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

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.

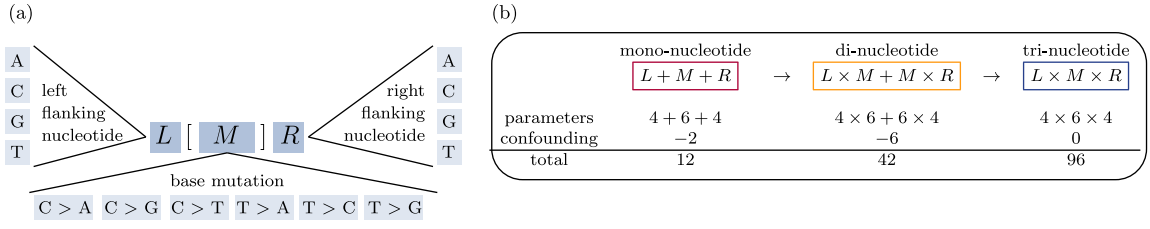


Figure 1: Graphical illustration of the parametrization of the mutation types. (a) The natural features for the mutation types are the left nucleotide L, right nucleotide R, and the base mutation M. (b) The three parametrizations we are analyzing in this paper for mutation types with one flanking nucleotide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

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

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 $h = (h_1, \ldots, h_T)$ 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

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

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

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

$$h_{\ell mr} = \frac{\exp\big(\beta^{L\times M\times R}_{\ell mr}\big)}{\sum_{\ell\in L}\sum_{m\in M}\sum_{r\in R} \exp\big(\beta^{L\times M\times R}_{\ell mr}\big)}, \tag{3}$$

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

$$h_{\ell mr} = \frac{\exp\big(\beta^M_m + \beta^L_\ell + \beta^R_r\big)}{\sum_{\ell\in L}\sum_{m\in M}\sum_{r\in R} \exp\big(\beta^M_m + \beta^L_\ell + \beta^R_r\big)}. \tag{4}$$

In order to avoid confounding we define $\beta^R_A = \beta^L_A = 0$. 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

$$
X = \begin{array}{c|cccccc|ccc|ccc}
 & \multicolumn{6}{c|}{\text{Mutation}} & \multicolumn{3}{c|}{\text{Left base}} & \multicolumn{3}{c}{\text{Right base}} \\
 & C>A & C>G & C>T & T>A & T>C & T>G & C & G & T & C & G & T \\
A[C>A]A & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
A[C>A]C & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
A[C>A]G & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\vdots & & & \vdots & & & & & \vdots & & & \vdots & \\
T[T>G]T & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1
\end{array} \tag{5}
$$

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

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

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	$L_2\times L_1 + L_1\times M\times R_1 + R_1\times R_2$	$96 + 12\times 2 = 120$

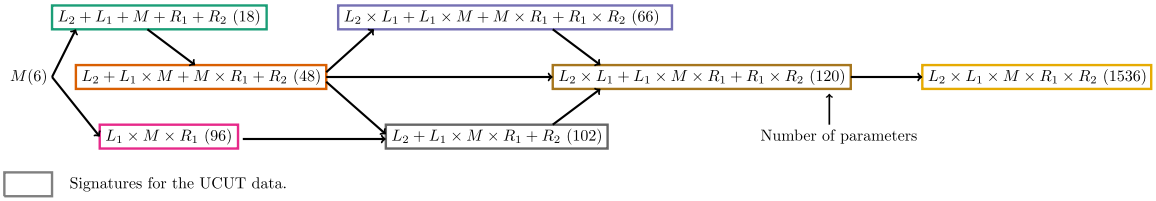


Figure 2: Factor diagram for the signatures used for the UCUT data set. The diagram shows the number of parameters for each signature and how the signatures are nested in each other.

