BMC Bioinformatics

RESEARCH Open Access

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

Marta Pelizzola[1*], Ragnhild Laursen[1] and Asger Hobolth[1]

*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

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 $V \in \mathbb{N}_0^{\mathrm{N}\times\mathrm{M}}$ into the product of two non-negative matrices $W \in \mathbb{R}_+^{\mathrm{N}\times\mathrm{K}}$ and $H \in \mathbb{R}_+^{\mathrm{K}\times\mathrm{M}}$ 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 $\mathrm{M} = 96$ (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]

$$V_{nm} \sim \mathrm{Po}((WH)_{nm}), \tag{1}$$

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.

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.

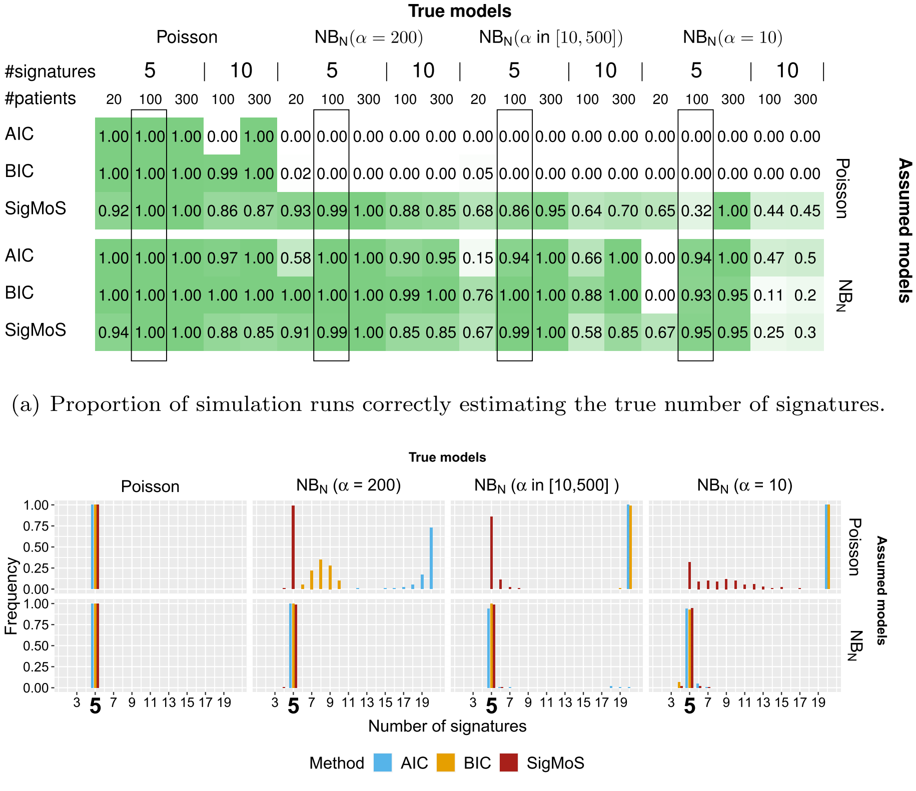
For simulated data we present a study on Negative Binomial simulated data with different levels of dispersion where results from AIC, BIC, `SigProfilerExtractor` [23], `SparseSignatures` [8], `SigneR` [13], `sigfit` [19] and `SignatureAnalyzer` [24] are compared with our proposed model selection procedure. 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.



True models

	Assumed models	Poisson					$\text{NB}_\text{N}(\alpha = 200)$					$\text{NB}_\text{N}(\alpha \text{ in } [10, 500])$					$\text{NB}_\text{N}(\alpha = 10)$				
#signatures		5			10		5			10		5			10		5			10	
#patients		20	100	300	100	300	20	100	300	100	300	20	100	300	100	300	20	100	300	100	300
AIC	Poisson	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BIC	Poisson	1.00	1.00	1.00	0.99	1.00	0.02	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SigMoS	Poisson	0.92	1.00	1.00	0.86	0.87	0.93	0.99	1.00	0.88	0.85	0.68	0.86	0.95	0.64	0.70	0.65	0.32	1.00	0.44	0.45
AIC	NB_N	1.00	1.00	1.00	0.97	1.00	0.58	1.00	1.00	0.90	0.95	0.15	0.94	1.00	0.66	1.00	0.00	0.94	1.00	0.47	0.5
BIC	NB_N	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.76	1.00	1.00	0.88	1.00	0.00	0.93	0.95	0.11	0.2
SigMoS	NB_N	0.94	1.00	1.00	0.88	0.85	0.91	0.99	1.00	0.85	0.85	0.67	0.99	1.00	0.58	0.85	0.67	0.95	0.95	0.25	0.3

