLSHAN, H. ABDOLLAHI, S. BEYRAMYSOLTAN, M. MAEDER, K. NEYMEYR, R. RAJKÓ, M. SAWALL, AND R. TAULER, *A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data*, Anal. Chim. Acta, 911 (2016), pp. 1–13.
- [10] R. C. HENRY AND B. M. KIM, *Extension of self-modeling curve resolution to mixtures of more than three components: Part 1. Finding the basic feasible region*, Chemometr. Intell. Lab. Syst., 8 (1990), pp. 205–216.
- [11] K. HUANG, N. D. SIDIROPOULOS, AND A. SWAMI, *Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition*, IEEE Trans. Signal Process., 62 (2013), pp. 211–224.
- [12] K. LANGE, E. C. CHI, AND H. ZHOU, *A brief survey of modern optimization for statisticians*, Int. Stat. Rev., 82 (2014), pp. 46–70.
- [13] H. LAURBERG, M. G. CHRISTENSEN, M. D. PLUMBLEY, L. K. HANSEN, AND S. H. JENSEN, *Theorems on positive data: On the uniqueness of NMF*, Comput. Intell. Neurosci., (2008), 764206.
- [14] W. H. LAWTON AND E. A. SYLVESTRE, *Self modeling curve resolution*, Technometrics, 13 (1971), pp. 617–633.
- [15] D. D. LEE AND H. S. SEUNG, *Algorithms for non-negative matrix factorization*, in Advances in Neural Information Processing Systems 13-Proceedings of the 2000 Conference, NIPS 2000, Neural Information Processing Systems Foundation, LA Jolla, CA, 2001, pp. 556–562.
- [16] C. MASON, M. MAEDER, AND A. WHITSON, *Resolving factor analysis*, Anal. Chem., 73 (2001), pp. 1587–1594.
- [17] S. MOUSSAOUI, D. BRIE, AND J. IDIER, *Non-negative source separation: Range of admissible solutions and conditions for the uniqueness of the solution*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’05) 2005 vol. 5, IEEE, 2005 pp. v–289.
- [18] K. NEYMEYR AND M. SAWALL, *On the set of solutions of the nonnegative matrix factorization problem*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1049–1069.
- [19] S. NIK-ZAINAL AND S. MORGANELLA, *Mutational Signatures in Breast Cancer: The Problem at the DNA Level*, Clin. Cancer Res., 23 (2017), pp. 2617–2629.
- [20] R. A. ROSALES, R. D. DRUMMOND, R. VALIERIS, E. DIAS-NETO, AND I. T. DA SILVA, *signeR: An empirical Bayesian approach to mutational signature discovery*, Bioinform., 33 (2017), pp. 8–16.
- [21] M. SAWALL, C. KUBIS, D. SELENT, A. BOERNER, AND K. NEYMEYR, *A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: Concepts and applications*, J. Chemometr., 27 (2013), pp. 106–116.
- [22] M. SAWALL AND K. NEYMEYR, *A ray casting method for the computation of the area of feasible solutions for multicomponent systems: Theory, applications and facpack-implementation*, Anal. Chim. Acta, 960 (2017), pp. 40–52.
- [23] Y. SHIRAIISHI, G. TREMMEL, S. MIYANO, AND M. STEPHENS, *A simple model-based approach to inferring and visualizing cancer mutation signatures*, PLoS Genet., 11 (2015), e1005657.
- [24] S. A. VAVASIS, *On the complexity of nonnegative matrix factorization*, SIAM J. Optim., 20 (2010), pp. 1364–1377.

Paper

B

**Model selection and robust inference of mutational
signatures using Negative Binomial non-negative matrix
factorization**

by Marta Pelizzola, Ragnhild Laursen and Asger Hobolth

Published in BMC bioinformatics

RESEARCH

Open Access



Model selection and robust inference of mutational signatures using Negative Binomial non-negative matrix factorization

Marta Pelizzola^{1*}, Ragnhild Laursen¹ and Asger Hobolth¹

*Correspondence:
marta@math.au.dk

¹ Department of Mathematics,
Aarhus University, Aarhus,
Denmark

Abstract

Background: The spectrum of mutations in a collection of cancer genomes can be described by a mixture of a few mutational signatures. The mutational signatures can be found using non-negative matrix factorization (NMF). To extract the mutational signatures we have to assume a distribution for the observed mutational counts and a number of mutational signatures. In most applications, the mutational counts are assumed to be Poisson distributed, and the rank is chosen by comparing the fit of several models with the same underlying distribution and different values for the rank using classical model selection procedures. However, the counts are often overdispersed, and thus the Negative Binomial distribution is more appropriate.

Results: We propose a Negative Binomial NMF with a patient specific dispersion parameter to capture the variation across patients and derive the corresponding update rules for parameter estimation. We also introduce a novel model selection procedure inspired by cross-validation to determine the number of signatures. Using simulations, we study the influence of the distributional assumption on our method together with other classical model selection procedures. We also present a simulation study with a method comparison where we show that state-of-the-art methods are highly overestimating the number of signatures when overdispersion is present. We apply our proposed analysis on a wide range of simulated data and on two real data sets from breast and prostate cancer patients. On the real data we describe a residual analysis to investigate and validate the model choice.

Conclusions: With our results on simulated and real data we show that our model selection procedure is more robust at determining the correct number of signatures under model misspecification. We also show that our model selection procedure is more accurate than the available methods in the literature for finding the true number of signatures. Lastly, the residual analysis clearly emphasizes the overdispersion in the mutational count data. The code for our model selection procedure and Negative Binomial NMF is available in the R package SigMoS and can be found at <https://github.com/MartaPelizzola/SigMoS>.

Keywords: Cancer genomics, Cross-validation, Model checking, Model selection, Mutational signatures, Negative Binomial, Non-negative matrix factorization, Poisson

AMS Classification: 92-08, 92-10, 62-08



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

1. Introduction

Introduction

Somatic mutations occur relatively often in the human genome and are mostly neutral. However, the accumulation of harmful mutations in a genome can lead to cancer. The summary of somatic mutations observed in a tumor is called a mutational profile and can often be associated with factors such as aging [1], UV light [2] or tobacco smoking [3]. A mutational profile is thus a mixture of mutational processes that are represented by mutational signatures. Several signatures have been identified from the mutational profiles and associated with different cancer types [4, 5]. The importance of mutational signatures thus lies in their association with the mutational processes causing cancer. Having more insights into the causes of cancer is a prerequisite for better understanding the role that genetics plays in the development of the disease and eventually also for discovering potential treatment.

A common strategy to derive the mutational signatures is non-negative matrix factorization [6–8]. Different approaches to estimate the signature and the exposure matrices from mutational count data have been extensively described in [9, 10].

Non-negative matrix factorization (NMF) is a factorization of a given matrix $V \in \mathbb{N}_0^{N \times M}$ into the product of two non-negative matrices $W \in \mathbb{R}_+^{N \times K}$ and $H \in \mathbb{R}_+^{K \times M}$ such that

$$V \approx WH.$$

The rank K of the lower-dimensional matrices W and H is much smaller than N and M .

In cancer genomics, the mutational matrix V contains the mutational counts for different patients, also referred to as mutational profiles. The number of rows N is the number of patients and the number of columns M is the number of different mutation types. In this paper we use the single-base-substitution-96 mutational context [11] where $M = 96$ (corresponding to the 6 base mutations when assuming strand symmetry times the 4 flanking nucleotides on each side, i.e. $4 \cdot 6 \cdot 4 = 96$). The matrix H consists of K mutational signatures defined by probability vectors over the different mutation types. In the matrix W , each row contains the weights of the signatures for the corresponding patient. In this context, the weights are usually referred to as the exposures of the different signatures.

To estimate W and H we need to choose a model and a rank K for the data V . These two decisions are highly related as the optimal rank of the data V is often chosen by comparing the fit under a certain model for many different values of K . The optimal K is then found using a model selection procedure such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or similar approaches described in “[Estimating the number of signatures](#)” section. Most methods used in the literature [6, 12, 13] for choosing the rank are based on the likelihood value, which depends on the assumed model. For mutational counts the usual model assumption is the Poisson distribution [6]

$$V_{nm} \sim \text{Po}((WH)_{nm}), \quad (1)$$

where W and H are estimated using the algorithm from [14] that minimizes the generalized Kullback–Leibler divergence. The algorithm is equivalent to maximum likelihood estimation, as the negative log-likelihood function for the Poisson model is equal to the generalized Kullback–Leibler up to an additive constant. We observe that this model assumption is often inadequate. In particular, we observe overdispersion in the

mutational counts, i.e. a situation where the variance in the data is greater than what is expected under the assumed model. This is a well known issue when modeling count data in biology [15].

We therefore suggest using a model where the mutational counts follow a Negative Binomial distribution that has an additional parameter to explain the overdispersion in the data. In recent years, this model is becoming more popular to model the dispersion in mutational counts [16, 17]. The Negative Binomial NMF is discussed in [18], where it is applied to recommender systems, and it has recently been used in the context of cancer mutations in [19–21]. In Lyu et al. [20] a supervised Negative Binomial NMF model is applied to mutational counts from different cancers which uses cancer types as metadata. Their aim is to obtain signatures with a clear etiology, which could be used to classify different cancer types. Vöhringer et al. [21] extends the analysis by including several genomic features and uses tensors instead of the mutational count matrix to account for the different features. Lastly, [19] applies Bayesian inference to extract mutational signatures and provide different probabilistic models for the signatures. Among the models implemented in this method also the Negative Binomial model is considered as a natural extension of the Poisson model.

For mutational count data, we extend the Negative Binomial NMF model by including patient specific dispersion which has not been included in the aforementioned works using the Negative Binomial model. The extended model is referred to as NB_N -NMF, where N is the number of dispersion parameters (equivalent to the number of patients). We investigate when and why NB_N -NMF is more suitable for mutational counts than the usual Poisson NMF (Po-NMF). In particular we evaluate the goodness of fit for mutational counts using a residual-based approach. Despite the recent efforts, we still believe, as it has also been mentioned in [22], that a great amount of research has been focusing on improving the performance of NMF algorithms given an underlying model and less attention has been directed to the choice of the underlying model given the data and application.

Since the number of signatures depends on the chosen distributional assumption, we suggest using NB_N -NMF and we also propose a novel model selection framework to choose the number of signatures. We show that our model selection procedure is more robust toward inappropriate model assumptions compared to classical methods (AIC and BIC) and other methods currently used in the literature such as SigProfilerExtractor [23], SparseSignatures [8], Signer [13], sigfit [19], and SignatureAnalyzer [24]. We use both simulated and real data to validate our proposed model selection procedure against other methods. We chose one classical data set and analyze it in “Breast cancer data” section and a larger data set from prostate cancer (Fig. 5). The latter is a subset of the available data from the Pan-Cancer Analysis of Whole Genomes (PCAWG) database [25], thus it corresponds to one of the largest available data sets for a single cancer type.

In comparison to the results published in [20] and in [21], our work is not exploiting the information coming from different cancer types or from different genomic features. However, we provide a patient specific dispersion component to account for the high variance between patients and derive the update steps for parameter estimation in the NB_N -NMF. Furthermore, we propose a model selection procedure which proves to be robust to model misspecification.

2. Results

We have implemented our methods in the R package SigMoS (Signatures Model Selection) that includes NB_N -NMF and the model selection procedure. The R package is available at <https://github.com/MartaPelizzola/SigMoS>. The package also contains the simulated and real data used in this paper.

Results

In this section we describe the results of our approach on both simulated and real data. Details on the method are provided in “Methods” section. In short, we propose a Negative Binomial model applied to mutational count data with a patient specific dispersion coefficient. The matrices W and H are estimated with a majorization–minimization (MM) procedure, and we propose to use Negative Binomial maximum likelihood estimation (MLE) for estimating the dispersion parameters. Additionally, we introduce a new algorithm based on cross-validation to estimate the number of signatures for a given data set.

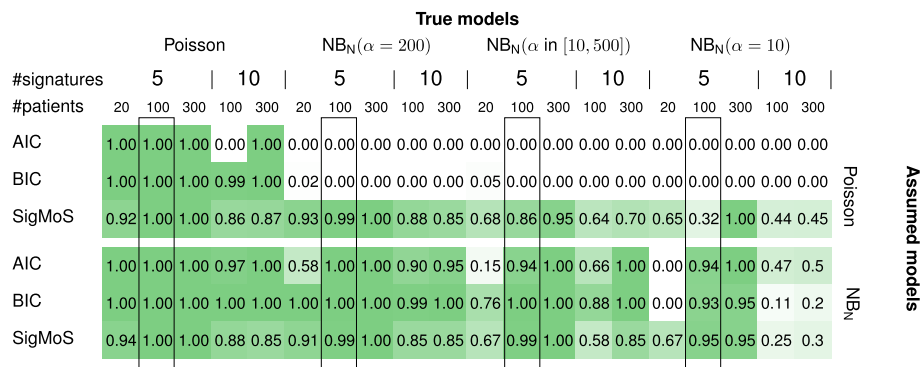
For simulated data we present a study on Negative Binomial simulated data with different levels of dispersion where results from AIC, BIC, SigProfilerExtractor [23], SparseSignatures [8], Signer [13], sigfit [19] and SignatureAnalyzer [24] are compared with our proposed model selection procedure. These results are discussed in “Simulation study” section, where we show that our method performs well and is robust to model misspecification. Our method is applied to the 21 breast cancer patients from [6] in “Breast cancer data” section, and to 286 prostate cancer patients from [25] in “Prostate cancer data” section. The goodness of fit of the different models are evaluated using a residual analysis that shows a clear overdispersion with the Poisson model. The use of residual plots to evaluate the goodness of fit is a common strategy in statistics; some examples can be found in [26, 27].

Simulation study

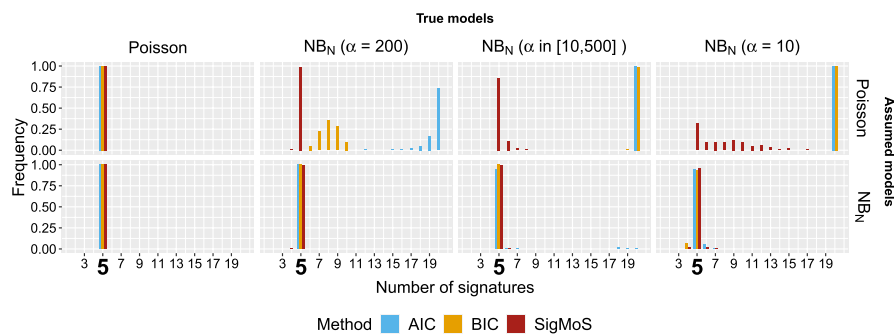
We simulated our data following the procedure of [8] using the signatures from [5]. We simulated 100 data sets for each scenario and varied the number of patients, the number of signatures and the model for the noise in the mutational count data. We considered 20, 100 and 300 patients and either 5 or 10 signatures following [28] which states that the number of common signatures in each organ is usually between 5 and 10. For each simulation run we use signature 1 and 5 from [5], as they have been shown to be shared across all cancer types, and then we sample at random three or eight additional signatures from this set. The exposures are simulated from a Negative Binomial model with mean 6000 and dispersion parameter 1.5 as in [8]. This choice is based on estimates from the real data in [29]. The mutational count data is then generated as the product of the exposure and signature matrix. Lastly, Poisson noise, Negative Binomial noise with dispersion parameter $\alpha \in \{10, 200\}$ or uniformly sampled in $[10, 500]$ are added to the mutational counts. The values of the patient specific dispersion are inspired from the data set in “Breast cancer data” section. A lower α is associated with higher dispersion, however the actual level of dispersion associated to a given α value depends on the absolute mutational counts as can be seen from the variance in Eq. (5). Therefore it is not possible to directly compare these values with the ones estimated for the real data.

Simulation results

The effect of the model assumption on the estimated number of signatures using AIC, BIC (see Eqs. (14) and (15)) and SigMoS as model selection procedures is shown in Fig. 1. Figure 1a summarizes results for all simulation studies and for each study. This figure displays the proportion of scenarios where the true number of signatures is correctly estimated from the different methods: the darker the green color the higher is this proportion. This shows that our proposed approach is estimating the number of signatures accurately and is much more robust to model misspecifications compared to AIC and BIC. For example, when the true model has a small dispersion of $\alpha = 200$ and the Poisson model is assumed, the difference between the performance of SigMoS and of AIC and BIC is already substantial. Here, AIC and BIC are never estimating the true number of signatures correctly, whereas our SigMoS procedure estimates the correct number of signatures in most cases ($\geq 85\%$). The table also shows that the higher the dispersion in the model, the harder it is to estimate the true number of signatures even when the correct model is specified.



(a) Proportion of simulation runs correctly estimating the true number of signatures.



(b) Estimated number of signatures.

Fig. 1 Results from AIC, BIC, and SigMoS based on Po-NMF and NB_N-NMF using simulated data. Each method is applied on different simulated data sets for four different types of noise: Poisson and Negative Binomial with dispersion parameter $\alpha = 10, 200$ and $\alpha \sim U(10, 500)$. **a** The proportion of simulation runs where the number of signatures is correctly estimated. The true number of signatures varies in $\{5, 10\}$ and the number of patients in $\{20, 100, 300\}$. The rectangular boxes highlight the results shown in **b**. The results are based on 100 simulation runs for scenarios with 20 and 100 patients and on 20 simulation runs for scenarios with 300 patients. **b** The estimated number of signatures in the range from 2 to 20 for 100 patients, where the true number of signatures is five

2. Results

Figure 1b depicts the actual estimated number of signatures in the range from 2 to 20 for the 100 data sets with 5 signatures and 100 patients. This clearly shows that the higher the overdispersion in the model, the more is the number of signatures overestimated. Assuming Poisson in the case of $\alpha = 200$ we see that AIC is already overestimating the number of signatures. Here, these additional signatures are needed to explain the noise that is not accounted for by the Poisson model. Having an even higher overdispersion makes both AIC and BIC highly overestimate the number of signatures to a value that is plausibly much higher than 20. Even high overdispersion does not influence our SigMoS procedure in the same way and our approach is still estimating the true number of signatures for a large proportion of the scenarios. Assuming the Negative Binomial model all of the three methods have a really high performance, as the Negative Binomial accounts for both low and high dispersion.

In the simulation study from Fig. 1b we also consider the accuracy of the MLE for the α value in the two scenarios where each patient has the same α . Our approach estimates the true α with high accuracy when the dispersion is high i.e. $\hat{\alpha} \in [9.21, 11.78]$ for $\alpha = 10$, α is slightly overestimated when the dispersion is low: for $\alpha = 200$ we find $\hat{\alpha} \in [225.8, 292.7]$. However, according to Fig. 1b this small bias does not affect the performance of our model selection procedure.

Method comparison

Several methods have been proposed in the literature for estimating the number of signatures in cancer data. In the following we present the results of a comparison between our method and four commonly used methods in the literature: `SigProfilerExtractor` [23], `SparseSignatures` [8], `SignatureAnalyzer` [24], `sigfit` [19], and `Signer` [13]. `SigProfilerExtractor` [23] extracts mutational signatures by applying NMF to 100 normalized Poisson resampled input matrices for different values for the number of signatures. The number of mutational signatures is then estimated by evaluating the stability of mutational signatures and choosing the solution with the lowest number of signatures among the stable solutions that describe the data well. `SparseSignatures` [8] provides an alternative cross-validation approach where the test set is defined by setting 1% of the entries in the count matrix to 0. Then NMF is iteratively applied to the modified count matrix and the entries are updated at each iteration. The resulting signature and exposure matrices are used to predict the entries of the matrix corresponding to the test set. `SignatureAnalyzer` [24], on the other hand, proposes a procedure where a Bayesian model is used and maximum a posteriori estimates are found with a majorize-minimization algorithm. `sigfit` [19] presents an R package providing different options for extracting and refitting signatures and exposures by Bayesian inference under different models. They propose a framework where a Multinomial, Normal, Poisson or Negative Binomial model (with mutation type specific dispersion parameter) can be used. The number of signatures is estimated using the elbow method by looking at changes in the accuracy of re-estimating the data with the extracted signatures and exposures. In our comparison we use the Poisson and Negative Binomial models within the `sigfit` package and refer to them as `sigfit-Po` and `sigfit-NB`. Lastly, with `Signer` [13] an empirical Bayesian approach based on BIC is used to estimate the number of mutational signatures.

For our method comparison, we run all methods on the simulated data from Fig. 1b. For each method and simulation setup we only allow the number of signatures to vary from two to eight due to the long running time of some of these methods.

Figure 2 shows that, when Poisson data are simulated almost all methods have a very good performance and can recover the true number of signatures in most of the simulations. The poor performance of `SparseSignatures` could be affected by not having a fixed background signature. Indeed, the improved performance of `SparseSignatures` when a background signature is included has also been shown in [8]. `sigfit-Po` is based on a more heuristic method and tends to underestimate the true number of signatures. When Negative Binomial noise is added to the simulated data with a moderate dispersion ($\alpha = 200$), `sigfit-Po`, `SignatureAnalyzer` and `Signer` have low power emphasizing the importance of correctly specifying the distribution for these methods, whereas our proposed approach (regardless of the distributional assumption), `sigfit-NB`, `SigProfilerExtractor` and `SparseSignatures` maintain good power. For patient specific dispersion also the power of `SparseSignatures` and `SigProfilerExtractor` decreases. Lastly, the power of `sigfit-NB` decreases for high dispersion ($\alpha = 10$): here the distributional assumptions are correctly specified, however this is a heuristic approach to estimate the number of signatures which tends to be less precise than `SigMoS`. Indeed, good performance is achieved with our proposed approach even under high dispersion if the correct distribution is assumed. These results demonstrate that `SigMoS` is accurate for detecting the correct number of signatures and it performs well also in situations with overdispersion compared to other methods.

For this set of simulations we also checked the quality of the estimated signatures. We sampled 10 runs for each scenario from Fig. 2 and calculated the cosine similarity between the estimated signatures and the true ones used for simulations. The results for all methods are shown in Fig. 3 where we display the average cosine similarity over 10 runs for each method and each scenario. For this study we fixed the number of signatures to five for all methods, which may favour methods such as `SignatureAnalyzer`, `sigfit` or `Signer` that usually overestimate the number of signatures. Nonetheless, these results also show that `SigMoS` combined with the Negative Binomial model and `sigfit-NB` are the methods that are able to retain the highest accuracy also with high levels of overdispersion (namely $\alpha = 10$). `SigProfilerExtractor` and `Signer` also show good accuracy especially when the overdispersion is low and under the Poisson model. These results, combined with

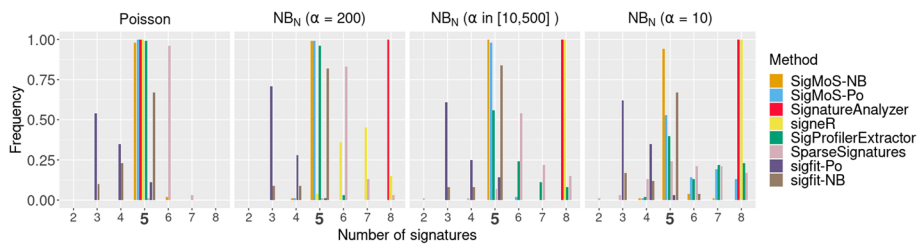


Fig. 2 Method comparison using simulated data. Each method is applied on the data sets from Fig. 1b and, for each data set, the value of the estimated number of signatures is kept. We test values for the number of signatures from two to eight for Poisson noise and Negative Binomial noise with $\alpha = \{10, 200\}$, and a patient specific dispersion parameter $\alpha \sim U(10, 500)$

2. Results

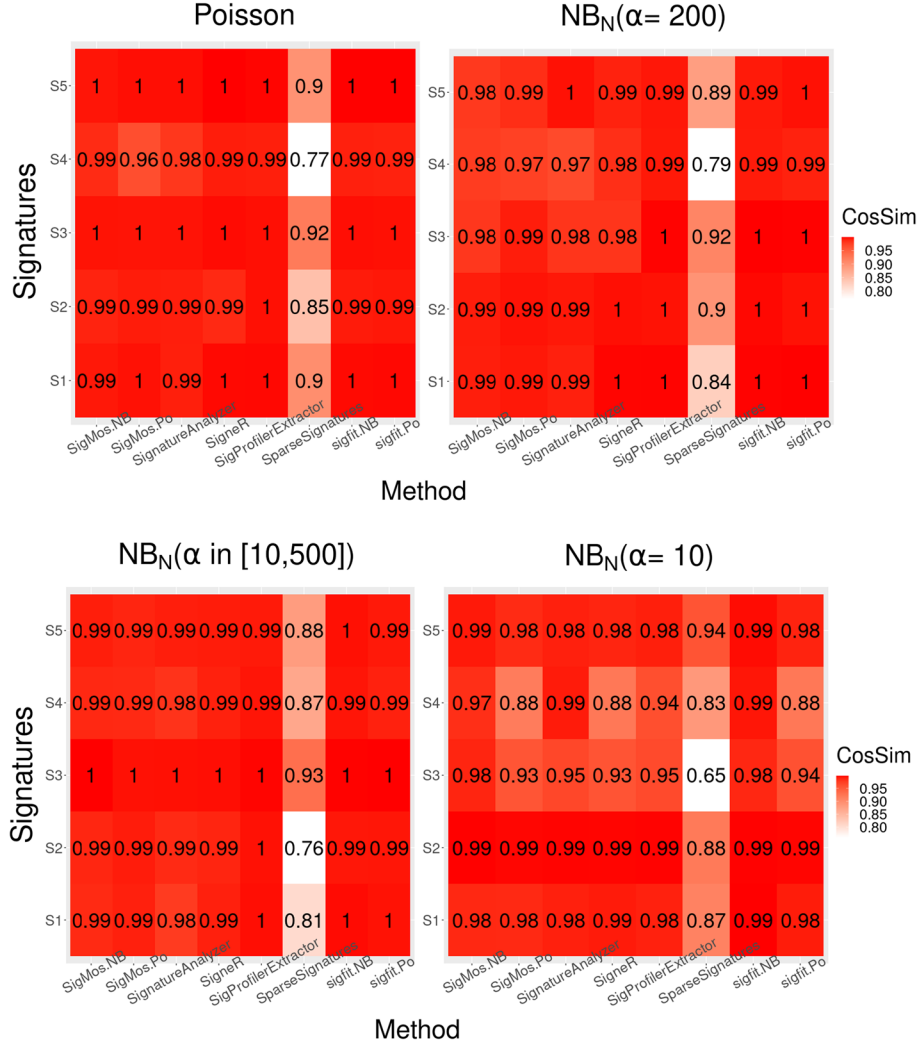


Fig. 3 Quality of estimated signatures using simulated data. Each method is applied on 10 randomly sampled data sets from Fig. 1b and, for each data set, the value of the estimated number of signatures is fixed to 5. We show the quality of the estimated signatures measured by cosine similarity for all methods with Poisson noise and Negative Binomial noise with $\alpha = 10, 200$, and a patient specific dispersion parameter $\alpha \sim U(10, 500)$

those in Fig. 2, show that for real data where the variance may be higher than the one accounted for under the Poisson model, using a Negative Binomial model is essential. Indeed, this distributional assumption leads to high accuracy in the estimated signatures and SigMoS combined with the Negative Binomial model is able to maintain high accuracy and also correctly infer the true number of signatures.

We additionally compared our method to an independent set of simulated data from [30]. Here, the authors propose an alternative cross-validation procedure for estimating the number of signatures and describe a method comparison where SigProfilerExtractor, SignatureAnalyzer and SignerR are included. We considered their 20 simulated data sets comprising of 200 patients and 9 signatures each and we run SigMoS under both the Negative Binomial and the Poisson model. The signatures used for this set of simulations have been taken from the PCAWG

breast cancer study [4] where two pairs of signatures are highly similar, namely signatures SBS1 and SBS5 as well as SBS2 and SBS13, and their exposures have been resampled jointly when generating the data. It is not surprising that our method often estimates less than 9 signatures (7 or 8 signatures are reconstructed in most of the scenarios). We compared these results to the ones in [30] where a method based on cross-validation is proposed to estimate the number of signatures. Here, an extensive method comparison is available showing the accuracy in estimating the true signatures. We provide similar results in Additional file 1: Figs. S1 and S2 where our method is run with Po-NMF and NB_N-NMF. Comparing these results to Fig. S9 in [30], we can see that most methods tend to estimate less than 9 signatures and that the accuracy of the signatures estimated by SigMoS is always higher or comparable to the ones estimated by the other methods.

These results indicate that our proposed approach is robust to different simulation set ups, has very good performance on a wide range of scenarios, and provides more accurate estimates of the underlying number of signatures and of the actual mutational signatures when compared to other methods available in the literature, suggesting that it will also be robust when applied to real data. Computational cost results for our method in terms of memory usage and time until convergence as a function of the number of patients are available in Additional file 1: Section S2. SigMoS runs on a standard laptop with Intel Core i7 processor in less than a few minutes and uses less than 25 gigabase of memory for data sets with up to 500 patients and 5 signatures. Both memory consumption and running time increase linearly with the number of patients, but even large data sets can be run fairly quickly on a standard laptop (for 1000 patients SigMoS used up to 100 GB and the running time went up to 7 min for the slowest scenarios).

Breast cancer data

This data set consists of the mutational counts from the 21 breast cancer patients that has previously been described and analyzed in several papers [6, 7, 12]. The data can be found through the link <ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl> from [11] and have been extensively analyzed in [4].

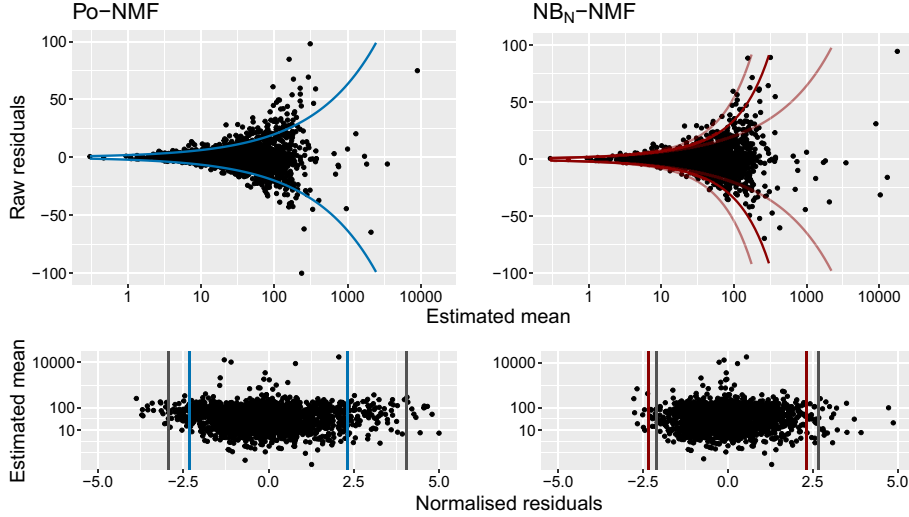
In Fig. 4a, we have applied SigMoS and BIC to choose the number of signatures for both Po-NMF and NB_N-NMF. We have included the BIC to compare with the SigMoS method as it provides similar results to the state-of-the-art methods. SigMoS indicates to use three signatures for both methods. This is in line with the results of our simulation study, where we show that our model selection is robust to model misspecification. According to BIC, six signatures are needed for Po-NMF whereas only three signatures should be used with NB_N-NMF which emphasizes the importance of a correct model choice when using BIC. In this framework and in “Prostate cancer data” section we compared SigMoS to BIC, as Fig. 1 shows that this is more robust than AIC. BIC is also often used as model selection criteria in the analysis of real data sets in the literature. We refer to “Method comparison” section for comparisons with other state-of-the-art methods.

For three signatures we show in Fig. 4b the corresponding raw residuals $R_{nm} = V_{nm} - (WH)_{nm}$ to determine the best fitting model. The residuals are plotted against the expected mean $(WH)_{nm}$, as the variance in both the Poisson and Negative

2. Results

Model selection procedure	Assumed models	
	Po-NMF	NB _N -NMF
SigMoS	3	3
BIC	6	3

(a) Estimated number of signatures.



(b) Model fit: residual analysis.

Fig. 4 Results for Po-NMF and NB_N-NMF applied to a data set with 21 breast cancer patients. **a** The optimal number of signatures estimated from SigMoS and BIC when using Po-NMF and NB_N-NMF. **b** The residual plots for Po-NMF and NB_N-NMF when assuming the estimated number of signatures from SigMoS i.e. 3 signatures in both cases. The lines in the top plot correspond to two times the expected variance under the chosen distributional assumption. As the NB_N-NMF holds 21 different expected variances, we have chosen to plot the median, minimum and maximum variance among the 21. The second plots show the normalized residuals. The vertical blue and red lines depict the theoretical quantiles and the gray lines show the observed quantiles

Binomial model depends on this value. The colored lines in the residual plots correspond to $\pm 2\sigma$ for the Poisson and the Negative Binomial distribution, respectively. The variance σ^2 can be derived from Eq. (5) for the Negative Binomial model and is equal to the mean for the Poisson model.

For Po-NMF we observe a clear overdispersion in the residuals, which suggests to use a Negative Binomial model. In the residual plot for the NB_N-NMF we see that the residuals have a much better fit to the variance structure, which is indicated by the colored lines. The quantile lines in the lower panel with normalized residuals again show that the quantiles from the NB_N-NMF are much closer to the theoretical ones, suggesting that the Negative Binomial model is better suited for this data. The patient specific dispersion is very diverse in this data as the α values for the first 20 patients are between 16 (very high dispersion) and 550 (moderate dispersion) and the last patient has $\alpha_{21} = 26083$.

We compare the signatures found by our method to the available signatures in the COSMIC database [5] downloaded from <https://cancer.sanger.ac.uk/cosmic>. We find that our three reconstructed signatures are similar to signatures SBS1, SBS2, SBS3.

The corresponding cosine similarities are reported in Table 1 and show high similarity between our reconstructed signatures and the ones from the COSMIC database especially for SBS2 and SBS3. Indeed, a cosine similarity of 0.8 has been used as threshold in [31] to group similar signatures, suggesting that SigMoS is able to identify relevant signatures in the COSMIC database. According to the results in [4] SBS1 and SBS2 are found across most cancer types and a large proportion of breast cancer samples showing these two signatures has been found. SBS3 has also been found in a large proportion of breast cancer samples and it also has high mutational burden in breast cancer tumors. SBS3 has also been associated to the BRCA1/2 mutation [4]. The validation of our signatures with the COSMIC database shows that in this case SigMoS can correctly infer signatures that have been proved to be strongly associated with breast cancer.

Prostate cancer data

We also considered a more recent data set from the Pan-Cancer Analysis of Whole Genomes (PCAWG) database [25] where 2782 patients from different cancer types are available. The mutational counts from the full PCAWG database can be found at <https://www.synapse.org/#!Synapse:syn11726620>. From this data set, we extracted mutational counts for all the 286 prostate cancer patients and used them directly for our analysis.