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

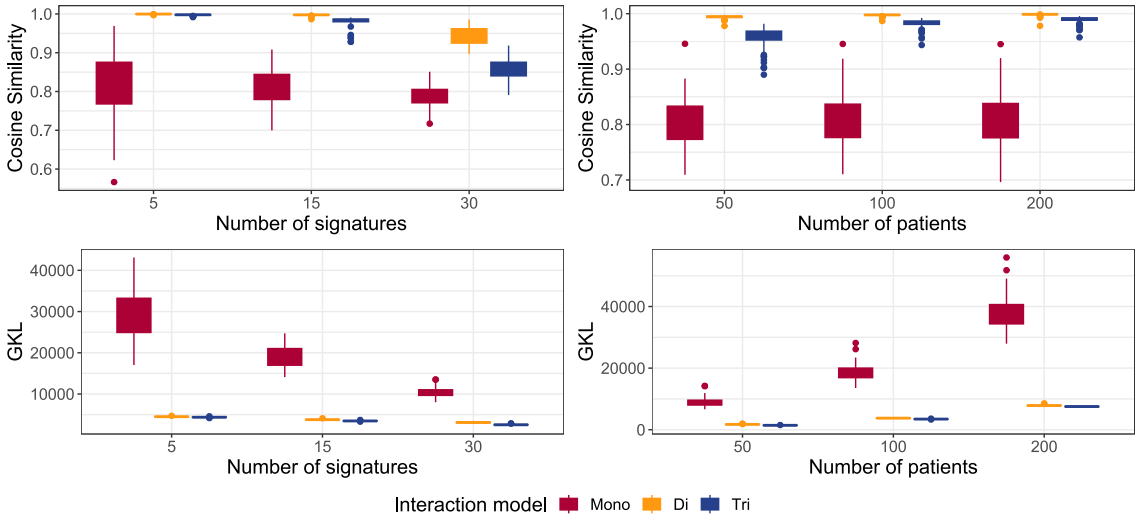


Figure 3: Simulating di-nucleotide signatures creating 100 different data sets for different number of patients and signatures. The figure both shows the reconstruction of the signatures through the average cosine similarity and the fit to data through the Generalized Kullback–Leibler divergence (GKL).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

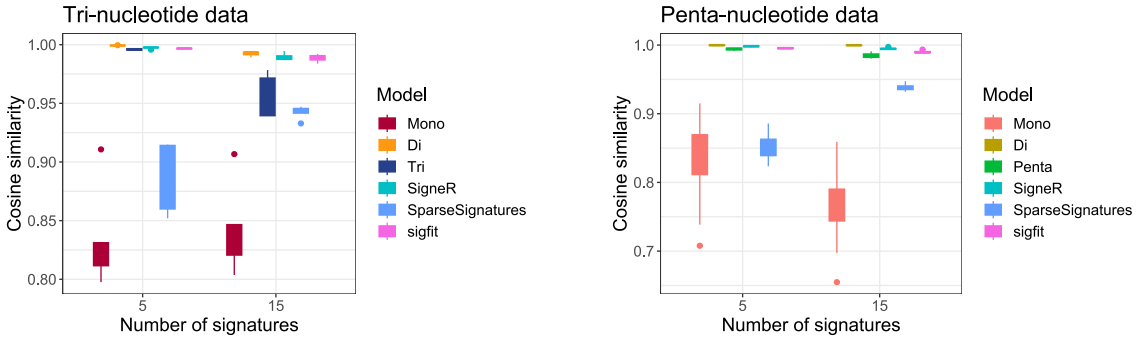


Figure 4: Comparing different methods for 10 datasets of 100 patients for 5 and 15 signatures. The methods `SigneR`, `SparseSignatures` and `sigfit` are run with their default implementaions. The two figures show the results for tri-nucleotide mutation types with only one flanking nucleotide *(left)* 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 $n_{\text{obs}} = T \times N = 96 \times 214 = 20,544$.