(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

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

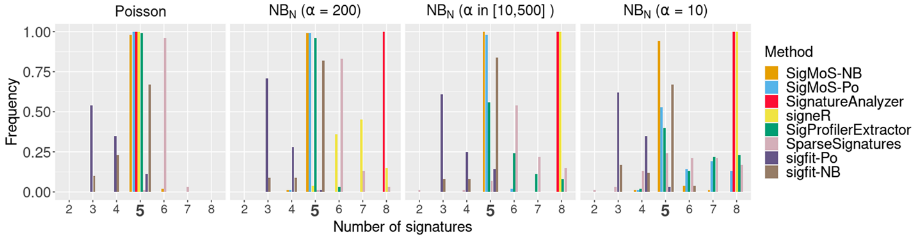


Fig. 2 Method comparison using simulated data. Each method is applied on the data sets from Fig. 1b and, for each data set, the value of the estimated number of signatures is kept. We test values for the number of signatures from two to eight for Poisson noise and Negative Binomial noise with $\alpha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

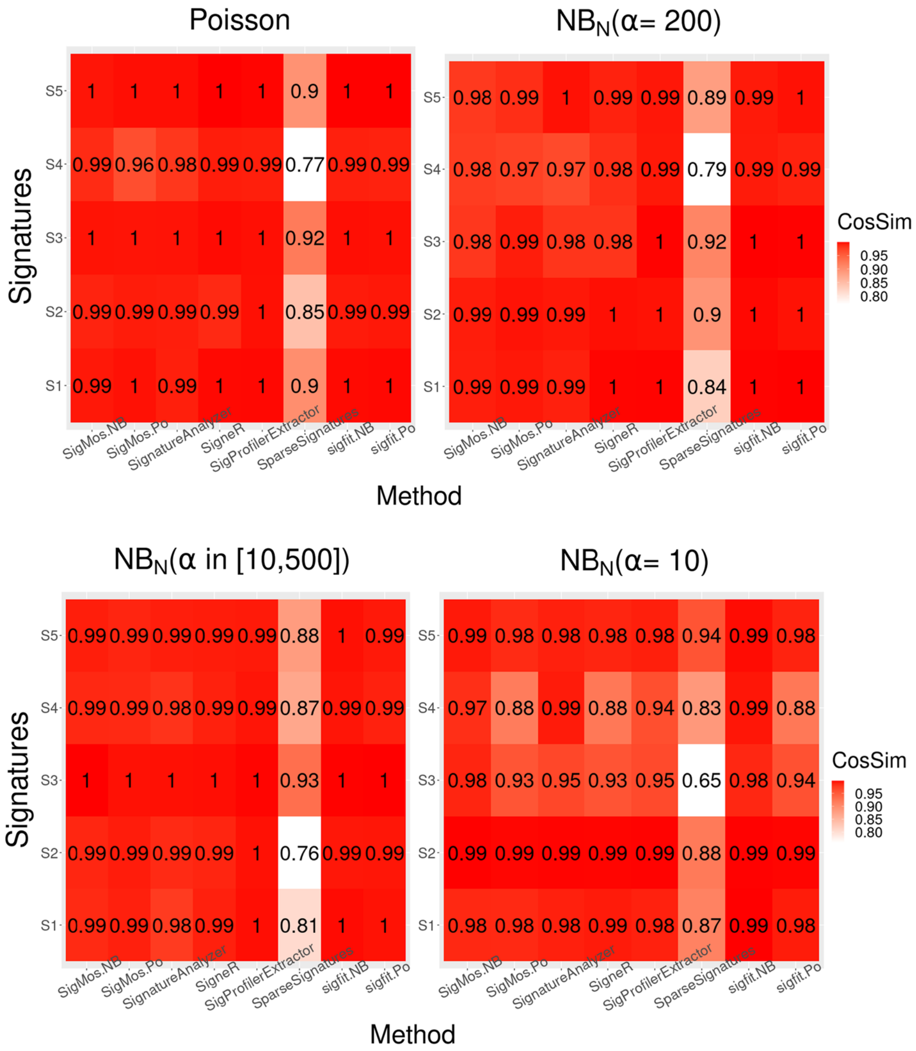


Fig. 3 Quality of estimated signatures using simulated data. Each method is applied on 10 randomly sampled data sets from Fig. 1b and, for each data set, the value of the estimated number of signatures is fixed to 5. We show the quality of the estimated signatures measured by cosine similarity for all 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 `SigneR` 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