We chose again both the Poisson and Negative Binomial as underlying distributions for the NMF and in both cases we applied SigMoS for determining the number of signatures. We present the results in Fig. 5. Figure 5a shows again that our model selection procedure is more stable under model misspecification compared to BIC: the estimated number of signatures is changing from 9 to 4 between the two model assumptions for BIC, but only from 6 to 5 for SigMoS. As for Fig. 4b, the residuals in Fig. 5b show that the NB_N -NMF model provides a much better fit to the data than the Po-NMF. The estimated values for the patient specific dispersion are $\alpha_n \in [1.4, 4279]$ with a median of 140 (corresponding to a quite large dispersion).

As for the previous section we compare our reconstructed signatures with the ones in the COSMIC database. Table 2 shows the cosine similarity between the signatures extracted by SigMoS and the most similar ones from the COSMIC repository. Here, NB_N -NMF provides much better results in terms of signatures estimation showing the importance of accounting for overdispersion. Indeed, NB_N -NMF finds signatures SBS1, SBS5, SBS8, SBS18, SBS37. These signatures are all largely present in prostate cancer either for their presence in many prostate tumor samples or for their contribution in terms of number of mutations per tumor or for both reasons combined. On the contrary, signatures SBS6 and SBS36 are not found in prostate cancer, showing that NB_N -NMF is more accurate.

Table 1 Cosine similarity for the breast cancer data set between the signatures extracted by SigMoS and the ones in the COSMIC database

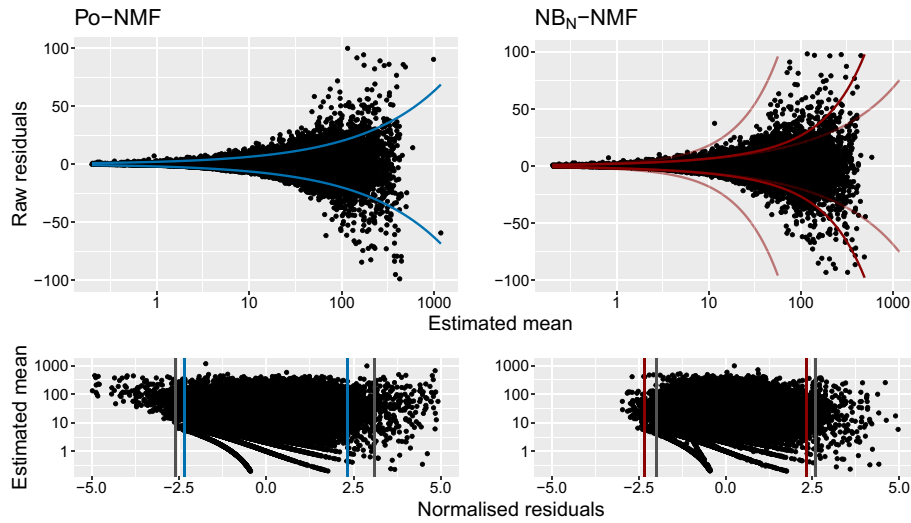
	SBS1	SBS2	SBS3
Po-NMF	0.65	0.76	0.79
NB_N -NMF	0.62	0.76	0.80

The COSMIC signature with the highest cosine similarity is shown for each signature estimated by SigMoS

3. Discussion

Model selection procedure	Assumed models	
	Po-NMF	NB _N -NMF
SigMoS	6	5
BIC	9	4

(a) Estimated number of signatures.



(b) Model fit: residual analysis.

Fig. 5 Results for Po-NMF and NB_N-NMF applied to a data set with 286 prostate cancer patients from the PCAWG database [25]. **a** The optimal number of signatures estimated from SigMoS and BIC when using Po-NMF and NB_N-NMF. **b** The residual plots for Po-NMF and NB_N-NMF when assuming the estimated number of signatures from SigMoS i.e. 5 and 6 signatures. The lines in the first plot correspond to two times the expected variance under the chosen distributional assumption. For NB_N-NMF, the colored lines in the top plot show the median, minimum and maximum variance among the patients. The bottom plots show the normalized residuals. The vertical blue and red lines depict the theoretical quantiles and the gray lines the observed quantiles

Table 2 Cosine similarity for the Prostate cancer data set between the signatures extracted by SigMoS and the ones in the COSMIC database

Table 2 Cosine similarity for the Prostate cancer data set between the signatures extracted by SigMoS and the ones in the COSMIC database

	<i>SBS1</i>	<i>SBS5</i>	<i>SBS6</i>	<i>SBS8</i>	<i>SBS18</i>	<i>SBS36</i>	<i>SBS37</i>	<i>SBS40</i>
Po-NMF	0.97	–	0.79	0.80	–	0.93	0.76	0.72
NB _N -NMF	0.96	0.67	–	0.84	0.67	–	0.79	–

The COSMIC signature with the highest cosine similarity is shown for each signature estimated by SigMoS. Signatures found in many prostate samples or having high mutation counts on prostate samples are highlighted in bold italic in the table

The COSMIC signature with the highest cosine similarity is shown for each signature estimated by SigMoS. Signatures found in many prostate samples or having high mutation counts on prostate samples are highlighted in colour in the table

Discussion

Mutational profiles from cancer patients are a widely used source of information and NMF is often applied to these data in order to identify signatures associated with cancer

types. We propose a new approach to perform the analysis and signature extraction from mutational count data where we emphasize the importance of validating the model using residual analysis, and we propose a robust model selection procedure.

We use the Negative Binomial model as an alternative to the commonly used Poisson model as the Negative Binomial can account for the high dispersion in the data. As a further extension of this model, we allow the Negative Binomial to have a patient specific variability component to account for heterogeneous variance across patients.

We propose a model selection approach for choosing the number of signatures. As we show in “[Simulation study](#)” section this method works well with both Negative Binomial and Poisson data, and it is a robust procedure for choosing the number of signatures. We note that the choice of the divergence measure for the *cost* function in Algorithm 2 is not trivial and may favor one or the other model and thus a comparison of the costs between different NMF methods is not possible. For example, in our framework, we use the Kullback–Leibler divergence which would favor the Poisson model. This means that a direct comparison between the cost values for Po-NMF and NB_N-NMF is not feasible. To check the goodness of fit and choose between the Poisson model and the Negative Binomial model we propose to use the residuals instead.

In Additional file 1: Section S4, we investigated the role of the cost function in our model selection by including the Frobenius norm and Itakura–Saito (IS) [32] divergence measure from [33], where the authors propose a fast implementation of the NMF algorithm with general Bregman divergence. In this investigation the cost function did not influence the optimal number of signatures. The only difference was how the cost values differed among the NMF methods, as each cost function favored the models differently. Therefore we chose to use the Kullback–Leibler divergence and compared the methods with the residual analysis.

Less signatures are found when accounting for overdispersion with the Negative Binomial model. Indeed, there is no need to have additional signatures explaining noise, which we assume is the case for the Poisson model. We show that the Negative Binomial model is more suitable and therefore believe the corresponding signatures are more accurate. This can be helpful when working with mutational profiles for being able to better associate signatures with cancer types and for a clearer interpretation of the signatures when analyzing mutational count data. For example, the recent results in [28] use a large data set with several different cancer types and show that there exists a set of common signatures that is shared across organs and a set of rare signatures that are only found with a sufficiently large sample size. To recover the common signatures the patients with unusual mutational profiles were excluded as they are introducing additional variance in the signature estimation procedure. We speculate that changing the Poisson assumption in this approach with the Negative Binomial distribution could provide a simpler and more robust way to extract common signatures. Indeed, the Negative Binomial model allows for more variability in the data and our simulation results and residual plots in “[Results](#)” section show that the Negative Binomial distribution is beneficial for stable signature estimation. In this work we have focused on single base substitutions, but the Negative Binomial NMF can be highly beneficial also for analyzing indels or other variant types. In [4] they discuss that mutational matrices corresponding to indels harbor more variation which means that more flexible models than the Poisson are needed in this situation.

4. Methods

The workflow for analyzing the data, and the procedures in Algorithms 1 and 2 are available in the R package SigMoS at <https://github.com/MartaPelizzola/SigMoS>.

Methods

This section is structured as follows: in “[Negative Binomial model for mutational counts](#)” section we describe the Negative Binomial model applied to mutational count data. Then we propose an extension where a patient specific dispersion coefficient is used. The majorization–minimization (MM) procedure for patient specific dispersion $\{\alpha_1, \dots, \alpha_N\}$ can be found in “[Patient specific NB_N-NMF](#)” section. In our application, we propose to use Negative Binomial maximum likelihood estimation (MLE) for α and $\{\alpha_n : 1 \leq n \leq N\}$ instead of the grid search adopted in [18]. The pseudocode shown in the initial steps of Algorithm 1 describes this approach for patient specific dispersion. For shared dispersion among all patients and mutation types we simply set $\alpha = \alpha_1 = \dots = \alpha_N$ in Algorithm 1. Lastly, in “[Estimating the number of signatures](#)” section we describe our proposed algorithm to estimate the number of signatures.

Negative Binomial model for mutational counts

In this section we argue why the Negative Binomial model in [18] is a natural model for the number of somatic mutations in a cancer patient. We start by illustrating the equivalence of the Negative Binomial to the more natural Beta-Binomial model as a motivation for our model choice.

Assume a certain mutation type can occur in τ triplets along the genome with a probability p . Then it is natural to model the mutational counts with a binomial distribution [34, 35]

$$V_{nm} \sim \text{Bin}(\tau, p). \quad (2)$$

However, [36] observed that the probability of a mutation varies along the genome and is correlated with both expression levels and DNA replication timing. We therefore introduce the Beta-Binomial model

$$\begin{aligned} V_{nm}|p &\sim \text{Bin}(\tau, p) \\ p &\sim \text{Beta}(\alpha, \beta), \end{aligned} \quad (3)$$

where the beta prior on the probability p models the heterogeneity of the probability of a mutation for the different mutation types due to the high variance along the genome. As p follows a Beta distribution, its expected value is $\mathbb{E}[p] = \alpha/(\alpha + \beta)$. For mutational counts, the number of triplets τ is extremely large and the probability of mutation p is very small. In the data described in [36] there are typically between 1 and 10 mutations per megabase with an average of 4 mutations per megabase ($\tau \approx 10^6$). This means $\mathbb{E}[p] = \alpha/(\alpha + \beta) \approx 4 \cdot 10^{-6}$ and thus, for mutational counts in cancer genomes we have that $\beta \gg \alpha$. As τ is large and p is small, the Binomial model is very well approximated by the Poisson model $\text{Bin}(\tau, p) \simeq \text{Pois}(\tau p)$. This distributional equivalence of Poisson and Binomial when τ is large and p is small is well known. This also means that the models (1) and (2) are approximately equivalent with $\tau p = (WH)_{nm}$.

The Beta and Gamma distributions are also approximately equivalent in our setting. Indeed, as $\beta \gg \alpha$, the Beta density can be approximated by the Gamma density in the following way

$$\begin{aligned}\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} &= \frac{p^{\alpha-1}}{\Gamma(\alpha)} (\beta-1+\alpha)(\beta-1+(\alpha-1)) \cdots (\beta-1)(1-p)^{\beta-1} \\ &\approx \frac{p^{\alpha-1}}{\Gamma(\alpha)} \beta^\alpha (e^{-p})^\beta.\end{aligned}$$

Therefore, for mutational counts, the model in (3) is equivalent to

$$\begin{aligned}V_{nm}|p &\sim \text{Po}(\tau p) \\ p &\sim \text{Gamma}(\alpha, \beta).\end{aligned}\tag{4}$$

Since the Negative Binomial model is a Gamma–Poisson model we can also write the model as

$$V_{nm} \sim \text{NB}\left(\alpha, \frac{\tau}{\beta + \tau}\right) \simeq \text{NB}\left(\alpha, \frac{\tau \mathbb{E}[p]}{\alpha + \tau \mathbb{E}[p]}\right) \simeq \text{NB}\left(\alpha, \frac{(WH)_{nm}}{\alpha + (WH)_{nm}}\right),$$

where the last parametrization is equivalent to the one in [18]. In the first distributional equivalence we use $\mathbb{E}[p] \approx \frac{\alpha}{\beta}$ and in the second we use $\tau \mathbb{E}[p] = (WH)_{nm}$. Compared to the Beta-Binomial model, the Negative Binomial model has one fewer parameter and is analytically more tractable. The mean and variance of this model are given by

$$\mathbb{E}[V_{nm}] = (WH)_{nm} \quad \text{and} \quad \text{Var}(V_{nm}) = (WH)_{nm} \left(1 + \frac{(WH)_{nm}}{\alpha}\right).\tag{5}$$

When $\alpha \rightarrow \infty$ above, the Negative Binomial model converges to the more commonly used Poisson model as $\text{Var}(V_{nm}) \downarrow (WH)_{nm}$. As shown in this section, the Negative Binomial model can be seen both as an extension of the Poisson model and as equivalent to the Beta-Binomial model. Thus, we opted to implement a Negative Binomial NMF model for mutational count data. More details on the approximation of the Negative Binomial to the Beta-Binomial distribution can also be found in [37].

Patient specific NB_N-NMF

In this section we describe our patient specific Negative Binomial non-negative matrix factorization NB_N-NMF model and the corresponding estimation procedure.

Gouvert et al. [18], Lyu et al. [20] and Vöhringer et al. [21] present a Negative Binomial model where α is shared across all observations. However, the probability of a mutation in (3) is highly variable across patients (see e.g. mutational burden in [28] and our discussion in “Breast cancer data” section), thus we extend the Negative Binomial NMF model from [18] by allowing patient specific dispersion. We noticed that the variability among different patients is usually much higher than the one among different mutation types, thus we decided to focus on patient specific dispersion.

The entries in V are modeled as

$$V_{nm} \sim \text{NB}\left(\alpha_n, \frac{(WH^T)_{nm}}{\alpha_n + (WH^T)_{nm}}\right),$$

where α_n is the dispersion coefficient of each patient, and the corresponding Gamma–Poisson hierarchical model can be rewritten as:

4. Methods

$$\begin{aligned} V_{nm}|a_{nm} &\sim \text{Po}(a_{nm}(WH)_{nm}) \\ a_{nm} &\sim \text{Gamma}(\alpha_n, \alpha_n). \end{aligned} \quad (6)$$

Here a_{nm} is the parameter responsible for the variability in the Negative Binomial model. Note that $\mathbb{E}[a_{nm}] = 1$ and $\text{Var}(a_{nm}) = 1/\alpha_n$.

Now we can write the Negative Binomial log-likelihood function with patent specific α_n

$$\begin{aligned} \ell(W, H; V) = \sum_{n=1}^N \sum_{m=1}^M \left\{ \log \left(\frac{\alpha_n + V_{nm} - 1}{\alpha_n} \right) + V_{nm} \log \left(\frac{(WH)_{nm}}{\alpha_n + (WH)_{nm}} \right) \right. \\ \left. + \alpha_n \log \left(1 - \frac{(WH)_{nm}}{\alpha_n + (WH)_{nm}} \right) \right\}, \end{aligned} \quad (7)$$

and recognize the negative of the log-likelihood function as proportional to the following divergence:

$$d_N(V||WH) = \sum_{n=1}^N \left\{ \sum_{m=1}^M V_{nm} \log \left(\frac{V_{nm}}{(WH)_{nm}} \right) - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + (WH)_{nm}} \right) \right\} \quad (8)$$

assuming fixed $\alpha_1, \dots, \alpha_N$. This is a divergence measure as $d_N(V||WH) = 0$ when $V = WH$ and $d_N(V||WH) > 0$ for $V \neq WH$. We can show this by defining $g(t) = (V_{nm} + t) \log((V_{nm} + t)/((WH)_{nm} + t))$ and realize $d_N(V||WH) = g(0) - g(\alpha) \geq 0$ because $g'(t) \leq 0$ with equality only when $V = WH$. The term $\log \left(\frac{\alpha_n + V_{nm} - 1}{\alpha_n} \right)$ in the likelihood is a constant we can remove and then we have added the constants $V_{nm} \log(V_{nm})$, $\alpha_n \log(\alpha_n)$ and $(V_{nm} + \alpha_n) \log(V_{nm} + \alpha_n)$.

Following the steps in [18], we will update W and H one at a time, while the other is assumed fixed. We will show the procedure for updating H using a fixed W and its current value H^t . First we construct a majorizing function $G(H, H^t)$ for $d_N(V||WH)$ with the constraint that $G(H, H) = d_N(V||WH)$. The first term in Eq. (8) can be majorized using Jensen's inequality leading to

$$\begin{aligned} d_N(V||WH) &= \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log \left(\frac{V_{nm}}{\sum_{k=1}^K W_{nk} H_{km}} \right) \right. \\ &\quad \left. - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km}} \right) \right\} \\ &\leq \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log V_{nm} - V_{nm} \sum_{k=1}^K \beta_k \log \frac{W_{nk} H_{km}}{\beta_k} \right. \\ &\quad \left. + (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + \sum_{k=1}^K W_{nk} H_{km}}{\alpha_n + V_{nm}} \right) \right\} \end{aligned} \quad (9)$$

where $\beta_k = W_{nk} H_{km}^t / \sum_{k=1}^K W_{nk} H_{km}^t$. The second term can be majorized with the tangent line using the concavity property of the logarithm:

$$\begin{aligned}
d_N(V||WH) &= \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log V_{nm} - V_{nm} \sum_{k=1}^K \beta_k \log \frac{W_{nk} H_{km}}{\beta_k} \right. \\
&\quad \left. + (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + \sum_{k=1}^K W_{nk} H_{km}}{\alpha_n + V_{nm}} \right) \right\} \\
&\leq \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log V_{nm} - V_{nm} \sum_{k=1}^K \beta_k \log \frac{W_{nk} H_{km}}{\beta_k} \right. \\
&\quad \left. + (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + (WH^t)_{nm}}{\alpha_n + V_{nm}} \right) \right. \\
&\quad \left. + \frac{W_{nm}}{\alpha_n + (WH^t)_{nm}} (H_{nm} - H_{nm}^t) \right\} = G(H, H^t).
\end{aligned} \tag{10}$$

Lastly, we need to show that $G(H, H) = d_N(V||WH)$. This follows from

$$\begin{aligned}
G(H, H) &= \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log V_{nm} - V_{nm} \sum_{k=1}^K \beta_k \log \frac{W_{nk} H_{km}}{\beta_k} \right. \\
&\quad \left. + (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + (WH)_{nm}}{\alpha_n + V_{nm}} \right) + \frac{W_{nm}}{\alpha_n + (WH)_{nm}} (H_{nm} - H_{nm}) \right\} \\
&= \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log V_{nm} - V_{nm} \sum_{k=1}^K \frac{W_{nk} H_{km}}{\sum_{k=1}^K W_{nk} H_{km}} \log \frac{W_{nk} H_{km}}{\frac{W_{nk} H_{km}}{\sum_{k=1}^K W_{nk} H_{km}}} \right. \\
&\quad \left. - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km}} \right) \right\} \\
&= \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log V_{nm} - V_{nm} \cdot 1 \cdot \log \left(\sum_{k=1}^K W_{nk} H_{km} \right) \right. \\
&\quad \left. - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km}} \right) \right\} \\
&= \sum_{n=1}^N \sum_{m=1}^M \left\{ V_{nm} \log \left(\frac{V_{nm}}{\sum_{k=1}^K W_{nk} H_{km}} \right) \right. \\
&\quad \left. - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km}} \right) \right\} \\
&= d_N(V||WH).
\end{aligned} \tag{11}$$

Having defined the majorizing function $G(H, H^t)$ in Eq. (10), we can derive the following multiplicative update for H :

$$H_{km}^{t+1} = H_{km}^t \frac{\sum_{n=1}^N \frac{V_{nm}}{(WH^t)_{nm}} W_{nk}}{\sum_{n=1}^N \frac{V_{nm} + \alpha_n}{(WH^t)_{nm} + \alpha_n} W_{nk}}. \tag{12}$$

Similar calculations can be carried out for W to obtain the following update:

4. Methods

$$W_{nk}^{t+1} = W_{nk}^t \frac{\sum_{m=1}^M \frac{V_{nm}}{(W^t H)_{nm}} H_{km}}{\sum_{m=1}^M \frac{V_{nm} + \alpha_n}{(W^t H)_{nm} + \alpha_n} H_{km}}. \quad (13)$$

It is straightforward to see that when $\alpha_n = \alpha$ for all $n = 1, \dots, N$ then the updates for W and H equal those in [18]. Additionally, as shown in [18] when $\alpha \rightarrow \infty$ the updates of the Po-NMF [14] are recovered.

In our application, we find maximum likelihood estimates (MLEs) of $\alpha_1, \dots, \alpha_N$ based on the Negative Binomial likelihood using Newton–Raphson together with the estimate of WH from Po-NMF. We opted for this more precise estimation procedure for $\alpha_1, \dots, \alpha_N$ instead of the grid search approach used in [18]. Final estimates of W and H are then found by minimizing the divergence in Eq. (8) by the iterative majorize-minimization procedure. The NB_N-NMF procedure is described in Algorithm 1 below. The model in [18, 20] is similar except $\alpha_1 = \dots = \alpha_N = \alpha$.

It is well known that NMF can result in non-unique solutions [38]. Following these findings on the non-uniqueness and the effect of different initializations, all our results are based on five random initializations for each NMF solution.

Algorithm 1 NB_N-NMF: Estimation of W , H and $\{\alpha_1, \dots, \alpha_N\}$

Input: V, K, ϵ

Output: $W, H, \{\alpha_1, \dots, \alpha_N\}$

- 1: $W^{Po}, H^{Po} \leftarrow$ apply Po-NMF to V with K signatures
 - 2: $\alpha_1, \dots, \alpha_N \leftarrow$ Negative Binomial MLE using W^{Po}, H^{Po} and V
 - 3: Initialize W^1, H^1 from a random uniform distribution
 - 4: **for** $i = 1, 2, \dots$ **do**
 - 5: $W_{nk}^{i+1} \leftarrow W_{nk}^i \frac{\sum_{m=1}^M \frac{V_{nm}}{(W^i H^i)_{nm}} H_{km}^i}{\sum_{m=1}^M \frac{V_{nm} + \alpha_n}{(W^i H^i)_{nm} + \alpha_n} H_{km}^i}$
 - 6: $H_{km}^{i+1} \leftarrow H_{km}^i \frac{\sum_{n=1}^N \frac{V_{nm}}{(W^{i+1} H^i)_{nm}} W_{nk}^{i+1}}{\sum_{n=1}^N \frac{V_{nm} + \alpha_n}{(W^{i+1} H^i)_{nm} + \alpha_n} W_{nk}^{i+1}}$
 - 7: **if** $|d_N(V || W^{i+1} H^{i+1}) - d_N(V || W^i H^i)| < \epsilon$ **then**
 - 8: **return** $W, H \leftarrow W^{i+1}, H^{i+1}$
 - 9: **end if**
 - 10: **end for**
-

Estimating the number of signatures

Estimating the number of signatures is a difficult problem when using NMF. More generally, estimating the number of components for mixture models or the number of clusters is a well known challenge in applied statistics.

Examples of the complexity of this problem can be found in the K -means clustering algorithm and in Gaussian mixture models where the number of clusters K has to be provided for the methods. A detailed description of these challenges can be found in [39]. Estimating the number of components is also a critical issue for mixed membership models. Some examples can be found in [40, 41].

Classical procedures to perform model selection are the AIC

$$\text{AIC} = -2 \ln L + 2n_{prm} \quad (14)$$

and the Bayesian Information Criterion (BIC)

$$\text{BIC} = -2 \ln L + \ln(n_{obs})n_{prm} \quad (15)$$

where $\ln L$ is the estimated log-likelihood value, n_{obs} is the number of observations and n_{prm} the number of parameters to be estimated. The two criteria attempt to balance the fit to the data (measured by $-2 \ln L$) and the complexity of the model (measured by the scaled number of free parameters). We have $n_{obs} = N$ where N is the number of patients, so $\ln(n_{obs}) > 2$ if $N \geq 8$, which means that in our context the number of parameters has a higher influence for BIC compared to AIC because real data sets always have at least tens of patients. Additionally, the structure of the mutational matrix V can lead to two different strategies for choosing n_{obs} when BIC is used. Indeed, the number of observations in this context can be set as the total number of counts (i.e. $N \cdot M$) or as the number of patients N , leading to an ambiguity in the definition of this criterion. Verity and Nichols [41] also presents results on the performance of AIC and BIC, where the power is especially low for BIC. AIC provides higher stability in the scenario from [41], however it does not seem suitable in our situation due to a small penalty term.

A very popular model selection procedure is cross-validation. In Gelman et al. [42] they compare various model selection methods including AIC and cross-validation. Here, the authors recommend to use cross-validation as they demonstrate that the other methods fail in some circumstances. In Luo et al. [43] they also show that cross-validation has better performance than the other considered methods, including AIC and BIC. Both papers evaluate the predictive fit to compare different methods.

Model selection for NMF

For NMF we propose an approach for estimating the rank which is highly inspired by cross-validation. As for classical cross-validation we split the patients in V in a training and a test set multiple times.

Since all the parameters in the model i.e. W and H are free parameters it means that the exposures for the patients in the test set are unknown from the estimation of the training set. The patients in the training set give an estimation of the signatures and the exposures of the patients in the training set. One could argue to fix the signatures from the training set and re-estimate exposures for the test set, but we observed that this lead to an overestimation of the test set.

Instead we have chosen to fix the exposures to the ones estimated from the full data. This means our evaluation on the test set is a combination of estimated signatures from the training set and exposures from the full data. The idea is to exploit the fact that the signature matrix should be robust to changes in the patients included in the training set. If the estimated signatures are truly explaining the main patterns in the data, then we expect the signatures obtained from the training set to be similar to the ones from the full data. Therefore the product of the exposures from the full data and the signatures from the training set should give a good approximation of the test set, if the number of signatures is appropriate. We tested this assumption on a real data set with hypermutated patients which may lead to patient specific signatures in Additional file 1:

4. Methods

Section S3 and we find that our method is robust to the removal of the hypermutated patient.

Inputs for the procedure are the data V , an NMF method, the number of signatures K , the number of splits into training and test J and the *cost* function. We evaluate the model for a range of values of K and then select the model with the lowest cost. The NMF methods we are using here are either Po-NMF from [14] or NB_N-NMF in Algorithm 1, but any NMF method could be applied.

A visualization of our model selection algorithm can be found in Fig. 6. First, we consider the full mutational matrix V and we apply the chosen NMF algorithm to obtain an estimate for both W and H . Afterwards, for each iteration, we sample 90% of the patients randomly to create the training set and determine the remaining 10% as our test set. We then apply the chosen NMF method to the mutational counts of the training set obtaining an estimate W_{train} and H_{train} .

Now, as for classical cross-validation, we want to evaluate our model on the test set. To evaluate the model here, we use the full data: indeed, we multiply the exposures relative to the patients in the test set estimated on the full data W_{test}^j times the corresponding signatures estimated from the training set H_{train}^j . As the order of the estimated signatures from the full data can be different to the one in the training set we reorder the exposures in W_{test}^j with respect to the signatures in H_{train}^j . We determine the order by calculating the cosine similarity between the signatures in H_{train}^j and those in H . We use the prediction of the test data to evaluate the model computing the distance between the true data V_{test}^j and their prediction $V_{predict}^j$ with a suitable *cost* function. This procedure is iterated J times leading to J cost values c_j , $j = 1, \dots, J$. The median of these values is calculated for each number of signatures K . We call this procedure SigMoS and summarize it in Algorithm 2. The optimal K is the one with the lowest cost. We use the generalized Kullback–Leibler divergence as a cost function and discuss the choice of cost function in “Discussion” section. We compare the influence of the model choice for our procedure to AIC and BIC. We also compare to SigProfilerExtractor, SignatureAnalyzer, Signer and SparseSignatures as these are recently introduced methods in the literature and examine the results from this comparison in “Simulation study” section.

Algorithm 2 SigMoS: Cost for a given number of signatures K for the count matrix V

Input: $V, K, J, cost$, NMF-method

Output: c_{median}

```

1:  $W, H \leftarrow$  apply the chosen NMF method to  $V$  with  $K$  signatures
2: for  $j = 1$  to  $J$  do
3:    $V_{train}^j \leftarrow$  mutational counts for the patients in the  $j^{th}$  training set
4:    $V_{test}^j \leftarrow V \setminus V_{train}^j$ 
5:    $W_{test}^j \leftarrow$  exposures from  $W$  for the patients in the test set
6:    $W_{train}^j, H_{train}^j \leftarrow$  apply the chosen NMF method to  $V_{train}^j$  with  $K$  signatures
7:    $c_j \leftarrow cost(V_{test}^j, W_{test}^j H_{train}^j)$ 
8: end for
9: return  $c_{median} \leftarrow median(c_1, \dots, c_J)$ 

```

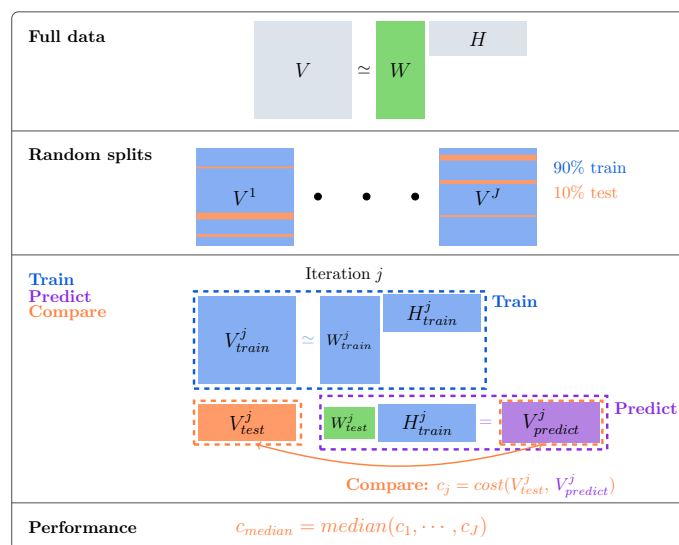


Fig. 6 SigMoS procedure for a given number of signatures K and a count matrix V . Pseudocode can be found in Algorithm 2

Code for method comparison

For SigProfilerExtractor we used the SigProfilerExtractor Python package with minimum_signatures equal to two, maximum_signatures equal to eight and opportunity_genome equal to “GRCh37”. For SparseSignatures we use the function nmfLassoCV with normalize_counts being set to FALSE and lambda_values_alpha and lambda_values_beta to zero. All the other parameters are set to their default values. When applying SignatureAnalyzer we used the following command `python SignatureAnalyzer- GPU.py --data f --prior_on_W L1 --prior_on_H L2 --output_dir d --max_iter 1000000 --tolerance 1e-7 --K0 8`. For Signer we used the default options.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-023-05304-1>.

Additional file 1. Supplementary material.

Acknowledgements

We would like to thank Simon Opstrup Drue and two anonymous reviewers for helpful comments and suggestions on an earlier version of this manuscript.

Author contributions

MP, RL and AH contributed to the design of the research. MP and RL wrote the code for software, simulations and data analysis and wrote the first version of the manuscript. All authors contributed to writing the manuscript, read and approved the manuscript.

Funding

The research reported in this publication is supported by the Novo Nordisk Foundation. MP acknowledges funding of the Austrian Science Fund (FWF Doctoral Program Vienna Graduate School of Population Genetics, DK W1225-B20).

Availability of data and materials

The code for our model selection procedure and Negative Binomial NMF and for the simulations is available in the R package SigMoS and can be found at <https://github.com/MartaPelizzola/SigMoS>. The real data used in “Breast cancer data” section are available at <ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl> from [11]. The real data used in “Prostate cancer data” section can be found at <https://www.synapse.org/#!Synapse:syn11726620>.