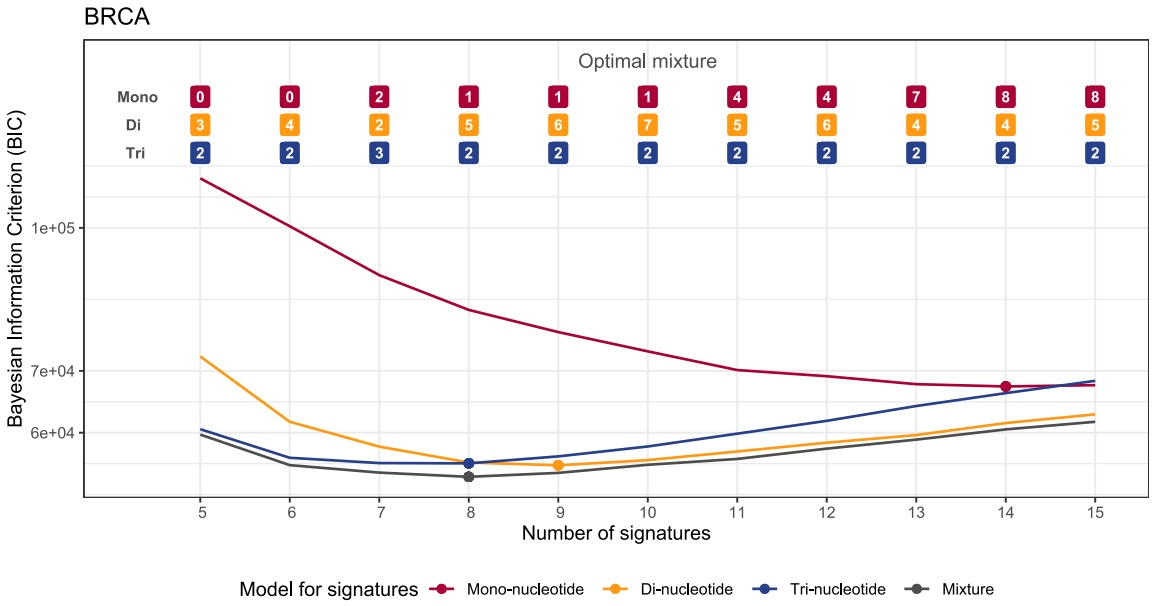


Figure 5: The Bayesian Information Criterion (BIC) for different number of signatures K for the BRCA dataset. The BIC is minimized for $K = 14$, $K = 9$ and $K = 8$ when all signatures are either mono-, di- or tri-nucleotide (dark red, yellow and blue curves). The BIC is minimized for $K = 8$ when the 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 parametrizations, but also the optimal mixture, where each signature can be any of the three parametrizations 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

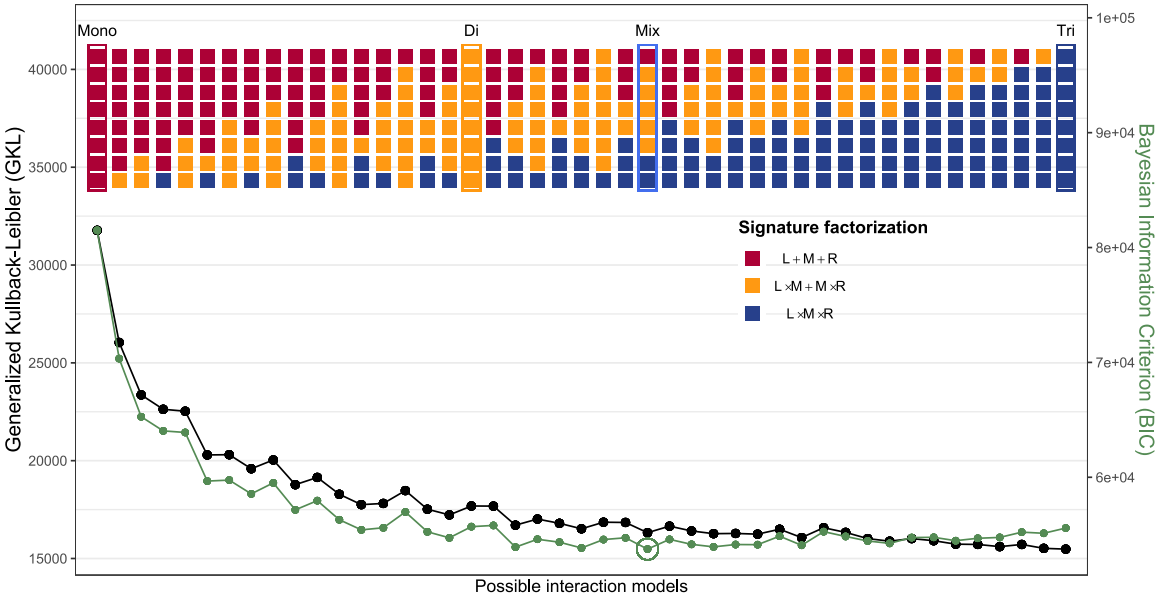


Figure 6: Fit to mutational count data from 214 breast cancer patients for all possible interaction models. The generalized Kullback–Leibler (GKL) and Bayesian Information Criterion (BIC) for all 45 models with $K = 8$ signatures. The models are ordered according to the total number of parameters for 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

