Using the Package NMF

Version 0.8.2

Renaud Gaujoux, renaud@cbio.uct.ac.za

June 13, 2012

This vignette presents the package NMF [**Gaujoux2010**], which implements a framework for Nonnegative Matrix Factorization (NMF) algorithms in R [**R**]. The objective is to provide an implementation of some standard algorithms, while allowing the user to easily implement new methods that integrate into the package's framework.

To cite the package *NMF* package in publications please use:
Gaujoux2010

The latest stable version of the *NMF* package package can be installed from any CRAN repository mirror, and loaded with the standard calls to `library`.

```
# Install
install.packages("NMF")
# Load
library(NMF)
```

Contents

1 Overview

1.1 Package features

This section provides a quick overview of the *NMF* package package's features. Section 2 provides more details, as well as sample code on how to actually perform common tasks in NMF analysis.

The *NMF* package package provides:

- 8 built-in algorithms;
- 4 built-in seeding methods;
- Single interface to perform all algorithms, and combine them with the seeding methods;
- Provides a common framework to test, compare and develop NMF methods;
- Accept custom algorithms and seeding methods;
- Plotting utility functions to visualize and help in the interpretation of the results;
- Transparent parallel computations;
- Optimized and memory efficient C++ implementations of the standard algorithms;
- Optional layer for bioinformatics using BioConductor [**Gentleman2004**];

1.2 Nonnegative Matrix Factorization

This section gives a formal definition for Nonnegative Matrix Factorization problems, and defines the notations used throughout the vignette.

Let X be a $n \times p$ non-negative matrix, (i.e with $x_{ij} \geq 0$, denoted $X \geq 0$), and $r > 0$ an integer. Non-negative Matrix Factorization (NMF) consists in finding an approximation

$$X \approx WH \ , \tag{1}$$

where W, H are $n \times r$ and $r \times p$ non-negative matrices, respectively. In practice, the factorization rank r is often chosen such that $r \ll \min(n, p)$. The objective behind this choice is to summarize and split the information contained in X into r factors: the columns of W.

Depending on the application field, these factors are given different names: basis images, metagenes, source signals. In this vignette we equivalently and alternatively use the terms *basis matrix* or *metagenes* to refer to matrix W, and *mixture coefficient matrix* and *metagene expression profiles* to refer to matrix H.

The main approach to NMF is to estimate matrices W and H as a local minimum:

$$\min_{W,H \geq 0} \underbrace{[D(X, WH) + R(W, H)]}_{=F(W,H)} \tag{2}$$

where

- D is a loss function that measures the quality of the approximation. Common loss functions are based on either the Frobenius distance

$$D : A, B \mapsto \frac{Tr(AB^t)}{2} = \frac{1}{2} \sum_{ij} (a_{ij} - b_{ij})^2,$$

  or the Kullback-Leibler divergence.

$$D : A, B \mapsto KL(A||B) = \sum_{i,j} a_{ij} \log \frac{a_{ij}}{b_{ij}} - a_{ij} + b_{ij}.$$

- R is an optional regularization function, defined to enforce desirable properties on matrices W and H, such as smoothness or sparsity [**Cichocki04**].

1.3 Algorithms

NMF algorithms generally solve problem (2) iteratively, by building a sequence of matrices (W_k, H_k) that reduces at each step the value of the objective function F. Beside some variations in the specification of F, they also differ in the optimization techniques that are used to compute the updates for (W_k, H_k).

For reviews on NMF algorithms see [**Berry06**, **Chu2004**] and references therein.

The *NMF* package package implements a number of published algorithms, and provides a general framework to implement other ones. Table 1 gives a short description of each one of the built-in algorithms:

The built-in algorithms are listed or retrieved with function `nmfAlgorithm`. A given algorithm is retrieved by its name (a `character` key), that is partially matched against the list of available algorithms:

```
# list all available algorithms
nmfAlgorithm()

##   brunet      lee   offset    nsNMF   ls-nmf   pe-nmf
## "brunet"    "lee" "offset"  "nsNMF" "ls-nmf" "pe-nmf"
##   snmf/r   snmf/l
## "snmf/r" "snmf/l"

# retrieve a specific algorithm: 'brunet'
nmfAlgorithm("brunet")

## <object of class:  NMFStrategyIterative >
## name:  brunet
## objective:  'KL'
## NMF model:  NMFstd
## <Iterative schema:>
## Update :  function (i, v, data, copy = FALSE, eps = .Machine$double.eps,
## Stop :  'connectivity'
## WrapNMF :  ''

# partial match is also fine
identical(nmfAlgorithm("br"), nmfAlgorithm("brunet"))

## [1] TRUE
```

1.4 Initialization: seeding methods

NMF algorithms need to be initialized with a seed (i.e. a value for W_0 and/or H_0[1]), from which to start the iteration process. Because there is no global minimization algorithm, and due to the problem's high dimensionality, the choice of the initialization is in fact very important to ensure meaningful results.

The more common seeding method is to use a random starting point, where the entries of W and/or H are drawn from a uniform distribution, usually within the same range as the target matrix's entries. This method is very simple to implement. However, a drawback is that to achieve stability one has to perform multiple runs, each with a different starting point. This significantly increases the computation time needed to obtain the desired factorization.

To tackle this problem, some methods have been proposed so as to compute a reasonable starting point from the target matrix itself. Their objective is to produce deterministic algorithms that need to run only once, still giving meaningful results.

[1] Some algorithms only need one matrix factor (either W or H) to be initialized. See for example the SNMF/R(L) algorithm of Kim and Park [4].

Key	Description
`brunet`	Standard NMF. Based on Kullback-Leibler divergence, it uses simple multiplicative updates from [5], enhanced to avoid numerical underflow. $$H_{kj} \leftarrow H_{kj} \frac{\left(\sum_l \frac{W_{lk} V_{lj}}{(WH)_{lj}}\right)}{\sum_l W_{lk}} \quad (3)$$ $$W_{ik} \leftarrow W_{ik} \frac{\sum_l [H_{kl} A_{il}/(WH)_{il}]}{\sum_l H_{kl}} \quad (4)$$ **Reference:** [**Brunet04**]
`lee`	Standard NMF. Based on euclidean distance, it uses simple multiplicative updates $$H_{kj} \leftarrow H_{kj} \frac{(W^T V)_{kj}}{(W^T W H)_{kj}} \quad (5)$$ $$W_{ik} \leftarrow W_{ik} \frac{(V H^T)_{ik}}{(W H H^T)_{ik}} \quad (6)$$ **Reference:** [5]
`nsNMF`	Non-smooth NMF. Uses a modified version of Lee and Seung's multiplicative updates for Kullback-Leibler divergence to fit a extension of the standard NMF model. It is meant to give sparser results. **Reference:** [**nsNMF2006**]
`offset`	Uses a modified version of Lee and Seung's multiplicative updates for euclidean distance, to fit a NMF model that includes an intercept. **Reference:** [1]
`pe-nmf`	Pattern-Expression NMF. Uses multiplicative updates to minimize an objective function based on the Euclidean distance and regularized for effective expression of patterns with basis vectors. **Reference:** [**Zhang2008**]
`snmf/r`, `snmf/l`	Alternating Least Square (ALS) approach. It is meant to be very fast compared to other approaches. **Reference:** [4]

Table 1: Description of the implemented NMF algorithms. The first column gives the key to use in the call to the `nmf` function.

For a review on some existing NMF initializations see [**Albright2006**] and references therein.

The *NMF* package package implements a number of already published seeding methods, and provides a general framework to implement other ones. Table 2 gives a short description of each one of the built-in seeding methods:

The built-in seeding methods are listed or retrieved with function `nmfSeed`. A given seeding method is retrieved by its name (a `character` key) that is partially matched against the list of available seeding methods:

```
# list all available seeding methods
nmfSeed()

## [1] "none"   "random" "ica"    "nndsvd"

# retrieve a specific method: 'nndsvd'
```

```
nmfSeed("nndsvd")

## <object of class:  NMFSeed >
## name:  nndsvd
## method:  <function>

# partial match is also fine
identical(nmfSeed("nn"), nmfSeed("nndsvd"))

## [1] TRUE
```

Key	Description
`ica`	Uses the result of an Independent Component Analysis (ICA) (from the `fastICA` package). Only the positive part of the result are used to initialize the factors.
`nnsvd`	Nonnegative Double Singular Value Decomposition. The basic algorithm contains no randomization and is based on two SVD processes, one approximating the data matrix, the other approximating positive sections of the resulting partial SVD factors utilizing an algebraic property of unit rank matrices. It is well suited to initialize NMF algorithms with sparse factors. Simple practical variants of the algorithm allows to generate dense factors. **Reference:** [2]
`none`	Fix seed. This method allows the user to manually provide initial values for both matrix factors.
`random`	The entries of each factors are drawn from a uniform distribution over $[0, max(V)]$, where V is the target matrix.