4. References

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 11 January 2023 Accepted: 25 April 2023

Published online: 08 May 2023

References

1. Risques RA, Kennedy SR. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet*. 2018;14(1): e1007108. <https://doi.org/10.1371/JOURNAL.PGEN.1007108>.
2. Shibai A, Takahashi Y, Ishizawa Y, Motooka D, Nakamura S, Ying B-W, Tsuru S. Mutation accumulation under UV radiation in *Escherichia coli*. *Sci Rep*. 2017;7(1):1–12. <https://doi.org/10.1038/s41598-017-15008-1>.
3. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, Campbell PJ, Vineis P, Phillips DH, Stratton MR. Mutational signatures associated with tobacco smoking in human cancer. *Science*. 2016;354(6312):618–22. <https://doi.org/10.1126/SCIENCE.AAG0299>.
4. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, Islam SMA, Lopez-Bigas N, Klimczak LJ, McPherson JR, Morganella S, Sabarinathan R, Wheeler DA, Mustonen V, Getz G, Rozen SG, Stratton MR. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94–101. <https://doi.org/10.1038/s41586-020-1943-3>.
5. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;47(D1):941–7. <https://doi.org/10.1093/NAR/GKY1015>.
6. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. 2013;3(1):264–259.
7. ...Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jönsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JWM, Aparicio SAJR, Borg Å, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR. Mutational processes molding the genomes of 21 breast cancers. *Cell*. 2012;149(5):979–93. <https://doi.org/10.1016/j.cell.2012.04.024>.
8. Lal A, Liu K, Tibshirani R, Sidow A, Ramazzotti D. De novo mutational signature discovery in tumor genomes using SparseSignatures. *PLoS Comput Biol*. 2021;17(6):1009119. <https://doi.org/10.1371/JOURNAL.PCBI.1009119>.
9. Baez-Ortega A, Gori K. Computational approaches for discovery of mutational signatures in cancer. *Brief Bioinform*. 2017;20(1):77–88. <https://doi.org/10.1093/bib/bbx082>.
10. Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLoS ONE*. 2019;14(9):0221235. <https://doi.org/10.1371/journal.pone.0221235>.
11. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale A-L, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500(7463):415–21.
12. Fischer A, Illingworth CJR, Campbell PJ, Mustonen V. EMu: Probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol*. 2013;14(4):1–10. <https://doi.org/10.1186/gb-2013-14-4-r39>.
13. Rosales RA, Drummond RD, Valieris R, Dias-Neto E, Da Silva IT. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics*. 2017;33(1):8–16. <https://doi.org/10.1093/bioinformatics/btw572>.
14. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401(6755):788–91. <https://doi.org/10.1038/44565>.
15. Bliss CI, Fisher RA. Fitting the negative binomial distribution to biological data. *Biometrics*. 1953;9(2):176. <https://doi.org/10.2307/3001850>.
16. Martincorena I, Raine K, Gerstung M, Dawson K, Haase K, Van Loo P, Davies H, Stratton M, Campbell P. Universal patterns of selection in cancer and somatic tissues. *Cell*. 2017;171(5):1029–104121. <https://doi.org/10.1016/J.CELL.2017.09.042>.
17. Zhang J, Liu J, McGillivray P, Yi C, Lochovsky L, Lee D, Gerstein M. NIMBus: a negative binomial regression based integrative method for mutation burden analysis. *BMC Bioinform* 2020 21:1. 2020;21(1):1–25. <https://doi.org/10.1186/S12859-020-03758-1>.
18. Gouvert O, Oberlin T, Fevotte C. Negative binomial matrix factorization. *IEEE Signal Process Lett*. 2020;27:815–9. <https://doi.org/10.1109/LSP.2020.2991613>.
19. Gori K, Baez-Ortega A. sigfit: flexible Bayesian inference of mutational signatures; 2018. <https://doi.org/10.1101/372896>

20. Lyu X, Garret J, Rätsch G, Lehmann KV. Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics*. 2020;36(Suppl-1):154–60. <https://doi.org/10.1093/BIOINFORMATICS/BTAA473>.
21. Vöhringer H, Hoeck AV, Cuppen E, Gerstung M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat Commun*. 2021;12(1):3628. <https://doi.org/10.1038/s41467-021-23551-9>.
22. Févotte C, Bertin N, Durrieu J. Nonnegative matrix factorization with the Itakura–Saito divergence: with application to music analysis. *Neural Comput*. 2009;21(3):793–830. <https://doi.org/10.1162/NECO.2008.04-08-771>.
23. Islam SMA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, He Y, Vella M, Wang J, Teague JW, Clapham P, Moody S, Senkin S, Li YR, Riva L, Zhang T, Gruber AJ, Steele CD, Otlu B, Khandekar A, Abbasi A, Humphreys L, Sylyukina N, Brady SW, Alexandrov BS, Pillay N, Zhang J, Adams DJ, Martincorena I, Wedge DC, Landi MT, Brennan P, Stratton MR, Rozen SG, Alexandrov LB. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics*. 2022;2(11): 100179. <https://doi.org/10.1016/j.xgen.2022.100179>.
24. Taylor-Weiner A, Aguet F, Haradhvala NJ, Gosai S, Anand S, Kim J, Ardlie K, Allen EMV, Getz G. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol*. 2019;20(1):1–5. <https://doi.org/10.1186/s13059-019-1836-7>.
25. Campbell PJ. Pan-cancer analysis of whole genomes. *Nature*. 2020;578(7793):82–93. <https://doi.org/10.1038/s41586-020-1969-6>.
26. Cook RD. Exploring partial residual plots. *Technometrics*. 1993;35(4):351–62. <https://doi.org/10.1080/00401706.1993.10485350>.
27. Miles J. Residual plot. 2014. <https://doi.org/10.1002/9781118445112.stat06619>.
28. Degasperis A, Zou X, Amarante TD, Martinez-Martinez A, Koh GCC, Dias JML, Heskin L, Chmelova L, Rinaldi G, Wang VYW, Nanda AS, Bernstein A, Momen SE, Young J, Perez-Gil D, Memari Y, Badja C, Shooter S, Czarnecki J, Brown MA, Davies HR, Nik-Zainal S, Ambrose JC, Arumugam P, Bevers R, Bleda M, Boardman-Pretty F, Boustred CR, Brittain H, Caulfield MJ, Chan GC, Fowler T, Giess A, Hamblin A, Henderson S, Hubbard TJP, Jackson R, Jones LJ, Kasperaviciute D, Kayikci M, Kousathanas A, Lahnstein L, Leigh SEA, Leong IUS, Lopez FJ, Maleady-Crowe F, McEntagart M, Minneci F, Moutsianas L, Mueller M, Murugaesu N, Need AC, O'Donovan P, Odhams CA, Patch C, Perez-Gil D, Pereira MB, Pullinger J, Rahim T, Rendon A, Rogers T, Savage K, Sawant K, Scott RH, Siddiq A, Sieghart A, Smith SC, Sosinsky A, Stuckey A, Tanguy M, Tavares ALT, Thomas ERA, Thompson SR, Tucci A, Welland MJ, Williams E, Witkowska K, Wood SM. Substitution mutational signatures in whole-genome sequenced cancers in the UK population. *Science*. 2022;376(6591):9283. <https://doi.org/10.1126/science.abl9283>.
29. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, Loo PV, Ju YS, Smid M, Brinkman AB, Morganella S, Aure MR, Lingjærde OC, Langerød A, Ringnér M, Ahn S-M, Boyault S, Brock JE, Brooks A, Butler A, Desmedt C, Dirix L, Dronov S, Fatima A, Foekens JA, Gerstung M, Hooijer GKJ, Jang SJ, Jones DR, Kim H-Y, King TA, Krishnamurthy S, Lee HJ, Lee J-Y, Li Y, McLaren S, Menzies A, Mustonen V, O'Meara S, Pauporté I, Pivrot X, Purdie CA, Raine K, Ramakrishnan K, Rodríguez-González FG, Romieu G, Sieuwerts AM, Simpson PT, Shepherd R, Stebbings L, Stefansson OA, Teague J, Tommasi S, Treilleux I, den Eynden GGV, Vermeulen P, Vincent-Salomon A, Yates L, Caldas C, van't Veer L, Tutt A, Knappskog S, Tan BKT, Jonkers J, Borg Å, Ueno NT, Sotiriou C, Viari A, Futreal PA, Campbell PJ, Span PN, Laere SV, Lakhani SR, Eyfjord JE, Thompson AM, Birney E, Stunnenberg HG, van de Vijver MJ, Martens JWM, Børresen-Dale A-L, Richardson AL, Kong G, Thomas G, Stratton MR. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;534(7605):47–54. <https://doi.org/10.1038/nature17676>.
30. Lee D, Wang D, Yang XR, Shi J, Landi MT, Zhu B. SUITOR: selecting the number of mutational signatures through cross-validation. *PLoS Comput Biol*. 2022;18(4):1009309. <https://doi.org/10.1371/journal.pcbi.1009309>.
31. Pei G, Hu R, Dai Y, Zhao Z, Jia P. Decoding whole-genome mutational signatures in 37 human pan-cancers by denoising sparse autoencoder neural network. *Oncogene*. 2020;39(27):5031–41. <https://doi.org/10.1038/s41388-020-1343-z>.
32. Févotte C, Idier J. Algorithms for nonnegative matrix factorization with the β -divergence. *Neural Comput*. 2011;23(9):2421–56. [arXiv:1010.1763](https://arxiv.org/abs/1010.1763).
33. Li L, Lebanon G, Park H. Fast Bregman divergence NMF using Taylor expansion and coordinate descent. In: *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. 2012.
34. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet*. 2014;46(11):1160–5. <https://doi.org/10.1038/NG.3101>.
35. Loichovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res*. 2015;43(17):8123–34. <https://doi.org/10.1093/NAR/GKV803>.
36. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortés ML, Auclair D, Saksena G, Voet D, Noble M, Dicara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CWM, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013;499(7457):214–8. <https://doi.org/10.1038/nature12213>.
37. Teerapabolarn K. Negative Binomial approximation to the Beta Binomial distribution. *Int J Pure Appl Math*. 2015;98(1):39–43. <https://doi.org/10.12732/ijpam.v98i1.5>.
38. Laursen R, Hobolth A. A sampling algorithm to compute the set of feasible solutions for non-negative matrix factorization with an arbitrary rank. *SIAM J Matrix Anal Appl*. 2022;43(1):257–73.
39. Gupta A, Datta S, Das S. Fast automatic estimation of the number of clusters from the minimum inter-center distance for k-means clustering. *Pattern Recogn Lett*. 2018;116:72–9. <https://doi.org/10.1016/J.PATREC.2018.09.003>.

4. References

40. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155(2):945–59.
41. Verity R, Nichols RA. Estimating the number of subpopulations (K) in structured populations. *Genetics*. 2016;203(4):1827–39. <https://doi.org/10.1534/genetics.115.180992>.
42. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Stat Comput*. 2013;24(6):997–1016. <https://doi.org/10.1007/S11222-013-9416-2>.
43. Luo Y, Al-Harbi K, Luo Y, Al-Harbi K. Performances of LOO and WAIC as IRT model selection methods. *Psychol Test Assess Model*. 2017;59(2):183–205.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

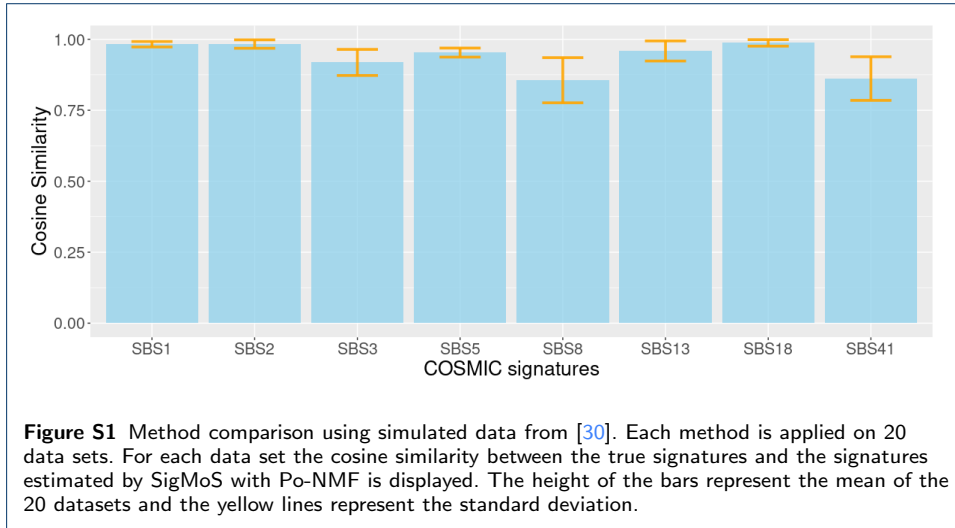
Learn more biomedcentral.com/submissions



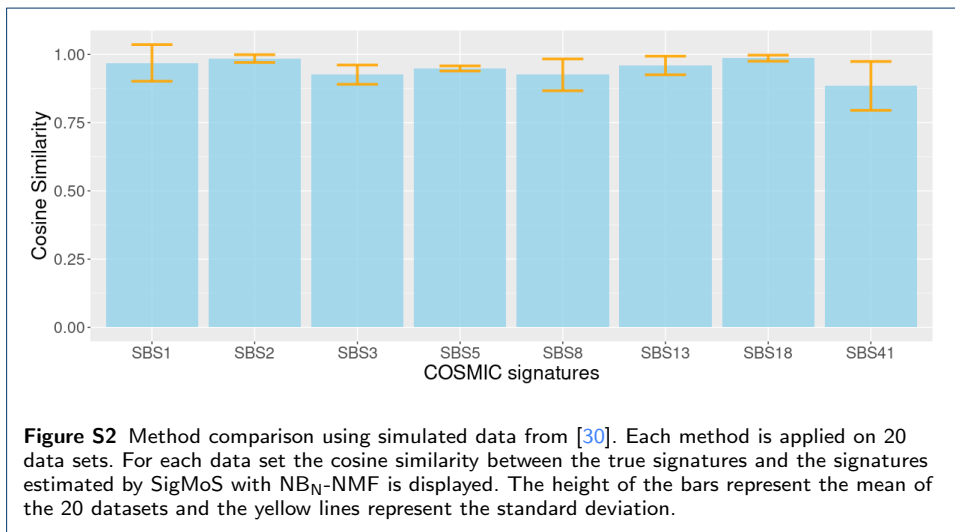
1 Supplementary material

2 S1 Results from simulated data in [30]

3 In this section we provide results from applying SigMoS to the data in [30]. Figures
4 S1 and S2 show the ability of SigMoS to recover the true mutational signatures
using Po-NMF and NB_N-NMF respectively.



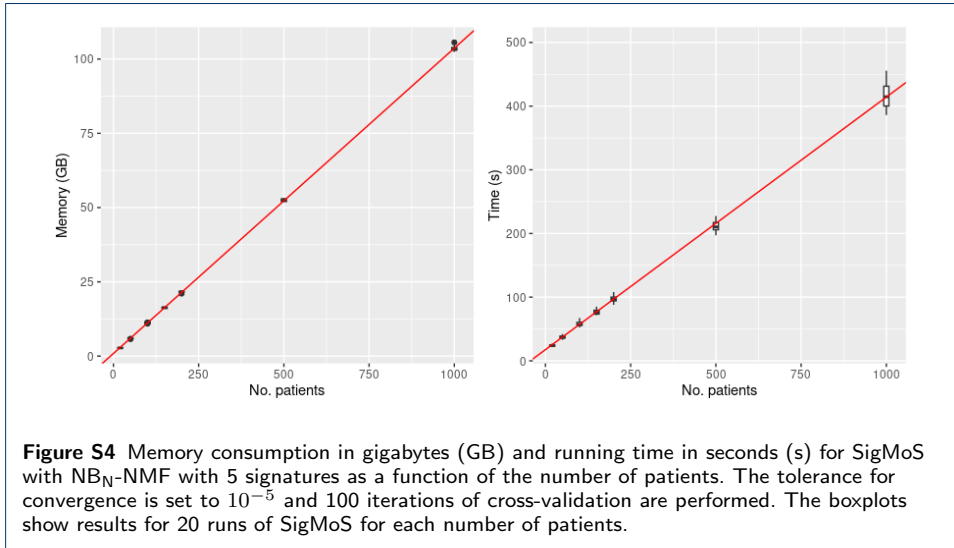
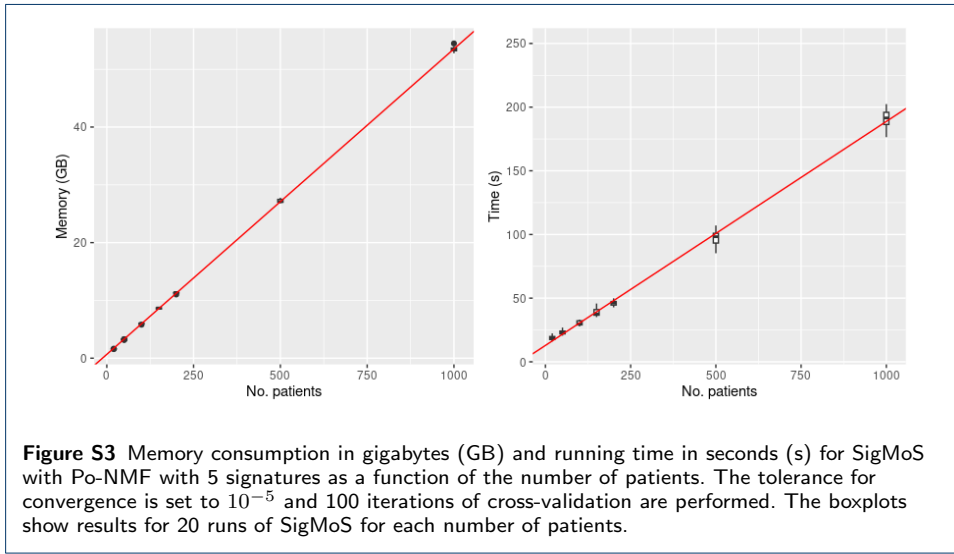
5



4. Supplementary material

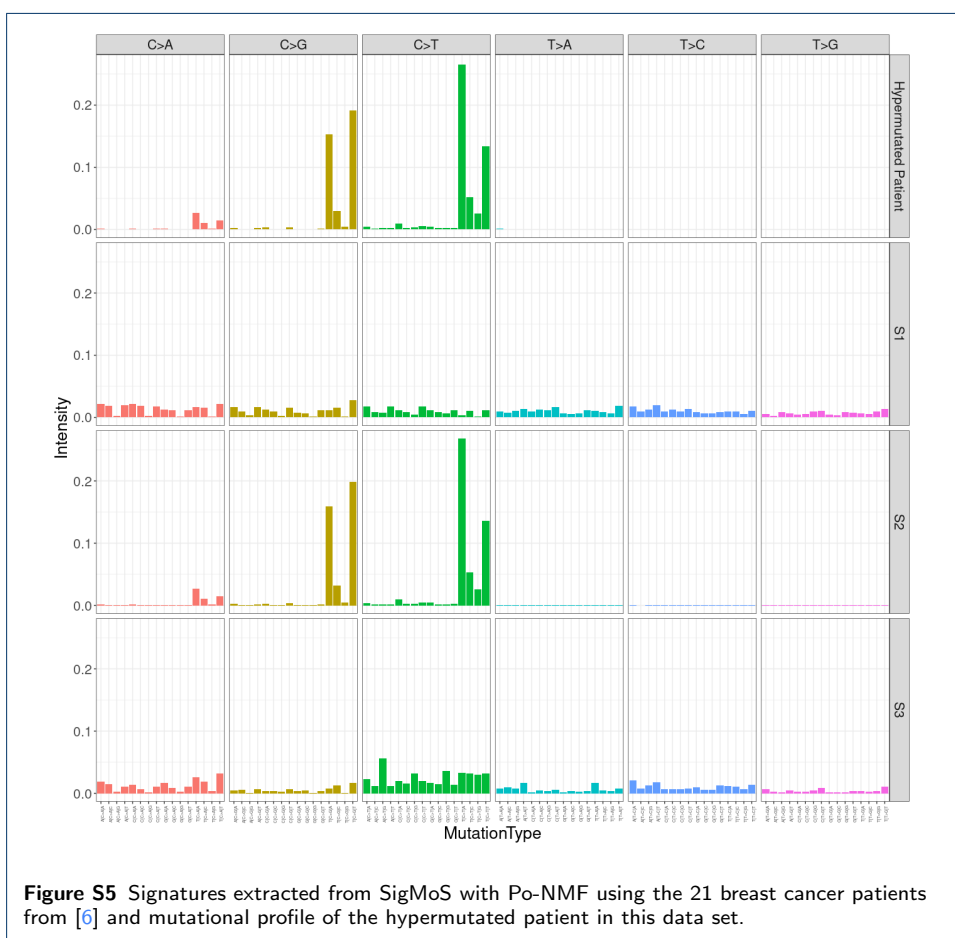
6 S2 Computational cost for SigMoS with Po-NMF and NB_N-NMF

7 In this section we present results on the memory consumption and running time
8 for SigMoS using Po-NMF and NB_N-NMF as a function of the number of patients.
9 We simulated data following the model in Section 2.1 and varying the number of
10 patients in $\{20, 50, 100, 150, 200, 500, 1000\}$. For each of these scenarios we run 20
11 simulations with 5 signatures. SigMoS is running in parallel for $J = 100$ iterations of
12 the model selection algorithm (see Figure 6). The memory consumption can thus be
13 decreased by a fraction a in exchange for an increase in time by the corresponding
14 multiplicative factor a . All analysis have been conducted on a standard laptop with
15 Intel Core i7 processor.



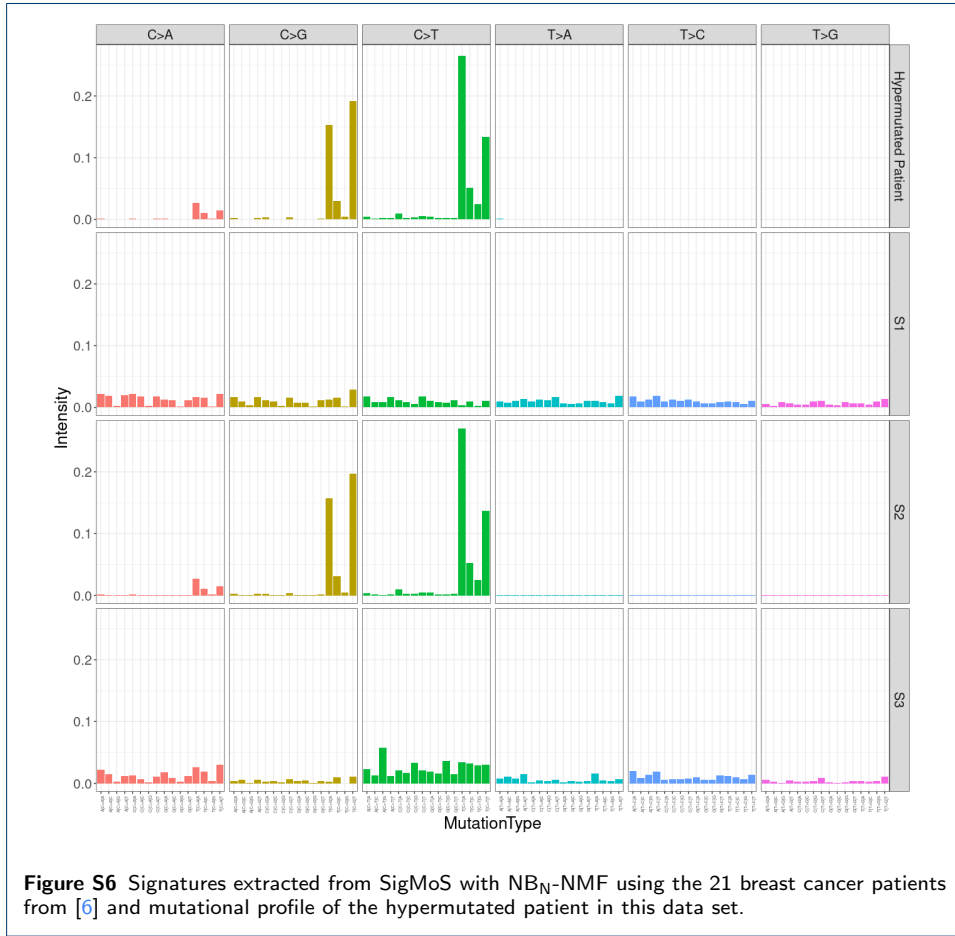
S3 Influence of hypermutated patients

We considered the real data set from [6] and applied both SigMoS with Po-NMF and with NB_N -NMF. First, we applied both methods to the 21 patients included in the data, and then we applied the same methods excluding the hypermutated patient (the last patient in the data). Figures S5 and S6 show the estimated mutational signatures from the Poisson and Negative Binomial models, respectively, and the normalized mutational profile of the hypermutated patient. The correlation between the profile of the hypermutated patients and the extracted signatures is (0.3002, 0.9999, 0.3203) with Po-NMF and (0.2135, 0.9999, 0.2013) under the Negative Binomial model.



When removing the hypermutated patient and running SigMoS again we obtain very similar signatures. The correlation between the hypermutated patient and the signatures indeed is (0.2453, 0.9894, 0.3095) under the Poisson model and (0.2755, 0.9903, 0.2906) with NB_N -NMF. This shows that the estimation of the signatures with SigMoS is not affected by the hypermutated patient and that our method is robust in this setting. It appears that the mutations of the hypermutated patient indeed originate from signature S2 that is shared among other patients as well. Thus our assumptions in the SigMoS method are not violated and the signatures can be correctly estimated also in this scenario.

4. Supplementary material



Figures S7 and S8 show the extracted exposures with Po-NMF and NB_N-NMF respectively. The top panels show results with all patients and the lower panels show results when the hypermutated patient is removed. This also emphasizes that the exposures to the reconstructed signatures are very similar (note the difference in the y-axis scale) in the two settings, suggesting a robust signature extraction.

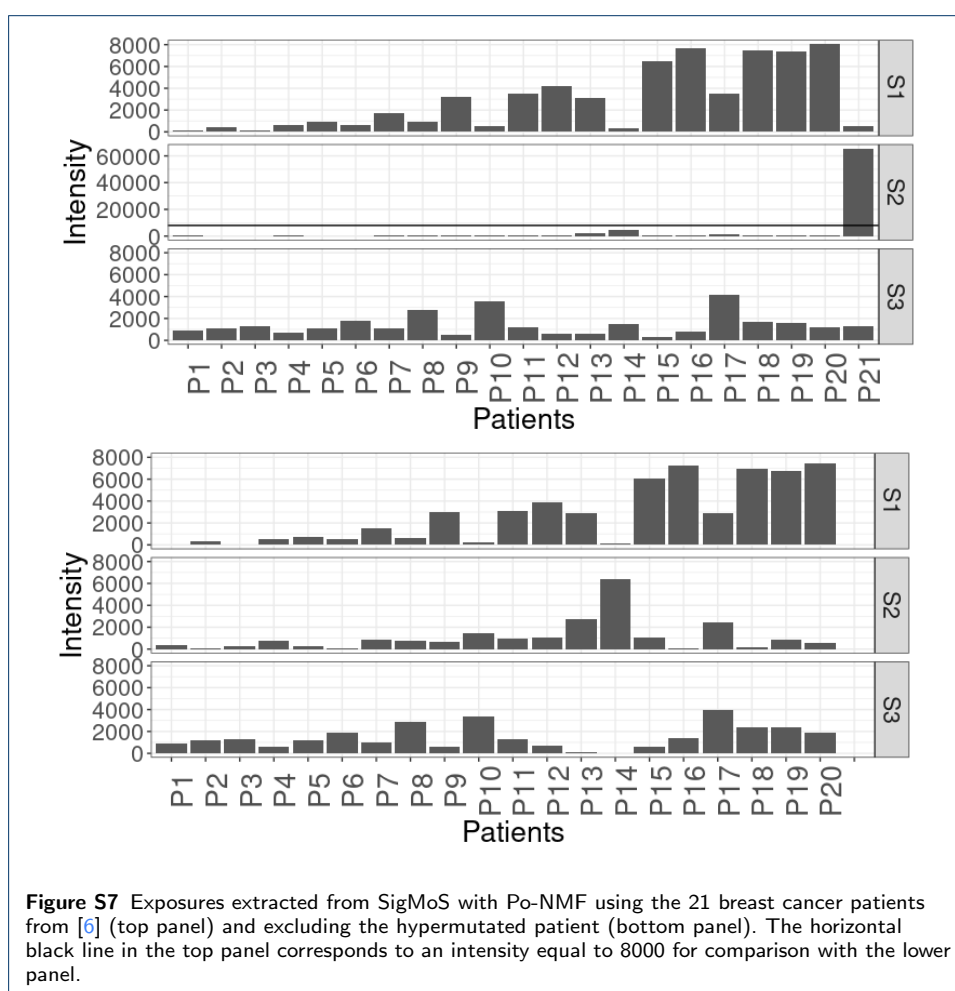
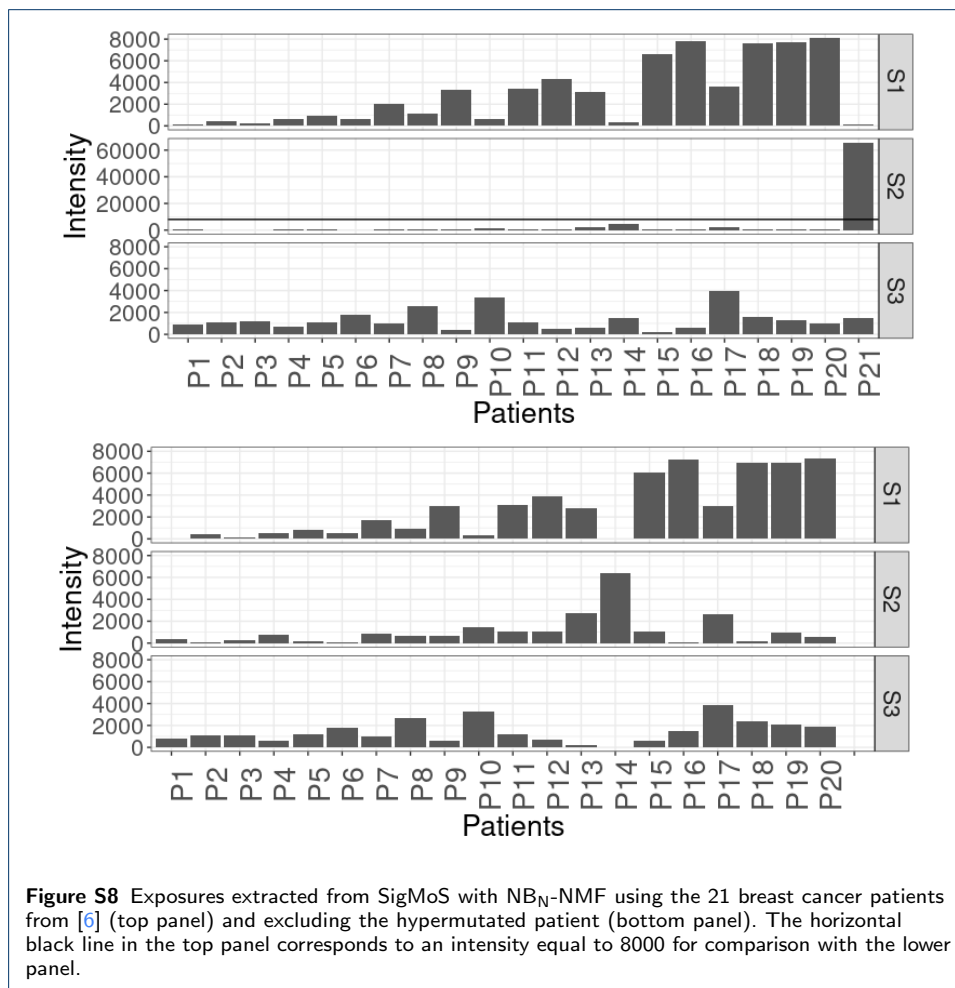


Figure S7 Exposures extracted from SigMoS with Po-NMF using the 21 breast cancer patients from [6] (top panel) and excluding the hypermutated patient (bottom panel). The horizontal black line in the top panel corresponds to an intensity equal to 8000 for comparison with the lower panel.

4. Supplementary material



40 S4 Influence of the choice of the cost function on SigMoS

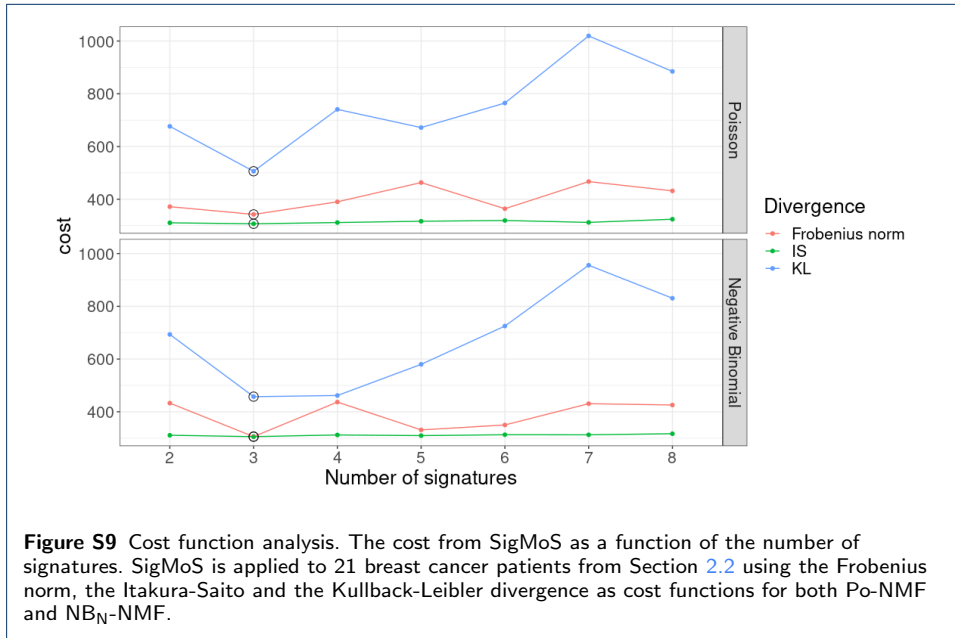
41 We note that the choice of the divergence measure for the cost function in Al-
 42 gorithm 2 is not trivial and may favor one or the other model. For example,
 43 in our framework, we may be biased towards the Poisson model when using the
 44 Kullback-Leibler (KL) divergence. Here, we test the influence of the choice of the
 45 cost functions in the SigMoS procedure. We consider the Itakura-Saito (IS) [32]
 46 divergence and the Frobenius norm as alternative cost functions for our proposed
 47 cross-validation algorithm. Both cost functions are also implemented in the SigMoS
 48 R package.

The KL divergence is given by equation (8) with $\alpha_1 = \dots = \alpha_N = 0$. Using the same notation as in Algorithm 2 with $B = V_{test}^j$ and $B^0 = H_{train}^j W_{test}^j$ the IS divergence can be expressed as:

$$d_{IS}(B, B^0) = \sum_{n=1}^N \sum_{m=1}^M \left\{ \frac{B_{nm}}{B_{nm}^0} - \log \frac{B_{nm}}{B_{nm}^0} \right\}.$$

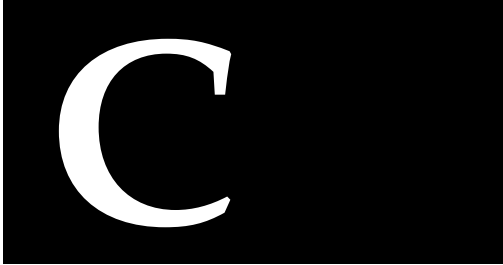
And the Frobenius norm is given by:

$$d_{FB}(B, B^0) = \sum_{n=1}^N \sum_{m=1}^M (B_{nm} - B_{nm}^0)^2.$$



49 We test our model selection procedure with the KL, IS and the Frobenius norm.
 50 We test the influence of the choice of the cost function on the same data described
 51 in Section 2.2 corresponding to the 21 breast cancer patients in [6]. The results
 52 of this analysis are shown in Figure S9 and demonstrate that our model selection
 53 framework is robust to different choices of the cost function. However, we observe
 54 that the minimum is more pronounced when using the KL divergence.
 55

Paper

A large, white, serif capital letter 'C' is centered within a solid black rectangular box. The box is positioned to the right of the word 'Paper'.

**Flexible model-based non-negative matrix factorization with
application to mutational signatures**

by Ragnhild Laursen, Lasse Maretty and Asger Hobolth

Published in Statistical Applications in Genetics and Molecular Biology

Ragnhild Laursen, Lasse Maretty and Asger Hobolth*

Flexible model-based non-negative matrix factorization with application to mutational signatures

<https://doi.org/10.1515/sagmb-2023-0034>

Received August 31, 2023; accepted April 3, 2024; published online May 16, 2024

Abstract: Somatic mutations in cancer can be viewed as a mixture distribution of several mutational signatures, which can be inferred using non-negative matrix factorization (NMF). Mutational signatures have previously been parametrized using either simple mono-nucleotide interaction models or general tri-nucleotide interaction models. We describe a flexible and novel framework for identifying biologically plausible parametrizations of mutational signatures, and in particular for estimating di-nucleotide interaction models. Our novel estimation procedure is based on the expectation–maximization (EM) algorithm and regression in the log-linear quasi–Poisson model. We show that di-nucleotide interaction signatures are statistically stable and sufficiently complex to fit the mutational patterns. Di-nucleotide interaction signatures often strike the right balance between appropriately fitting the data and avoiding over-fitting. They provide a better fit to data and are biologically more plausible than mono-nucleotide interaction signatures, and the parametrization is more stable than the parameter-rich tri-nucleotide interaction signatures. We illustrate our framework in a large simulation study where we compare to state of the art methods, and show results for three data sets of somatic mutation counts from patients with cancer in the breast, Liver and urinary tract.