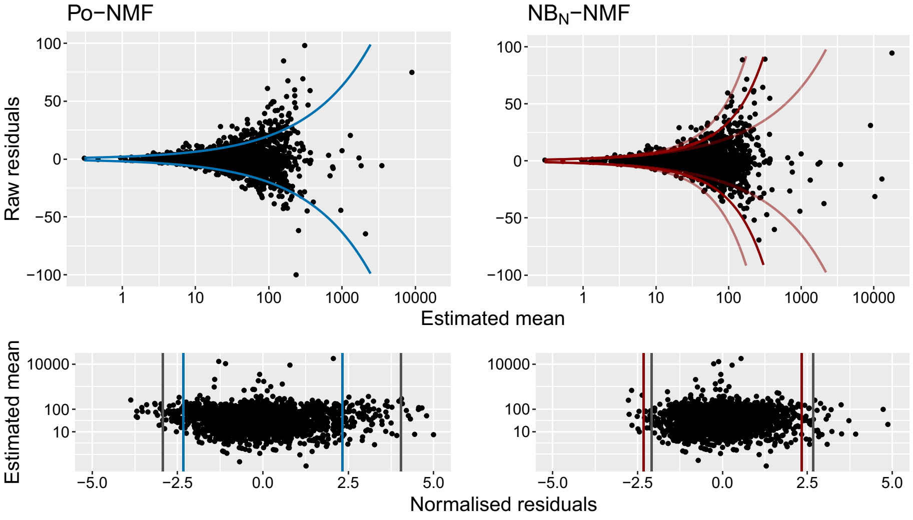
This data set consists of the mutational counts from the 21 breast cancer patients that has previously been described and analyzed in several papers [6, 7, 12]. The data can be found through the link ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl from [11] and have been extensively analyzed in [4].

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

	Assumed models	
Model selection procedure	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.

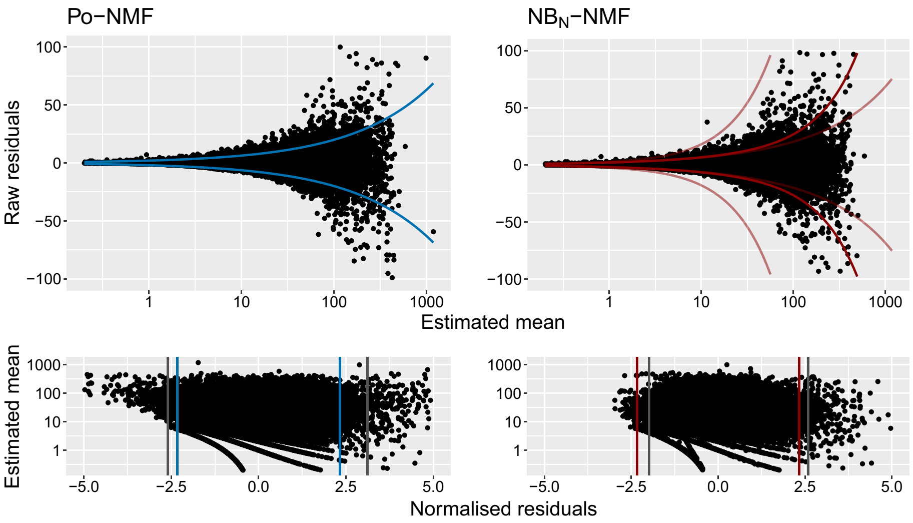
Table 1 Cosine similarity for the breast cancer data set between the signatures extracted by SigMoS and the ones in the COSMIC database

	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

	Assumed models	
Model selection procedure	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

	SBS1	***SBS5***	***SBS6***	***SBS8***	***SBS18***	**SBS36**	***SBS37***	***SBS40***
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.