Table 2: Description of the implemented seeding methods to initialize NMF algorithms. The first column gives the key to use in the call to the `nmf` function.

1.5 How to run NMF algorithms

Method `nmf` provides a single interface to run NMF algorithms. It can directly perform NMF on object of class `matrix` or `data.frame` and `ExpressionSet` – if the `Biobase` package is installed. The interface has four main parameters:

```
nmf(x, rank, method, seed, ...)
```

`x` is the target `matrix`, `data.frame` or `ExpressionSet` [2]

`rank` is the factorization rank, i.e. the number of columns in matrix W.

`method` is the algorithm used to estimate the factorization. The default algorithm is given by the package specific option `'default.algorithm'`, which defaults to `'brunet'` on installation [**Brunet04**].

`seed` is the seeding method used to compute the starting point. The default method is given by the package specific option `'default.seed'`, which defaults to `'random'` on initialization (see method `?rnmf` for details on its implementation).

See also `?nmf` for details on the interface and extra parameters.

[2] `ExpressionSet` is the base class for handling microarray data in BioConductor, and is defined in the `Biobase` package.

1.6 Performances

Since version 0.4, some built-in algorithms are optimized in C++, which results in a significant speed-up and a more efficient memory management, especially on large scale data.

The older R versions of the concerned algorithms are still available, and accessible by adding the prefix `'.R#'` to the algorithms' access keys (e.g. the key `'.R#offset'` corresponds to the R implementation of NMF with offset [1]). Moreover they do not show up in the listing returned by the `nmfAlgorithm` function, unless argument `all=TRUE`:

```
nmfAlgorithm(all = TRUE)

##      brunet      brunet         lee         lee      offset
## ".R#brunet"    "brunet"    ".R#lee"       "lee" ".R#offset"
##      offset       nsNMF       nsNMF      ls-nmf      pe-nmf
##    "offset"  ".R#nsNMF"     "nsNMF"    "ls-nmf"    "pe-nmf"
##      snmf/r      snmf/l
##    "snmf/r"    "snmf/l"


# to get all the algorithms that have a secondary R version
nmfAlgorithm(version = "R")

##      brunet         lee      offset       nsNMF
## ".R#brunet"    ".R#lee" ".R#offset"  ".R#nsNMF"
```

Table 3 shows the speed-up achieved by the algorithms that benefit from the optimized code. All algorithms were run once with a factorization rank equal to 3, on the Golub data set which contains a 5000×38 gene expression matrix. The same numeric random seed (`seed=123456`) was used for all factorizations. The columns C and R show the elapsed time (in seconds) achieved by the C++ version and R version respectively. The column *Speed.up* contains the ratio R/C.

```
# retrieve all the methods that have a secondary R version
meth <- nmfAlgorithm(version = "R")
meth <- c(names(meth), meth)
meth

##                                                      brunet
##    "brunet"       "lee"    "offset"     "nsNMF" ".R#brunet"
##         lee      offset       nsNMF
##    ".R#lee" ".R#offset"  ".R#nsNMF"


# load the Golub data
data(esGolub)

# compute NMF for each method
res <- nmf(esGolub, 3, meth, seed = 123456)

## Compute NMF method 1 ...
## OK
## Compute NMF method 2 ...
## OK
## Compute NMF method 3 ...
## OK
## Compute NMF method 4 ...
```

```
## OK
## Compute NMF method 5 ...
## OK
## Compute NMF method 6 ...
## OK
## Compute NMF method 7 ...
## OK
## Compute NMF method 8 ...
## OK


# extract only the elapsed time
t <- sapply(res, runtime)[3, ]
```

	C	R	Speed.up
brunet	5.06	10.98	2.17
lee	4.75	11.98	2.52
offset	6.05	16.53	2.73
nsNMF	9.39	16.47	1.75

Table 3: Performance speed up achieved by the optimized C++ implementation for some of the NMF algorithms.

1.7 How to cite the package NMF

To view all the package's bibtex citations, including all vignette(s) and manual(s):

```
# plain text
citation("NMF")

# or to get the bibtex entries
toBibtex(citation("NMF"))
```

2 Use case: Golub dataset

We illustrate the functionalities and the usage of the *NMF* package package on the – now standard – Golub dataset on leukemia. It was used in several papers on NMF [**Brunet04**, **Gao2005**] and is included in the *NMF* package package's data, wrapped into an `ExpressionSet` object. For performance reason we use here only the first 200 genes. Therefore the results shown in the following are not meant to be biologically meaningful, but only illustrative:

```
data(esGolub)
esGolub

## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 5000 features, 38 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: ALL_19769_B-cell ALL_23953_B-cell ...
```

```
##     AML_7 (38 total)
##   varLabels: Sample ALL.AML Cell
##   varMetadata: labelDescription
## featureData
##   featureNames: M12759_at U46006_s_at ... D86976_at
##     (5000 total)
##   fvarLabels: Description
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

```
esGolub <- esGolub[1:200, ]

# remove the uneeded variable 'Sample' from the phenotypic
# data
esGolub$Sample <- NULL
```

Note: To run this example, the `Biobase` package from BioConductor is required.

2.1 Single run

2.1.1 Performing a single run

To run the default NMF algorithm on data `esGolub` with a factorization rank of 3, we call:

```
# default NMF algorithm
res <- nmf(esGolub, 3)
```

Here we did not specify either the algorithm or the seeding method, so that the computation is done using the default algorithm and is seeded by the default seeding methods. These defaults are set in the package specific options `'default.algorithm'` and `'default.seed'` respectively.

See also Sections 2.2 and 2.3 for how to explicitly specify the algorithm and/or the seeding method.

2.1.2 Handling the result

The result of a single NMF run is an object of class `NMFfit`, that holds both the fitted NMF model and data about the run:

```
res
```

```
## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFstd >
##   features: 200
##   basis/rank: 3
##   samples: 38
##  # Details:
##   algorithm:  brunet
##   seed:  random
##   RNG: 403, 624, 452835868, ... [ed7ba52c9c2666ca159b185949fd9d73]
##   distance metric:  'KL'
##   residuals:  543536
##   Iterations: 510
```

```
##   Timing:
##      user  system elapsed
##     0.424   0.008   0.432
```

The fitted model can be retrieved via method `fit`, which returns an object of class `NMF`:

```
fit(res)

## <Object of class: NMFstd >
## features: 200
## basis/rank: 3
## samples: 38
```

The estimated target matrix can be retrieved via the generic method `fitted`, which returns a – generally big – `matrix`:

```
V.hat <- fitted(res)
dim(V.hat)

## [1] 200  38
```

Quality and performance measures about the factorization are computed by method `summary`:

```
summary(res)

## Length  Class   Mode
##      1 NMFfit     S4


# More quality measures are computed, if the target matrix
# is provided:
summary(res, target = esGolub)

## Length  Class   Mode
##      1 NMFfit     S4
```

If there is some prior knowledge of classes present in the data, some other measures about the unsupervised clustering's performance are computed (purity, entropy, ...). Here we use the phenotypic variable `Cell` found in the Golub dataset, that gives the samples' cell-types (it is a factor with levels: T-cell, B-cell or `NA`):

```
summary(res, class = esGolub$Cell)

## Length  Class   Mode
##      1 NMFfit     S4
```

The basis matrix (i.e. matrix W or the metagenes) and the mixture coefficient matrix (i.e matrix H or the metagene expression profiles) are retrieved using methods `basis` and `coef` respectively:

```
# get matrix W
w <- basis(res)
dim(w)

## [1] 200   3
```

```
# get matrix H
h <- coef(res)
dim(h)

## [1]  3 38
```

If one wants to keep only part of the factorization, one can directly subset on the `NMF` object on features and samples (separately or simultaneously). The result is a `NMF` object composed of the selected rows and/or columns:

```
# keep only the first 10 features
res.subset <- res[1:10, ]
class(res.subset)