Keywords: cancer genomics; expectation-maximization (EM) algorithm; interaction terms; mutational signatures; non-negative matrix factorization (NMF); Poisson regression

JEL Classification: Primary: 62; Secondary: 62F10; 62F30; 62H12; 62P10; 68T05; 92B20

1 Introduction

The mutation rate at a particular site in the genome often depends on both the left and right flanking nucleotides. Hwang and Green (2004) analysed a 1.7 mega-base alignment of 19 mammalian species, and perhaps the most striking observation was a much elevated mutation rate for $C > T$ mutations when the right flanking nucleotide is a G . The elevated rate reflects deamination of methyl cytosine. The CG -methylation-deamination process was the main focus in the neighbour-dependent models described in Arndt et al. (2003) and Hobolth (2008). Furthermore, longer contextual patterns have recently been shown to impact the mutation rates induced by ultraviolet light (Lindberg et al. 2019).

Analyses of somatic mutations in cancer patients have increased our basic understanding of the mutational processes operating in human cancer (Alexandrov et al. 2020). For example, mutational signatures from tobacco

*Corresponding author: **Asger Hobolth**, Department of Mathematics, Aarhus University, Aarhus, Denmark, E-mail: asger@math.au.dk
Ragnhild Laursen, Department of Mathematics, Aarhus University, Aarhus, Denmark, E-mail: ragnhild@math.au.dk
Lasse Maretty, Department of Clinical Medicine and Bioinformatics Research Center, Aarhus University, Aarhus, Denmark, E-mail: lasse.maretty@clin.au.dk

1. Introduction

smoking (Alexandrov et al. 2016) and UV-light (e.g. Shen et al. 2020) have been identified. Furthermore, mutational signatures can be used as biomarkers for drug sensitivity (Levatić et al. 2022) and deciding the diagnosis and treatment of cancer patients (Nik-Zainal and Morganella 2017). A simple parametrization of mutational signatures is essential to achieve statistically stable estimation, easier interpretation of signatures, and the possibility of including more flanking nucleotides than just the nearest neighbors.

Our method is a flexible framework for parametrizing mutational signatures by biologically plausible interaction terms. The framework makes it possible to greatly reduce the number of parameters while still maintaining a good fit to the data. The mutational signatures from Alexandrov et al. (2013) and Shiraishi et al. (2015) constitute two extremes in our framework. We view signatures as a composition of interactions between the mutation type M and the left and right flanking nucleotides L and R as shown in Figure 1.

In this context, the general model from Alexandrov et al. (2013) with 96 mutation types includes all tri-nucleotide interaction terms, and the independence model from Shiraishi et al. (2015) has no interaction terms between the mutation and the flanking nucleotides i.e. mono-nucleotide interaction terms. Using classical factor analysis notation we can write the general model as $L \times M \times R$ and the mono-nucleotide model as $L + M + R$. We propose a model that reaches the middle-ground between the complex model of Alexandrov et al. (2013) and the simple model of Shiraishi et al. (2015). Our model includes di-nucleotide interaction terms between the mutation type and flanking nucleotides and can be written $L \times M + M \times R$. We also investigate combinations of the parametrizations for mutational signatures. Our novel and flexible estimating procedure is based on the EM-algorithm and a quasi-Poisson log-linear model for optimizing the free parameters.

In a simulation study with changing number of signatures and patients we show that the di-nucleotide model strikes a good balance between maintaining a good fit to the data and reconstruction of the underlying true signatures. We also compare our framework to state of the art methods for 96 mutation types with one flanking nucleotide as well as 1536 mutation types with two flanking nucleotides. We find that the di-nucleotide model reconstructs the true signatures very well, and compares favorable to three other methods for mutational signature extraction; signeR (Rosales et al. 2017), SparseSignatures (Lal et al. 2021a) and sigfit (Gori and Baez-Ortega 2018).

We also analyze three data sets of somatic mutations in cancer patients. The first data set is from breast cancer patients with 96 mutation types. We analyze the 214 breast cancer patients from Alexandrov et al. (2020), and we refer to this data set as BRCA. We show that many of the recovered signatures can be parametrized by the simpler di-nucleotide or even mono-nucleotide parametrization. In a bootstrap and downsampling experiment we also show how simpler parametrizations give a better reconstruction of both the exposures and the signatures.

The second data set is from 260 Liver cancer patients with 96 mutation types from Alexandrov et al. (2020). For this data set we again see that many of the recovered signatures can be explained by much simpler parametrizations. The signatures found for the di-nucleotide model is also more similar to the COSMIC signatures identified for Liver cancer in Alexandrov et al. (2020) compared to the mono- and tri-nucleotide models.

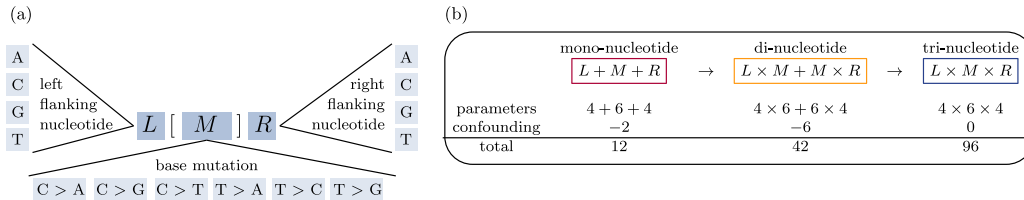


Figure 1: Graphical illustration of the parametrization of the mutation types. (a) The natural features for the mutation types are the left nucleotide L , right nucleotide R , and the base mutation M . (b) The three parametrizations we are analyzing in this paper for mutation types with one flanking nucleotide at each side.

The third data set is from urothelial carcinoma of the upper urinary tract (Hoang et al. 2013) from 26 patients with 1536 mutation types. These mutation types include two flanking nucleotides to each side of the base mutation. This data was also analysed by Shiraishi et al. (2015), and we refer to the data as UCUT. We find that the di-nucleotide interaction models fit the data substantially better than the mono-nucleotide models and are statistically much more stable than the full penta-nucleotide model.

In general, our analyses validate the relevance of our flexible framework for mutational signatures. The di-nucleotide signatures provide a better fit to the data and are biologically more plausible than mono-nucleotide signatures, and the parametrization is more stable than the parameter-rich higher-order signatures that include all interaction terms.

Our paper is organized as follows. In Section 2 we describe non-negative matrix factorization and parametrization of a mutational signature in terms of interactions between the nucleotides in the mutation type. Section 3 includes a simulation study and analyzes the BRCA, Liver and UCUT data sets. Maximum likelihood estimation is carried out using a novel combination of the expectation-maximization algorithm (Dempster et al. 1977) and regression in the quasi-Poisson model (e.g. McCullagh and Nelder 1989), and is described in detail in Section 4. The paper ends with a general discussion about parametrization and model selection for mutational signatures (Section 5). The data and code for reproducing the results and figures are available at https://github.com/ragnhildlaursen/paramNMF_ms.

2 Determining the mutational signatures

Mutational signatures are derived from mutational counts using an unsupervised method called non-negative matrix factorization (NMF). In this section we first explain NMF in general terms and afterwards how parametrization of the mutational signatures is included in the framework.

2.1 Non-negative matrix factorization

Given a data matrix $V \in \mathbb{N}_+^{N \times T}$, the main aim of non-negative matrix factorization (NMF) is to find a factorization WH , where the product of the non-negative exposure (sometimes also called weight or loading) matrix $W \in \mathbb{R}_+^{N \times K}$ and the non-negative signature matrix $H \in \mathbb{R}_+^{K \times T}$ provide a good approximation of the data matrix, i.e.

$$V \approx WH. \quad (1)$$

In our application N is the number of cancer patients, T is the number of mutation types, and each entry V_{nt} is the total number of somatic cancer mutations of type t in patient n . The non-negative weight matrix W is of size $N \times K$, and the non-negative mutational signature matrix H is of size $K \times T$. Each of the K signatures is a discrete probability distribution of length T , i.e. has $T - 1$ free non-negative parameters that sum to at most one. The rank K of the factorization is most often one or more magnitudes smaller than the minimum of N and T . For the BRCA data set, for example, we have the number of signatures K around 6–10, number of patients $N = 214$, and number of mutation types $T = 96$.

In general, the number of observations is $N \times T$ and the number of free parameters is $N \times K$ for the weight matrix and $K \times (T - 1)$ for the signature matrix. With $N = 214$ patients and $K = 8$ signatures the number of observations $N \times T = 214 \times 96 = 20,544$ are estimated using $N \times K + K \times (T - 1) = 214 \times 8 + 8 \times 95 = 1712 + 760 = 2472$ free parameters. Thus, in general, this approach has a large number of free parameters compared to the size of the data matrix. These considerations suggest that parametrizing a mutational signature is fruitful.

2.2 Parametrization of a mutational signature

We parametrize each mutational signature $h = (h_1, \dots, h_T)$ by the mutation type as a function of the base mutation M , the flanking left base L and the flanking right base R as shown in Figure 1(a). The number of mutations is 12 without strand-symmetry, and 6 with strand-symmetry. Each flanking nucleotide can be one of the four types

2. Determining the mutational signatures

A, C, G or T. The different factors are thus the left neighbour L (4 categories), the right neighbour R (4 categories), and the mutation type M (6 or 12 categories). In all of the following we assume strand-symmetry, so that M has 6 categories.

We model the mutational signatures with a log-linear parametrization given by

$$h_t = \frac{\exp((X\beta)_t)}{\sum_{t=1}^T \exp((X\beta)_t)}, \quad t = 1, \dots, T, \quad (2)$$

where X has dimension $T \times S$ and is the design matrix that describe the common factors among the different mutation types and $\beta \in \mathbb{R}^S$ is a vector of S parameters for the different factors. This framework therefore makes it possible to choose any type of parametrization for the signatures through the designmatrix X . In Section 2.2.1 we consider parametrizations for 96 mutation types (one flanking nucleotide at each side of the mutation). We consider the general tri-nucleotide interaction model $L \times M \times R$, the simple mono-nucleotide model $L + M + R$ and the di-nucleotide interaction model $L \times M + M \times R$. In Section 2.2.2 we consider parametrizations for 1536 mutation types (two flanking nucleotides at each side of the mutation). We consider the general penta-nucleotide interaction model $L_2 \times L_1 \times R \times R_1 \times R_2$, the simple mono-nucleotide model $L_2 + L_1 + M + R_1 + R_2$, and a suite of models in-between such as the full di-nucleotide interaction model $L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$. We explain in detail the parametrizations and corresponding design matrix in the next two subsections.

2.2.1 One flanking nucleotide at each side of the mutation

A summary of the three parametrizations for mutational signatures with 96 mutation types is provided in Figure 1(b). We consider parametrizations with no interaction between nucleotides (mono-nucleotide signatures), interaction between neighboring nucleotides (di-nucleotide signatures) and general interaction (tri-nucleotide signatures).

The mutational signature h with one flanking nucleotide at each side is a vector of length $T = 4 \times 6 \times 4 = 96$ indexed by ℓmr . Following classical factorial analysis of variance we specify the general tri-nucleotide interaction model from Alexandrov et al. (2013) by $L \times M \times R$. The model can be written as

$$h_{\ell mr} = \frac{\exp(\beta_{\ell mr}^{L \times M \times R})}{\sum_{\ell \in L} \sum_{m \in M} \sum_{r \in R} \exp(\beta_{\ell mr}^{L \times M \times R})}, \quad (3)$$

where m describes the six base mutation, and ℓ and r describe the four possible flanking nucleotides to the left or right of the base mutation. This gives $S = T = 4 \times 6 \times 4 = 96$ different parameters in the β vector and $X = I_T$ is the $T \times T$ identity matrix in the general formulation (2).

The mono-nucleotide interaction model $L + M + R$ of Shiraishi et al. (2015) takes the form

$$h_{\ell mr} = \frac{\exp(\beta_m^M + \beta_\ell^L + \beta_r^R)}{\sum_{\ell \in L} \sum_{m \in M} \sum_{r \in R} \exp(\beta_m^M + \beta_\ell^L + \beta_r^R)}. \quad (4)$$

In order to avoid confounding we define $\beta_A^R = \beta_A^L = 0$. Therefore, we have $S = 6 + 4 + 4 - 2 = 12$ remaining parameters in the β vector, which is a substantial reduction from the original model with 96 parameters. The corresponding 96×12 design matrix X takes the form

$$\begin{aligned}
& \begin{array}{c} \text{Mutation} \\ \begin{array}{cccccc} C > A & C > G & C > T & T > A & T > C & T > G \end{array} \end{array} \quad \begin{array}{c} \text{Left base} \\ \begin{array}{ccc} C & G & T \end{array} \end{array} \quad \begin{array}{c} \text{Right base} \\ \begin{array}{ccc} C & G & T \end{array} \end{array} \\
X = \begin{array}{c} A[C > A]A \\ A[C > A]C \\ A[C > A]G \\ \vdots \\ T[T > G]T \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{array}{c} \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vdots & & \\ 0 & 0 & 1 \end{array} \\ \begin{array}{ccc} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & & \\ 0 & 0 & 1 \end{array} \end{array} \quad (5)
\end{aligned}$$

We propose the di-nucleotide interaction signature $L \times M + M \times R$ given by

$$h_{\ell mr} = \frac{\exp(\beta_m^M + \beta_{\ell m}^{L \times M} + \beta_{mr}^{M \times R})}{\sum_{\ell \in L} \sum_{m \in M} \sum_{r \in R} \exp(\beta_m^M + \beta_{\ell m}^{L \times M} + \beta_{mr}^{M \times R})}. \quad (6)$$

In order to avoid confounding we define $\beta_{Am}^{L \times M} = \beta_{mA}^{M \times R} = 0$ for all the six possible base mutations $m \in \{C > A, C > G, C > T, T > A, T > C, T > G\}$. This signature therefore has a total of $S = 4 \times 6 + 4 \times 6 - 6 = 42$ parameters and is a biologically plausible alternative between the simple mono-nucleotide multiplicative signature of Shiraishi et al. (2015) and the complex tri-nucleotide interaction signature of Alexandrov et al. (2013). From the mutational pattern of spontaneous cytosine deamination in CpG contexts, we know that some processes are dependent on only one neighbouring nucleotide (Arndt et al. 2003). Results for the models with one flanking nucleotide at each side of the mutation are shown for the breast and Liver cancer patients in Section 3.2 and 3.3, respectively.

2.2.2 Two flanking nucleotides at each side of the mutation

In Table 1 and Figure 2 we give an overview of the factorizations with two flanking nucleotides at each side and how they are nested in each other.

Shiraishi et al. (2015) considers higher-order context dependencies where the mutation types include four flanking bases, which gives five different factors L_2, L_1, M, R_1 and R_2 . The number of mutation types in this case is $T = 4^2 \times 6 \times 4^2 = 6 \times 4^4 = 1536$ and the number of parameters in the mono-nucleotide model with two flanking neighbours on each side of the mutation is $3 + 3 + 6 + 3 + 3 = 6 + 3 \times (2 \times 2) = 18$.

Table 1: Parametrizations of a mutational signature with two flanking nucleotides at each side. We consider two categories of di-nucleotide interaction models. The first category has interaction between the flanking nucleotide and the mutation. The second category has interaction between the two nearest neighbours.

Signature	Factorization	Number of parameters
Mono-nucleotide	$L_2 + L_1 + M + R_1 + R_2$	$6 + 3 \times 4 = 18$
Di-nucleotide	$L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$	$42 + 12 \times 2 = 66$
Tri-nucleotide	$L_1 \times M \times R_1$	$6 \times 4^2 = 96$
Penta-nucleotide	$L_2 \times L_1 \times M \times R_1 \times R_2$	$6 \times 4^4 = 1536$
Di- and mono-nucleotide	$L_2 + L_1 \times M + M \times R_1 + R_2$	$42 + 3 \times 2 = 48$
Tri- and mono-nucleotide	$L_2 + L_1 \times M \times R_1 + R_2$	$96 + 3 \times 2 = 102$
Tri- and di-nucleotide	$L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$	$96 + 12 \times 2 = 120$

3. Results

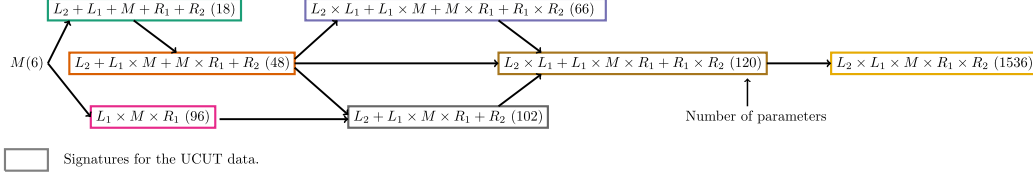


Figure 2: Factor diagram for the signatures used for the UCUT data set. The diagram shows the number of parameters for each signature and how the signatures are nested in each other.

Our framework is very flexible, and we are able to analyse combinations of mono-, di- and tri-nucleotide interaction terms within a signature. For example, we consider the signatures $L_2 + L_1 \times M + M \times R_1 + R_2$, $L_2 + L_1 \times M \times R_1 + R_2$, and $L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$. These three signatures are combinations of mono-, di- and tri-nucleotide interactions. Results for applying these models to the UCUT data are provided in Section 3.4.

3 Results

This section includes a simulation study to compare the different parametrizations and afterwards an analysis of three real data sets. In the simulation study we vary both the number of signatures and the number of patients. For the real data sets we analyze two of the largest PCAWG tumor data sets: the BRCA data set and the Liver cancer data set. We compare the retrieved signatures with the identified COSMIC signatures from Alexandrov et al. (2020). The third real data set includes two flanking nucleotides in the mutation type and is the same data analyzed in Shiraishi et al. (2015). We determine the optimal number of signatures, compare and evaluate the various parametrizations, and use parametric bootstrap and downsampling to investigate statistical robustness and stability of the signatures.

The most appropriate statistical model can be determined by several methods that are balancing between a good fit to the data and avoiding over-fitting, and the choice depends on the application of the model (e.g. Shmueli 2010). In this paper we use the Bayesian Information Criterion (BIC) given by

$$\text{BIC} = n_{\text{prm}} \log n_{\text{obs}} - 2\ell(W, H; V) \equiv n_{\text{prm}} \log n_{\text{obs}} + 2\text{GKL},$$

where n_{prm} is the number of parameters, n_{obs} is the number of observations, $\ell(W, H; V)$ is the log-likelihood function from (8), GKL is the generalized Kullback–Leibler divergence from (9), and \equiv means that the statement is true up to an additive constant. Appropriate models have a small BIC because they represent a good balance between model complexity (measured in terms of the number of parameters) and goodness of fit (measured in terms of the negative log-likelihood).

3.1 Simulation study

In this simulation study we are simulating signatures having the di-nucleotide parametrization. The exposure for the different signatures are simulated using a negative binomial model with mean 1000 and dispersion parameter 1.5 following Lal et al. (2021a). The data sets are then constructed as the matrix product of the exposures and the signatures. At last Poisson noise has been added to all the data sets. In Figure 3 we are both changing the number of signatures and the number of patients included in the dataset. We observe that if the true mutational signatures are di-nucleotide interaction signatures, then the di-nucleotide model is always superior to the simple mono-nucleotide or general tri-nucleotide model for any number of signatures or patients. Additionally we observe that the di-nucleotide model maintain a good fit to data even though the number of parameters is greatly reduced.

In Figure 4 we compare our method to other state of the art methods that has also been implemented in R. This includes `signer` (Rosales et al. 2017), `SparseSignatures` (Lal et al. 2021a) and `sigfit` (Gori and

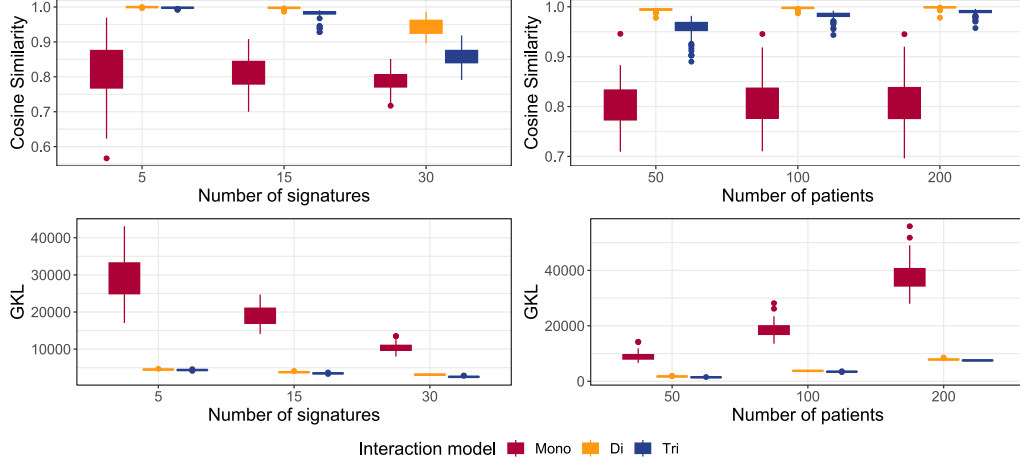


Figure 3: Simulating di-nucleotide signatures creating 100 different data sets for different number of patients and signatures. The figure both shows the reconstruction of the signatures through the average cosine similarity and the fit to data through the Generalized Kullback–Leibler divergence (GKL). The number of patients is fixed to 100, when the number of signatures varies (*left*) and the number of signatures is fixed to 15, when the number of patients varies (*right*).

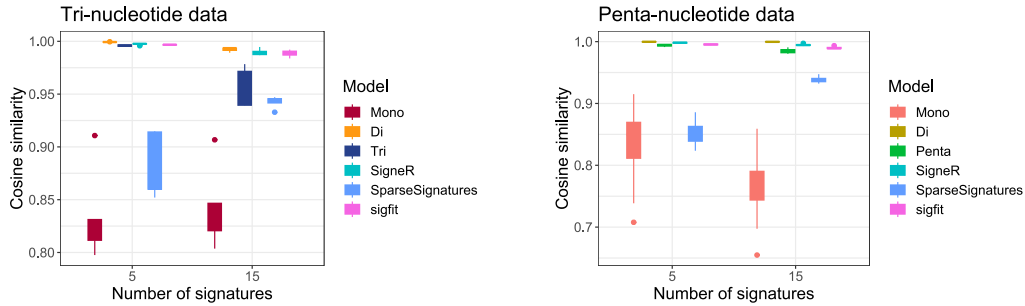


Figure 4: Comparing different methods for 10 datasets of 100 patients for 5 and 15 signatures. The methods SigneR, SparseSignatures and sigfit are run with their default implementations. The two figures show the results for tri-nucleotide mutation types with only one flanking nucleotide (*left*) and the results for the penta-nucleotide mutation types with two flanking nucleotides (*right*).

Baez-Ortega 2018). We compare these methods with the three models from our framework; the mono- and di-nucleotide parametrization and the regular NMF with no parametrization. The regular NMF is called tri-nucleotide when the mutation types has one flanking nucleotide and penta-nucleotide when the mutation type has two flanking nucleotides. We have only conducted this for 10 datasets with 5 or 15 signatures as many of the methods are very time consuming. Again we can clearly see that when the true mutational signatures are di-nucleotide signatures the di-nucleotide model has the best performance among all the methods.

3.2 Analysis of BRCA

Recall that the breast cancer data set has $T = 96$ mutation types and $N = 214$. The number of observations for the data set is $n_{\text{obs}} = T \times N = 96 \times 214 = 20,544$.

3. Results

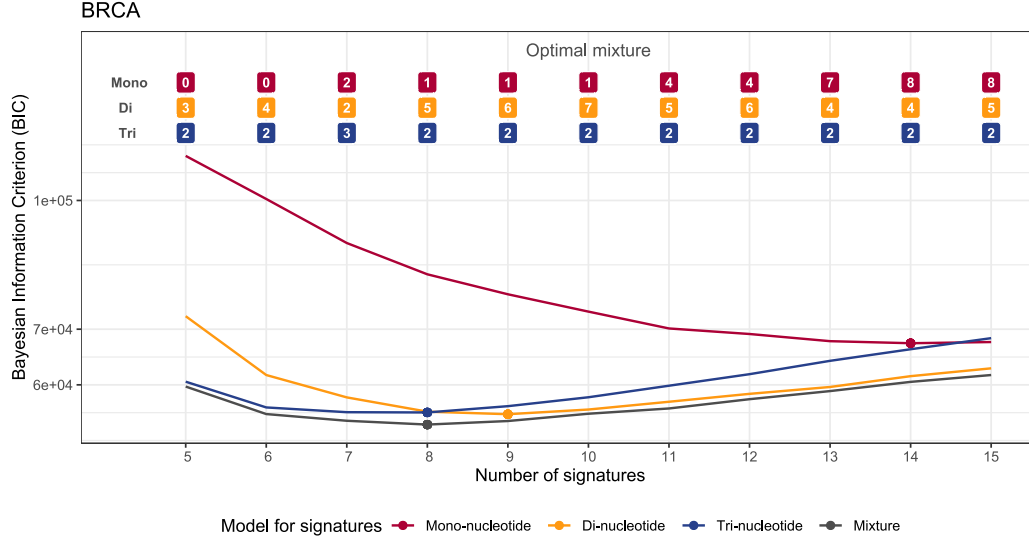


Figure 5: The Bayesian Information Criterion (BIC) for different number of signatures K for the BRCA dataset. The BIC is minimized for $K = 14$, $K = 9$ and $K = 8$ when all signatures are either mono-, di- or tri-nucleotide (dark red, yellow and blue curves). The BIC is minimized for $K = 8$ when the parametrization of signatures is free (dark curve). The top shows the optimal mixture of signature parametrizations for each number of signatures K . For example, the optimal mixture for $K = 8$ signatures consists of 1 mono-nucleotide, 5 di-nucleotide and 2 tri-nucleotide signatures.

3.2.1 Choosing the number of signatures and parametrization

In Figure 5 we plot the BIC for different types of parameterizations. We plot the BIC for models where all signatures are either mono-, di- or tri-nucleotide parameterizations, but also the optimal mixture, where each signature can be any of the three parameterizations from Figure 1(b). The mono-nucleotide model has an optimal number of signatures at $K = 14$, which is much higher than the $K = 8$ signatures that are optimal for both the mixture model and the exclusive tri-nucleotide model. The optimal number of signatures is $K = 9$ when all signatures are of the di-nucleotide type. Even though there are much fewer parameters in the mixture model compared to the exclusive tri-nucleotide model, the optimal number of signatures is identical. In the analysis of the signatures we therefore choose to fix the number of signatures at $K = 8$.

We allow a flexible parametrization of type $L \times M \times R$, $L \times M + M \times R$, and $L + M + R$ for each of the $K = 8$ signatures. We could investigate $3^8 = 6561$ models, but the models are only identifiable up to permutation (see the beginning of Section 4); this results in 45 different models. For the 45 models, Figure 6 shows the Generalized Kullback–Leibler divergence (GKL) and the Bayesian Information Criterion (BIC). The models are ordered according to the number of free parameters. The EM-algorithm can get stuck in local maxima of the likelihood function, so we start the algorithm by running 100 different initializations for 500 iterations and identify the maximum. From that maximum we then continue iterating until convergence. This procedure of starting the algorithm multiple times and running for a few iterations is recommended by Biernacki et al. (2003) who tested many different ways of running the EM-algorithm to escape local maxima and identify the global maximum likelihood value.

We observe a steep decrease in GKL when the mono-nucleotide assumption is relaxed, and one or more signatures are allowed to contain di-nucleotide or even tri-nucleotide interactions. This indicates that only applying mono-nucleotide signatures is biologically too restrictive. The mixture model with the smallest BIC (Mix in Figure 6) has one mono-nucleotide signature, five di-nucleotide signatures and two tri-nucleotide interaction

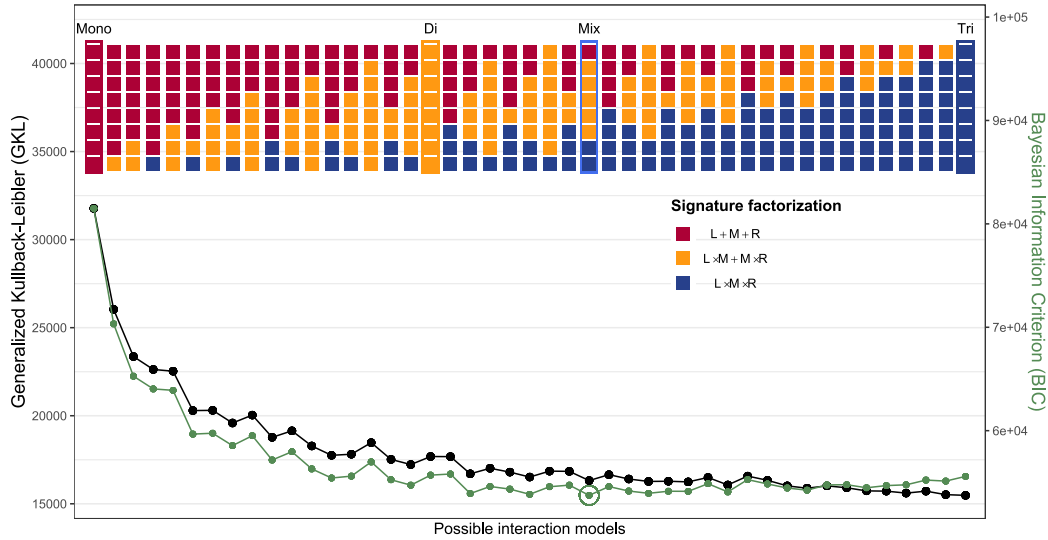


Figure 6: Fit to mutational count data from 214 breast cancer patients for all possible interaction models. The generalized Kullback–Leibler (GKL) and Bayesian Information Criterion (BIC) for all 45 models with $K = 8$ signatures. The models are ordered according to the total number of parameters for the 8 signatures; e.g. $8 \times 12 = 96$ for the sole mono-nucleotide model and $8 \times 96 = 768$ for the sole tri-nucleotide model. The model with the smallest BIC is indicated, and consists of two tri-nucleotide signatures, five di-nucleotide signatures and one mono-nucleotide signature.

signature. The fit to the data is too poor for the independent model, and the general model has too many free parameters. This is even more evident when we look at the robustness of the signatures; this is the topic for the next section.

3.2.2 Model validation and stability of signatures

In Figure 7, we show the eight signatures for the four different models marked in Figure 6. Each row corresponds to a model, and the signatures are matched for comparison. For the mixture model the parametrization is ordered according to Figure 6, which means signature 1 has a mono-nucleotide parametrization, signature 2 to 6 have a di-nucleotide parametrization and the last two have a tri-nucleotide parametrization. We observe that the signatures are very similar across the mixture, di- and tri-nucleotide models, whereas the mono-nucleotide model differs more from the others.

We validated the inferred signatures by matching to the signatures from version 3 of the Catalogue Of Somatic Mutations In Cancer (COSMIC) database (<https://cancer.sanger.ac.uk/cosmic>) with the highest cosine similarity. Notice that signature 4 is matched with SBS39 for the mono- and di-nucleotide parametrization and with SBS3 for the mixture and tri-nucleotide parametrization. All the models have a cosine similarity above 0.8 to the COSMIC signatures except the mono-nucleotide model for signature 1, 5 and 8. All of the COSMIC signatures we have matched is equivalent to the ones recovered for the same breast cancer data in Alexandrov et al. (2020). This includes all the six signatures (SBS1, SBS2, SBS3, SBS5, SBS13 and SBS18) that was included in more than half of the tumors.

This indicates that many of the COSMIC signatures can be parametrized by a much simpler di-nucleotide parametrization and a few can even be explained by mono-nucleotide parametrization. The ten most important interactions for these eight COSMIC signatures are shown in Figure 8. The top interactions are found with forward selection, where we include the interaction making the largest increase in the cosine similarity to the underlying true signature. The coefficient for each interaction is determined as the average over all the

Inferred signatures for BRCA

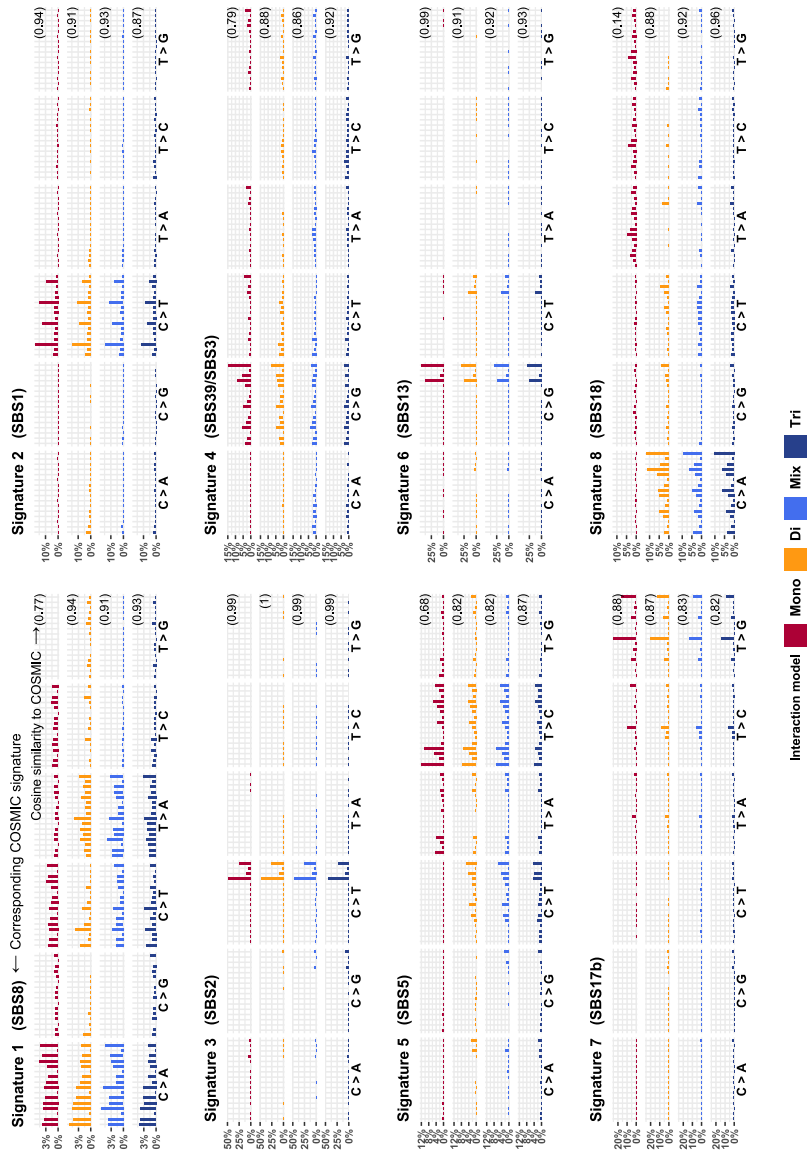


Figure 7: Inferred signatures for the BRCA data set. Comparison of the eight signatures for the four highlighted models in Figure 6. The four models are three parametrizations where all eight signatures are mono-nucleotide, di-nucleotide or tri-nucleotide, and for the mixture model the parametrization is ordered in the following way: one mono-nucleotide, five di-nucleotide and at last two tri-nucleotide parametrizations.