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 $\mathrm{NB_N}$-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]

$$V_{nm} \sim \mathrm{Bin}(\tau, p). \tag{2}$$

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

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

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 $\beta >> \alpha$. As τ is large and p is small, the Binomial model is very well approximated by the Poisson model $\mathrm{Bin}(\tau, p) \backsimeq \mathrm{Pois}(\tau p)$. 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 $\beta >> \alpha$, 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

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

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 $\alpha \to \infty$ 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:

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

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

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

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

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

assuming fixed $\alpha_1, \ldots, \alpha_N$. 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 $\log\binom{\alpha_n + V_{nm} - 1}{\alpha_n}$ 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

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

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

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

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

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

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

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

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

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

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

In our application, we find maximum likelihood estimates (MLEs) of $\alpha_1, \ldots, \alpha_N$ 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 $\alpha_1, \ldots, \alpha_N$ 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 $\alpha_1 = \cdots = \alpha_N = \alpha$.

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 $\{\alpha_1, \ldots, \alpha_N\}$

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

$$\text{AIC} = -2 \ln L + 2n_{prm} \tag{14}$$

and the Bayesian Information Criterion (BIC)

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

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 $n_{obs} = \mathrm{N}$ where N is the number of patients, so $\ln(n_{obs}) > 2$ if $\mathrm{N} \geq 8$, 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:

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^j_{test} times the corresponding signatures estimated from the training set H^j_{train}. 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^j_{test} with respect to the signatures in H^j_{train}. We determine the order by calculating the cosine similarity between the signatures in H^j_{train} and those in H. We use the prediction of the test data to evaluate the model computing the distance between the true data V^j_{test} and their prediction $V^j_{predict}$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: $c_{median}$
$W, H \leftarrow$ apply the chosen NMF method to $V$ with K signatures
for $j = 1$ to $J$ do
    $V^j_{train} \leftarrow$ mutational counts for the patients in the $j^{th}$ training set
    $V^j_{test} \leftarrow V \setminus V^j_{train}$
    $W^j_{test} \leftarrow$ exposures from $W$ for the patients in the test set
    $W^j_{train}, H^j_{train} \leftarrow$ apply the chosen NMF method to $V^j_{train}$ with K signatures
    $c_j \leftarrow cost(V^j_{test}, W^j_{test} H^j_{train})$
end for
return $c_{median} \leftarrow median(c_1, \dots, c_J)$
```

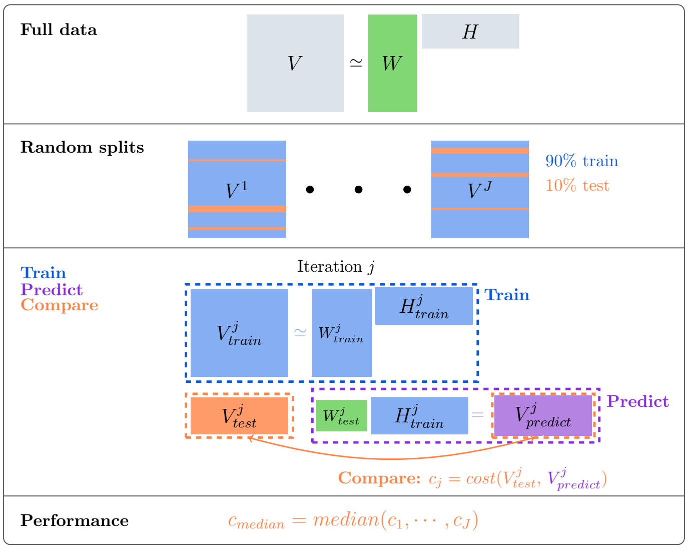

Full data
$V \simeq W$
H
Random splits
V^1
• • •
V^J
90% train
10% test
Train
Predict
Compare
Iteration j
$V^j_{train} \simeq W^j_{train}$
H^j_{train}
Train
V^j_{test}
W^j_{test} $H^j_{train} = V^j_{predict}$
Predict
Compare: $c_j = cost(V^j_{test}, V^j_{predict})$
Performance
$c_{median} = median(c_1, \cdots, c_J)$

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.

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, Børresen-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, Syulyukina 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. Degasperi 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, Broeks 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, Pivot 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.
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. Lochovsky 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.

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.