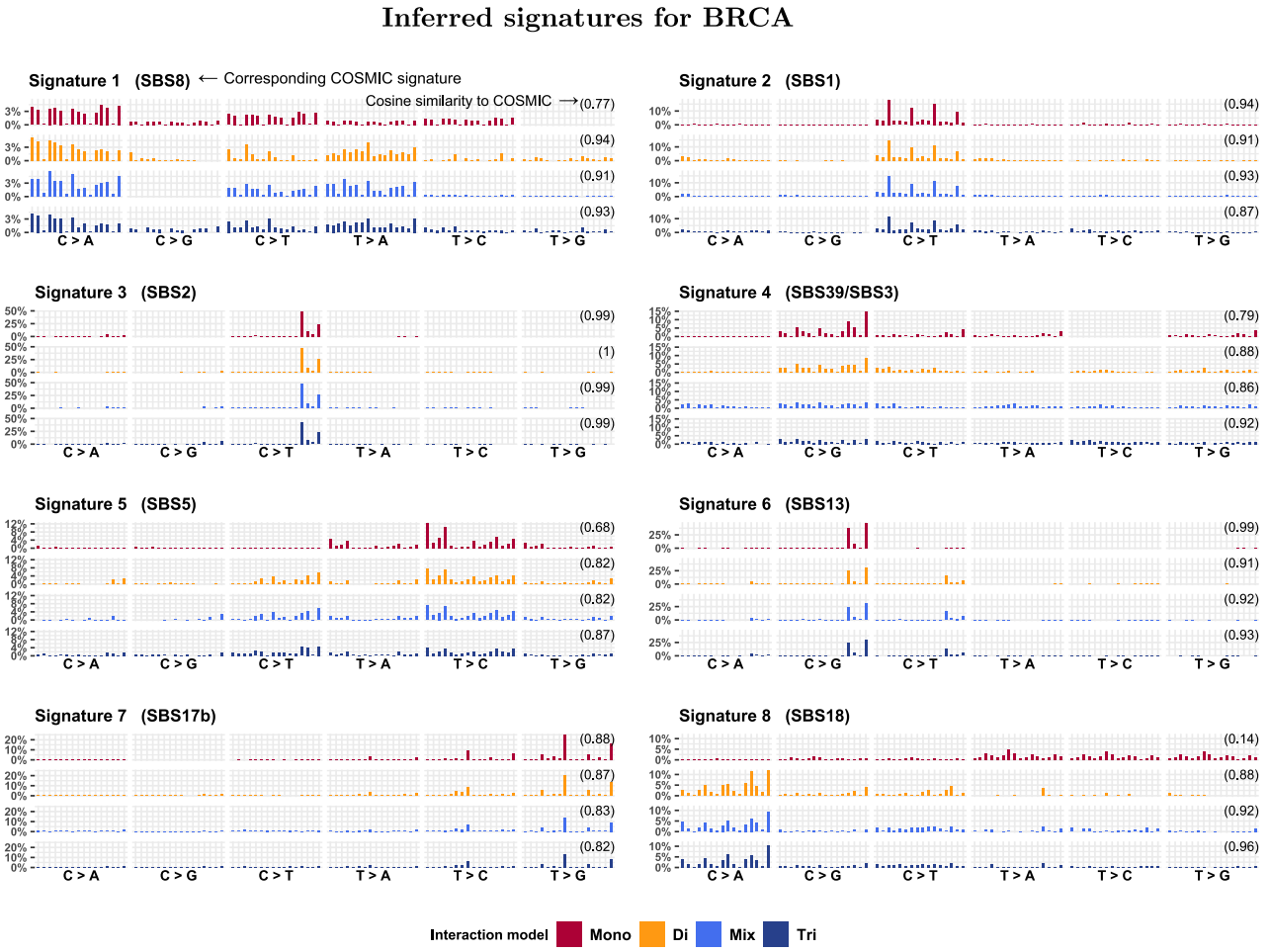


Figure 7: Inferred signatures for the BRCA data set. Comparison of the eight signatures for the four highlighted models in Figure 6. The four models are three parametrizations where all eight signatures are mono-nucleotide, di-nucleotide or tri-nucleotide, and for the mixture model the parametrization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

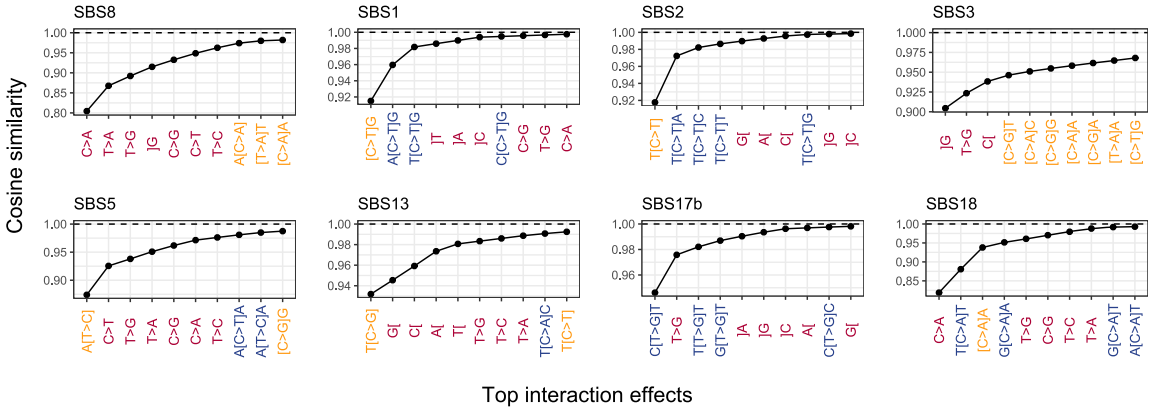


Figure 8: The top interactions for the eight COSMIC signatures found for the BRCA dataset. The top interactions are found with forward selection from the interaction making the largest increase in the cosine similarity to the COSMIC signature.