3. Results

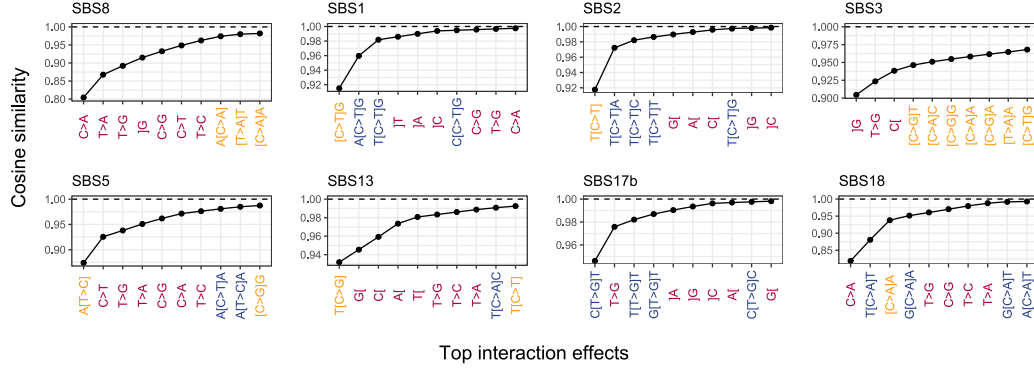


Figure 8: The top interactions for the eight COSMIC signatures found for the BRCA dataset. The top interactions are found with forward selection from the interaction making the largest increase in the cosine similarity to the COSMIC signature.

mutation types including that specific interaction. The figure again supports that many of the most important interactions are mono- or di-nucleotide interactions. This figure also supports the results for the mixture model, where SBS8 is parametrized with the mono-nucleotide model as the top seven interactions are from the mono-nucleotide model. Similarly the mixture model parametrized SBS17b and SBS18 with the tri-nucleotide model, which is shown by the many top tri-nucleotide interactions. The rest of the signatures were parametrized by the di-nucleotide model, which are mostly driven by one or two important di-nucleotide interactions.

In order to investigate the statistical stability of the signatures we use parametric bootstrap. For a given model with an estimate of the count matrix $\hat{W}\hat{H}$ we simulate 50 data sets from the Poisson model (7). For each of the simulated data sets we re-estimate the exposures and signatures and use cosine similarity to investigate how close the re-estimated signatures are to the true signatures under the specific model. In Figure 9 we show the cosine similarity for reconstructing the signatures from the parametric bootstrap procedure.

The mono-nucleotide model has very stable signatures as the cosine similarity is consistently high, but the signatures are also rather different from the signatures in the other models, and they are giving a substantially worse fit to the data. In contrast, the exclusive tri-nucleotide model generally provides a good fit to the data, but due to the many parameters in the model, the bootstrap variability is generally higher than for the other

Parametric bootstrap for BRCA

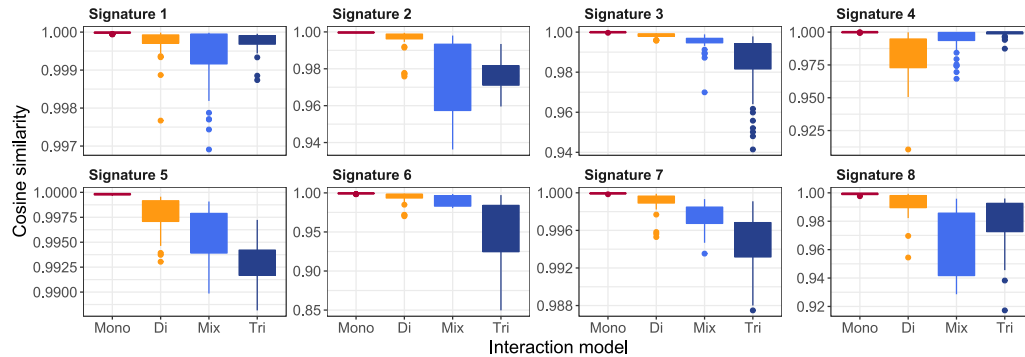


Figure 9: The cosine similarity for reconstructing the signatures with parametric bootstrapping for the BRCA data.

3. Results

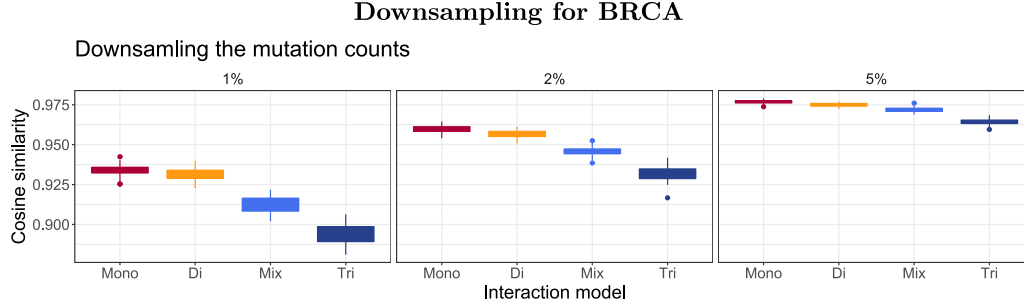


Figure 10: The mean cosine similarity between the recovered exposures from down-sampled BRCA data compared to the exposures from the original BRCA data.

models. Our new exclusive di-nucleotide model and mixture model reach the middle ground between these two extremes. The mixture model shows more bootstrap variability than the di-nucleotide model, but the mixture model also gives a better fit to the data.

Finally, we use down-sampling to investigate the stability of the exposures for the different parametrizations of the signatures. We again compare the four different models Mono (8 mono-nucleotide interaction signatures), Di (8 di-nucleotide interaction signatures), Tri (8 tri-nucleotide interaction signatures) and Mix (1 mono-nucleotide, 5 di-nucleotide and 2 tri-nucleotide interaction signatures). We fix the eight signatures to the values obtained from the full data and down-sample to 1 percent, 2 percent or 5 percent of the total original mutation counts. We repeat the downsampling 50 times. In each experiment we then re-estimate the exposures for the eight signatures of the four interaction models by minimizing the generalized Kullback–Leibler divergence. In Figure 10 we show the mean cosine similarity between the original and re-estimated exposures from the down-sampled data for the four different models. We observe that the exposures for the di-nucleotide model are better recovered than the exposures for the tri-nucleotide model. In general, we observe that a simpler parametrization gives a more robust estimation of the exposures. This feature could be important if the exposures are used in the clinic for deciding upon diagnosis or treatment of cancer patients.

3.3 Analysis of Liver data

In this section we analyse 260 Liver cancer patients from the PCAWG tumors with the three models, where all the signatures are parametrized with either mono-, di- or tri-nucleotide interactions. The results for these models are shown in Figure 11 together with the mixture model, where each signature can be any of the three parametrizations. When running all the possible mixture models for different number of signatures we see that the models with the smallest BIC include both di-nucleotide signatures and even mono-nucleotide signatures (Figure 11(a)). In addition, we see in Figure 11(b) that the di-nucleotide and mixture model are identifying more of the COSMIC signatures that were found for Liver cancer in Alexandrov et al. (2020). The top interaction effects for many of these COSMIC signatures also include many mono- or di-nucleotide interactions, which again shows that simpler parametrizations can be used to explain many COSMIC signatures (Figure 11(c)).

3.4 Analysis of UCUT data

The UCUT data contains information about the two flanking bases at each side. The UCUT count matrix has $T = 6 \times 4^4 = 1536$ mutation types and $N = 26$ patients. The data consists of 14,715 somatic mutations, and the number of non-zero entries in the count matrix is $n_{\text{obs}} = 5260$.

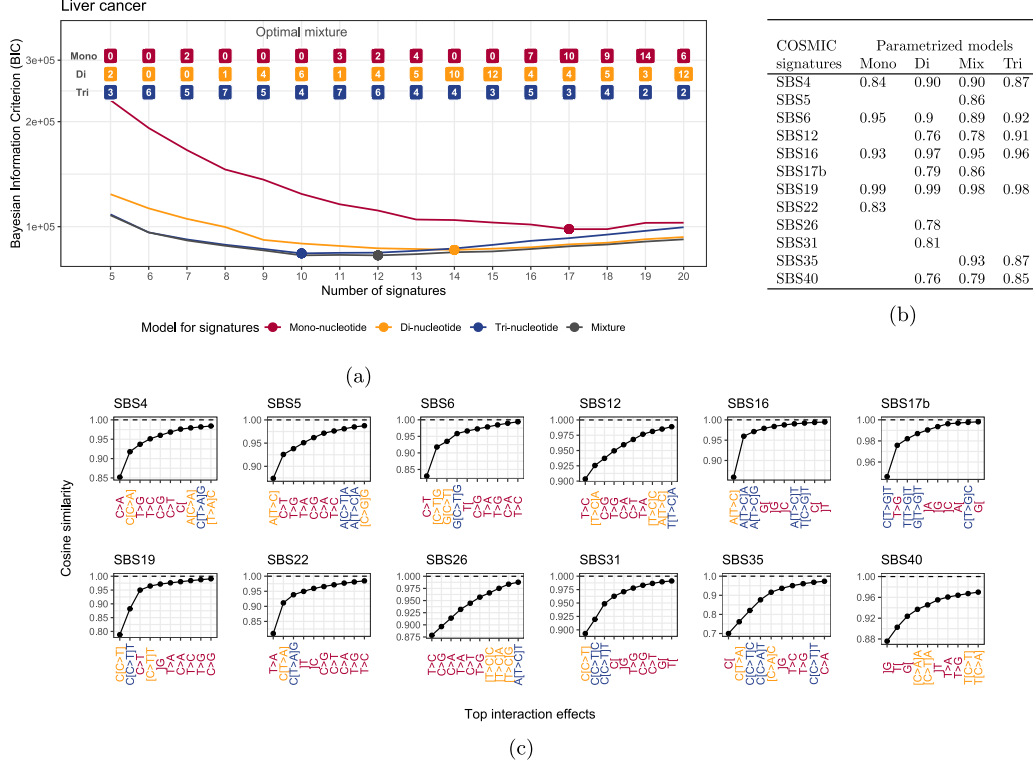


Figure 11: Analysis of the Liver data set. (a) The bayesian information criteria (BIC) for changing number of signatures K . This is shown for four different models; the red, orange and blue lines are where all the signatures are parametrized with mono-, di-og tri-nucleotide signatures, respectively. The grey line shows the BIC for the optimal mixture of the three different parametrizations. In the top it is shown how many of the signatures that are parametrized with each of the three different parametrizations. (b) Fixing the number of signatures at 12, the figure shows the match to the COSMIC signatures identified for Liver cancer in Alexandrov et al. (2020). The number is the cosine similarity and it is only shown if the value was above 0.75. (c) The top ten interactions for the COSMIC signatures recovered for the Liver data set. The top interactions are found with forward selection from the interaction making the largest increase in the cosine similarity to the COSMIC signature.

3.5 Choosing the number of signatures and parametrization

For the UCUT with two flanking nucleotides at each side of the mutation we have also found the optimal number of signatures for different number of parametrizations in Figure 12. Recall the possible parametrizations from Table 1. Three parametrizations are not included in the plot because they were never part of the optimal mixture. We also decided to remove the full penta-nucleotide model from the plot because the BIC was extremely high due to the many parameters. The optimal number of signatures for the penta-nucleotide model was therefore also only one signature. Again, we see that a simpler parametrization gives a higher optimal number of signatures to model the data. We chose to fix the number of signatures at $K = 2$ to follow Shiraishi et al. (2015) and this is also the optimal number of signatures for the di-nucleotide model.

We firstly consider the seven models shown in Table 2, where both signatures have the same parametrization. The table summarizes the number of parameters n_{prm} , model complexity $n_{\text{prm}} \log n_{\text{obs}}$, model fit GKL, and the differences between the model selection measure BIC and the smallest obtained BIC.

The penta-nucleotide interaction signature $L_2 \times L_1 \times M \times R_1 \times R_2$ has 1536 parameters (recall Table 1), and this many parameters inevitably results in over-fitting for the UCUT data set. This model is included as a

3. Results

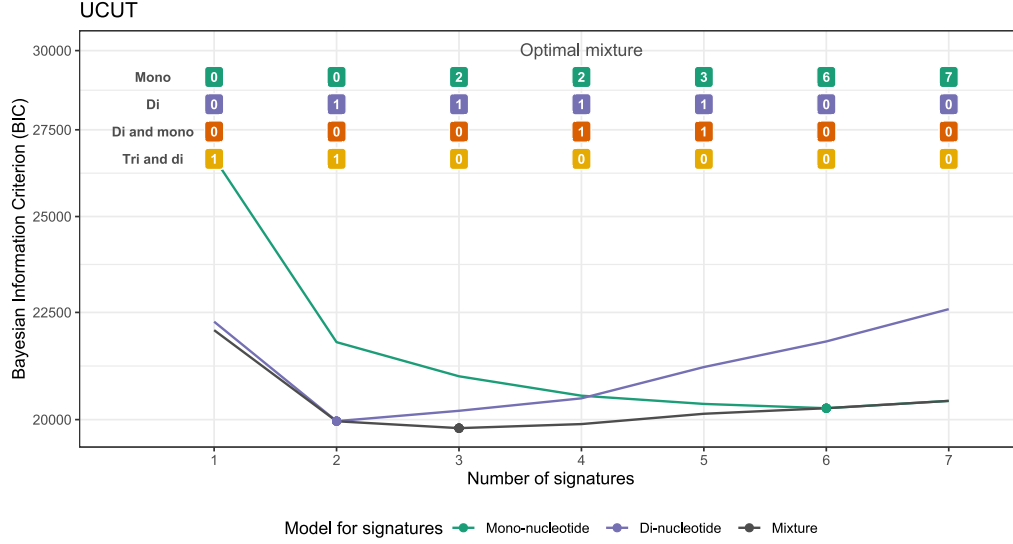


Figure 12: The Bayesian Information Criterion (BIC) for different number of signatures K to find the optimal number of signatures for the UCUT dataset. The top shows the optimal mixture of signature parametrizations for each number of signatures K .

Table 2: Summary statistics for the seven basic models for the UCUT data where both signatures have the same parametrization. The models are ordered according to their GKL value. The number of signatures is $K = 2$ and the number of observations is $n_{\text{obs}} = 5260$. At last the mixture model with the smallest BIC is also depict, which all the other BIC values are compared to.

Model for the two signatures	Number of parameters n_{prm}	Model complexity $n_{\text{prm}} \log n_{\text{obs}}$	Fit to data GKL	Model selection ΔBIC
$L_2 + L_1 + M + R_1 + R_2$	$2 \times 18 = 36$	308	10,422	2116
$L_1 \times M \times R_1$	$2 \times 96 = 192$	1645	10,182	2972
$L_2 + L_1 \times M + M \times R_1 + R_2$	$2 \times 48 = 96$	823	9788	1363
$L_2 + L_1 \times M \times R_1 + R_2$	$2 \times 102 = 204$	1748	9438	1588
$L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$ (a)	$2 \times 66 = 132$	1131	9008	111
$L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$ (b)	$2 \times 120 = 240$	2056	8658	336
$L_2 \times L_1 \times M \times R_1 \times R_2$	$2 \times 1536 = 3072$	26,321	6982	21,249
Mixture of signature (a) and (b)	$120 + 66 = 186$	1594	8721	0

control to show that the full parametrization gives an extremely high BIC value compared to the other models. A parametrization with much fewer parameters is needed for inferring robust signatures, and the mono-nucleotide interaction signatures $L_2 + L_1 + M + R_1 + R_2$ from Shiraishi et al. (2015) was originally developed for this purpose. Here, we also consider a di-nucleotide signature of the type $L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$, and three signatures that have a combination of interaction terms $L_2 + L_1 \times M + M \times R_1 + R_2$, $L_2 + L_1 \times M \times R_1 + R_2$ and $L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$. Finally, we include the tri-nucleotide signature $L_1 \times M \times R_1$ to investigate whether the two immediate flanking nucleotides are sufficient for explaining the probability of a somatic cancer mutation.

We observe that two immediate flanking nucleotides (one at each side) are not sufficient for explaining the mutation patterns: the $L_1 \times M \times R_1$ model has the same poor fit to data as the mono-nucleotide model despite having more than five times as many parameters. The four models $L_2 + L_1 \times M + M \times R_1 + R_2$, $L_2 + L_1 \times M \times R_1 + R_2$, $L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$ and $L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$ all show a relatively good fit

to the data, but the $L_2 + L_1 \times M \times R_1 + R_2$ model is penalized for the many parameters. Finally, the $L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$ and $L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$ model have a superior fit to the data compared to the other models, and does not contain too many parameters. We note that these two models are the only models with di-nucleotide interaction between the two left flanking nucleotides (both models contain the term $L_2 \times L_1$) and the two right flanking nucleotides (the term $R_1 \times R_2$), and conclude that these interaction terms are important for quantifying the probability of a somatic mutation in this cancer type.

We also consider parametrizations of the signature matrix where the two signatures have different parametrizations. The GKL and BIC for 16 different combinations of the seven parametrizations is summarized in Figure 13. Here, we have ordered the models by the GKL value as this automatically groups the different signature parametrizations. We have only included the penta-nucleotide signature once at last, as it gives extremely high BIC values due to the many parameters in the model.

Similar to our finding for the BRCA data set, we observe that two mono-nucleotide signatures $L_2 + L_1 + M + R_1 + R_2$ give a poor fit to the data. We emphasize that two tri-nucleotide signatures $L_1 \times M \times R_1$ or a mixture of the two all have a poor fit to the data, which means the information about the flanking nucleotides two positions away from the mutation is important. We find that a mixture between the two parametrizations $L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$ and $L_2 \times L_1 + L_1 \times M \times R_1 + R_1 \times R_2$ fits the data very well despite the rather few parameters; this mixture model has the smallest BIC value.

In Figure 14 the two signatures are visualized for the Mono, Di, Mix and Penta model. For the mixture model, signature 1 is described by the tri- and di-nucleotide interactions and signature 2 only by the di-nucleotide interactions. In the original study in Hoang et al. (2013) they identify signature 1 as a novel mutation signature that predominantly contains $T > A$ substitutions at CpTpG site caused by aristolochic acids. This is also reflected in Figure 15, where the top interaction is the CpTpG site. This single tri-nucleotide interaction is the likely the reason why the best parametrization for the signature includes tri-nucleotide interactions.

3.5.1 Model comparisons and stability of signatures

The cosine similarities for reconstructing the signatures from parametric bootstrap show that the penta-nucleotide signatures are much worse at reconstructing the same signatures (Figure 16). Again, this indicates

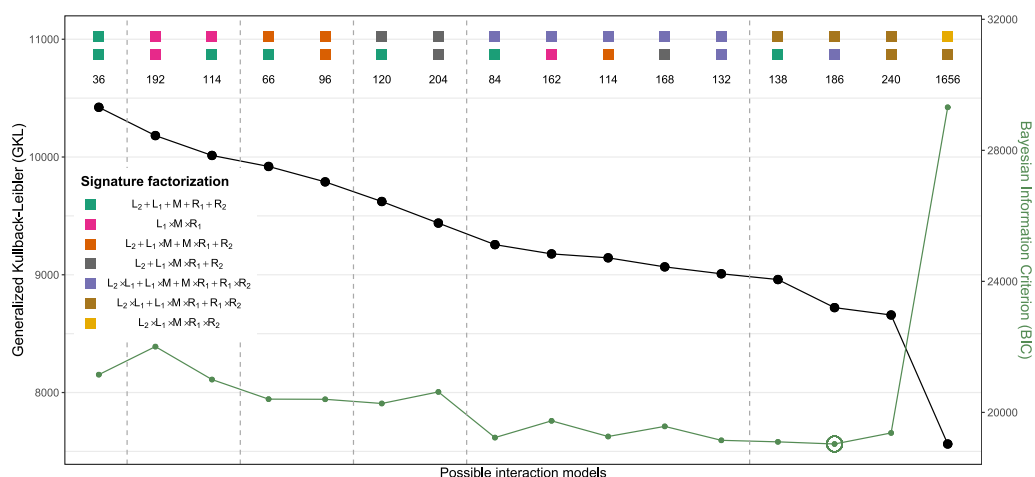


Figure 13: The Generalized Kullback–Leibler for 16 models with two signatures for the UCUT data set. The models are ordered according to GKL values, which also order the models by the first signature.

3. Results

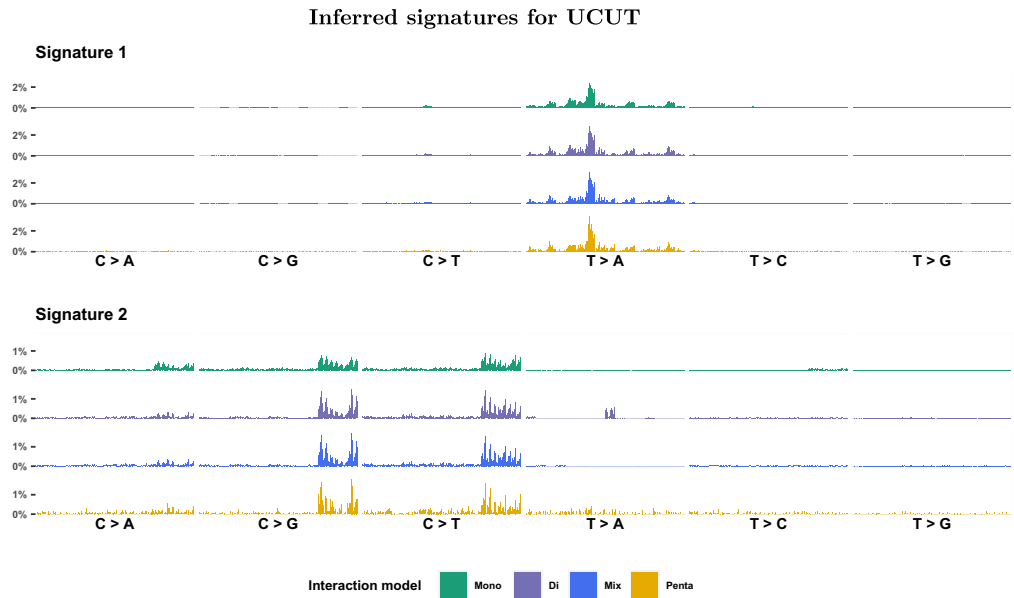


Figure 14: Inferred signatures for the UCUT data set. Comparison of the two signatures for the Mono, Di, Mix and Penta models.

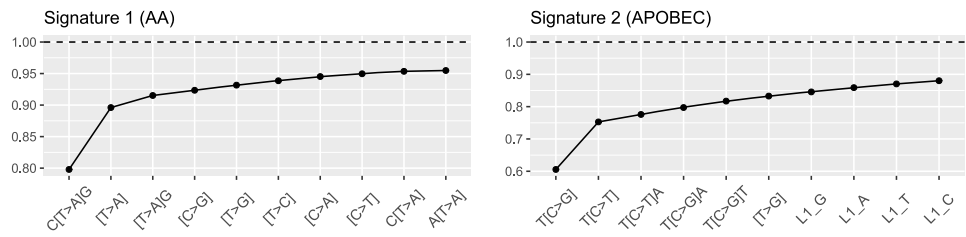


Figure 15: The top ten interactions that is increasing the cosine similarity to the retrieved signatures.

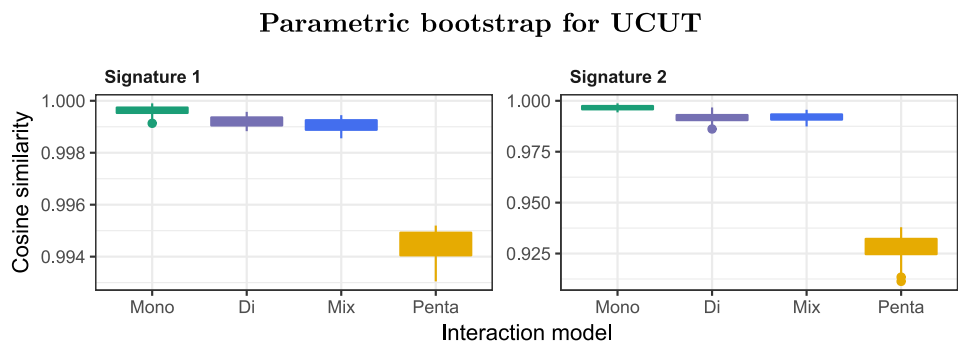


Figure 16: The cosine similarity for reconstructing the signatures with parametric bootstrap for the UCUT data.

the problem with too many parameters in the model. On the other hand, the model with two di-nucleotide signatures and the mixture model is almost as stable as the mono-nucleotide signatures, but gives a much better fit to data.

These findings demonstrate the relevance of our flexible framework for mutational signatures. The di-nucleotide signatures provide a better fit to the data and are biologically more plausible than mono-nucleotide signatures, and the parametrization is more stable than the parameter-rich signatures with interaction terms higher than or equal to three. The ability to allow a combination of signatures is also advantageous.

4 Methods

In this section we describe the EM-algorithm for estimating the parameters in non-negative matrix factorization. We first describe the EM-algorithm for the traditional model where the only constraints on the exposure matrix W and signature matrix H in the matrix factorization are that the entries must be non-negative (e.g. Cemgil 2009). Second, we extend the EM-algorithm to the situation where the signatures are parametrized according to (2).

For mutational count data it is natural to assume that each entry is Poisson distributed

$$V_{nt} \sim \text{Pois}((WH)_{nt}), \quad n = 1, \dots, N, \quad t = 1, \dots, T. \quad (7)$$

The data log-likelihood is then, up to an additive constant, given by

$$\ell(W, H; V) = \sum_{n=1}^N \sum_{t=1}^T \{V_{nt} \log((WH)_{nt}) - (WH)_{nt}\}, \quad (8)$$

and we determine W and H by maximizing the data log-likelihood. The details are provided in Section 4. Maximization of the data log-likelihood is identical to minimizing the generalized Kullback–Leibler (GKL) divergence

$$\text{GKL} = \text{GKL}(W, H; V) = \sum_{n=1}^N \sum_{t=1}^T \{V_{nt} \log V_{nt} - V_{nt} \log((WH)_{nt}) - V_{nt} + (WH)_{nt}\}. \quad (9)$$

This follows as the negative data log-likelihood is proportional to the GKL up to an additive constant. The factorization is clearly not unique up to permutation and scaling. Indeed, if W and H are non-negative and A is a $K \times K$ permutation matrix, we have that WA and $A^{-1}H$ are non-negative and $WH = W(AA^{-1})H = (WA)(A^{-1}H)$. The permutation issue is taken into account by a potential re-ordering of the mutational signatures and their corresponding weights. If A is a diagonal matrix with positive entries we also have that WA and $A^{-1}H$ are non-negative and $WH = (WA)(A^{-1}H)$. The scaling issue can be solved by normalizing the signatures in H such that they sum to one, i.e. by choosing $A = \text{diag}(d_1, \dots, d_K)$ as the diagonal matrix with entries $d_k = \sum_{t=1}^T H_{kt}$, $k = 1, \dots, K$, on the diagonal. We refer to Laursen and Hobolth (2022) for a general discussion of the NMF non-uniqueness problem and a general procedure to determine the set of feasible solutions.

The data log-likelihood (8) is analytically intractable, but we can view the problem as a missing data problem where the missing information is the assignment of each mutation to a signature. If this information was available, then a likelihood analysis would be easy, and therefore the EM-algorithm (Dempster et al. 1977) applies.

4.1 EM-algorithm for traditional non-negative matrix factorization

Given a data matrix $V \in \mathbb{N}_+^{N \times T}$ the aim of NMF is to find a non-negative factorization WH , where $W \in \mathbb{R}_+^{N \times K}$ and $H \in \mathbb{R}_+^{K \times T}$ approximates of our data V i.e. $V \approx WH$. The rank K of the factorization is often chosen magnitudes smaller than the minimum of N and T . A larger K obviously gives a better fit, but would potentially overfit the data. In traditional NMF all the entries in W and H are free parameters that need to be estimated. Later we will reduce the number of free parameters in H , but first we describe the traditional estimation of W and H .

A challenge with the likelihood function in (8) is that it is only convex in either W or H , but not in both matrices together. This means we cannot find a closed form solution for the maximum likelihood estimates of W and H , and instead we use the EM-algorithm. For the EM-algorithm we introduce the latent variables

$$Z_{nkt} \sim \text{Pois}(W_{nk}H_{kt})$$

which is the mutational count from each of the K signatures for each observation, such that the total number of mutations for a cancer patient n of a certain type t is given by

$$V_{nt} = \sum_{k=1}^K Z_{nkt} \sim \text{Pois}((WH)_{nt}).$$

4. Methods

The complete log-likelihood is given by

$$\ell(W, H; Z) = \sum_{n=1}^N \sum_{t=1}^T \sum_{k=1}^K \{Z_{nkt} \log(W_{nk}H_{kt}) - W_{nk}H_{kt} - \log(Z_{nkt}!)\} \quad (10)$$

$$\equiv \sum_{k=1}^K \sum_{t=1}^T \left(\sum_{n=1}^N Z_{nkt} \right) \log(H_{kt}) + \sum_{k=1}^K \sum_{n=1}^N \left\{ \left(\sum_{t=1}^T Z_{nkt} \right) \log(W_{nk}) - W_{nk} \right\} \quad (11)$$

where we use that signatures are probability distributions that sum to one, $\sum_{t=1}^T H_{kt} = 1$, and \equiv means that the statement is true up to the additive constant $\sum_{n=1}^N \sum_{t=1}^T \sum_{k=1}^K \log(Z_{nkt})$.

E-step: For fixed values W^i and H^i this step finds the expected value of the latent variables $\{Z_{nkt}\}$ conditional on the data V . The distribution of $\{Z_{nkt}\}$ conditional on their sum is given by the multinomial distribution

$$(Z_{n1t}, \dots, Z_{nkt}) | V_{nt} = \sum_{k=1}^K Z_{nkt} \sim \text{Multi} \left(V_{nt}, \frac{1}{(WH)_{nt}} (W_{n1}H_{1t}, \dots, W_{nk}H_{kt}) \right),$$

which implies that

$$\mathbb{E}_{W^i, H^i}[Z_{nkt} | V] = \mathbb{E}_{W^i, H^i}[Z_{nkt} | V_{nt}] = V_{nt} \frac{W_{nk}^i H_{kt}^i}{(W^i H^i)_{nt}}.$$

Replacing $\{Z_{nkt}\}$ with their expected values $\mathbb{E}_{W^i, H^i}[Z_{nkt} | V]$ gives the expected complete log-likelihood

$$Q(W, H | W^i, H^i) = \sum_{k=1}^K \sum_{t=1}^T \left(\sum_{n=1}^N \mathbb{E}_{W^i, H^i}[Z_{nkt} | V] \right) \log(H_{kt}) \quad (12)$$

$$+ \sum_{k=1}^K \sum_{n=1}^N \left\{ \left(\sum_{t=1}^T \mathbb{E}_{W^i, H^i}[Z_{nkt} | V] \right) \log(W_{nk}) - W_{nk} \right\} \quad (13)$$

M-step: The first term of the expected complete log-likelihood (12) is recognised as K independent multinomial log-likelihood functions and the second term (13) is recognised as $N \times K$ Poisson log-likelihoods. Maximum of the expected complete log-likelihood with respect to W and H is therefore given by

$$H_{kt}^{i+1} = \frac{\sum_{n=1}^N \mathbb{E}_{W^i, H^i}[Z_{nkt} | V]}{\sum_{t=1}^T \sum_{n'=1}^N \mathbb{E}_{W^i, H^i}[Z_{n'kt} | V]} = \frac{\sum_{n=1}^N V_{nt} \frac{W_{nk}^i H_{kt}^i}{(W^i H^i)_{nt}}}{\sum_{t=1}^T \sum_{n'=1}^N V_{n't} \frac{W_{nk}^i H_{kt}^i}{(W^i H^i)_{n't}}} \quad (14)$$

and

$$W_{nk}^{i+1} = \sum_{t=1}^T \mathbb{E}_{W^i, H^i}[Z_{nkt} | V] = \sum_{t=1}^T V_{nt} \frac{W_{nk}^i H_{kt}^i}{(W^i H^i)_{nt}}. \quad (15)$$

The expected value of $\{Z_{nkt}\}$ from the E-step is also inserted, which means these updates include both steps of the EM-algorithm to find the optimal estimates W and H . The entire EM-algorithm with initialization and stopping criteria to obtain the optimal parameters is summarized in Algorithm 1. The updates are written in vector form for H and matrix form for W . Note that \otimes and division means entry wise multiplication and division, the vector $\mathbf{1}$ is of length T and consists only of ones, W_k is the k 'th column of W , and H_k is the k 'th row of H . We stop the EM-algorithm when the data log-likelihood after a full update of W and H is smaller than a threshold ϵ .

Algorithm 1: General EM-algorithm to estimate exposures W and signatures H .

Given data matrix V , rank K and threshold ϵ .