## [1] "NMFfit"
## attr(,"package")
## [1] "NMF"

dim(res.subset)

## [1] 10 38  3

# keep only the first 10 samples
dim(res[, 1:10])

## [1] 200  10   3

# subset both features and samples:
dim(res[1:20, 1:10])

## [1] 20 10  3
```

2.1.3 Extracting metagene-specific features

In general NMF matrix factors are sparse, so that the metagenes can usually be characterized by a relatively small set of genes. Those are determined based on their relative contribution to each metagene.

Kim and Park [4] defined a procedure to extract the relevant genes for each metagene, based on a gene scoring schema.

The NMF package implements this procedure in methods `featureScore` and `extractFeature`:

```
# only compute the scores
s <- featureScore(res)
summary(s)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0001  0.0163  0.0549  0.1190  0.1210  1.0000


# compute the scores and characterize each metagene
s <- extractFeatures(res)
str(s)

## List of 3
```

```
##  $ : Named int [1:8] 39 74 2 91 167 190 103 174
##   ..- attr(*, "names")= chr [1:8] "K01911_at" "M54992_at" "U46006_s_at" "D61391_at" ...
##  $ : Named int [1:13] 94 1 112 42 8 64 96 182 59 41 ...
##   ..- attr(*, "names")= chr [1:13] "M23178_s_at" "M12759_at" "U64998_at" "M22324_at" ...
##  $ : Named int [1:5] 43 120 128 130 129
##   ..- attr(*, "names")= chr [1:5] "M16336_s_at" "U76421_at" "U76189_at" "Z69881_at" ...
##  - attr(*, "method")= chr "kim"
```

2.2 Specifying the algorithm

2.2.1 Built-in algorithms

The *NMF* package package provides a number of built-in algorithms, that are listed or retrieved by function `nmfAlgorithm`. Each algorithm is identified by a unique name. The following algorithms are currently implemented (cf. Table 1 for more details):

```
nmfAlgorithm()
```

```
##   brunet      lee   offset    nsNMF   ls-nmf   pe-nmf
## "brunet"    "lee" "offset"  "nsNMF" "ls-nmf" "pe-nmf"
##   snmf/r   snmf/l
## "snmf/r" "snmf/l"
```

The algorithm used to compute the NMF is specified in the third argument (`method`). For example, to use the NMF algorithm from Lee and Seung [5] based on the Frobenius euclidean norm, one make the following call:

```
# using Lee and Seung's algorithm
res <- nmf(esGolub, 3, "lee")
algorithm(res)
```

```
## [1] "lee"
```

To use the Nonsmooth NMF algorithm from [**nsNMF2006**]:

```
# using the Nonsmooth NMF algorithm with parameter
# theta=0.7
res <- nmf(esGolub, 3, "ns", theta = 0.7)
algorithm(res)
```

```
## [1] "nsNMF"
```

```
fit(res)
```

```
## <Object of class: NMFns >
## features: 200
## basis/rank: 3
## samples: 38
## theta: 0.7
```

Or to use the PE-NMF algorithm from [**Zhang2008**]:

```
# using the PE-NMF algorithm with parameters alpha=0.01,
# beta=1
```

```
res <- nmf(esGolub, 3, "pe", alpha = 0.01, beta = 1)
res

## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFstd >
##   features: 200
##   basis/rank: 3
##   samples: 38
##  # Details:
##   algorithm:  pe-nmf
##   seed:  random
##   RNG: 403, 270, 726613627, ... [9f02c5e72ae1205c96d271bdf793721e]
##   distance metric:  <function>
##   residuals:  67.36
##   parameters: alpha=0.01, beta=1
##   Iterations: 2000
##   Timing:
##      user  system elapsed
##     1.868   0.140   1.956
```

2.2.2 Custom algorithms

The *NMF* package package provides the user the possibility to define his own algorithms, and benefit from all the functionalities available in the NMF framework. There are only few contraints on the way the custom algorithm must be defined. See the details in Section 3.1.1.

2.3 Specifying the seeding method

The seeding method used to compute the starting point for the chosen algorithm can be set via argument `seed`. Note that if the seeding method is deterministic there is no need to perform multiple run anymore.

2.3.1 Built-in seeding methods

Similarly to the algorithms, the `nmfSeed` function can be used to list or retrieve the built-in seeding methods. The following seeding methods are currently implemented:

```
nmfSeed()

## [1] "none"   "random" "ica"    "nndsvd"
```

To use a specific method to seed the computation of a factorization, one simply passes its name to `nmf`:

```
res <- nmf(esGolub, 3, seed = "nndsvd")
res

## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFstd >
##   features: 200
##   basis/rank: 3
##   samples: 38
```

```
##  # Details:
##   algorithm:  brunet
##   seed:  nndsvd
##   RNG: 403, 360, -1643107813, ... [db0a30b7ce8968b8eba695bc5c8aa9a1]
##   distance metric:  'KL'
##   residuals:  547144
##   Iterations: 1090
##   Timing:
##      user  system elapsed
##     0.892   0.000   0.894
```

2.3.2 Numerical seed

Another possibility, useful when comparing methods or reproducing results, is to set the random number generator (RNG) by passing a numerical value in argument `seed`. This value is used to set the state of the RNG, and the initialization is performed by the built-in seeding method `'random'`. When the function `nmf` exits, the value of the random seed (`.Random.seed`) is restored to its original state – as before the call.

In the case of a single run (i.e. with `nrun=1`), the default is to use the current RNG, set with the R core function `set.seed`. In the case of multiple runs, the computations use RNGstream, as provided by the core RNG "L'Ecuyer-CMRG" [**Lecuyer2002**], which generates multiple independent random streams (one per run). This ensures the complete reproducibility of any given set of runs, even when their computation is performed in parallel. Since RNGstream requires a 6-length numeric seed, a random one is generated if only a single numeric value is passed to `seed`. Moreover, single runs can also use RNGstream by passing a 6-length seed.

```
# single run and single numeric seed
res <- nmf(esGolub, 3, seed = 123456)
RNGinfo(res)

##  RNG kind:  Mersenne-Twister / Inversion
##  RNG state: NMFfit [399e26db0f730cbce5a1ac5f54cbc902]


# multiple runs and single numeric seed
res <- nmf(esGolub, 3, seed = 123456, nrun = 2)
RNGinfo(res)

##  RNG kind:  L'Ecuyer-CMRG / Inversion
##  RNG state: NMFfitX1 [164c524c5aff23154fffa2b238fd805d]


# single run with a 6-length seed
res <- nmf(esGolub, 3, seed = rep(123456, 6))
RNGinfo(res)

##  RNG kind:  L'Ecuyer-CMRG / Inversion
##  RNG state: NMFfit [48741154645afd08fae61f410b9a5677]
```

NB: To show the RNG changes happening during the computation use `.options='v4'` to turn on verbosity at level 4.
In versions prior 0.6, one could specify option `restore.seed=FALSE` or `'-r'`, this option is now deprecated.

2.3.3 Fixed factorization

Yet another option is to completely specify the initial factorization, by passing values for matrices W and H:

```
# initialize a 'constant' factorization based on the target
# dimension
init <- nmfModel(3, esGolub, W = 0.5, H = 0.3)
head(basis(init))

##                [,1] [,2] [,3]
## M12759_at       0.5  0.5  0.5
## U46006_s_at     0.5  0.5  0.5
## X70083_at       0.5  0.5  0.5
## X03100_cds2_at  0.5  0.5  0.5
## L32976_at       0.5  0.5  0.5
## M19878_s_at     0.5  0.5  0.5


# fit using this NMF object
res <- nmf(esGolub, 3, seed = init)
```

2.3.4 Custom function

The *NMF* package package provides the user the possibility to define his own seeding method, and benefit from all the functionalities available in the NMF framework. There are only few contraints on the way the custom seeding method must be defined. See the details in Section 3.2.

2.4 Multiple runs

When the seeding method is stochastic, multiple runs are usually required to achieve stability or a resonable result. This can be done by setting argument `nrun` to the desired value. For performance reason we use `nrun=5` here, but a typical choice would lies between 100 and 200:

```
res.multirun <- nmf(esGolub, 3, nrun = 5)
res.multirun

## <Object of class: NMFfitX1 >
##   Method: brunet
##   Runs:  5
##   RNG:
##    407, 2060366693, -1979266846, -1467721989, 23029248, 1469230593, 1095525326
##   Total timing:
##    user  system elapsed
##   1.540   0.092   2.132
```

By default, the returned object only contains the best fit over all the runs. That is the factorization that achieved the lowest approximation error (i.e. the lowest objective value). Even during the computation, only the current best factorization is kept in memory. This limits the memory requirement for performing multiple runs, which in turn allows to perform more runs.

The object `res.multirun` is of class `NMFfitX1` that extends class `NMFfit`, the class returned by single NMF runs. It can therefore be handled as the result of a single run and benefit from all the methods defined for single run results.

If one is interested in keeping the results from all the runs, one can set the option `keep.all=TRUE`:

```
# explicitly setting the option keep.all to TRUE
res <- nmf(esGolub, 3, nrun = 5, .options = list(keep.all = TRUE))
res

## <Object of class: NMFfitXn >
##   Method: brunet
##   Runs:  5
##   RNG:
##    407, 1997333926, -492612625, 469522788, -1828607723, 1996453586, -1167273941
##   Total timing:
##    user  system elapsed
##   2.996   0.192   2.459
##   Sequential timing:
##    user  system elapsed
##   2.664   0.040   3.593
```

```
# or using letter code 'k' in argument .options
nmf(esGolub, 3, nrun = 5, .options = "k")
```

In this case, the result is an object of class `NMFfitXn` that also inherits from class `list`.

Note that keeping all the results may be memory consuming. For example, a 3-rank `NMF` fit[3] for the Golub gene expression matrix (5000×38) takes about 18Kb[4].

2.4.1 Parallel computing on multi-core machines (non-Windows)

To speed-up the analysis whenever possible, the *NMF* package package implements transparent parallel computations when run on multi-core machines. It uses the `foreach` framework developed by REvolution Computing [**foreach**], together with the related `doMC` parallel backend [**doMC**] – based on the `multicore` package – to make use of all the CPUs available on the system. Each core will simultaneously perform part of the runs. Therefore, the required memory increases linearly with the number of cores used. When only the best run is of interest, the memory usage is optimized by using shared memory and mutex objects from the packages `bigmemory` and `synchronicity`, to only keep the current best factorization.

IMPORTANT NOTE: because it uses the `multicore` package, parallel computation over multi-cores is available only for Unix and Mac machines. The parallel computation is based on the `doMC` and `multicore` packages, so the same care should be taken as stated in the vignette of `doMC`:

> *... it usually isn't safe to run doMC and multicore from a GUI environment. In particular, it is not safe to use doMC from R.app on Mac OS X. Instead, you should use doMC from a terminal session, starting R from the command line.*

Therefore, the `nmf` function does not allow to run multicore computation from the MacOS X GUI.

The default parallel backend used by the `nmf` function is defined by the package specific option `'backend'`, which defaults to `'mc'` – for `doMC`. The backend can also be set on runtime via argument `'.pbackend'`.

There are two other runtime options, `parallel` and `parallel.required`, that can be passed via argument `.options`, to control the behaviour of the parallel computation (see below).

A call for multiple runs will be computed in parallel if one of the following condition is satisfied:

[3] i.e. the result of a single NMF run with rank equal 3.

[4] This size might change depending on the architecture (32 or 64 bits)

- call with option `'P'` or `parallel.required` set to TRUE (note the upper case in `'P'`). In this case, if for any reason the computation cannot be run in parallel (packages requirements, OS, ...), then an error is thrown. Use this mode to force the parallel execution.
- call with option `'p'` or `parallel` set to TRUE. In this case if something prevents a parallel computation, the factorizations will be done sequentially.
- a valid parallel backend is specified in argument `.pbackend`. For the moment it can either be the string `'mc'` or a single `numeric` value specifying the number of core to use. Unless option `'P'` is specified, it will run using option `'p'` (i.e. try-parallel mode).

NB: The number of processors to use can also be specified in the runtime options as e.g. `.options='p4'` or `.options='P4'` – to ask or request 4 CPUs.

Examples

The following exmaples are run with `.options='v2'` which turn on verbosity at level 2. However *Sweave* do not show all the messages. The user is therefore encouraged to run these commands on his machine to see the internal differences of each call.

```
# the default call will try to run in parallel using all
# the cores => will be in parallel if all the requirements
# are satisfied
nmf(esGolub, 3, nrun = 5, .opt = "v")

## NMF algorithm: 'brunet'
## Multiple runs: 5
## Mode: parallel(2/2 core(s))
## Runs: ... DONE
## System time:
##    user  system elapsed
##   1.424   0.112   2.282
## <Object of class: NMFfitX1 >
##   Method: brunet
##   Runs:  5
##   RNG:
##    407, 479565376, 313168193, 758619662, -1499576521, -1114385908, 588631965
##   Total timing:
##    user  system elapsed
##   1.424   0.112   2.282
```

```
# request a certain number of cores to use => no error if
# not possible
nmf(esGolub, 3, nrun = 5, .opt = "vp8")

# force parallel computation: use option 'P'
nmf(esGolub, 3, nrun = 5, .opt = "vP")

# require an improbable number of cores => error
nmf(esGolub, 3, nrun = 5, .opt = "vP200")
```

2.4.2 High Performance Computing on a cluster

To achieve further speed-up, the computation can be run on an HPC cluster. In our tests we used the `doMPI` package to perform 100 factorizations using hybrid parallel computation on 4 quadri-core machines – making use of all the cores computation on each machine.

The scripts used to launch and run the factorizations can be found in file *mpi.R* in the package's *examples* directory:

```
file.show(system.file("examples/mpi.R", package = "NMF"))

# and
file.show(system.file("examples/mpi_run.sh", package = "NMF"))
```

The script file *mpi.R* contains some extra code to log and trace the computation. Reducing it to the essential gives the following piece of code:

```
## 0. Create and register an MPI cluster
library(doMPI)
cl <- startMPIcluster()
registerDoMPI(cl)
library(NMF)

## 1. Schedule the runs accross the workers
nrun <- 100
nworker <- getDoParWorkers()
ntasks <- rep(round(nrun/nworker), nworker)
# allocate remainder runs
if ((remain <- nrun%%nworker) > 0) ntasks[1:remain] <- ntasks[1:remain] +
    1

## 2. Send the jobs to the workers using a foreach loop
t <- system.time({
    res <- foreach(i = 1:getDoParWorkers(), n = ntasks, .packages = c("NMF",
        "doMC", "Biobase")) %dopar% {

        # each worker run its factorizations in parallel Note: only
        # the best result is kept
        data(esGolub)
        nmf(esGolub, 3, "brunet", nrun = n, .opt = "p")
    }
})
## 3. reduce the result and save it in a file
res <- NMFfitX(res, runtime.all = t)
save(res, file = "result.RData")

## 4. Shutdown the cluster and quit MPI
closeCluster(cl)
mpi.quit()
```

Passing the following shell script to *qsub* should launch the execution on a Sun Grid Engine HPC cluster, with OpenMPI. Some adaptation might be necessary for other queueing systems.

```
#!/bin/bash
#$ -cwd
#$ -q opteron.q
#$ -pe mpich 5
echo "Got $NSLOTS slots. $TMP/machines"

orterun -v -n $NSLOTS -hostfile $TMP/machines R --slave -f mpi.R
```

2.4.3 Forcing sequential execution

When running on a single core machine, *NMF* package package has no other option than performing the multiple runs sequentially, one after another. This is done via the `sapply` function.

On multi-core machine, one usually wants to perform the runs in parallel, as it speeds up the computation (cf. Section 2.4.1). However in some situation (e.g. while debugging), it might be useful to force the sequential execution of the runs. This can be done via the option `'-p'` or `'p1'` by setting the parallel backend to `NULL`, `'seq'` or `''`:

```
# force sequential computation by sapply: use option '-p'
# or .pbackend=''
nmf(esGolub, 3, nrun = 5, .opt = "v-p")
nmf(esGolub, 3, nrun = 5, .opt = "v", .pbackend = "")

# or use the SEQ backend of foreach: .pbackend=NULL or
# 'seq'
nmf(esGolub, 3, nrun = 5, .opt = "v", .pbackend = NULL)
nmf(esGolub, 3, nrun = 5, .opt = "v", .pbackend = "seq")
nmf(esGolub, 3, nrun = 5, .opt = "vp1")
```

2.5 Estimating the factorization rank

A critical parameter in NMF is the factorization rank r. It defines the number of metagenes used to approximate the target matrix. Given a NMF method and the target matrix, a common way of deciding on r is to try different values, compute some quality measure of the results, and choose the best value according to this quality criteria.

Several approaches have then been proposed to choose the optimal value of r. For example, [**Brunet04**] proposed to take the first value of r for which the cophenetic coefficient starts decreasing, [3] suggested to choose the first value where the RSS curve presents an inflection point, and [**Frigyesi2008**] considered the smallest value at which the decrease in the RSS is lower than the decrease of the RSS obtained from random data.

The *NMF* package package provides functions to help implement such procedures and plot the relevant quality measures. Note that this can be a lengthy computation, depending on the data size. Whereas the standard NMF procedure usually involves several hundreds of random initialization, performing 30-50 runs is considered sufficient to get a robust estimate of the factorization rank [**Brunet04**, 3]. For performance reason, we perform here only 10 runs for each value of the rank.