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. Similarily 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

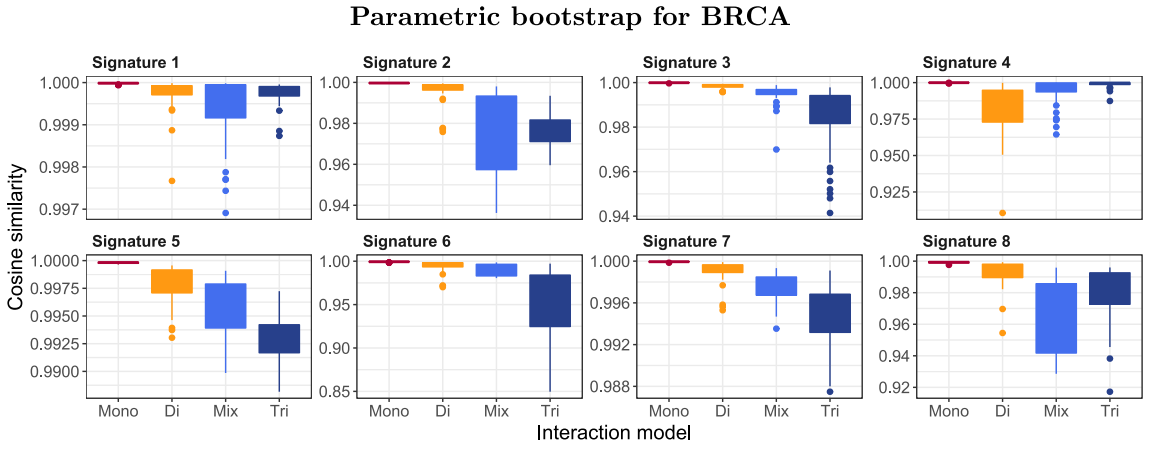


Figure 9: The cosine similarity for reconstructing the signatures with parametric bootstrapping for the BRCA data.

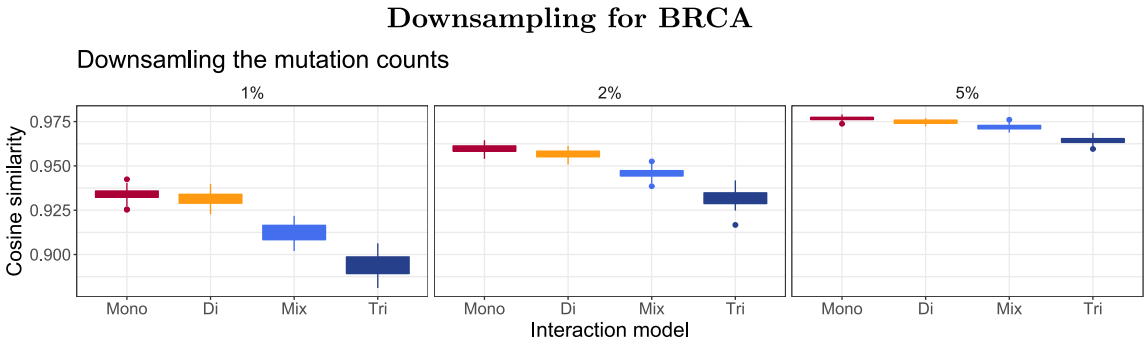


Figure 10: The mean cosine similarity between the recovered exposures from down-sampled BRCA data compared to the exposures from the original BRCA data.

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.

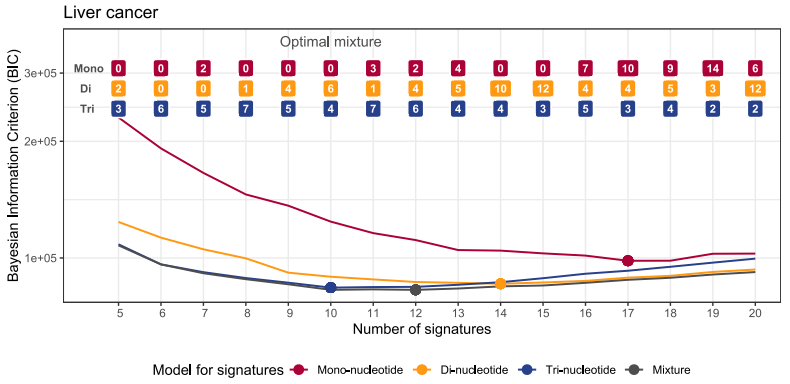
Finally, we use down-sampling to investigate the stability of the exposures for the different parametrizations of the signatures. We again compare the four different models Mono (8 mono-nucleotide interaction signatures), Di (8 di-nucleotide interaction signatures), Tri (8 tri-nucleotide interaction 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 by 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