Initialize W^1 and H^1 with random entries.

for $i = 1, 2, 3, \dots$ **do**

for $k = 1, \dots, K$ **do**

 Update each signature

$$H_k^{i+1} = \frac{H_k^i \otimes ((W_k^i)' \frac{V}{W^i H^i})}{1' (H_k^i \otimes ((W_k^i)' \frac{V}{W^i H^i}))} \quad (16)$$

end

 Update exposures

$$W^{i+1} = W^i \otimes \left(\frac{V}{W^i H^i} (H^i)' \right)$$

stop if $\frac{\ell(W^{i+1}, H^{i+1}; Z) - \ell(W^i, H^i; Z)}{\ell(W^{i+1}, H^{i+1}; Z)} < \epsilon$

end

4.2 EM-algorithm for parametric non-negative matrix factorization

Another parametrization of the signatures H_1, \dots, H_K requires a change in update (14) which was based on maximizing (12). The parametrization of the signatures are given by the design matrices X_1, \dots, X_K . Recall that the number of mutations from a specific signature for each observation is given by the latent variables $\{Z_{nkt}\}$. We observe that we again have K independent multinomial log-likelihood terms that we can maximize separately. Define

$$Y_{kt}^i = \sum_{n=1}^N \mathbb{E}_{W^i, H^i} [Z_{nkt} | V],$$

which is the expected number of mutations at the i th iteration for signature k of type t . We now suppress the superscript i and subscript k by introducing the simple notation $y_t = Y_{kt}^i$ and $h_t = H_{kt}$. In parallel to (12) we need to maximize

$$\sum_{t=1}^T y_t \log(h_t)$$

with respect to β where we set

$$h_t = \frac{\exp((X\beta)_t)}{\sum_{t=1}^T \exp((X\beta)_t)}, \quad (17)$$

and again we have suppressed the dependency on k in both X and β . Instead of estimating β in this model, we use the 'Poisson Trick' (see e.g. Lee et al. 2017 or Section 6.4 in McCullagh and Nelder 1989). The 'Poisson Trick' means that the log-linear Poisson model

$$\log(y_t) = (X\beta)_t, \quad t = 1, \dots, T, \quad (18)$$

is equivalent to the multinomial response model with probabilities given by (17). We therefore determine the maximum likelihood estimate of β by fitting the log-linear Poisson model instead of the multinomial response model. The full EM-algorithm is presented in matrix form in Algorithm 2.

4. Methods

Algorithm 2: Parametric EM-algorithm to estimate exposures W and signatures H .

Given data matrix V , rank K , design matrices X_1, \dots, X_K , and threshold ϵ .

Initialize W^1 and H^1 with random entries.

for $i = 1, 2, 3, \dots$ **do**

for $k = 1, \dots, K$ **do**

 Update each signature

$$\mathbf{y}_k^i = H_k^i \otimes \left((W_k^i)' \frac{V}{W^i H^i} \right)$$

 Fit the log-linear Poisson regression

$$\log(\mathbf{y}_k^i) = X_k \beta_k^i \quad (19)$$

 for estimating β_k^i and set

$$H_k^{i+1} = \frac{\exp(X_k \hat{\beta}_k^i)}{\mathbf{1}' \exp(X_k \hat{\beta}_k^i)}$$

end

 Update exposures

$$W^{i+1} = W^i \otimes \left(\frac{V}{W^i H^i} (H^i)' \right)$$

stop if $\frac{\ell(W^{i+1}, H^{i+1}; Z) - \ell(W^i, H^i; Z)}{\ell(W^{i+1}, H^{i+1}; Z)} < \epsilon$

end

Estimation of β in (18) is obtained by fitting the log-linear Poisson model using the Newton-Raphson method, and for clarity we provide the details. The log-likelihood function for the Poisson model with design matrix X of dimension $T \times S$, parameter vector β of length S and data vector $y = (y_1, \dots, y_T)$ of length T is given by

$$\ell(\beta; y, X) \equiv \sum_{t=1}^T \{y_t(X\beta)_t - \exp((X\beta)_t)\}.$$

A closed form solution for the maximum likelihood estimate is in general not available, but we can use the Newton-Raphson method. The gradient and the Hessian of the log-likelihood function are

$$\frac{\partial \ell}{\partial \beta} = X' \{y - \exp(X\beta)\} \quad \text{and} \quad \frac{\partial^2 \ell}{\partial \beta' \partial \beta} = -X' A X,$$

where $A = A(\beta)$ is a diagonal matrix of dimension $T \times T$ with $\exp\left(\sum_{s=1}^S X_{ts} \beta_s\right)$, $t = 1, \dots, T$, on the diagonal. The Newton-Raphson update is given by

$$\beta^{i+1} = \beta^i + (X' A^i X)^{-1} X' \{y - \exp(X\beta^i)\},$$

where $A^i = A(\beta^i)$, which can be re-written as

$$\begin{aligned} \beta^{i+1} &= (X' A^i X)^{-1} X' A^i [X\beta^i + (A^i)^{-1} \{y - \exp(X\beta^i)\}] \\ &= (X' A^i X)^{-1} X' A^i v^i, \end{aligned}$$

where

$$v^i = X\beta^i + (A^i)^{-1} \{y - \exp(X\beta^i)\}.$$

This means that the update is the solution to the weighted least square problem

$$\beta^{i+1} = \arg \min_{\beta} \|(A^i)^{1/2} (v - X\beta)\|^2.$$

In our implementation in *R* we call the built-in method to solve the weighted least squares problem.

To accelerate the EM-algorithm we have both made a version that uses the *R* package SQUAREM (Du and Varadhan 2020) and another version implemented in C++. To escape local minimum of the divergence function we typically start the algorithm 100 or even 500 times and run each of them for 100 or 500 iterations before we identify a minimum, which was recommended in Biernacki et al. (2003). We then let the identified minimum iterate until convergence.

5 Discussion

We have presented new biologically plausible parametrizations of mutational signatures. The parametrization is based on interaction terms between neighbouring nucleotides. In general we find that the di-nucleotide interaction signature strikes a good balance between a satisfactory fit to our data and statistically stable and robust signatures. Importantly, our framework also allows a mixture of parametrizations for the signature matrix in non-negative matrix factorization. This makes the parametrization of the signature matrix very flexible because we allow each signature to have its own parametrization. We also identify the most important interaction effects for many of the COSMIC signatures, which in many cases is mono- or di-nucleotide interactions. Specifically we show the exact interactions that is driving the signatures.

Our main goal has been statistical robustness and interpretation of the signatures, and this is achieved by biologically plausible constraints on the parameters: we allow each signature to contain mono-, di-, tri-nucleotide or higher-order interaction terms. An alternative to the constraints imposed by interaction terms is to impose sparseness on the signatures in the spirit of Lal et al. (2021a). We believe that robust signatures obtained via constraints on the interaction terms is biologically more plausible than robust signatures obtained via sparseness constraints.

In general the number of mutation types is $T = 6 \times 4^{2n}$ when n bases are considered upstream and downstream of the mutated site. The number of mutation types T (and signature parameters in the general model) thus increases exponentially with the number of neighbouring nucleotides. There are $6 + 3 \times (2n) = 6(1 + n)$ parameters in the mono-nucleotide model, i.e. a linear increase in the number of parameters. In this paper we introduce di-nucleotide models that include interactions between neighbors given by $L_1 \times M + M \times R_1 + \sum_{i=1}^{n-1} (L_{i+1} \times L_i + R_i \times R_{i+1})$. This model results in $42 + 12 \times 2 \times (n - 1) = 6(3 + 4n)$ parameters. Thus, our di-nucleotide signatures are also linear in the number of flanking nucleotides.

We have focused on finding a single parametrization for each signature where interpretation is easy. This is useful when the aim is to recover the true underlying biological mechanisms that cause the various signatures (e.g. UV-light or tobacco smoking). Model averaging over different parametrizations for a signature would make sense if the goal is a statistically robust signature where interpretation is less important (e.g. classification of a genomic region based on the mutation profiles). The BIC values are rather similar for many of the models, suggesting that model averaging could be useful. Another extension of our model would be to change the poisson assumption of the data to the negative binomial model, as it has been shown to be better suited for mutational counts (Pelizzola et al. 2023).

Our flexible framework also allows inclusion of other factors known to have an impact on somatic mutations such as replication timing (Woo and Li 2012), expression level (Lawrence et al. 2013) or general conservation of the position when compared to other species (Bertl et al. 2018). Epigenetic data could be included in our model as an independent feature.

Acknowledgments: We thank Camilla Provstgaard Kudahl and Maiken Bak Poulsen for valuable initial results and discussions. We are grateful to Marta Pelizzola and Gustav Alexander Poulsgaard for helpful comments on an earlier version of the manuscript. We also want to thank the two anonymous reviewers for many constructive and helpful comments and suggestions for improving the presentation and analyses.

Research ethics: Not applicable.

Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Competing interests: The authors state no conflict of interest.

Research funding: Novo Nordisk Foundation grant number 22OC0079957.

Data availability: github.com/ragnhildlaursen/paramNMF_ms.

5. References

References

- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 3: 246–259.
- Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science* 354: 618–622.
- Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* 578: 94–101.
- Arndt, P.F., Burge, C.B., and Hwa, T. (2003). DNA sequence evolution with neighbor-dependent mutation. *J. Comput. Biol.* 10: 313–322.
- Bertl, J., Guo, Q., Juul, M., Besenbacher, S., Nielsen, M.M., Hornshøj, H., Pedersen, J.S., and Hobolth, A. (2018). A site specific model and analysis of the neutral somatic mutation rate in whole-genome cancer data. *BMC Bioinf.* 19: 147.
- Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Stat. Data Anal.* 41: 561–575.
- Cemgil, A.T. (2009). Bayesian inference for non-negative matrix factorisation models. *Comput. Intell. Neurosci.* 2009: 785152.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Methodol.* 39: 1–38.
- Du, Y. and Varadhan, R. (2020). SQUAREM: an R package for off-the-shelf acceleration of EM, MM and other EM-like monotone algorithms. *J. Stat. Software* 92: 1–41.
- Gori, K. and Baez-Ortega, A. (2018). sigfit: flexible bayesian inference of mutational signatures, *bioRxiv*, pp. 372896.
- Hoang, M.L., Chen, C.-H., Sidorenko, V.S., He, J., Dickman, K.G., Yun, B.H., Moriya, M., Niknafs, N., Douville, C., Karchin, R., et al. (2013). Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci. Transl. Med.* 5: 197.
- Hobolth, A. (2008). A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *J. Comput. Graph. Stat.* 17: 138–162.
- Hwang, D.G. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.* 101: 13994–14001.
- Lal, A., Liu, K., Tibshirani, R., Sidow, A., and Ramazzotti, D. (2021a). De novo mutational signature discovery in tumor genomes using sparsesignatures. *PLoS Comput. Biol.* 17: e1009119.
- Laursen, R. and Hobolth, A. (2022). A sampling algorithm to compute the set of feasible solutions for nonnegative matrix factorization with an arbitrary rank. *SIAM J. Matrix Anal. Appl.* 43: 257–273.
- Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499: 214–218.
- Lee, J.Y.L., Green, P.J., and Ryan, L.M. (2017). On the 'Poisson Trick' and its extensions for fitting multinomial regression models, *arXiv: 1707.08538*.
- Levatić, J., Salvadores, M., Fuster-Tormo, F., and Supek, F. (2022). Mutational signatures are markers of drug sensitivity of cancer cells. *Nat. Commun.* 13: 2926.
- Lindberg, M., Boström, M., Elliott, K., and Larsson, E. (2019). Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light. *Proc. Natl. Acad. Sci. U. S. A.* 116: 20411–20417.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, 2nd ed. Chapman & Hall, New York.
- Nik-Zainal, S. and Morganella, S. (2017). Mutational signatures in breast cancer: the problem at the DNA level. *Clin. Cancer Res.* 23: 2617–2629.
- Pelizzola, M., Laursen, R., and Hobolth, A. (2023). Model selection and robust inference of mutational signatures using negative binomial non-negative matrix factorization. *BMC Bioinf.* 24: 187.
- Rosales, R.A., Drummond, R.D., Valieris, R., Dias-Neto, E., and Da Silva, I.T. (2017). signer: an empirical bayesian approach to mutational signature discovery. *Bioinformatics* 33: 8–16.
- Shen, Y., Ha, W., Zeng, W., Queen, D., and Liu, L. (2020). Exome sequencing identifies novel mutation signatures of UV radiation and trichostatin A in primary human keratinocytes. *Sci. Rep.* 10: 4943.
- Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet.* 11: e1005657.
- Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* 25: 289–310.
- Woo, Y.H. and Li, W.-H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* 3: 1004.

Paper

D

**Integration of opportunities and parametrized signatures to
improve mutational signatures estimation**

by Raghild Laursen, Marta Pelizzola, Lasse Maretty and Asger Hobolth

Paper draft

Integration of opportunities and parametrized signatures to improve mutational signatures estimation

Ragnhild Laursen¹, Marta Pelizzola¹, Lasse Maretty², and Asger Hobolth¹

¹Department of Mathematics, Aarhus University. ² Bioinformatics Research Center.
Emails: {marta, ragnhild, asger}@math.au.dk, lasse.maretty@protonmail.com

September 10, 2024

Abstract

Mutational signatures describe the pattern of mutations over the different mutation types. Each mutation type is determined by a base substitution and the flanking nucleotides to the left and right of that base substitution. Due to the widespread interest in mutational signatures, several efforts have been put in developing methods for robust and stable signature estimation. Here, we combine different extension of the standard method to estimate mutational signatures. This includes using the negative binomial model, parametrizing the signatures and add opportunities to the analysis. We show that combining these extensions give more robust mutational signatures. In particular we highlight the importance of including mutational opportunities and parametrizing the signatures when mutation types describe an extended sequence context, with two and three flanking nucleotides to each side of the base substitution. Lastly, we also show that opportunities highly increases the predictive power for unknown data.

Keywords: cancer genomics, mutational signatures, parametrizing, opportunities, Negative Binomial, non-negative matrix factorization.

1 Introduction

Mutations observed in cancer genomes are known to be generated by combinations of different mutational processes acting on the genome. The collection of mutations observed in one cancer patient is often referred to as the mutational catalogue of the patient. In this context, mutations are usually defined taking into account the base substitution and its right and left flanking nucleotides, i.e. using the single-base-substitution-96 (SBS-96) mutational context (Alexandrov et al., 2013a).

Nowadays the genome of numerous cancer patients is sequenced and their mutational catalogues are available in public datasets such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) database (Campbell et al., 2020). From these large collections of sequenced data from patients it is possible to obtain mutational signatures. These are probability vectors over the possible mutation types and each

1. Introduction

describe the pattern of a mutational process operative in the cancer genomes. Cancer is mostly driven by few mutations in the genome, thus mutational signatures are highly important for distinguishing driver mutations from other processes taking place in the genome. A thorough understanding of mutational signatures and their associations with DNA repair deficiencies or exposures to certain agents can therefore play a key role for defining treatments for cancer patients (Caruso et al., 2017) or developing prevention strategies (Zhang et al., 2021). Well known examples of mutational processes leaving distinctive signatures are e.g. aging (Risques and Kennedy, 2018), UV light (Shibai et al., 2017) or tobacco smoking (Alexandrov et al., 2016). Other examples and aetiologies of different mutational signatures can be found in Tate et al. (2019).

Mutational signatures are usually identified using non-negative matrix factorization (NMF) (Alexandrov et al., 2013b; Lyu et al., 2020; Lal et al., 2021; Pelizzola et al., 2023). NMF takes as input a matrix of mutational profiles for different patients and factorizes it into the product of two non-negative matrices. In our framework, these are a matrix of mutational signatures and a matrix of weights representing the contribution of each signature to the total mutation counts of the different patients. Different approaches to estimate the signatures from mutational count data are reviewed in (Baez-Ortega and Gori, 2017; Omichessan et al., 2019).

The mutation rate varies along the genome and depends on the sequence context around each site (Lindberg et al., 2019; Dietlein et al., 2020; Bethune et al., 2022). Thus, it is essential to include the sequence context information around the base substitution when defining mutation types for the inference of mutational signatures. The SBS-96 mutational context has been commonly used in the literature to define the mutation types for the mutational count matrix used as input in NMF. This definition can be extended by including two, three or more flanking nucleotides at each side of the base substitution to better capture the information carried out by the sequence context of each mutation.

Extending the sequencing context used to define the mutation type leads to a much larger number of parameters than with the SBS-96 context which leads to more sparse data sets. Furthermore, sets of nucleotide sequences occur with very heterogeneous rates along the genome, which leaves different opportunities for the mutation types to occur. The opportunity of a mutation type describes the number of sites in the genome where that specific mutation type can occur. A higher opportunity for a specific mutation type gives a higher probability of observing a mutation from that type.

The disparity in opportunities for different mutation types can be observed in the SBS-96 context, but becomes even more pronounced when looking at extended sequence contexts with two or three flanking nucleotides at each side. If the information contained in the opportunities is incorporated in the framework for estimating mutational signatures, the inferred signatures can be generalized regardless of the considered sequence composition (e.g. between whole genome and exome sequencing). Furthermore, by taking into account the sequence context in a genome, we obtain a more accurate estimation of the relative contribution of the different mutational processes.

Recent work has shown that a Negative Binomial model is better suited for

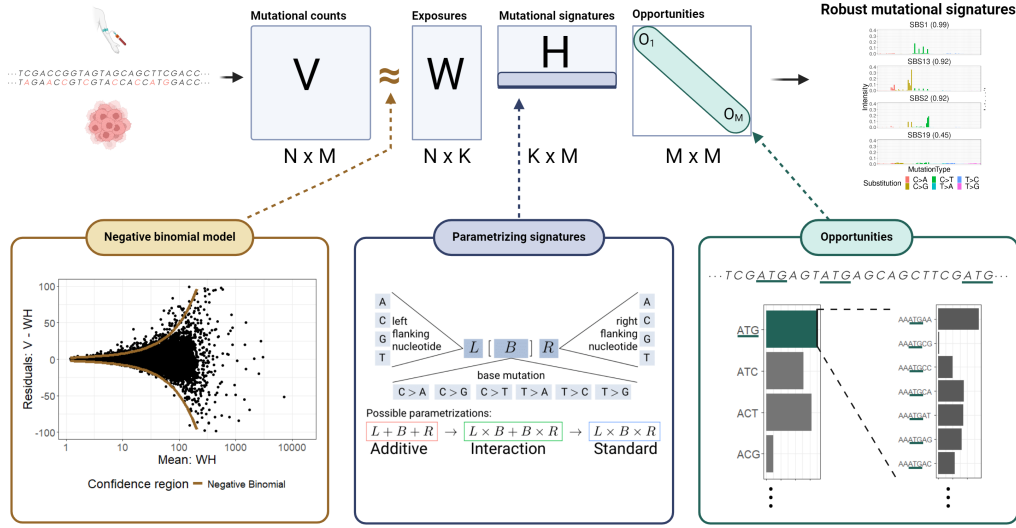


Figure 1: A graphical representation of the methods incorporated in the model to obtain robust mutational signatures. It includes the negative binomial distributional assumption of the counts, a parametrization of the signatures and the inclusion of the opportunity for each mutation type.

modeling mutational counts data (Lyu et al., 2020; Pelizzola et al., 2023). However, mutational opportunities have not been included in Negative Binomial NMF models to estimate mutational signatures and their importance in extended sequence contexts still needs to be explored. Ignoring opportunities in NMF can lead to an overestimation of the relative contribution of certain types of mutations to the overall mutational burden of the cancer genome. This can in turn lead to inaccurate conclusions about the underlying mutational processes. Mutational opportunities have already been used in Fischer et al. (2013) and Gori and Baez-Ortega (2018) in connection to the classical Poisson model in the SBS-96 sequence context.

In this paper, we combine different extensions of standard NMF to estimate more robust mutational signatures. Firstly, we use the Negative Binomial distribution for the NMF framework to account for overdispersion in the data. Then, we derive new update rules for NMF where we include the mutational opportunities in the model directly. At last, we considerably reduce the number of parameters in the model by parametrizing the mutational signatures as described in Laursen et al. (2024). This leads to more robust and stable signatures and avoids overfitting. Our model thereby combines Negative Binomial NMF, opportunities and parametrized signatures and shows its importance in estimating mutational signatures when more than one flanking nucleotide is used to define the sequence context. A graphical representation of our approach is shown in Figure 1 and further elaborated in Section 2.

The paper is structured in the following way: Section 2 starts by introducing the mathematical details of NMF and then Sections 2.2 and 2.3 describe our model with opportunities and parametrization, respectively. In Section 3, we show that the combination of a Negative Binomial NMF model with the opportunities and

2. Methods

parametrized signatures leads to robust signature estimation and generalizes well to unseen data. In order to show this, we consider two data sets from the PCAWG database (Campbell et al., 2020). Results on the breast cancer data set are discussed in Section 3.1 and results on the liver cancer data set are summarized in Section 3.2. The *R* implementation of our proposed method and the code for reproducing our results are available at <https://github.com/ragnhildlaursen/ParOpp>.

2 Methods

Consider a mutational count data set $V \in \mathbb{N}_0^{N \times M}$ where N is the number of patients and M is the number of mutation types. Mutational signatures are commonly derived by NMF (Lee and Seung, 1999) which factorizes V into the product of two non-negative matrices $W \in \mathbb{R}_+^{N \times K}$ and $H \in \mathbb{R}_+^{K \times M}$ such that

$$V \approx WH.$$

In cancer genomics, the rows in H represent the mutational signatures defined by probability vectors over the different mutation types, i.e. H is normalized such that its rows sum to one. The matrix W includes the exposures, where each row contains the exposure of each mutational signature for the corresponding patient.

In the SBS-96 mutational context (Alexandrov et al., 2013a) $M = 96$, corresponding to the 6 base mutations when assuming strand symmetry times the 4 flanking nucleotides on each side, i.e. $4 \cdot 6 \cdot 4 = 96$. As we want to go beyond this definition we consider also $M = 1536$ and $M = 24576$, where sequence contexts of length 5 and 7 are used, respectively. Here, the number of parameters in our model clearly increases considerably. The classical model used when estimating mutational signatures has been proposed in Lee and Seung (1999) and first applied to mutational count data in Alexandrov et al. (2013a), where the data is assumed to follow a Poisson distribution:

$$V_{nm} \sim \text{Pois}((WH)_{nm}). \quad (1)$$

In order to estimate W and H , we need to minimize the Kullback-Leibler divergence (GKL) :

$$d_{Po}(V||WH) = \sum_{n=1}^N \sum_{m=1}^M \{V_{nm} \log V_{nm} - V_{nm} \log((WH)_{nm}) - V_{nm} + (WH)_{nm}\}, \quad (2)$$

which is proportional to the Poisson likelihood.

Recently, Gori and Baez-Ortega (2018), Lyu et al. (2020), Vöhringer et al. (2021) and Pelizzola et al. (2023) proposed an alternative model based on the Negative Binomial distribution to account for overdispersion in the mutational counts. We introduce this model in Section 2.1 and extend it in Sections 2.2 and 2.3.

2.1 Negative Binomial NMF

This section summarizes the Negative Binomial NMF (NB-NMF) model with a patient specific dispersion coefficient. Here,

$$V_{nm} \sim \text{NB} \left(\alpha_n, \frac{(WH)_{nm}}{\alpha_n + (WH)_{nm}} \right),$$

where the mean and variance are given by

$$\mathbb{E}[V_{nm}] = (WH)_{nm} \quad \text{and} \quad \mathbb{V}[V_{nm}] = (WH)_{nm} \left(1 + \frac{(WH)_{nm}}{\alpha_n} \right).$$

The parameter α_n is the dispersion coefficient for each patient. Small values of α_n correspond to high dispersion in the data. On the contrary, when $\alpha_n \rightarrow \infty$, the Negative Binomial model converges to the more commonly used Poisson model in Equation (1).

It can be shown that the Negative Binomial likelihood is proportional to the following divergence:

$$d_{NB}(V||WH) = \sum_{n=1}^N \left\{ \sum_{m=1}^M V_{nm} \log \left(\frac{V_{nm}}{(WH)_{nm}} \right) - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + (WH)_{nm}} \right) \right\}. \quad (3)$$

Estimation of W and H are derived using a Majorization-Minimization (MM) algorithm on this divergence measure. Here, we propose to further extend the NB-NMF for extracting mutational signatures by including mutational opportunities in the model and parametrizing the mutational signatures especially for large mutational contexts.

2.2 NB-NMF with opportunities

Mutational opportunities correspond to the fraction of sites in the genome where a mutation type can occur. In Figure 2 the mutational opportunity for each mutation type is plotted against the mean number of mutations per patient for the corresponding mutation type. Patients from the breast and liver cancer data sets in (Campbell et al., 2020) are used in this figure and each panel refers to one of the different mutational contexts. Figure 2 demonstrates that the correlation between the observed counts and the opportunities increases when moving from the classical SBS-96 context to larger context sizes. Thus, accounting for the opportunities is needed when modeling the mutational counts and even more for the extended nucleotide contexts. Signature reconstruction from mutational count data with opportunities will provide a more accurate estimation of the relative contributions of different mutational processes.

To incorporate opportunities into the model we first view the mutational counts as coming from a binomial distribution as seen in Weinhold et al. (2014); Lochovsky et al. (2015).

$$V_{nm} \sim \text{Bin}(O_m, p_{nm}). \quad (4)$$

Here, we can think of the opportunity O_m as the number of possible sites for a mutation to occur and p_{nm} as the probability of success that a mutation does occur

2. Methods

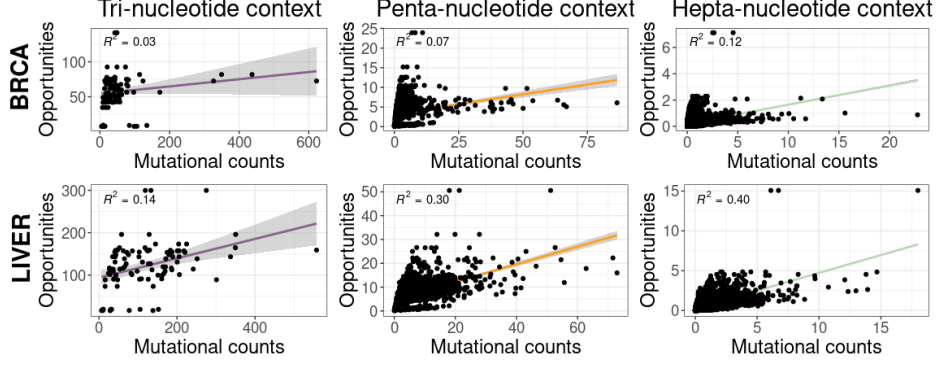


Figure 2: The average mutational count across patients for each mutation type is plotted against the corresponding mutational opportunities for the BRCA (top panel) and liver cancer (bottom panel) patients in (Campbell et al., 2020). The mutational contexts with 1 (left), 2 (middle) and 3 (right) flanking nucleotides on each side are shown for each cancer type.

of type m for patient n . If $m = A[T > C]G$, then O_m would be the number of times the triplet ATG is observed in a genome and p_{nm} would denote the probability that this specific mutation occurs at the sites ATG in patient n .

Since O_m is large and p_{nm} is small, we can approximate this model with the Poisson model where

$$\text{Bin}(O_m, p_{nm}) \simeq \text{Pois}(p_{nm}O_m) = \text{Pois}((WH)_{nm}O_m). \quad (5)$$

As O_m is fixed we can write p_{nm} as the factorization $(WH)_{nm}$ that needs to be estimated. This is an extension of the model presented in Equation (1), where $(WH)_{nm}$ is replaced by $WH_{nm}O_m$. The Poisson model can be further extended to the Negative Binomial model by allowing additional dispersion in the model.

The Negative Binomial model including opportunities has the following mean and variance:

$$\mathbb{E}[V_{nm}] = (WH)_{nm}O_m \quad \text{and} \quad \mathbb{V}[V_{nm}] = (WH)_{nm}O_m \left(1 + \frac{(WH)_{nm}O_m}{\alpha_n} \right).$$

where α_n is the dispersion coefficient for each patient as before and O_m is the opportunity for mutation type m .

Under this model, we would like to maximize the following Negative Binomial log-likelihood function

$$\ell(W, H; V) = \sum_{n=1}^N \sum_{m=1}^M \left\{ \log \binom{\alpha_n + V_{nm} - 1}{\alpha_n} + V_{nm} \log \left(\frac{(WH)_{nm}O_m}{\alpha_n + (WH)_{nm}O_m} \right) + \alpha_n \log \left(1 - \frac{(WH)_{nm}O_m}{\alpha_n + (WH)_{nm}O_m} \right) \right\}$$

which is equivalent to minimizing the following divergence measure:

$$d_{NBO}(V||WH) = \sum_{n=1}^N \left\{ \sum_{m=1}^M V_{nm} \log \left(\frac{V_{nm}}{(WH)_{nm}O_m} \right) - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + (WH)_{nm}O_m} \right) \right\}.$$

Notice, that this is almost equivalent to the divergence in (3), where $(WH)_{nm}$ is simply replaced by $(WH)_{nm}O_m$. To derive the updates for W and H we use the

MM algorithm and follow the same rationale as Gouvert et al. (2020) and Pelizzola et al. (2023).

The resulting multiplicative updates for H_{km} and W_{nk} when including opportunities are as follows:

$$H_{km}^{t+1} = H_{km}^t \frac{\sum_{n=1}^N \frac{V_{nm}}{(WH^t)_{nm} O_m} W_{nk}}{\sum_{n=1}^N \frac{V_{nm} + \alpha_n}{(WH^t)_{nm} O_m + \alpha_n} W_{nk}}. \quad (6)$$

and

$$W_{nk}^{t+1} = W_{nk}^t \frac{\sum_{m=1}^M \frac{V_{nm}}{(W^t H)_{nm} O_m} H_{km}}{\sum_{m=1}^M \frac{V_{nm} + \alpha_n}{(W^t H)_{nm} O_m + \alpha_n} H_{km}}. \quad (7)$$

The complete derivations leading to these updates are found in Appendix A.

2.3 NB-NMF with opportunities and parametrized signatures

Opportunities are particularly important when working with extended context, where more flanking nucleotides to the base mutation is included as emphasized in Figure 2. Here, the number of sites for different groups of nucleotides becomes more and more heterogeneous. Laursen et al. (2024) have described a framework to parametrize mutational signatures for an arbitrary large nucleotide context to avoid overfitting and obtain a more stable set of signatures, which are also easier to interpret. The signatures are parametrized by the natural features of the mutation type, which considers the base substitution and each of the flanking nucleotides as separate variables. We are incorporating this framework into the NB-NMF model with opportunities to avoid overfitting the data and obtain more robust signatures.

We consider three main models for the signatures: the standard model from Alexandrov et al. (2013a), the additive model from Shiraishi et al. (2015) and the interaction model included in (Laursen et al., 2024). The models are illustrated in Figure 1 when one flanking nucleotide is included. In the standard model all interactions between nucleotides are included. Let f be the number of flanking nucleotides considered to each side of the base mutation. In the standard model, the number of parameters is $D = 6 \cdot 4^{2f}$, which means it increases exponentially with the number of flanking nucleotides. The additive model does not include any interaction term and the number of parameters can be calculated as a linear function of the number of flanking nucleotides: $D = 6 + 2f \cdot (4 - 1) = 6 + 6f$. Lastly, in the interaction model the number of parameters is larger than for the additive model as interactions for neighboring nucleotides are considered. Though, the number of parameters still remain linear in the number of flanking nucleotides: $D = 6 \cdot 3 + 2f(3 \cdot 4) = 18 + 24f$. All three models can be incorporated into the framework of Laursen et al. (2024), where each mutational signature is parametrized using the log-linear link function

$$H_k = H_k(\beta_k) = \frac{1}{C_k} \exp(X_k \beta_k),$$

where $C_k = \mathbf{1}' \exp(X_k \beta_k)$ is a normalizing constant and X_k has dimension $M \times D$, where M is the number of mutation types and D is the number of parameters in the

3. Results

parametrization. The entries in X_k are fixed dependent on the specified parametrization of the mutation types. The values in β_k are estimated using quasi-poisson regression. More details on parametrizing mutational signatures are available in Laursen et al. (2024).

Algorithm 1 in Appendix B shows the estimation of the factorization under the Negative Binomial model including both parametrization and opportunities. For each iteration t we do the following: Update the mutational signatures under the NB-NMF model with opportunities in (6), fit a log-linear Poisson model to the signatures, update the signatures according to the estimated $\hat{\beta}_k$ and specified parametrization of the design matrix X_k and lastly update the exposure matrix W with the update rule from the NB-NMF model with opportunities in (7).

3 Results

In this section we illustrate the results of the NB-NMF model with opportunities and parametrization for two data sets from the PCAWG database (Campbell et al., 2020). In this database 2782 patients from different cancer types are available and the mutational counts can be found at <https://www.synapse.org/#!Synapse:syn11726620>. We consider a data set of breast cancer patients in Section 3.1 and one of liver cancer patients in Section 3.2. For both examples we show how the inclusion of opportunities and the parametrization improves the model for different number of signatures and parametrizations. Lastly, for the breast cancer data set we illustrate that using the opportunities greatly increases the predictive power of the model. These results can be found in Section 3.1.2.

In our results, we express the cost in estimating the data in terms of the generalized Kullback Leibler (GKL) divergence (see Equation (2)) and we use the Bayesian Information Criterion (BIC)

$$\text{BIC} = -2 \ln L + \ln(n_{\text{obs}})n_{\text{prm}} \quad (8)$$

to determine which models are more appropriate under the different scenarios. Here $\ln L$ is the log-likelihood from the Negative Binomial model, n_{obs} is number of observations and n_{prm} the number of parameters to be estimated. A model is thus penalized if it does not have good fit and/or if it has a large number of parameters. Models achieving a good balance between the fit and the number of parameters will result in the best BIC values.

3.1 Results on the breast cancer data set

In this section we extract mutational counts for the 120 breast cancer patients from the PCAWG database and use them for our analysis.

The results in Figure 3 show the BIC against the number of signatures used for estimation. These results show that the additive (Shiraishi et al., 2015) and the interaction models can achieve a good fit to data while keeping the number of parameters low. Additionally, for the parametrized models, including the opportunities (solid lines) returns lower BIC values. This is especially true for the penta and hepta-nucleotide contexts, where the difference between the cost of the models

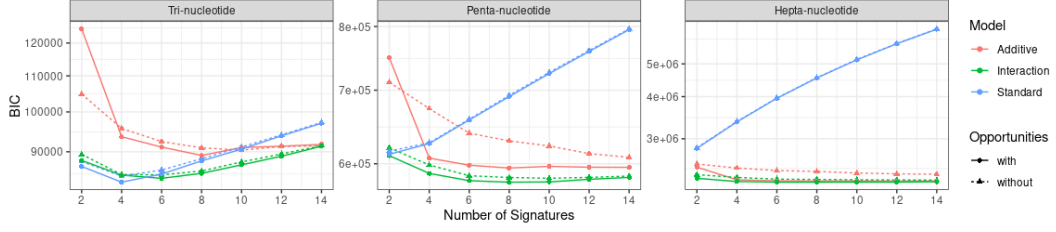


Figure 3: The influence of parametrizing and including opportunities on the BRCA data set. The BIC in log-scale plotted against the number of signatures.

with and without opportunities becomes larger. These results also emphasize that the inclusion of the opportunities is essential in larger contexts as the BIC for the standard model shows consistent overfitting of this model to the data.