```
# perform 10 runs for each value of r in range 2:6
estim.r <- nmf(esGolub, 2:6, nrun = 10, seed = 123456)
```

```
plot(estim.r)
```

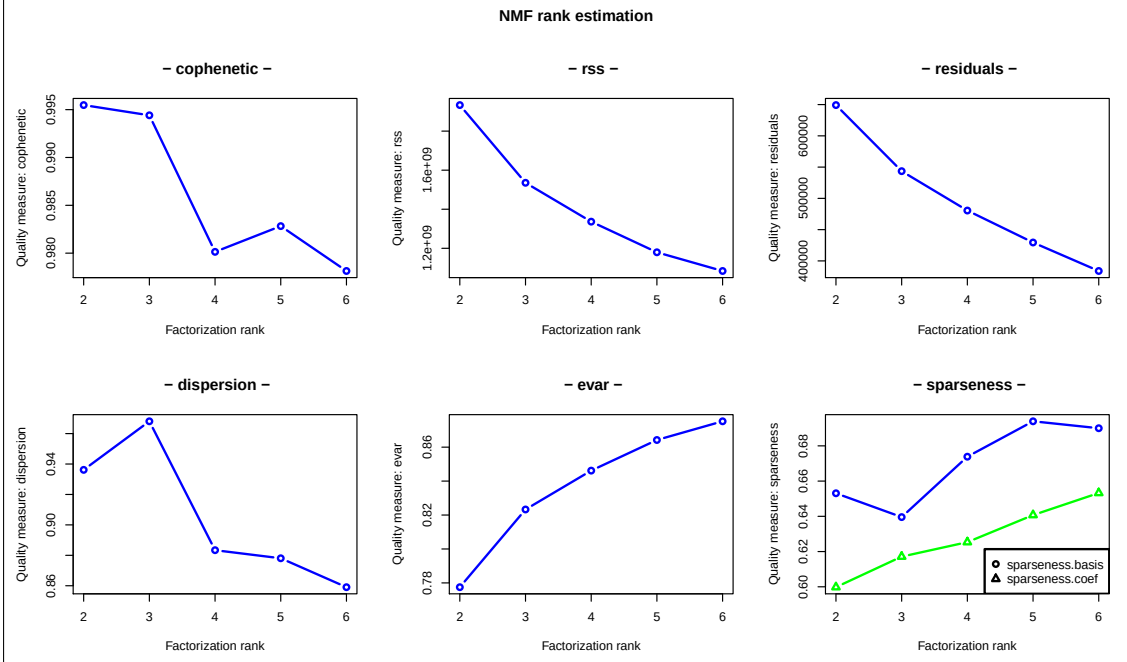

Figure 1: Estimation of the rank: Quality measures computed from 10 runs for each value of r.

The result is a S3 object of class `NMF.rank`, that contains a `data.frame` with the quality measures in column, and the values of r in row. It also contains a list of the consensus matrix for each value of r.

All the measures can be plotted at once with the method `plot` (Footnote 5), and the function `consensusmap` generates heatmaps of the consensus matrix for each value of the rank. In the context of class discovery, it is useful to see if the clusters obtained correspond to known classes. This is why in the particular case of the Golub dataset, we added annotation tracks for the two covariates available ('Cell' and 'ALL.AML'). Since we removed the variable 'Sample' in the preliminaries, these are the only variables in the phenotypic `data.frame` embedded within the `ExpressionSet` object, and we can simply pass the whole object to argument `annCol` (Figure 1). One can see that at rank 2, the clusters correspond to the ALL and AML samples respectively, while rank 3 separates AML from ALL/T-cell and ALL/B-cell[5].

2.5.1 Overfitting

Even on random data, increasing the factorization rank would lead to decreasing residuals, as more variables are available to better fit the data. In other words, there is potentially an overfitting problem.

In this context, the approach from [**Frigyesi2008**] may be useful to prevent or detect overfitting as it takes into account the results for unstructured data. However it requires to compute the quality measure(s) for the random data. The *NMF* package package provides a function that shuffles the original data, by permuting the rows of each column, using each time a different permutation. The rank estimation procedure can then be applied to the randomized data, and the "random" measures added to the plot for comparison (Section 2.5.1).

[5]Remember that the plots shown in Figure 1 come from only 10 runs, using the 200 first genes in the dataset, which explains the somewhat not so clean clusters. The results are in fact much cleaner when using the full dataset

```
consensusmap(estim.r, annCol = esGolub)
```

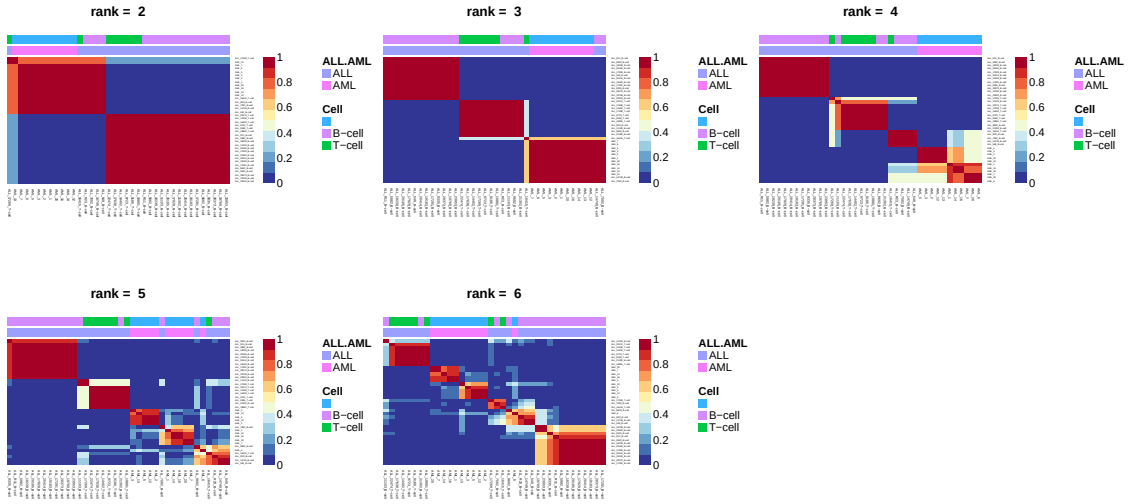

Figure 2: Estimation of the rank: Consensus matrices computed from 10 runs for each value of r.