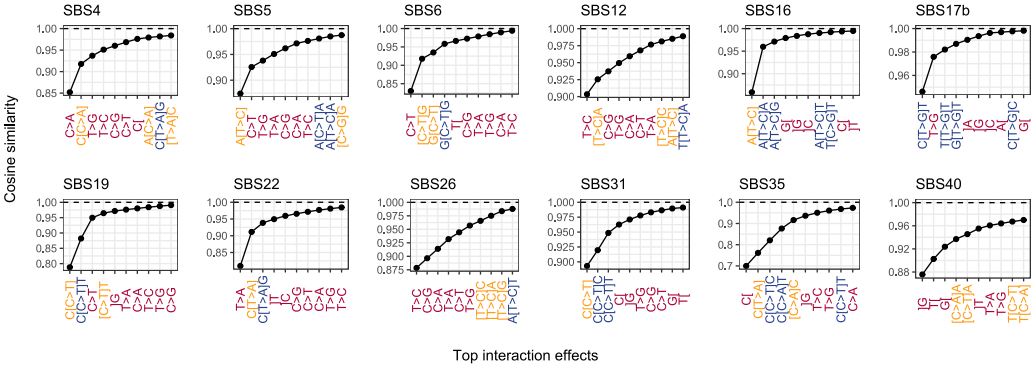
The UCUT data contains information about the two flanking bases at each side. The UCUT count matrix has $T = 6 \times 4^4 = 1536$ mutation types and $N = 26$ patients. The data consists of 14.715 somatic mutations, and the number of non-zero entries in the count matrix is $n_{\text{obs}} = 5260$.



(a)

COSMIC signatures	Parametrized models Mono	Di	Mix	Tri
SBS4	0.84	0.90	0.90	0.87
SBS5			0.86	
SBS6	0.95	0.9	0.89	0.92
SBS12		0.76	0.78	0.91
SBS16	0.93	0.97	0.95	0.96
SBS17b		0.79	0.86	
SBS19	0.99	0.99	0.98	0.98
SBS22	0.83			
SBS26		0.78		
SBS31		0.81		
SBS35			0.93	0.87
SBS40		0.76	0.79	0.85

(b)



(c)

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 parametrisations. 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

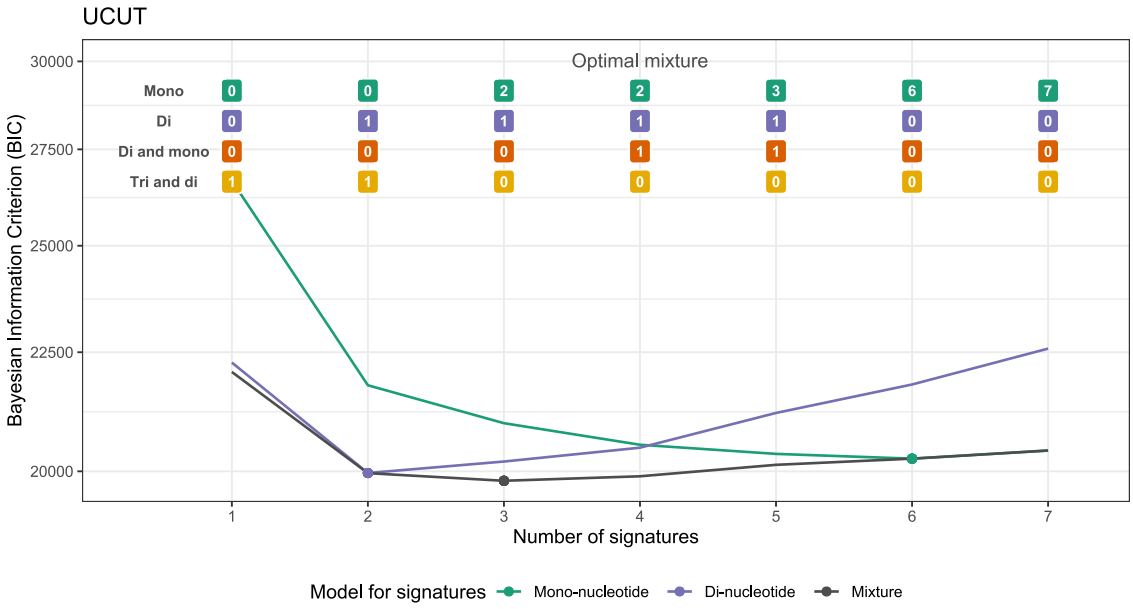


Figure 12: The Bayesian Information Criterion (BIC) for different number of signatures K to find the optimal number of signatures for the UCUT dataset. The top shows the optimal mixture of signature parametrizations for each number of signatures K.