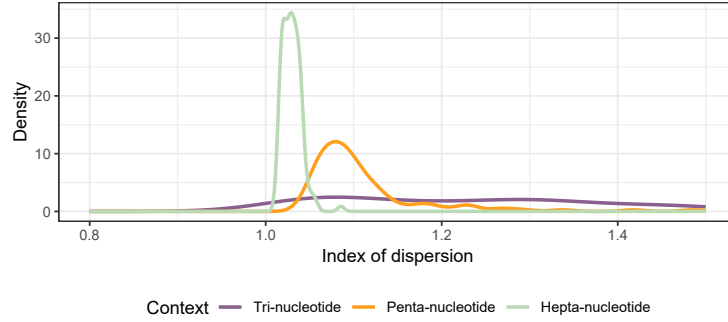


Figure 4: Distribution of the dispersion index $\frac{\mathbb{V}[V_{nm}]}{\mathbb{E}[V_{nm}]}$ for the NB-NMF model with four signatures.

When modeling data in a penta-nucleotide or hepta-nucleotide context the number of parameters becomes large very quickly with the standard model, whereas an additive or interaction parametrization can keep the number of parameters under control and still provide a very good fit.

Figure 4 shows the distribution of the dispersion index given by

$$\frac{\mathbb{V}[V_{nm}]}{\mathbb{E}[V_{nm}]} = 1 + \frac{(WH)_{nm}O_m}{\alpha_n}. \quad (9)$$

It is illustrated for the different sizes of nucleotide context from Figure 3 in the scenario where four signatures is assumed. This dispersion index in (9) is exactly the value that scales up the variance in the negative binomial distribution compared to the Poisson distribution. In Figure 4 we see that the dispersion index is persistently larger than one, which support that the negative binomial model are more suited for mutational counts. Though, the dispersion index is decreasing with the larger context, which could be caused by the natural lower counts for each mutation type.

The procedure to estimate the dispersion parameters α is equivalent to the one outlined in Pelizzola et al. (2023) with the inclusion of the opportunities. We obtain maximum likelihood estimates of α based on the Negative Binomial likelihood using

3. Results

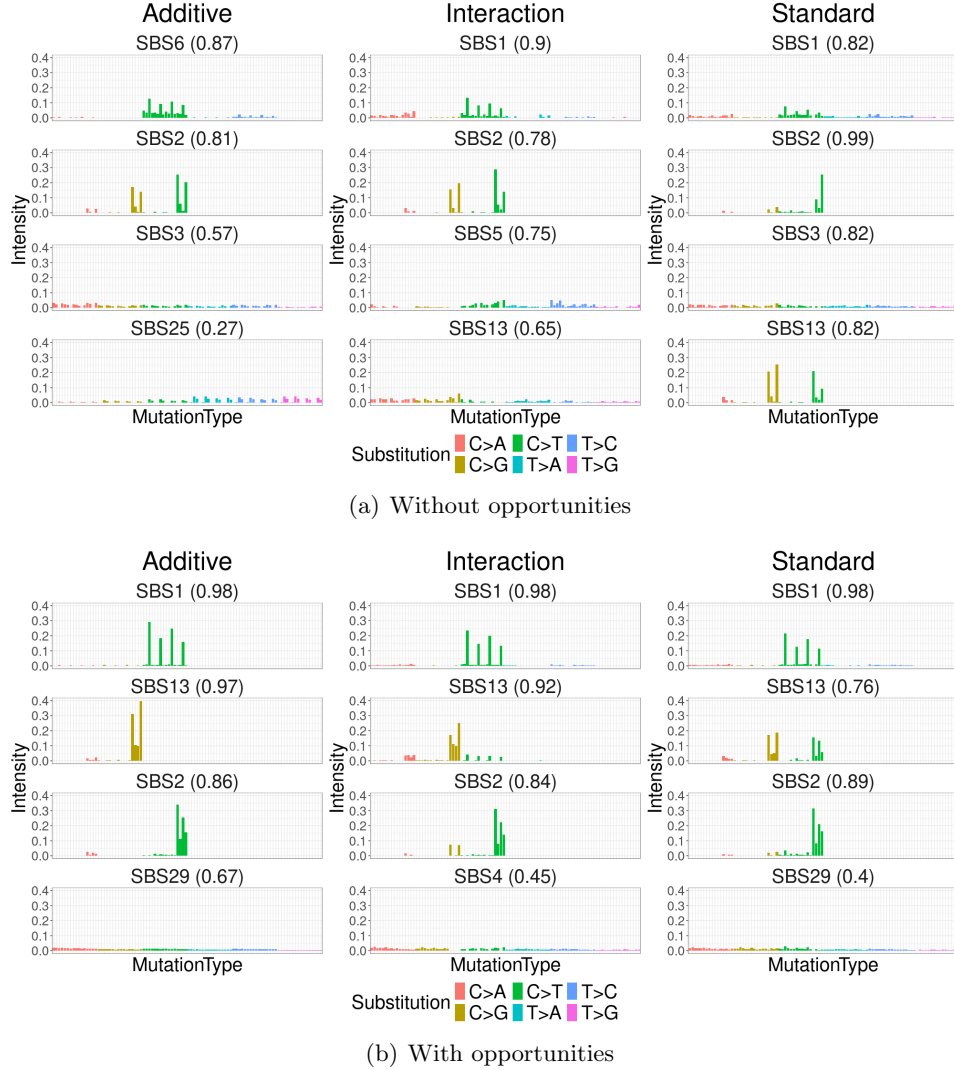


Figure 5: Cosine similarity between the constructed signatures from the additive, interaction and standard model without (a) and with (b) opportunities to the COSMIC signatures for the SBS-96 context.

the Newton-Raphson method with the estimate of WH from the Poisson NMF with opportunities.

In Figure 5 we have visualized the signatures for rank four for the three different parametrizations, with and without opportunities. These are compared to the results presented in Alexandrov et al. (2020) which also suggest that there are mainly 4 active signatures in breast cancer patients, namely SBS1, SBS2, SBS5, and SBS13. Furthermore, SBS3 and SBS18 are also moderately present in breast cancer patients.

We identified the most similar signatures from the Catalogue Of Somatic Mutations In Cancer (COSMIC) version 3. Figure 5 shows the estimated signatures from the different models together with the cosine similarities to the most similar COSMIC signatures. The signatures from the COSMIC database are available at

<https://cancer.sanger.ac.uk/cosmic>.

We show results without the inclusion of opportunities in Figure 5(a). Here, the interaction and standard models can reconstruct major signatures of breast cancer patients with high accuracy: the interaction model can reconstruct SBS1, SBS2, SBS5, and SBS13, whereas under the standard model SBS3 is reconstructed instead of SBS5. The average cosine similarity of the standard model (0.86) is larger than the one for the interaction model (0.77). The additive model is not able to estimate the four major signatures of breast cancer patients and has lower cosine similarity in this scenario.

On the contrary, when opportunities are included the cosine similarity between the estimated signatures from both parametrized models with opportunities is much higher. All three models reconstruct signatures SBS1, SBS2, and SBS13 with high cosine similarity. Considering these three signatures, the average cosine similarity for the additive model is 0.94, for the interaction model is 0.91, and for the standard model is 0.88. This shows that the parametrized models with opportunities can achieve very high accuracy in reconstructing signatures with a much simpler model. The last signature is poorly reconstructed for all scenarios under a model with opportunities: indeed, the estimation of flat signatures such as SBS5 in this case becomes harder and thus the accuracy is much lower here. A combination of the different parametrizations for the mutational signatures as suggested in Laursen et al. (2024) seems to be an effective way to decrease the number of parameters and maintaining high accuracy.

3.1.1 Estimation with and without opportunities

To further understand in which cases the opportunities improve the estimation of the data we have included Figure 6. This figure shows the difference in the poisson likelihood of the approximation of the data with and without opportunities. Each point illustrate the average difference for a specific mutation type as a function of its corresponding opportunity. It shows that the model with opportunities gives better prediction of data points with a high opportunity in the Penta and Hepta context.

In both Figure 3 and 6 there seem to be no difference between the standard model with and without opportunities. This is due to equivalence in the model estimation besides the scaling of the opportunities. As the mean is simply scaled in the model with opportunities it would be possible to obtain the same estimation by scaling the mean parameter afterwards.

The greater difference observed in the parametrized models for models with and without opportunities arises from the fact that these models involve parametrizing the signatures during estimation. Including opportunities will change the signature vectors that are being parametrized in the estimation. In Figure 2 it is shown that the opportunities are correlated with the mutational count and it is therefore expected to reduce variability in the signatures. As a result, it becomes easier to parametrize the signatures when opportunities are considered, leading to improved estimation in this case.

3. Results

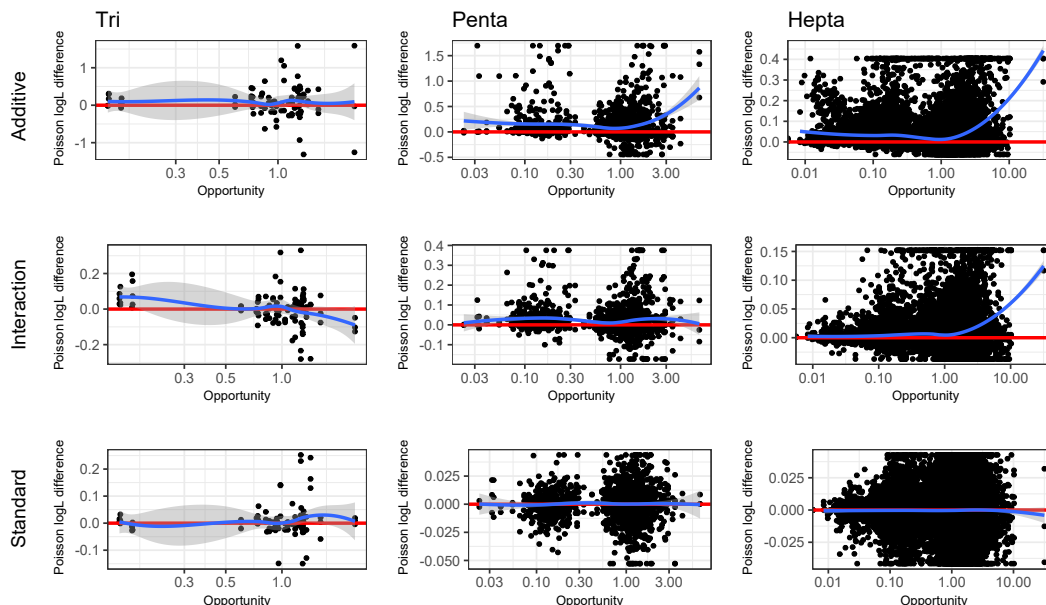


Figure 6: The difference in the Poisson likelihood between the approximations of the data with and without opportunities, such that positive values means better prediction with opportunities. Each point represent the results for a mutation type as a function of its corresponding opportunity. The red line illustrate the zero line and the blue line is a generalization of a moving average estimated using the LOESS method.

3.1.2 Increase in predictive power

In this section we use our estimated signatures to predict new data. We show that if the signatures are estimated with opportunities then the fit to the new data is better. Furthermore, simple models such as the additive and interaction models provide better prediction of new data as they avoid overfitting to the training data.

For our analysis we partitioned the genome into ten subsets to create ten data sets to be used for prediction. Each mutation is randomly assigned to one of the ten subsets. We then performed 10-fold cross-validation where the mutational signatures are estimated from a training set, where one of the ten data sets are left out and then prediction is preformed on the left out test set.

Figure 7 shows the results of this analysis for the Tri-, Penta- and Hepta-nucleotide contexts. Parametrizing the signatures will provide better estimation in larger context with 5 or 7 nucleotides, where we also see that the standard model are being overparametrized as the error is increasing with the number of signatures.

3.2 Results on the liver cancer data set

We consider here a data set with 260 liver cancer patients from the PCAWG database. Similar conclusions can be drawn also for this dataset. Here we only show results for the hepta-nucleotide context.

Figure 3.2 shows results on the influence of parametrizing and including the

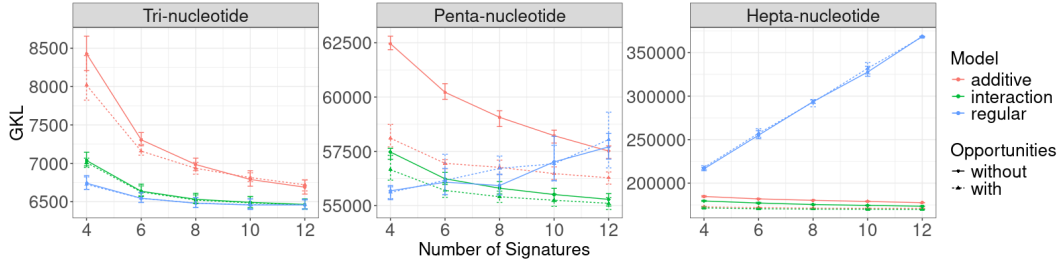


Figure 7: The influence of parametrizing and including opportunities on predicting the mutational counts in the BRCA data set.

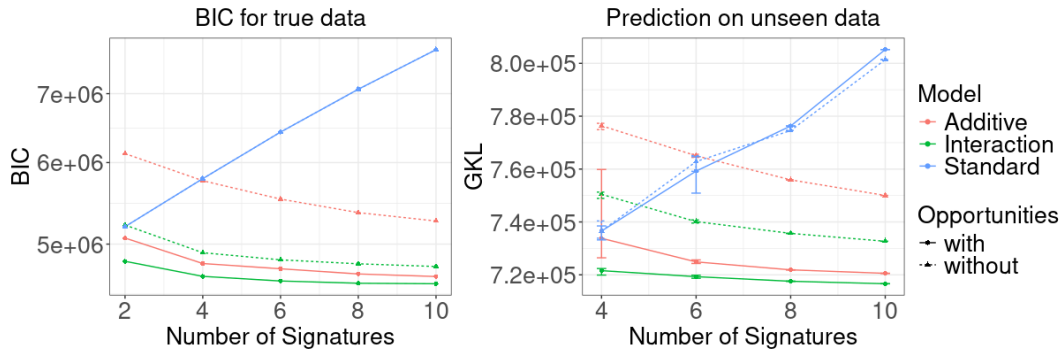


Figure 8: Results for the liver cancer data set for the hepta-nucleotide context. The influence of parametrizing and including opportunities in reconstructing the original data on the left and in predicting on unseen data on the right. The GKL divergence plotted against the number of signatures.

opportunities in estimating the original data and in predicting on unseen data where the GKL divergence is plotted as a function of the number of signatures. With these results we show that parametrizing the signatures is a satisfactory compromise between decreasing the number of parameters and retaining good performance of the method. Decreasing the number of parameters is essential when looking at the hepta-nucleotide context and in this context including the opportunities provides again lower GKL in the parametrized models. We also obtained similar results on the prediction on unseen data as those discussed in section 3.1.2, and we can show again that parametrized model with opportunities provide much better prediction on unseen data. The signatures estimated by these models are more robust than the one estimated by the standard NMF model and generalize better to unseen data.

4 Discussion

The mutation rate in a genome is heavily influenced by its sequence context and the count of each context in the genome. This is referred to as the opportunity of a mutation type and is often overlooked when analyzing mutational count data. Understanding mutational patterns requires considering the opportunity, as it pro-

4. Appendices

vides additional insights into the underlying processes leading to mutations. In light of this, we propose an extension of Non-negative Matrix Factorization tailored to provide robust signatures in extended sequence contexts.

Our approach integrates mutational opportunities into the negative binomial model, enabling more accurate estimation of exposures and signatures while accounting for the heterogeneous values of opportunities within extended sequence contexts. We augment this with a parameterization technique, as introduced by (Laursen et al., 2024), aimed at reducing overfitting by minimizing the number of parameters. This combined model demonstrates the necessity of incorporating mutational opportunities and parametrizing the signatures for effective performance in extended nucleotide contexts.

The incorporation of mutational opportunities offers a significant advantage, particularly in scenarios with limited data availability. For instance, in cases where only exome sequencing is feasible due to resource constraints, less data is available to estimate mutational signatures compared to whole genome sequencing. Our method ensures robust signatures from exome sequencing data capable of fitting whole genome sequencing data as well.

Moreover, considering extended sequence contexts and incorporating mutational opportunities facilitates investigating the reasons behind the excess of certain types of mutations. These discrepancies can be attributed to various factors such as damage repair mechanisms. By accurately measuring overrepresented mutations, our approach enables the study of these mechanisms in greater detail. Furthermore, the inclusion of mutational opportunities is crucial for estimating signatures from indels data, as deletions and insertions often occur in certain repeats of the genome (Streisinger et al., 1966).

Developing models that yield robust signatures is essential to predict accurate exposure for new data and interpreting signatures clearly. Our utilization of mutational opportunities and parametrization represents a methodological enhancement, providing a new tool for extracting more stable mutational signatures.

Acknowledgement

MP and AH acknowledges funding of the Novo Nordisk Foundation (Grant number NNF21OC0069105).

Appendices

A Derivation of NB-NMF multiplicative updates with opportunities

For the MM algorithm, we construct a majorizing function $G(H, H^t)$ for $d_{NBO}(V||WH)$ with the constraint that $G(H, H) = d_{NBO}(V||WH)$. Using Jensen's inequality and replacing $\log\left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km} O_m}\right)$ with the tangent line in H^t because of the concavity property of the logarithm we obtain:

$$\begin{aligned}
d_{NBO}(V||WH) &= \sum_{n=1}^N \sum_{m=1}^M V_{nm} \log \left(\frac{V_{nm}}{\sum_{k=1}^K W_{nk} H_{km} O_m} \right) \\
&\quad - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km} O_m} \right) \\
&\leq \sum_{n=1}^N \sum_{m=1}^M V_{nm} \log V_{nm} - V_{nm} \sum_{k=1}^K \gamma_k \log \left(\frac{W_{nk} H_{km} O_m}{\gamma_k} \right) \\
&\quad + (\alpha_n + V_{nm}) \left[\log \left(\frac{\alpha_n + (WH^t)_{nm} O_m}{\alpha_n + V_{nm}} \right) \right. \\
&\quad \left. + \frac{W_{nk}}{\alpha_n + (WH^t)_{nm} O_m} (H_{km} - H_{km}^t) O_m \right] \\
&= G(H, H^t).
\end{aligned} \tag{10}$$

where $\gamma_k = W_{nk} H_{km}^t O_m / \sum_{k=1}^K W_{nk} H_{km}^t O_m$. Lastly, it can easily be shown that $G(H, H) = d_{NBO}(V||WH)$ as follows:

$$\begin{aligned}
G(H, H) &= \sum_{n=1}^N \sum_{m=1}^M V_{nm} \log V_{nm} \\
&\quad - V_{nm} \sum_{k=1}^K \frac{W_{nk} H_{km} O_m}{\sum_{k=1}^K W_{nk} H_{km} O_m} \log \left(\frac{W_{nk} H_{km} O_m}{\left(\frac{W_{nk} H_{km} O_m}{\sum_{k=1}^K W_{nk} H_{km} O_m} \right)} \right) \\
&\quad - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km} O_m} \right) \\
&= \sum_{n=1}^N \sum_{m=1}^M V_{nm} \log \left(\frac{V_{nm}}{\sum_{k=1}^K W_{nk} H_{km} O_m} \right) \\
&\quad - (\alpha_n + V_{nm}) \log \left(\frac{\alpha_n + V_{nm}}{\alpha_n + \sum_{k=1}^K W_{nk} H_{km} O_m} \right) \\
&= d_{NBO}(V||WH)
\end{aligned} \tag{11}$$

Having defined the majorizing function $G(H, H^t)$ in (10) we can derive the multiplicative updates for H_{km} and W_{nk} which leads to the update rules in (6) and (7)

4. Appendices

B Parametric NMF with opportunities

Algorithm 1: Parametric NMF with opportunities.

Given data matrix V , rank K , design matrices X_1, \dots, X_K , opportunity vector $O = (O_1, \dots, O_M)$ and threshold ϵ .

Estimate dispersion parameters:

Get W^{Po}, H^{Po} by applying poisson NMF updates to V with K signatures

Estimate $\alpha_1, \dots, \alpha_N$ by Negative Binomial MLE using W^{Po}, H^{Po} and V

Initialize W^1 and H^1 with random entries.

for $t = 1, 2, 3, \dots$ **do**

for $k = 1, \dots, K$ **do**

 Update each signature

$$H_k^t = H_k^t \otimes \frac{((W_k^t)' \frac{V}{W^t H^t}) \frac{1}{O}}{\left((W_k^t)' \frac{V + \alpha_n}{W^t H^t O + \alpha_n} \right)}$$

 Fit the log-linear Poisson regression

$$\log(H_k^t) = X_k \beta_k^t \tag{12}$$

 for estimating β_k^t and set

$$H_k^{t+1}(\hat{\beta}_k^t) = \frac{\exp(X_k \hat{\beta}_k^t)}{\mathbf{1}' \exp(X_k \hat{\beta}_k^t)}$$

end

 Update exposures

$$W^{t+1} = W^t \otimes \frac{\left(\frac{V}{W^t H^t(\hat{\beta}^t)} (H^t(\hat{\beta}^t))' \right)}{\left(\frac{V + \alpha_n}{W^t H^t(\hat{\beta}^t) O + \alpha_n} (H^t(\hat{\beta}^t))' \right) \frac{1}{O}}$$

stop if $\frac{\ell(W^{t+1}, H^{t+1}(\hat{\beta}^t); Z) - \ell(W^t, H^t(\hat{\beta}^{t-1}); Z)}{\ell(W^{t+1}, H^{t+1}(\hat{\beta}^t); Z)} < \epsilon$

end

References

- Alexandrov, L. B., Ju, Y. S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P. J., Vineis, P., Phillips, D. H., and Stratton, M. R. (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622.
- Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Getz, G., Rozen, S. G., and Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101.
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013a). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421.
- Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):264–259.
- Baez-Ortega, A. and Gori, K. (2017). Computational approaches for discovery of mutational signatures in cancer. *Briefings in Bioinformatics*, 20(1):77–88.
- Bethune, J., Kleppe, A., and Besenbacher, S. (2022). A method to build extended sequence context models of point mutations and indels. *Nature Communications*, 13(1).
- Campbell et al. (2020). Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93.
- Caruso, D., Papa, A., Tomao, S., Vici, P., Panici, P. B., and Tomao, F. (2017). Niraparib in ovarian cancer: results to date and clinical potential. *Therapeutic Advances in Medical Oncology*, 9(9):579–588.
- Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., Lander, E. S., Allen, E. M. V., and Sunyaev, S. R. (2020). Identification of cancer driver genes based on nucleotide context. *Nature Genetics*, 52(2):208–218.
- Fischer, A., Illingworth, C. J., Campbell, P. J., and Mustonen, V. (2013). EMu: Probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biology*, 14(4):1–10.
- Gori, K. and Baez-Ortega, A. (2018). sigfit: flexible bayesian inference of mutational signatures. *arXiv*.
- Gouvert, O., Oberlin, T., and Fevotte, C. (2020). Negative Binomial Matrix Factorization. *IEEE Signal Processing Letters*, 27:815–819.

4. References

- Lal, A., Liu, K., Tibshirani, R., Sidow, A., and Ramazzotti, D. (2021). De novo mutational signature discovery in tumor genomes using SparseSignatures. *PLOS Computational Biology*, 17(6):e1009119.
- Laursen, R., Maretty, L., and Hobolth, A. (2024). Flexible model-based non-negative matrix factorization with application to mutational signatures. *Statistical Applications in Genetics and Molecular Biology*, 23(1):20230034.
- Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.
- Lindberg, M., Boström, M., Elliott, K., and Larsson, E. (2019). Intragenomic variability and extended sequence patterns in the mutational signature of ultraviolet light. *Proceedings of the National Academy of Sciences*, 116(41):20411–20417.
- Lochovsky, L., Zhang, J., Fu, Y., Khurana, E., and Gerstein, M. (2015). LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic acids research*, 43(17):8123–8134.
- Lyu, X., Garret, J., Rättsch, G., and Lehmann, K. V. (2020). Mutational signature learning with supervised negative binomial non-negative matrix factorization. *Bioinformatics*, 36(Suppl.1):i154–i160.
- Omichessan, H., Severi, G., and Perduca, V. (2019). Computational tools to detect signatures of mutational processes in DNA from tumours: A review and empirical comparison of performance. *PLOS ONE*, 14(9):e0221235.
- Pelizzola, M., Laursen, R., and Hobolth, A. (2023). Model selection and robust inference of mutational signatures using negative binomial non-negative matrix factorization. *BMC Bioinformatics*, 24(1).
- Risques, R. A. and Kennedy, S. R. (2018). Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genetics*, 14(1).
- Shibai, A., Takahashi, Y., Ishizawa, Y., Motooka, D., Nakamura, S., Ying, B.-W., and Tsuru, S. (2017). Mutation accumulation under UV radiation in *Escherichia coli*. *Scientific Reports*, 7(1):1–12.
- Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLOS Genetics*, 11(12):e1005657.
- Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M. (1966). Frameshift mutations and the genetic code. *Cold Spring Harbor Symposia on Quantitative Biology*, 31(0):77–84.
- Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupp, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., Stefancsik, R., Thompson, S. L., Wang, S., Ward, S., Campbell, P. J., and Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947.

- Vöhringer, H., Hoeck, A. V., Cuppen, E., and Gerstung, M. (2021). Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nature Communications*, 12(1).
- Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, 46(11):1160–1165.
- Zhang, T., Joubert, P., Ansari-Pour, N., Zhao, W., Hoang, P. H., Lokanga, R., Moye, A. L., Rosenbaum, J., Gonzalez-Perez, A., Martínez-Jiménez, F., Castro, A., Muscarella, L. A., Hofman, P., Consonni, D., Pesatori, A. C., Kebede, M., Li, M., Gould Rothberg, B. E., Peneva, I., Schabath, M. B., Poeta, M. L., Costantini, M., Hirsch, D., Heselmeyer-Haddad, K., Hutchinson, A., Olanich, M., Lawrence, S. M., Lenz, P., Duggan, M., Bhawsar, P. M. S., Sang, J., Kim, J., Mendoza, L., Saini, N., Klimczak, L. J., Islam, S. M. A., Otlu, B., Khandekar, A., Cole, N., Stewart, D. R., Choi, J., Brown, K. M., Caporaso, N. E., Wilson, S. H., Pommier, Y., Lan, Q., Rothman, N., Almeida, J. S., Carter, H., Ried, T., Kim, C. F., Lopez-Bigas, N., Garcia-Closas, M., Shi, J., Bossé, Y., Zhu, B., Gordenin, D. A., Alexandrov, L. B., Chanock, S. J., Wedge, D. C., and Landi, M. T. (2021). Genomic and evolutionary classification of lung cancer in never smokers. *Nature Genetics*, 53(9):1348–1359.

Paper

E

**A simple extension of non-negative matrix factorization to
find structures and spatially variable genes in multiple tissues**

by Ragnhild Laursen and Barbara E Engelhardt

Paper draft

A simple extension of non-negative matrix factorization to find structures and spatially variable genes in multiple tissues

Ragnhild Laursen^{1*} and Barbara E Engelhardt^{2,3}

¹Department of Mathematics, Aarhus University, Denmark ² Gladstone Institutes, San Francisco, CA, United States ⁴ Department of Biomedical Data Science, Stanford University, Stanford, USA

Abstract

The gene programs within different cell types will be correlated in cellular neighborhoods due to shared signaling environments. Methods to identifying these spatial gene programs may be powerful, but currently do not scale to existing data sets; approximations including sparse selection of cells or bagging together genes by using principal components as the cellular features reduce the precision of these complex programs. To better identify these multi-cellular microenvironments with shared gene programs for large-scale spatial genomics data, we developed a method that combines nonnegative matrix factorization (NMF) with Gaussian smoothing across cells in space. Our method, neighborhood NMF, identifies anti-tumor hubs and other known and unknown localized interactions among diverse cell types. Our spatially-aware dimension reduction method has many advantages over currently available methods, including the ability to run on very large scale data with thousands of features, including multiple tissue samples, with around a million cells at once. In three diverse spatial gene expression data sets, we benchmark our method against related methods and expert-annotated samples. We identify known and new structures in the tissue and other microenvironments, including anti-tumor hubs and inflammatory hubs, and we illustrate the important information gained from using the highest resolution experimental data for analyses.

Keywords: Non-negative matrix factorization (NMF), Gaussian smoothing, spatial hubs, Merfish

1 Introduction

There is an evolving interest in understanding the different tissue structures and microenvironments in human tissues, which often consists of many different genes and cell types working together to create location-specific functions and processes. Recently, various experimental technologies have been developed to analyze spatial

*Email: ragnhild@math.au.dk

1. Introduction

transcriptomics, where each observation both include a gene expression profile and a corresponding spatial location in the tissue [9, 1, 14]. These new technologies have made it possible to enhance our understanding of tissue structure and localized functions. Together with the development of the technology and associated spatial transcriptomics data sets, there has been a need to develop computational methods to analyze these types of data, in particular, to include the spatial information of each observation into the gene expression analyses. If the spatial information is ignored in the analysis, one can often only identify the different cell types in the tissue just as for disassociated single cell sequencing assays; the inclusion of the spatial information allows recovery of multi-cellular tissue structures and neighborhoods.

Current methods to analyze spatial transcriptomics data can roughly be split into two groups. One group of methods uses graph neural network models (GNNs) [6, 16, 5, 4] and the other group of methods uses probabilistic graphical models (PGMs) [19, 2, 8, 3, 17, 15]. The GNNs often have much faster computational time and therefore scale better to larger datasets, where the PGMs are more robust and interpretable. Many of the PGMs *bag* cells together, where each observation (e.g., cell at a single location) is an average of its own gene expression levels and the expression levels of its nearest neighbors. This means that much of the information from each individual cell is lost, including cell type information. Moreover, to reduce the computational time of current methods, genes are often consolidated into meta genes (or *eigengenes*) using PCA or another dimension reduction technique, which makes it difficult to recover distinct gene programs active in different parts of a tissue. Thus, a gap exists in developing a rigorous probabilistic model that will scale to existing data without resorting to approximations and data simplifications.

Here, we introduce a new method, neighborhood nonnegative matrix factorization (NNMF), that preserves the information of each cell and its complete gene expression profile, while remaining computationally tractable. This is possible because the spatial information of each cell is incorporated into the fast multiplicative updates of nonnegative matrix factorization (NMF) as an additional multiplicative update step. The method scales to millions of cells because of the simplicity of the estimation procedure, which only consist of multiplicative updates. The increasing scale of large spatial datasets including millions of cells, thousands of markers, and multiple samples makes tractable methods necessary, but it is important to maintain the complexity of the biology and the interpretability of the results. We show that NNMF preserves the gene expression for every cell during analysis, which makes it possible to identify the gene signatures active in each cell; moreover, we show that spatially-associated genes, tissue microenvironments, and shared neighborhoods across samples are naturally recovered without additional post processing steps.

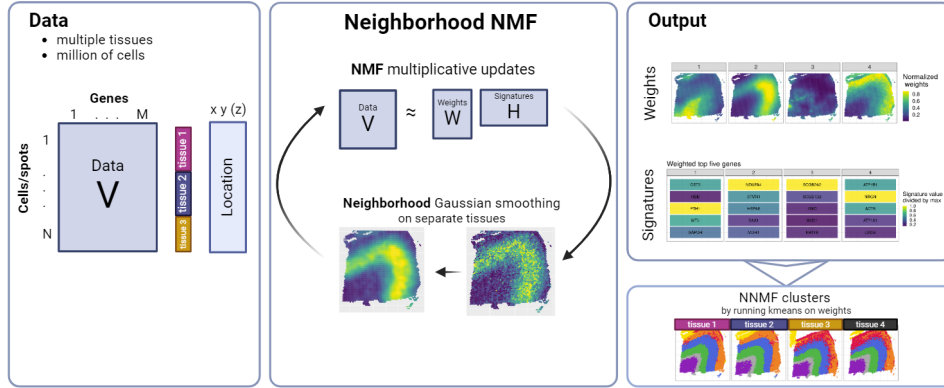


Figure 1: Overview of Neighborhood NMF

1.1 Overview of neighborhood nonnegative matrix factorization (NNMF)

Our method, neighborhood nonnegative matrix factorization (NNMF), is a simple and fast method that gives interpretable results for spatial genomics data. We explain the method for spatial single-cell transcriptomics data, but it can be applied to any type of data with a count vector and an associated 1D, 2D, 3D, or 4D location for each vector, including time-series single-cell RNA-sequencing data. NNMF decomposes a count matrix—here a matrix V of M gene transcript counts for each of N cells—into an $N \times K$ weight matrix W , and a $K \times M$ gene signature matrix H as shown in Figure 1. The dimension K is the assumed number of gene signatures active in the data, which is chosen magnitudes smaller than both M and N . We can interpret the rows of H to show which genes contribute to each of the K signatures and the columns of W as the signature weight across cells. The term *signature* therefore both refer to a specific set of genes and an associated weight across cells.

The signatures are recovered by combining the NMF multiplicative updates used to fit the underlying probabilistic Poisson model together with gaussian smoothing.

In NNMF, Gaussian smoothing through those multiplicative updates incorporates the location of the cells to encourage the gene weights to be correlated between neighboring cells (Figure 1; full details in Methods). With these smoothed signatures, we then run K-means clustering to label cells with similar signatures; if we only used the top ranked signature to label each cell, we would instead identify cells with one shared predominant gene signature, possibly marking a single cell type instead of a heterogeneous region with a collective function.

An advantage of our method is that we do not bag cells together, but maintain the distinct gene expression profile for each cell during the analysis, even when that profile contains a large number of zeros. Moreover, we do not subset the cells during fitting, but we fit each cell in these large datasets. NNMF outputs a predetermined

2. Results

number of gene signatures and their corresponding weights for each of the cells individually. Because we use probabilistic graphical models, we are able to identify the genes that compose each of the signatures and corresponding clusters, allowing us to give them meaningful labels. The gene signatures represent spatially variable gene programs that are active in different tissue regions. The factor matrix can be clustered and studied to identify regions of interest, including microenvironments and other cellular collectives.

2 Results

To show the behavior and performance of our method, we applied it to three spatial transcriptomics data sets, which are summarized in Figure 2A. The 10x Visium human brain data is useful to show how NNMF can be applied to multiple tissue samples and to compare results from our method against other related methods, and also against expert labels on the tissue itself. In the MERFISH mouse brain data the observations are single cells with annotated celltypes. This makes it possible to study how the celltypes are structured in the different spatial structures. Additionally this data set can also show how NNMF scales to large numbers of cells in three dimensions. The MERFISH human colorectal cancer data shows how NNMF scales to large numbers of cells across multiple samples and identifies known and new tumor-immune microenvironments in these samples.