```
# plot measures on same graph
plot(estim.r, ref = estim.r.random)
```

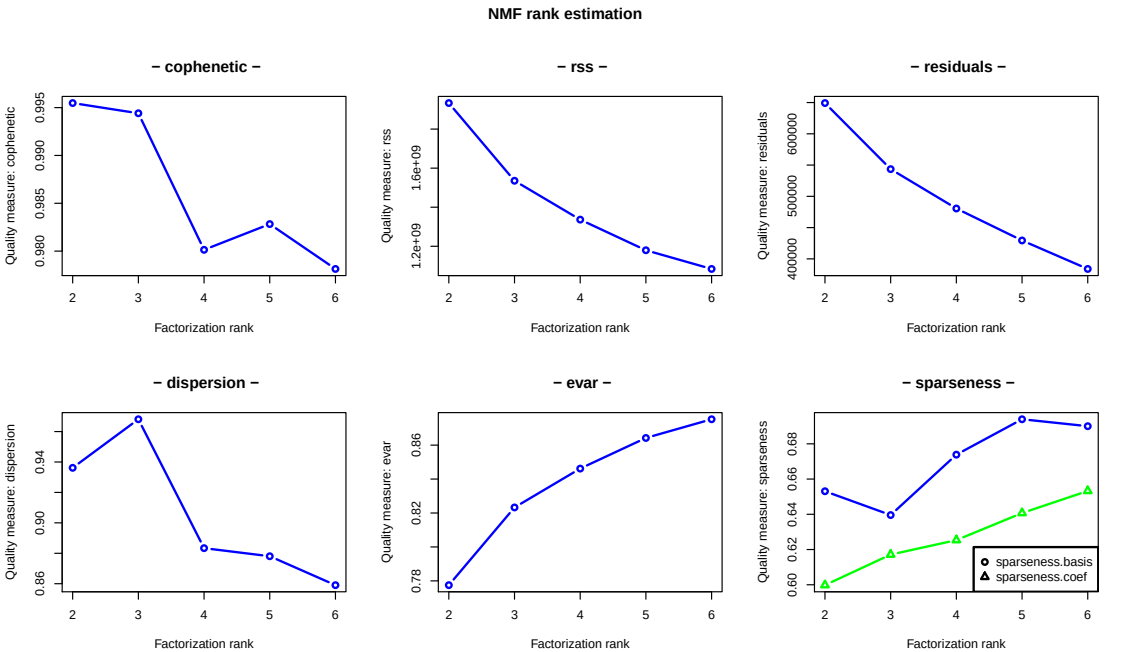

Figure 3: Estimation of the rank: Comparison of the quality measures with those obtained from randomized data. The curves for the actual data are in blue and green, those for the randomized data are in red and pink. The estimation is based on Brunet's algorithm.

2.6 Visualization methods

Error track

If the NMF computation is performed with error tracking enabled – using argument `.options` – the trajectory of the objective value is computed during the fit. This computation is not enabled by default as it induces some overhead.

```
# run nmf with .option='t'
res <- nmf(esGolub, 3, .options = "t")
# or with .options=list(track=TRUE)
```

The trajectory can be plot with the method `plot` (Figure 4):

Heatmaps

The methods `basismap`, `coefmap` and `consensusmap` provide an easy way to visualize respectively the resulting basis matrix (i.e. metagenes), mixture coefficient matrix (i.e. metaprofiles) and the consensus matrix, in the case of multiple runs. It produces pre-configured heatmaps based on the function `aheatmap`. The default heatmaps produced by these functions are shown in Figures 5 and 6. They can be customized in many different ways (colours, annotations, labels). See the dedicated vignette *"NMF: generating heatmaps"* or the help pages `?coefmap` and `?aheatmap` for more information.

The rows of the basis matrix often carry the high dimensionality of the data: genes, loci, pixels, features, etc... The function `basismap` extends the use of argument `subsetRow` (from `aheatmap`) to the specification of a feature selection method. In Figure 5 we simply used `subsetRow=TRUE`, which subsets the rows using the method described in [4], to only keep the basis-specific features (e.g. the metagene-specific genes). We refer to the relevant help pages `?basismap` and `?aheatmap` for more details about other possible values for this argument.

NB: Note how the function `aheatmap`, the underlying heatmap engine provided with the package NMF, makes it is possible to combine several heatmaps on the same plot, using the standard layout call `par(mfrow=...)`. It is also directly compatible with both `layout(...)` and viewports from `grid` graphics. The plotting context is automatically detected behind the scene, and a correct behaviour is achieved thanks to the package `gridBase` [**gridBase**].

In the case of multiple runs the function `consensusmap` plots the consensus matrix, i.e. the average connectivity matrix across the runs (see results in Figure 6 for a consensus matrix obtained with 100 runs of Brunet's algorithm on the complete Golub dataset):

2.7 Comparing algorithms

To compare the results from different algorithms, one can pass a list of methods in argument `method`. To enable a fair comparison, a deterministic seeding method should also be used. Here we fix the random seed to 123456.

```
# fit a model for several different methods
res.multi.method <- nmf(esGolub, 3, list("brunet",
    "lee", "ns"), seed = 123456, .options = "t")

## Compute NMF method 1 ...
## OK
## Compute NMF method 2 ...
## OK
```

(Figure 6).

```
plot(res)
```

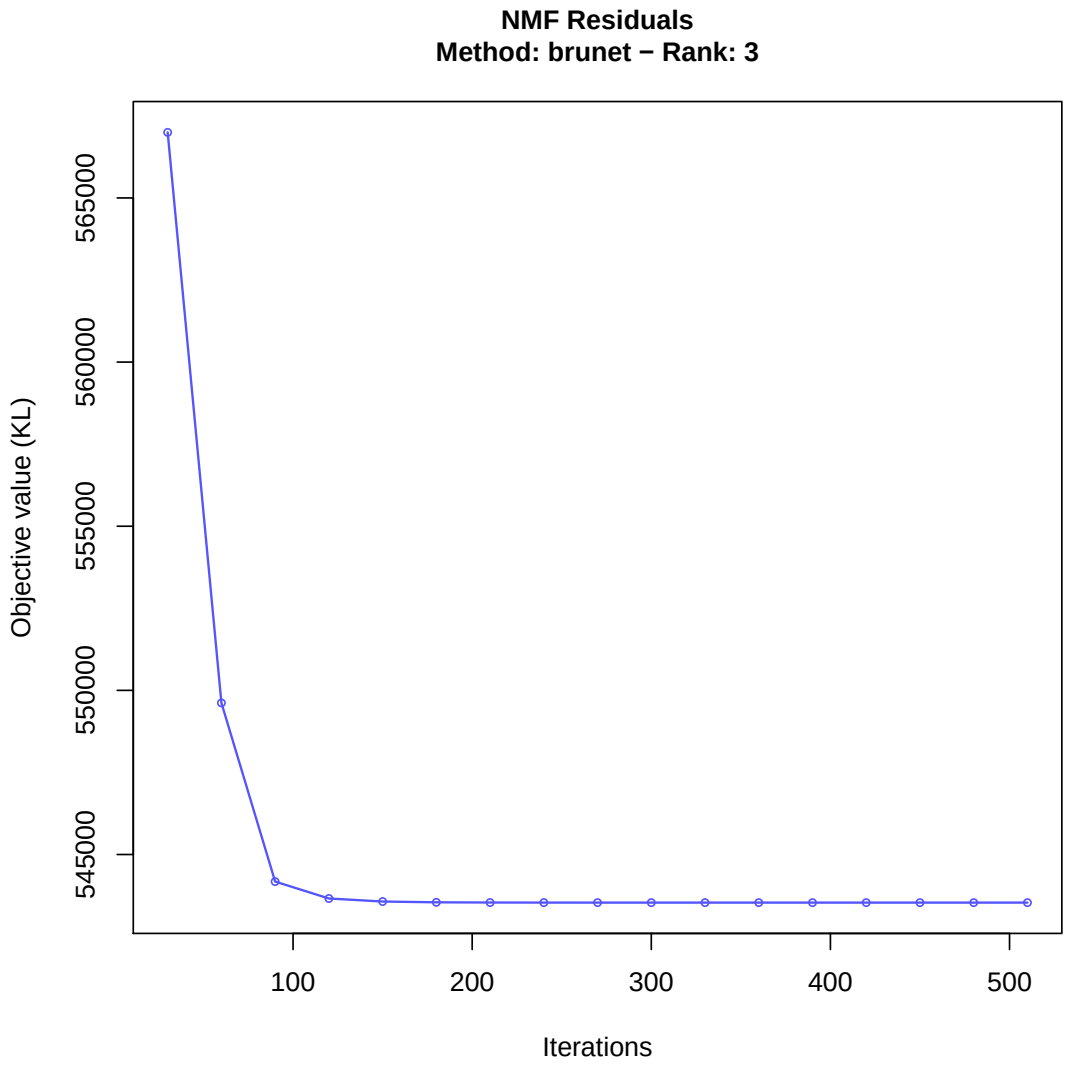

Figure 4: Error track for a single NMF run