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 $\triangle$BIC
$L_2 + L_1 + M + R_1 + R_2$	$2 \times 18 = 36$	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 parmetrizations 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 aristolocthic 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

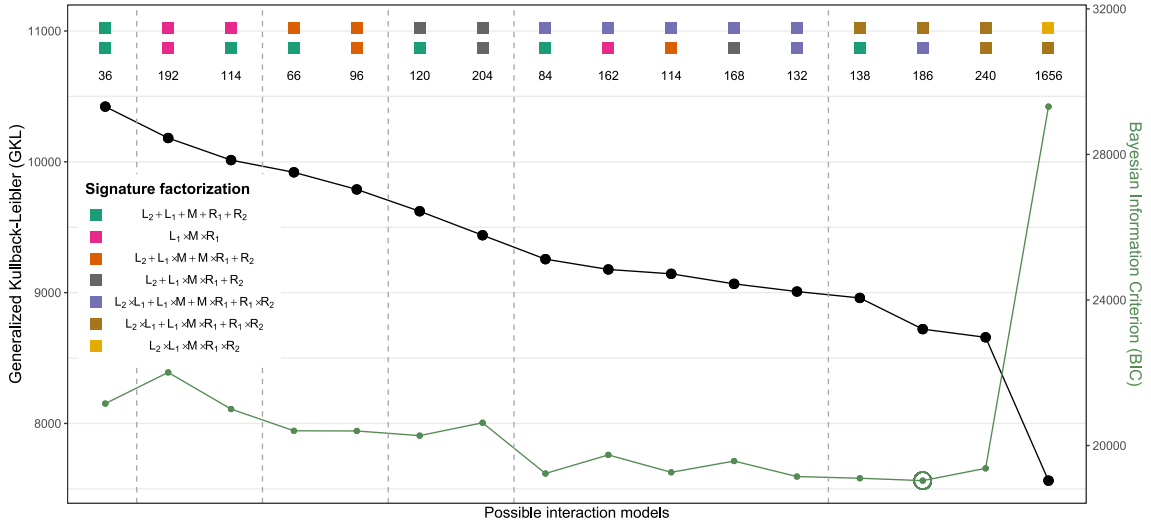


Figure 13: The Generalized Kullback–Leibler for 16 models with two signatures for the UCUT data set. The models are ordered according to GKL values, which also order the models by the first signature.

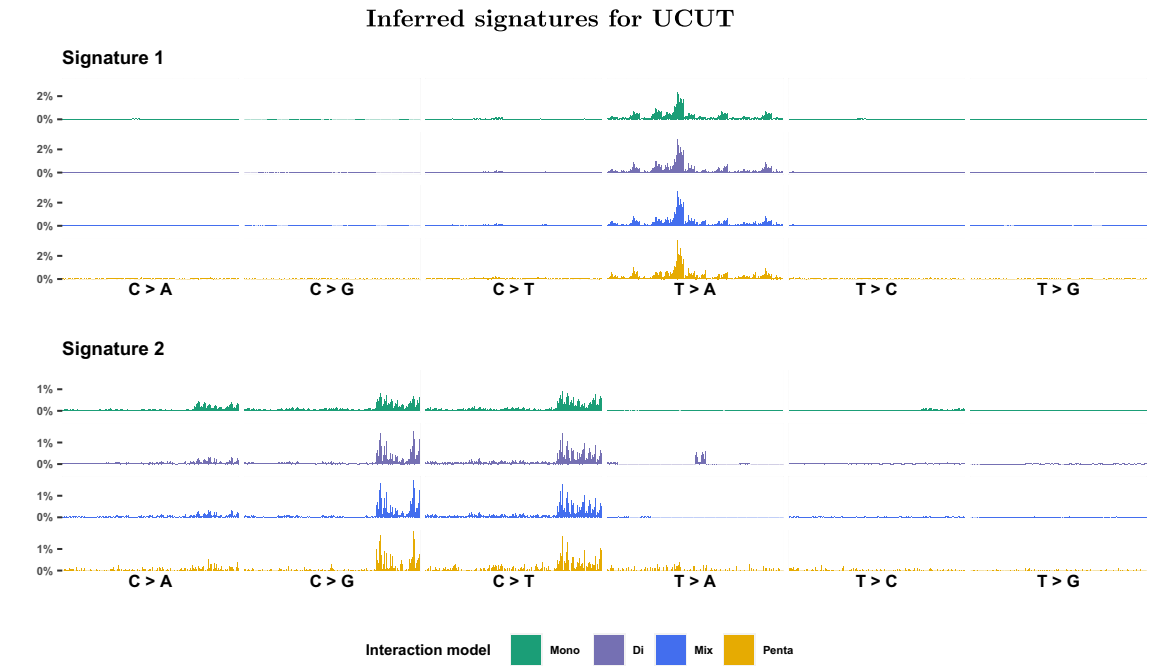


Figure 14: Inferred signatures for the UCUT data set. Comparison of the two signatures for the Mono, Di, Mix and Penta models.