A

Dataset	# Cells	# Genes	# Slices	Running time in minutes		
				NNMF	BASS	MENDER
DLPFC	14,364	6482	4	7	22	2
mouseMERFISH	50,627	154	8	4	81	1.5
ColonCancer	1,858,418	500	2	426 (\approx 7 hours)	-	-

B

Method	Framework	Program	Multiple samples	Gene signatures
NNMF	NMF + GP smoothing	R	✓	✓
NSF	NMF + GP prior + inducing points	Python	-	✓
BayesSpace	PCA reduced data (eigengenes) + PGM	R	-	-
Spatial LDA	Bag of cells + LDA	Python	-	✓
BASS	Bag of cells + PCA reduced data + Leiden clustering	R	✓	-
SpiceMix	NMF + HMRF	Python	-	✓
SOTIP	Bag of cells + PCA reduced data + Leiden clustering	Python	✓	-
SpaGCN	GNN	Python	(✓)	-
MENDER	GNN	Python	✓	-
CytoCommunity	GNN	Python	✓	-
STAGATE	GNN	Python	✓	-

Figure 2: **A) Overview of datasets and computational time.** **B) Overview of spatially-aware dimension reduction methods.** NMF: nonnegative matrix factorization; GP: Gaussian process; PCA: principal components analysis; HMRF: hidden Markov random field; PGM: probabilistic graphical model; GNN: graph neural network. *Multiple samples*: does the method handle multiple samples? *Gene signatures*: are gene signatures recovered from the method?

There already exist a number of methods in the literature to analyze spatial transcriptomics data and to understand how our method stands out compared to the current literature we created a general overview of the current methods in Table 2B.

Notice, that very few of the PGMs are able to run on multiple tissues at once, except for BASS and SOTIP. Though, these two methods both are bagging cells together and reducing the feature space of genes with PCA before identifying tissue structures. This means the expression of each specific gene is lost and need to be recovered by different post processing steps. So to our knowledge there are currently no other methods that runs on multiple sample and identify gene signatures.

We chose to compare our method to BASS and MENDER, which can both run on multiple samples. A further benchmarking of current methods in the literature can be found in [18], where they are comparing 14 different methods including BASS [8]. In [18] they are benchmarking on the same first two datasets included below, where BASS is among the best performing. MENDER [16] is one of the most recently developed methods that are highlighting speed and scalability, which is also one of the main advantages of our method. Another interesting thing to note is that most methods are implemented in Python, where our method stands out as one of the few that are available in R. And in particular the only package in R together with BASS that can run on multiple samples. As shown in [12] there is two times as many downstream analysis for spatial transcriptomics conducted in R compared to Python, which makes it essential to produce more and efficient packages for R as well.

In Table 2A there is a summary of the datasets included in this paper together with the running times of NNMF, BASS and MENDER. Our method is clearly faster than BASS, but MENDER is the fastest. Notice, that the NNMF method is much faster one the second dataset compared to the first even though there are many more observation, but this is because the number of genes are greatly reduced. As our method preserves the gene expression of each cell the running times will both depend on the number of cells and the number of genes included in the dataset. MENDER is also faster on the second dataset because the celltypes are known and it therefore only clusters based on the celltype annotations. As BASS is using a PCA reduction of the genes the speed mostly depend on the number of cells in the data, which is why the time approximately scales with the number of cells. In the following sections is shown results from NNMF on the three datasets described in Table 2A together with the results of BASS and MENDER for the first two datasets.

2.1 10x Visium data from human brain

The first data set we analyzed is the spatial transcriptomics dataset from the human dorsolateral prefrontal cortex (DLPFC) [10] created using the 10x Visium technology. Each observation represents a spot in a grid of the tissue, so each observation

2. Results

may include multiple cells of different types. The dataset is commonly used to verify the performance of de novo tissue structure labeling and is included in most of the papers accompanying the related methods in Table 2. Here, we analyzed the four tissues 151673 – 151676. We filtered out genes that was expressed in less than ten percent of the samples, which kept 6482 of the 33538 genes. We did apply our method to the data including all the genes, but this took much longer to run as our method preserves the information of all the genes in the data. The results in Figure 3A show that the reduction in the number of genes had no effect on the recovered tissue structures, which aligns well with the annotations.

In Figure 3A the results from applying NNMF to the DLPFC data is compared to results from BASS and MENDER. Specifically, we clustered the NNMF weights for the signatures resulting from running NNMF with rank 10 using K-means with $K = 7$, and we labeled each of the clusters with a single color. BASS uses cell bagging with eigengenes and Leiden clustering to identify the regions. MENDER uses a graph neural network. We found that NNMF replicated the results in the expert-annotated samples about as well as BASS, but in much less time. We found that MENDER was not able to annotate the samples well at all, besides the outer- and inner-most layers, despite the speed advantage. These results suggest that our approach identifies meaningful functional signatures and is competitive with the best existing methods.

Next, we wanted to study the cluster composition for the K-means clustering of our NNMF signatures (Figure 3B). Five of the seven the clusters include primarily a single signature (clusters 3-7); the remaining two clusters contain two or three approximately equally weighted primary signatures. Conversely, we see that the signatures 2, 4, 5 and 7 are spread across several clusters. The rest of the signatures are also mainly enriched in a single cluster, which shows that some of the signatures itself also recover the structure in the tissue. But because some of the signatures are present in several clusters it is essential to cluster the signatures to identify the cellular collectives with similar sets of gene signatures, possibly with different proportions.

We next studied the gene signatures themselves to determine how they were represented across the sample. While most of these gene signatures appear to capture cellular processes specific to a single region of the sample, the remainder (signatures 2, 4 and 7) of the signatures capture more general, diffuse processes that span regions of the sample, suggesting that these signatures are not region specific (Figure 3C). They appear to be present in cells across the entire sample except possibly the innermost layer. With the weights normalized, the gene signature with the largest value is less spatially consistent than the clustered annotation. These results suggest that gene signatures from NNMF on their own are not easily interpretable as region annotations.

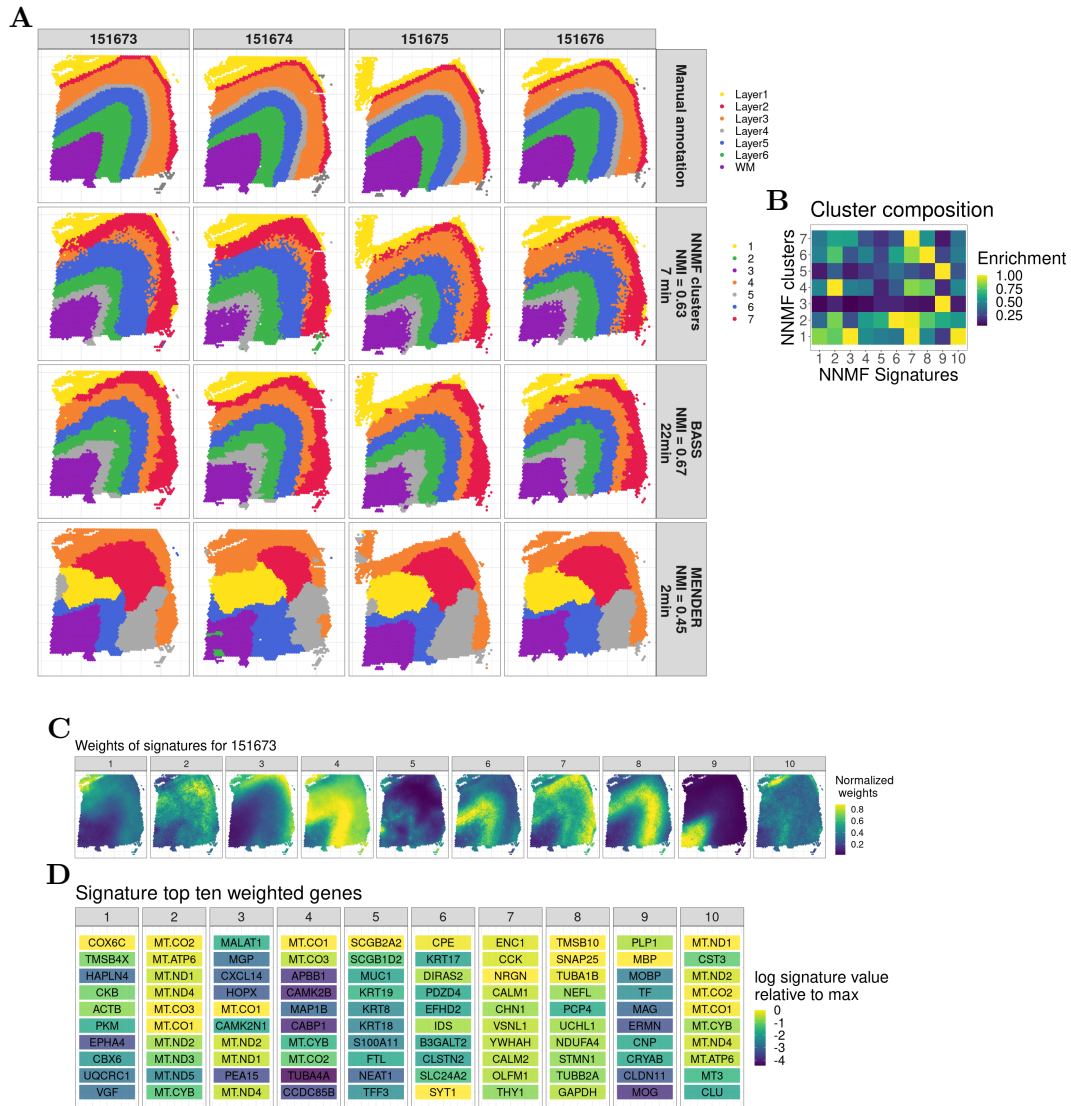


Figure 3: Results on the human dorsolateral prefrontal cortex (DLPFC) data. A) Annotations for (columns) each of the four samples for (rows) expert annotations, NMF clusters, BASS, and MENDER. B) NMF signature enrichment in each of the seven clusters annotated in A. C) NMF signature weights for a single sample. D) NMF signatures with the top ten genes for each signature; colors represent weight drop off relative to max weight.

2. Results

We next studied the genes with the ten largest weights for each gene signature (Figure 3D). These are identified by first scaling the original gene weights and afterwards identifying the top ten genes. The scaling is applied to identify more unique genes in the signatures similar to TF-IDF, which is further described in the Method section. The color gradient in Figure 3D indicate the log of the original weight of the specific gene.

With this key, we can understand what cellular processes these gene signatures capture. For example, signature 3, including *MALAT1* and *CXCL14* captures neuroinflammation. NMF and its parts-based decomposition appears useful for identifying sets of genes jointly participating in functional processes in subsets of cells. These ranked gene lists give us good insight into the cellular process documented in these gene signatures.

2.2 MERFISH data of mouse brain

The second dataset consists of eight adjacent tissue sections of the posterior part of the mouse hypothalamus from bregma -0.29mm to 0.06mm (Figure 4A). The dataset comes from the first animal in a prior analysis [11], which was created using the MERFISH technology. We refer the dataset as mouseMERFISH. For this technology each observation represents a single cell, and the observations are not on a grid as in the prior data. As the tissue sections are located adjacent to one another, we chose to analyze the slices in three dimensions. Each tissue section was aligned by the their minimum location on the x axis and maximum location on the y axis, and the Bregma is included as a z coordinate by multiplying it by 1000 to match the same μm scale as in each slice (Figure 4C).

The results from applying NNMf to the 3D mouseMERFISH dataset can be seen in Figure 4B and C. The results are constructed similar to the dataset above, but here we ran NNMf with rank 15 and afterwards used K-means with $K = 10$. We also recovered the results of BASS and MENDER in Figure 4D, for which we also assumed ten clusters. Comparing the results of the different methods to the illustrations in Figure 4A we see that NNMf is better at capturing the two parts of the periventricular nucleus and the circles on the top compared to BASS and MENDER. Both in Figure 4B and C it shows how the different clusters are changing and connected through the slices.

To further understand the different clusters we can take a look at the underlying gene signatures in Figure 4E together with their connection to the clusters described in Figure 4F. From Figure 4F we can see that many of the clusters mainly include a single signature except for clusters 4, 5 and 7, which are a combination of many different signatures. From Figure 4H, that shows the weights of the first four signatures, it is clear to see how some gene signatures (2 and 4) have a clear boundary of specific subsets where they are active, where signature 3 has more of a smooth

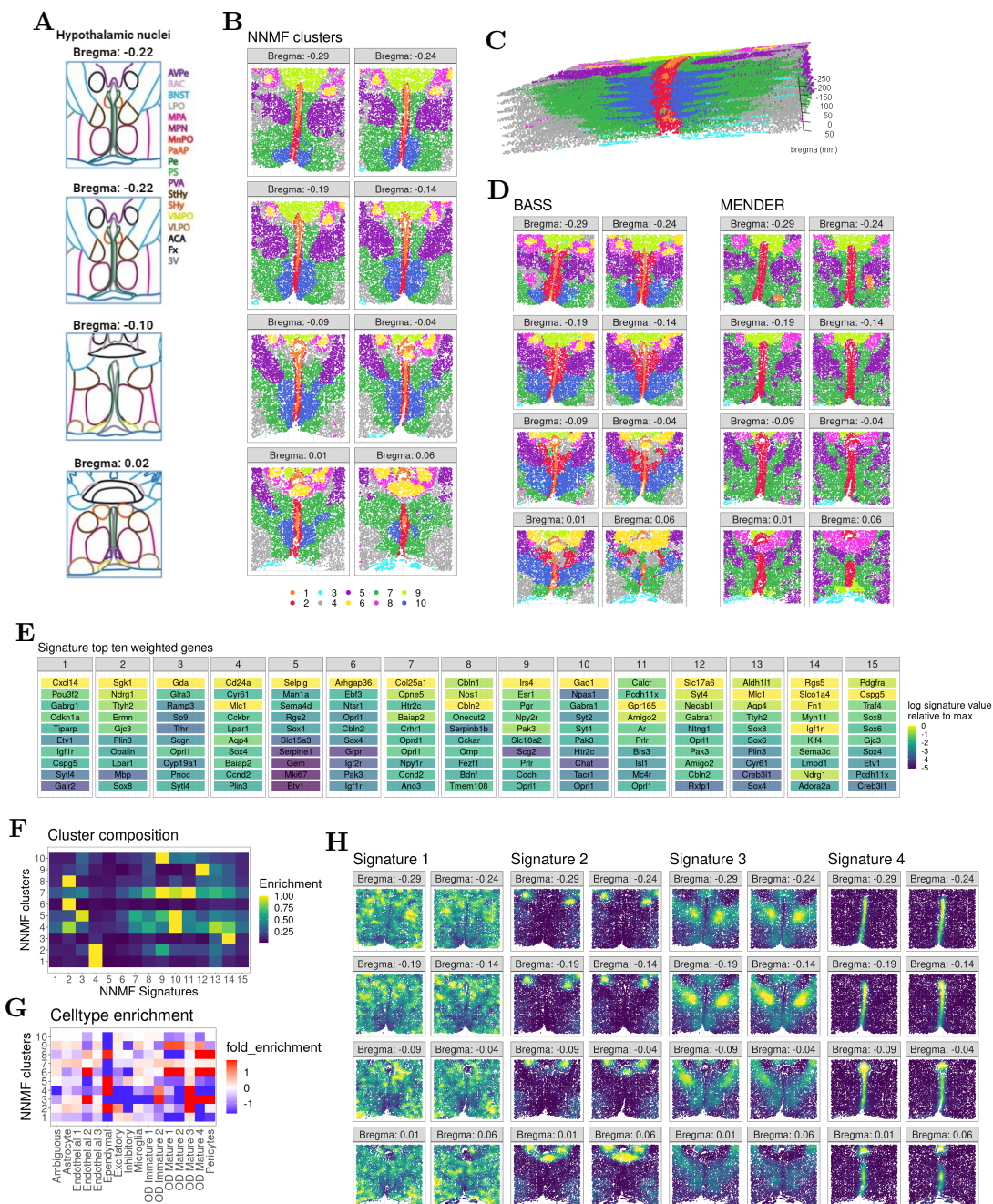


Figure 4: Results on the mouse hypothalamic preoptic region data generated using the MERFISH technology. Results are shown for the posterior eight slices from bregma -0.29 to 0.06. **A)** Illustration of mouse hypothalamic nuclei at different bregma from the original study [11]. The imaged regions are colored according to the legend on the right. The nuclei abbreviations in the legend are further described in the original study. **B)** The ten NNMF clusters that are constructed after running k-means on the resulting 15 signature weights. **C)** The 3D alignment of the different slices, where the scale are equal in all dimensions. **D)** The resulting clusters from BASS and MENDER from assuming ten clusters. **E)** The weighted top ten genes of the different signatures colored by their original weight in the corresponding signatures. **F)** The enrichment of the NNMF signatures (**E**) in the different NMF clusters (**B**). **G)** The fold enrichment of cell types in the different NNMF clusters (**B**). **H)** The weights for signature 1-4 over the different tissue slices, where yellow illustrates high values and dark blue low values.

3. Method

boundary and signature 1 is spread over the whole tissue. We see that the clusters 1 and 2, which capture the periventricular nucleus are completely enriched by signature 4, including *Mlc1* which suggest to be an integral membrane transporter. We also see from Figure 4G that especially cluster two have a high enrichment of ependymal and mature celltypes which aligns with the results found in [11].

2.3 MERFISH data of colon cancer

The last data set is also constructed using the MERFISH technology and includes expression of 500 genes in ≈ 1.9 million cells spread over two slices from two different tissues of colon cancer. Results on these data highlight the ability of NNMF to scale to a large numbers of cells across multiple samples with shared structures. The clusters shown in Figure 5A are found from running NNMF with 20 signatures and afterwards applying k-means with $K = 10$ on the weights of the signatures to find NNMF clusters annotated in the ten different colors.

For example signature 12 is an inflammatory region, because the top genes include *FOS*, *IL1B* and *PTGS2* (Figure 5D). The region of the signature are shown in Figure 5E and from Figure 5B and C it shows that this region is highly enriched by mastocytes, fibroblast and monocytes.

Signature 5 is also interesting because the top gene is *VEGFA*, which is usually upregulated in many known tumors. It is a member of the PDGF/VEGF growth factor family, and induces proliferation and migration of vascular endothelial cells.

Signature 10 and 17 are enriched in the same cluster 1 (Figure 5B), that are enriched by a combination of many different celltypes (Figure 5C) including T-cells. This suggest that cluster 1 is an anti-tumor region in the tissue.

3 Method

Nonnegative matrix factorization (NMF) takes a nonnegative data matrix V of dimension $N \times M$ and represents it by two smaller matrices of a much lower rank $K \ll \min(N, M)$, such that

$$V \approx WH$$

where W and H has dimension $N \times K$ and $K \times M$, respectively.

In the setting of spatially resolved transcriptomics data, then the V contains the gene expression counts of the different cells, which means N is the different cells and M is the number of genes. Each row in the matrix W will represent the weights for the different gene signatures for a specific cell and the rows in H will represent the different gene signatures, which are distributions over the genes. We therefore let the rows of H sum to one, which will remove the scaling ambiguity of NMF.

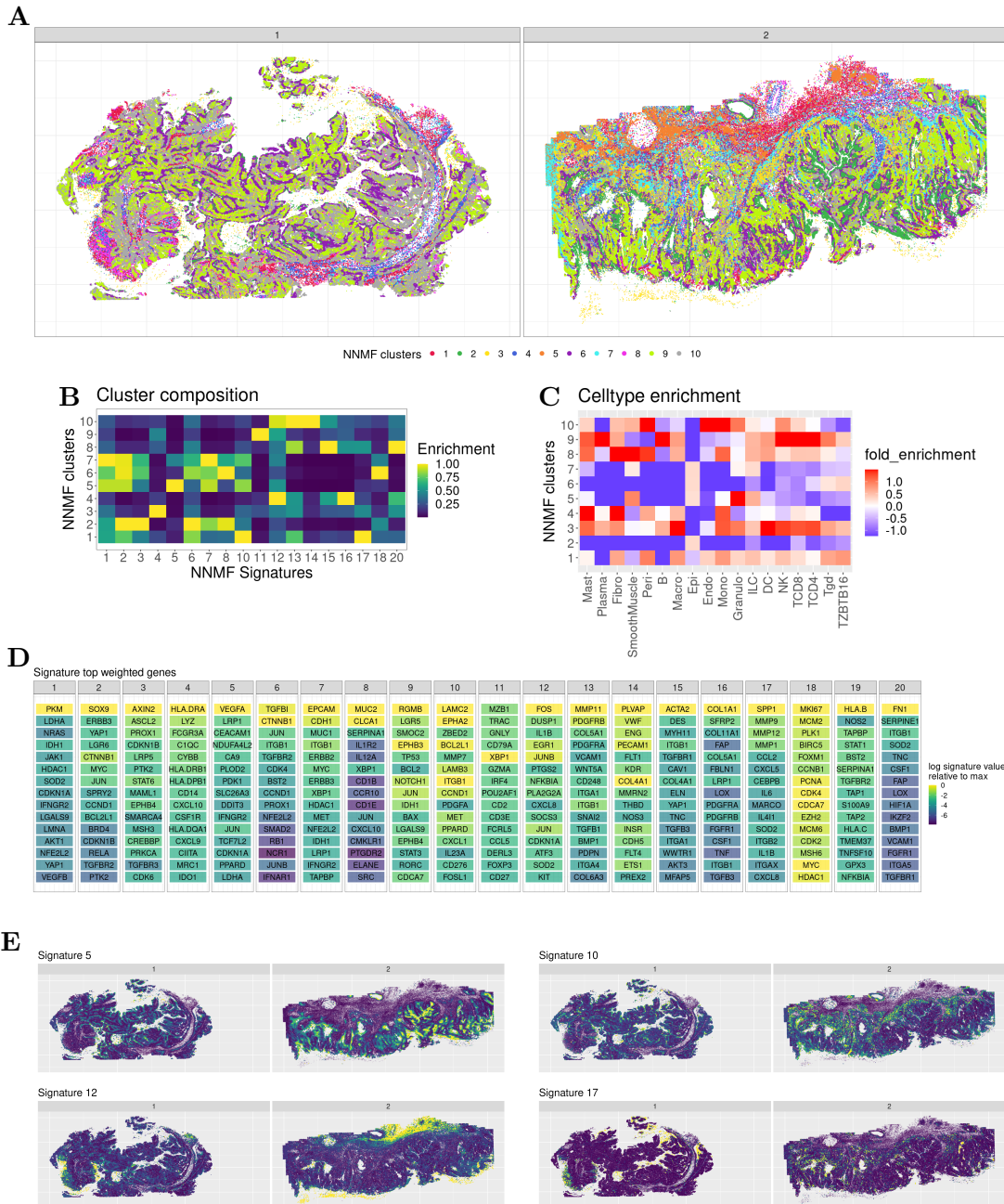


Figure 5: **Results for two colon cancer tissue samples.** **A)** Ten clusters constructed from applying k-means to the retrieved NNMf signature weights. **B)** The enrichment of the NNMf signatures (**D**) in the different NNMf clusters (**A**). **C)** The fold enrichment of celltypes in the NNMf clusters (**A**). **D)** The top 15 genes of the 20 retrieved signatures from NNMf. **E)** The signature weights over both tissues for the signatures 5, 10, 12 and 17.

3. Method

As the matrix consist of counts it is natural to assume a Poisson distribution for the entries of the matrix. The Poisson assumption is equivalent to recovering the factorization by minimizing the generalized kullback leibler (GKL), which is given by

$$D(V||WH) = \sum_{n=1}^N \sum_{m=1}^M \{V_{nm} \log V_{nm} - V_{nm} \log((WH)_{nm}) - V_{nm} + (WH)_{nm}\}. \quad (1)$$

Besides the gene counts each cell also have a location, which can be written as a matrix $X \in \mathbb{R}^{N \times d}$. The dimension d is typically 2 or 3. The goal is to make close cells share similar signatures to identify spatial microenvironments and structures in the tissue. This means we want a high correlation of the weights for near by cells, which can be described in terms of a Gaussian kernel function

$$\text{cor}(W_{i,:}, W_{j,:}) \approx \exp(-\frac{\|x_i - x_j\|_2^2}{\phi^2}) = (S_\phi(X))_{ij},$$

where x_i represents the location of cell i and ϕ is the length scale. The equation says the closer two cells are, the closer should the weights be i.e. higher correlation. Specifically, the matrix $S_\phi(X)$ represents the desired neighborhood correlation of the weights for the cells. The goal is to minimize the error function in (1), while imposing higher correlation of the weights for nearby cells. This is attained by adding Gaussian smoothing to the standard NMF multiplicative updates [7]. The updates for our method NNMF are therefore given by

$$H \leftarrow \text{rnorm} \left(H \otimes \left(W^T \frac{V}{WH} \right) \right) \quad (2)$$

$$W \leftarrow W \otimes \left(\frac{V}{WH} H^T \right) \quad (3)$$

$$W \leftarrow \text{rnorm}(S_\phi(X))W \quad (4)$$

where $\text{rnorm}(\cdot)$ define a matrix transformation that row normalize. This means $\text{rnorm}(S_\phi(X))$ is a normalized version of $S_\phi(X)$, where the rows sum to one to assure a neighborhood average. Here, the updates (2) and (3) are the standard NMF multiplicative updates under the Poisson model, where (4) is the additional update that incorporate the information of neighboring cells.

The simple extension of the standard updates comes from the fact that given a multivariate normal $Z \in \mathbb{R}^N$, where the covariance is the identity I_N we know that $S_\phi(X)Z$ has covariance $S_\phi(X)I_N S_\phi(X)^T = S_\phi(X)^2$. As the kernel $S_\phi(X)$ contain higher values for neighboring points we thereby know that the correlation will increase for neighboring points. In the middle of Figure 1 the effects of smoothing W once is illustrated, where it is clear that the neighboring correlation is increased. The following update of the gene signatures H will now correct for these spatial changes in W to construct more spatial gene signatures. The next update of W

will then update W correspondingly to minimize the GKL and again followed by a smoothing update to assure spatial weights. The results of these updates are therefore spatial weights W and corresponding gene signatures, that recovers the spatial gene structures in the tissue. To recover the NNMF clusters standard k-means is applied to the weights.

3.1 Initialization

As the multiplicative updates only assure a local minimum it is standard to initialize multiple times to increase the chance of a global minimum, which is also applied here. We initialize with a *warm start* by run standard NMF [7] 50-100 iterations for a set of initializations and choose the factorization with the smallest GKL to initialize NNMF. The number of initializations depends on the size of the dataset, but we use a minimum of 3 initializations.

3.2 Length scale

The length scale ϕ is estimated from the data by a grid search to find the size of neighborhood that best predict each of the cells. For a given vector $(\phi_i)_{i=1}^p$ of potential length scales the prediction error of each point based on the neighborhood is calculated in the following way

$$\text{Error}_i = \sum_{n,m} (V - \text{rnorm}(S_N(\phi_i) - I_N)V)_{nm}.$$

The optimal lengthscale is then chosen as the ϕ_i that had the corresponding smallest error value Error_i . This length scale is then fixed through the analysis.

3.3 Multiple slices and batching of large datasets

A feature of the package that makes it possible to run on multiple slices or very large slices is that one can specify a batch grouping of the dataset. In this case our method only calculates a neighborhood correlation $S_N^b(\phi)$ for each batch b , which makes it possible to run on multiple slices and larger datasets. This also means that the update step in (4) is preformed as a for loop over each batch b separately in the following way:

$$W^b \leftarrow \text{rnorm}(S_N^b(\phi))W^b \quad (5)$$

where W^b represent the weights of W that correspond to the observations in batch b .

In the case of multiple separate slices then the batch groups would specify which slice the different observations belong to. When separate slices do not have any spatial interaction it also makes sense only to calculate a neighborhood correlation

3. Method

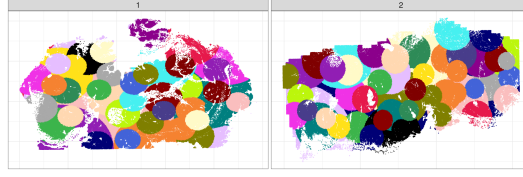


Figure 6: Illustration of batches with 20,000 cells constructed by the `groupondist()` function for the two colon cancer samples. Each color will represent a batch of cells.

matrix $S_\phi(X)$ for each slice. The batches of the first DLPFC dataset was therefore simply the four different slices 151673 - 151676.

In the mouse MERFISH data set the different slices actually did have spatial interaction and it therefore made sense to consider it as one large dataset in three dimensions. Though, calculating $S_\phi(X)$ for all the 50,627 cells together would require to calculate and save a matrix of $50,627^2 = 2,563,093,129$ entries, so instead we decided to split the data into two batches. One batch of slices for bregma above -0.1 and another for the rest below -0.1, such that four slices were included in both. This left 24,594 and 26,033 cells in each batch, respectively. This meant the total number of entries in $S_N^{<0.1}(\phi)$ and $S_N^{>0.1}(\phi)$ that now needed to be calculated and saved instead were $24,594^2 + 26,033^2 = 1,282,581,925$ which is around half of the size as having all the data together in one batch.

There is many ways to choose these batches and we have also created a function `groupondist()` that creates batches of observations close in space. This function was used on the two colon cancer samples to split the samples into batches of size 20,000 as these two slices include 840,405 and 1,018,013 cells, respectively. An example of batches is shown in Figure 6, where each color represent a batch of 20,000 cells. Notice that batches have different sizes because the density of cells differ over the tissue. The batches are constructed by recursively choosing a random cell in the tissue not already included in a batch and creating a new batch of its 20,000 nearest cells not already included in a batch, which of course also will include that cell itself. This is continued until the remaining cells are less than 20,000, which will be the last batch in the tissue.

The size of the batches can be determined by the user and should be made as large as possible where your server is able to save the matrices in memory. Even though the optimal solution would be to have all the cells together in one batch the option of batching makes it possible to incorporate spatial information to the many datasets that are evolving with millions of cells in each slice.

3.4 Computational efficiency

Besides the batching described above to reduce the memory usage, there is several things that makes our algorithm computationally powerfull. First of all it utilizes the

fast multiplicative updates from [7] and the spatial correlation matrix is incorporated without having to take the inverse, which is otherwise often required for spatial deconvolution [15]. This means that all the updates in our algorithm are vectorized to large matrix operations, which makes it highly efficient. To further improve the speed of our R package NNMF we have used Rcpp to implement the core algorithm in C++. All of these things combined makes our algorithm scalable to datasets of millions of cells.

3.5 Weighting of genes

Instead of simply showing the genes with the highest values in each of the gene signatures as the top genes we are weighting them to recover unique genes in the different gene signatures in a similar spirit to TF-IDF. Our weighting scheme is very similar to the one used in [13]. Given a signature k and a gene i , the new weighted signature value is given by:

$$W_{ki}^{new} = W_{ki} \cdot \log \left(1 + \frac{W_{ki}}{\max_{j \neq k} W_{ji}} \right)$$

This weighting will scale up a the weight of a gene if it is uniquely expressed in a single gene signature and in a similar way scale down the weights of genes that are expressed in several gene signatures.

3.6 Avoiding celltype bias

Before our analyses we normalize the gene expression count for each cell to be the median gene count across cells to avoid a bias towards certain celltypes. There are a huge difference in the size of different celltypes, which influence the gene count for different celltype. To avoid a bias in which cells that are fitted we therefore chose to normalize each cell to have the same gene count.

4 Discussion

Our method NNMF can analyze spatial transcriptomics data of million of cells to recover spatial gene signatures of tissue structure and multicellular microenviroments. The method only relies on the gene expression count matrix and the associated locations of each cell and is implemented as an R package that can be downloaded from github.com/ragnhildlaursen/NNMF. The method applies to multiple samples and millions of cells because the model is a simple extension of the fast multiplicative NMF updates from [7].

NNMF gives intepretable results of data as it preserves the gene expression of each individual cell, contrary to most method (including BASS and MENDER) in the literature today that are either bagging cells together or reducing the genes to a

4. References

fewer number of *eigengenes* or *metagenes* with PCA. To our knowledge NNMF is the only method that can run on multiple slices and millions of cells while preserving the gene expression of each individual cell. This makes it possible to recover the specific groups of genes that are active in the different part of the tissue and recover the activity of the different gene signatures for each cell individually.

After running k-means on the weights of the gene signatures we recover NNMF clusters that recover tissue structures comparable to other methods in the literature such as BASS or MENDER. Some of the gene signatures are only enriched in single structures of the tissue, where others are active over the whole tissue.

References

- [1] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090, 2015.
- [2] Z. Chen, I. Soifer, H. Hilton, L. Keren, and V. Jojic. Modeling multiplexed images with spatial-lda reveals novel tissue microenvironments. *Journal of Computational Biology*, 27(8):1204–1218, 2020.
- [3] B. Chidester, T. Zhou, S. Alam, and J. Ma. Spicemix enables integrative single-cell spatial modeling of cell identity. *Nature genetics*, 55(1):78–88, 2023.
- [4] K. Dong and S. Zhang. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature communications*, 13(1):1739, 2022.
- [5] J. Hu, X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li. Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, 18(11):1342–1351, 2021.
- [6] Y. Hu, J. Rong, Y. Xu, R. Xie, J. Peng, L. Gao, and K. Tan. Unsupervised and supervised discovery of tissue cellular neighborhoods from cell phenotypes. *Nature Methods*, 21(2):267–278, 2024.
- [7] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- [8] Z. Li and X. Zhou. Bass: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome biology*, 23(1):168, 2022.
- [9] E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, and L. Cai. Single-cell in situ rna profiling by sequential hybridization. *Nature methods*, 11(4):360–361, 2014.

- [10] K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.
- [11] J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416):eaau5324, 2018.
- [12] L. Moses and L. Pachter. Museum of spatial transcriptomics. *Nature methods*, 19(5):534–546, 2022.
- [13] K. Pelka, M. Hofree, J. H. Chen, S. Sarkizova, J. D. Pirl, V. Jorgji, A. Bejnood, D. Dionne, H. G. William, K. H. Xu, et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18):4734–4752, 2021.
- [14] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.
- [15] F. W. Townes and B. E. Engelhardt. Nonnegative spatial factorization applied to spatial genomics. *Nature methods*, 20(2):229–238, 2023.
- [16] Z. Yuan. Mender: fast and scalable tissue structure identification in spatial omics data. *Nature Communications*, 15(1):207, 2024.
- [17] Z. Yuan, Y. Li, M. Shi, F. Yang, J. Gao, J. Yao, and M. Q. Zhang. Sotip is a versatile method for microenvironment modeling with spatial omics data. *Nature Communications*, 13(1):7330, 2022.
- [18] Z. Yuan, F. Zhao, S. Lin, Y. Zhao, J. Yao, Y. Cui, X.-Y. Zhang, and Y. Zhao. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nature Methods*, 21(4):712–722, 2024.
- [19] E. Zhao, M. R. Stone, X. Ren, J. Guenthoer, K. S. Smythe, T. Pulliam, S. R. Williams, C. R. Uytingco, S. E. Taylor, P. Nghiem, et al. Spatial transcriptomics at subspot resolution with bayesspace. *Nature biotechnology*, 39(11):1375–1384, 2021.