```
layout(cbind(1, 2))
# basis components
basismap(res, subsetRow = TRUE)
# mixture coefficients
coefmap(res)
```

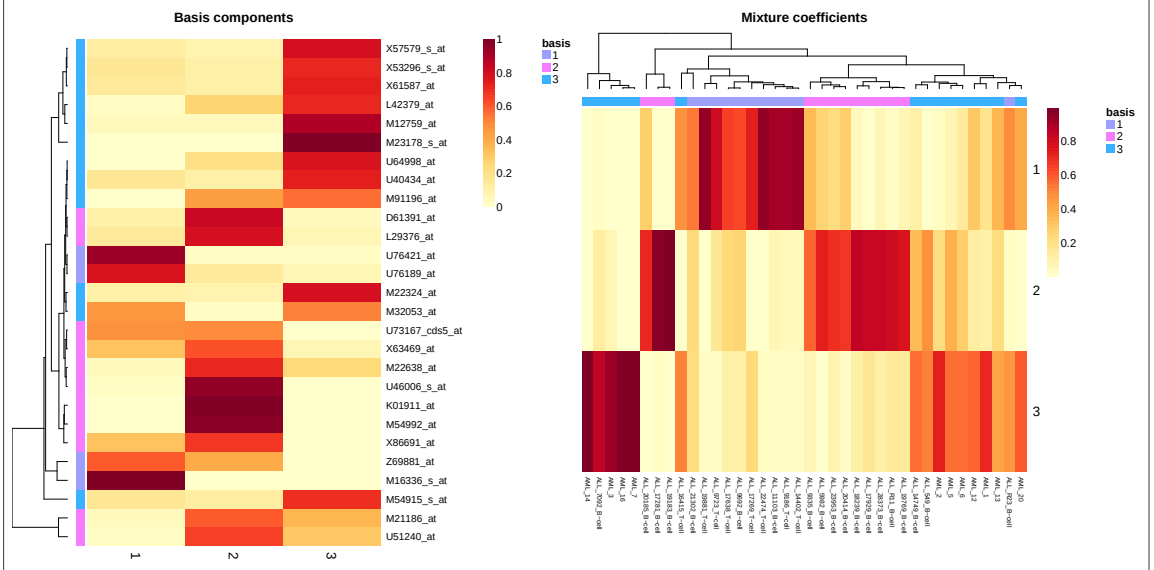

Figure 5: Heatmap of the basis and the mixture coefficient matrices. The rows of the basis matrix were selected using the default feature selection method – described in [4].

```
## Compute NMF method 3 ...
## OK
```

Passing the result to method `compare` produces a `data.frame` that contains summary measures for each method. Again, prior knowledge of classes may be used to compute clustering quality measures:

```
compare(res.multi.method)

##        method   seed rng    metric rank sparseness.basis
## brunet brunet random   1        KL    3           0.6393
## lee       lee random   1 euclidean    3           0.7269
## nsNMF   nsNMF random   1        KL    3           0.6777
##        sparseness.coef residuals niter   cpu cpu.all nrun
## brunet          0.6218    543536   510 0.464   0.464    1
## lee             0.4466 673513121  1780 1.364   1.364    1
## nsNMF           0.7350    585106   970 1.256   1.256    1


# If prior knowledge of classes is available
compare(res.multi.method, class = esGolub$Cell)

##        method   seed rng    metric rank sparseness.basis
## brunet brunet random   1        KL    3           0.6393
## lee       lee random   1 euclidean    3           0.7269
```

```
## nsNMF   nsNMF random   1        KL    3           0.6777
##        sparseness.coef purity entropy residuals niter   cpu
## brunet          0.6218 0.8158  0.3927    543536   510 0.464
## lee             0.4466 0.5789  0.7231 673513121  1780 1.364
## nsNMF           0.7350 0.7895  0.4691    585106   970 1.256
##        cpu.all nrun
## brunet   0.464    1
## lee      1.364    1
## nsNMF    1.256    1
```

Because the computation was performed with error tracking enabled, an error plot can be produced by method `plot` (Figure 7). Each track is normalized so that its first value equals one, and stops at the iteration where the method's convergence criterion was fulfilled.

3 Extending the package

We developed the *NMF* package package with the objective to facilitate the integration of new NMF methods, trying to impose only few requirements on their implementations. All the built-in algorithms and seeding methods are implemented as strategies that are called from within the main interface method `nmf`.

The user can define new strategies and those are handled in exactly the same way as the built-in ones, benefiting from the same utility functions to interpret the results and assess their performance.

3.1 Custom algorithm

3.1.1 Using a custom algorithm

To define a strategy, the user needs to provide a `function` that implements the complete algotihm. It must be of the form:

```
my.algorithm <- function(x, seed, param.1, param.2) {
    # do something with starting point ...

    # return updated starting point
    return(seed)
}
```

Where:

target is a `matrix`;

start is an object that inherits from class `NMF`. This `S4` class is used to handle NMF models (matrices `W` and `H`, objective function, etc...);

param.1, param.2 are extra parameters specific to the algorithms;

The function must return an object that inherits from class `NMF`.

For example:

```
my.algorithm <- function(x, seed, scale.factor = 1) {
    # do something with starting point ...  for example: 1.
    # compute principal components
    pca <- prcomp(t(x), retx = TRUE)
```

```
    # 2. use the absolute values of the first PCs for the
    # metagenes Note: the factorization rank is stored in
    # object 'start'
    factorization.rank <- nbasis(seed)
    metagenes(fit(seed)) <- abs(pca$rotation[, 1:factorization.rank])
    # use the rotated matrix to get the mixture coefficient use
    # a scaling factor (just to illustrate the use of extra
    # parameters)
    metaprofiles(fit(seed)) <- t(abs(pca$x[,
1:factorization.rank]))/scale.factor

    # return updated data
    return(seed)
}
```

To use the new method within the package framework, one pass `my.algorithm` to main interface `nmf` via argument `method`. Here we apply the algorithm to some matrix `V` randomly generated:

```
n <- 50
r <- 3
p <- 20
V <- syntheticNMF(n, r, p, noise = TRUE)
```

```
nmf(V, 3, my.algorithm, scale.factor = 10)

## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFstd >
##   features: 50
##   basis/rank: 3
##   samples: 20
##  # Details:
##   algorithm:  nmf_1c395e0ddb5c
##   seed:  random
##   RNG: 403, 173, 1191562961, ... [0d87fde5bf092083bd059bcc33d58a3e]
##   distance metric:  'euclidean'
##   residuals:  554.9
##   parameters: scale.factor=10
##   Timing:
##      user  system elapsed
##     0.004   0.000   0.004
```

3.1.2 Using a custom distance measure

The default distance measure is based on the euclidean distance. If the algorithm is based on another distance measure, this one can be specified in argument `objective`, either as a `character` string corresponding to a built-in objective function, or a custom `function` definition:

```
# based on Kullback-Leibler divergence
nmf(V, 3, my.algorithm, scale.factor = 10, objective = "KL")

## <Object of class: NMFfit >
```

```
##  # Model:
##   <Object of class: NMFstd >
##   features: 50
##   basis/rank: 3
##   samples: 20
##  # Details:
##   algorithm:  nmf_1c392f035299
##   seed:  random
##   RNG: 403, 383, 1191562961, ... [2f2572f5c6b643d37859d97079951964]
##   distance metric:  'KL'
##   residuals:  1539
##   parameters: scale.factor=10
##   Timing:
##      user  system elapsed
##     0.004   0.000   0.003

# based on custom distance metric
nmf(V, 3, my.algorithm, scale.factor = 10, objective = function(target,
    x) {
    (sum((target - fitted(x))^4))^{
        1/4
    }
})

## Error: unused argument(s) (scale.factor = 10)
```

3.1.3 Defining algorithms for mixed sign data

All the algorithms implemented in the *NMF* package package assume that the input data is nonnegative. However, some methods exist in the litterature that work with relaxed constraints, where the input data and one of the matrix factors (W or H) are allowed to have negative entries (eg. semi-NMF [**Ding2008**, **Roux2008**]). Strictly speaking these methods do not fall into the NMF category, but still solve constrained matrix factorization problems, and could be considered as NMF methods when applied to non-negative data. Moreover, we received user requests to enable the development of semi-NMF type methods within the package's framework. Therefore, we designed the *NMF* package package so that such algorithms – that handle negative data – can be integrated. This section documents how to do it.

By default, as a safe-guard, the sign of the input data is checked before running any method, so that the `nmf` function throws an error if applied to data that contain negative entries [6]. To extend the capabilities of the *NMF* package package in handling negative data, and plug mixed sign NMF methods into the framework, the user needs to specify the argument `mixed=TRUE` in the call to the `nmf` function. This will skip the sign check of the input data and let the custom algorithm perform the factorization.

As an example, we reuse the previously defined custom algorithm[7]:

```
# put some negative input data
V.neg <- V
V.neg[1, ] <- -1
```

[6] Note that on the other side, the sign of the factors returned by the algorithms is never checked, so that one can always return factors with negative entries.

[7] As it is defined here, the custom algorithm still returns nonnegative factors, which would not be desirable in a real example, as one would not be able to closely fit the negative entries.