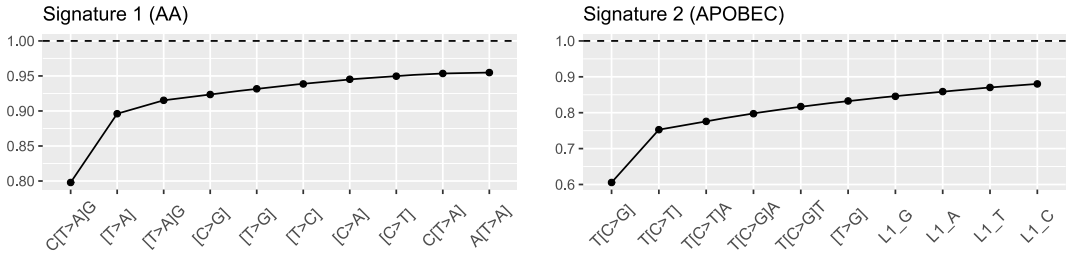


Figure 15: The top ten interactions that is increasing the cosine similarity to the retreived signatures.

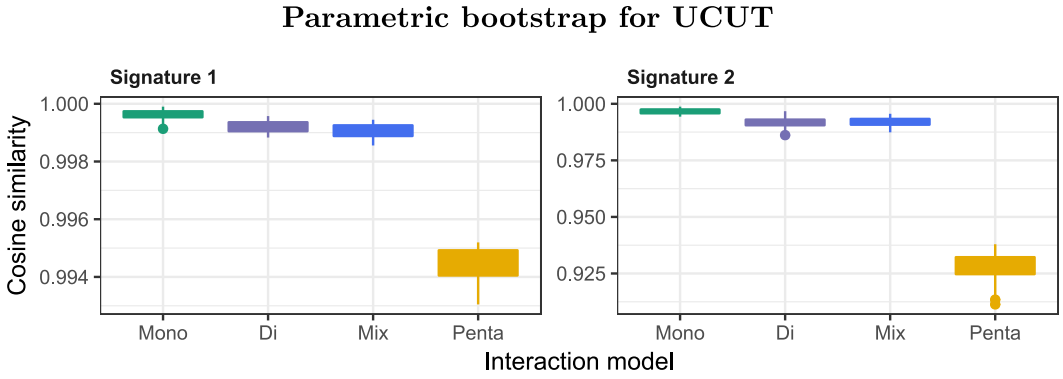


Figure 16: The cosine similarity for reconstructing the signatures with parametric bootstrap for the UCUT data.

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

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

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

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

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

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

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

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

The complete log-likelihood is given by

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

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

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 $\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\log(Z_{nkm}!)$.

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

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

which implies that

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

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

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

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

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

$$H^{i+1}_{kt} = \frac{\sum_{n=1}^{N}\mathbb{E}_{W^i,H^i}[Z_{nkt}|V]}{\sum_{t=1}^{T}\sum_{n'=1}^{N}\mathbb{E}_{W^i,H^i}\left[Z_{n'kt}|V\right]} = \frac{\sum_{n=1}^{N} V_{nt}\frac{W^i_{nk}H^i_{kt}}{(W^iH^i)_{nt}}}{\sum_{t=1}^{T}\sum_{n'=1}^{N} V_{n't}\frac{W^i_{n'k}H^i_{kt}}{(W^iH^i)_{n't}}} \tag{14}$$

and

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

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 $i = 1, 2, 3, \ldots$ **do**
  for $k = 1, \ldots, K$ **do**
    Update each signature

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

  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 $H_1, \ldots H_K$ requires a change in update (14) which was based on maximizing (12). The parametrization of the signatures are given by the design matrices $X_1, \ldots X_K$. 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

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

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

$$\log(y_t) = (X\beta)_t, \quad t = 1, \ldots, T, \tag{18}$$

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.

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

Given data matrix V, rank K, design matrices $X_1, \ldots, X_K$, and threshold ϵ.
Initialize W^1 and H^1 with random entries.
for $i = 1, 2, 3, \ldots$ **do**
  for $k = 1, \ldots, K$ **do**
    Update each signature

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

    Fit the log-linear Poisson regression

$$\log(\boldsymbol{y}_k^i) = X_k \beta_k^i \tag{19}$$

    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 $y = (y_1, \ldots, y_T)$ 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 $\exp\left(\sum_{s=1}^{S} X_{ts} \beta_s \right)$, $t = 1, \ldots, T$, 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

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

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

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

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.

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.