```
# this generates an error
err <- try(nmf(V.neg, 3, my.algorithm, scale.factor = 10))
err

## [1] "Error : NMF::nmf - Input matrix x contains some negative entries.\n"
## attr(,"class")
## [1] "try-error"
## attr(,"condition")
## <simpleError: NMF::nmf - Input matrix x contains some negative entries.>


# this runs my.algorithm without error
nmf(V.neg, 3, my.algorithm, mixed = TRUE, scale.factor = 10)

## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFstd >
##   features: 50
##   basis/rank: 3
##   samples: 20
##  # Details:
##   algorithm:  nmf_1c392ba152d
##   seed:  random
##   RNG: 403, 593, 1191562961, ... [7ed64fa772a4b1e68cebc0bfed3cb018]
##   distance metric:  'euclidean'
##   residuals:  561.4
##   parameters: scale.factor=10
##   Timing:
##      user  system elapsed
##     0.004   0.000   0.003
```

3.1.4 Specifying the NMF model

If not specified in the call, the NMF model that is used is the standard one, as defined in Equation (1). However, some NMF algorithms have different underlying models, such as non-smooth NMF [**nsNMF2006**] which uses an extra matrix factor that introduces an extra parameter, and change the way the target matrix is approximated.

The NMF models are defined as S4 classes that extends class `NMF`. All the available models can be retreived calling the `nmfModel()` function with no argument:

```
nmfModel()

## <Object of class: NMFstd >
## features: 0
## basis/rank: 0
## samples: 0
```

One can specify the NMF model to use with a custom algorithm, using argument `model`. Here we first adapt a bit the custom algorithm, to justify and illustrate the use of a different model. We use model `NMFOffset` [1], that includes an offset to take into account genes that have constant expression levels accross the samples:

```
my.algorithm.offset <- function(x, seed, scale.factor = 1) {
    # do something with starting point ...  for example: 1.
    # compute principal components
    pca <- prcomp(t(x), retx = TRUE)

    # retrieve the model being estimated
    data.model <- fit(seed)

    # 2. use the absolute values of the first PCs for the
    # metagenes Note: the factorization rank is stored in
    # object 'start'
    factorization.rank <- nbasis(data.model)
    metagenes(data.model) <- abs(pca$rotation[, 1:factorization.rank])
    # use the rotated matrix to get the mixture coefficient use
    # a scaling factor (just to illustrate the use of extra
    # parameters)
    metaprofiles(data.model) <- t(abs(pca$x[,
1:factorization.rank]))/scale.factor

    # 3. Compute the offset as the mean expression
    data.model@offset <- rowMeans(x)

    # return updated data
    fit(seed) <- data.model
    seed
}
```

Then run the algorithm specifying it needs model `NMFOffset`:

```
# run custom algorithm with NMF model with offset
nmf(V, 3, my.algorithm.offset, model = "NMFOffset",
    scale.factor = 10)

## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFOffset >
##   features: 50
##   basis/rank: 3
##   samples: 20
##   offset: [ 0.3677 0.9383 1.347 1.019 0.6695 ... ]
##  # Details:
##   algorithm:  nmf_1c395d140086
##   seed:  random
##   RNG: 403, 179, -963168695, ... [850b670d66e22c6d00f5d79570d1ce8a]
##   distance metric:  'euclidean'
##   residuals:  334.3
##   parameters: scale.factor=10
##   Timing:
##      user  system elapsed
##     0.004   0.000   0.003
```

3.2 Custom seeding method

The user can also define custom seeding method as a function of the form:

```
# start: object of class NMF target: the target matrix
my.seeding.method <- function(model, target) {

    # use only the largest columns for W
    w.cols <- apply(target, 2, function(x) sqrt(sum(x^2)))
    metagenes(model) <- target[, order(w.cols)[1:nbasis(model)]]

    # initialize H randomly
    metaprofiles(model) <- matrix(runif(nbasis(model) * ncol(target)),
        nbasis(model), ncol(target))

    # return updated object
    return(model)
}
```

To use the new seeding method:

```
nmf(V, 3, "snmf/r", seed = my.seeding.method)

## <Object of class: NMFfit >
##  # Model:
##   <Object of class: NMFstd >
##   features: 50
##   basis/rank: 3
##   samples: 20
##  # Details:
##   algorithm:  snmf/r
##   seed:  NMF.seed.1c39147b624f
##   RNG: 403, 439, -963168695, ... [49267d84a430d6788491e4513dd6c28c]
##   distance metric:  <function>
##   residuals:  195.6
##   Iterations: 105
##   Timing:
##      user  system elapsed
##     0.456   0.000   0.458
```

4 Advanced usage

4.1 Package specific options

The package specific options can be retieved or changed using the `nmf.getOption` and `nmf.options` functions. These behave similarly as the `getOption` and `nmf.options` base functions:

```
# show default algorithm and seeding method
nmf.options("default.algorithm", "default.seed")

# retrieve a single option
nmf.getOption("default.seed")

# All options
nmf.options()
```

Currently the following options are available:

Option	Default value	Description
default.algorithm	brunet	Default NMF algorithm used by the nmf function when argument method is missing. The value should the key of one of the available NMF algorithms. See ?nmfAlgorithm.
track.interval	30	Number of iterations between two points in the residual track. This option is relevant only when residual tracking is enabled. See ?nmf.
error.track	FALSE	Toggle default residual tracking. When TRUE, the nmf function compute and store the residual track in the result – if not otherwise specified in argument .options. Note that tracking may significantly slow down the computations.
default.seed	random	Default seeding method used by the nmf function when argument seed is missing. The value should the key of one of the available seeding methods. See ?nmfSeed.
backend	mc	Default parallel backend used used by the nmf function when argument .pbackend is missing. Currently the following values are supported: 'mc' for multicore, 'seq' for sequential, '' for sapply.
verbose	FALSE	Toggle verbosity.
debug	FALSE	Toggle debug mode, which is an extended verbose mode.

Table 4:

5 Session Info

- R version 2.15.0 (2012-03-30), i686-pc-linux-gnu
- Locale: LC_CTYPE=en_ZA.UTF-8, LC_NUMERIC=C, LC_TIME=en_ZA.UTF-8, LC_COLLATE=en_ZA.UTF-8, LC_MONETARY=en_ZA.UTF-8, LC_MESSAGES=en_ZA.UTF-8, LC_PAPER=C, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_ZA.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, utils
- Other packages: bigmemory 4.2.11, Biobase 2.16.0, BiocGenerics 0.2.0, codetools 0.2-8, colorspace 1.1-1, digest 0.5.2, doParallel 1.0.1, doRNG 1.4.1, foreach 1.4.0, gridBase 0.4-5, iterators 1.0.6, knitr 0.6, NMF 0.8.2, pkgmaker 0.5.4, RColorBrewer 1.0-5, registry 0.2, stringr 0.6, synchronicity 1.0.13, xtable 1.7-0
- Loaded via a namespace (and not attached): compiler 2.15.0, evaluate 0.4.2, formatR 0.4, highlight 0.3.1, parser 0.0-14, plyr 1.7.1, Rcpp 0.9.10, tools 2.15.0

References

[1] Liviu Badea. "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization." In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 290 (2008), pp. 267–78. ISSN: 1793-5091. URL: http://www.ncbi.nlm.nih.gov/pubmed/18229692.

[2] C Boutsidis and E Gallopoulos. "SVD based initialization: A head start for nonnegative matrix factorization". In: *Pattern Recognition* 41.4 (2008), pp. 1350–1362. ISSN: 00313203. DOI: 10.1016/j.patcog.2007.09.010. URL: http://linkinghub.elsevier.com/retrieve/pii/S0031320307004359.

[3] Lucie N Hutchins et al. “Position-dependent motif characterization using non-negative matrix factorization.” In: *Bioinformatics (Oxford, England)* 24.23 (2008), pp. 2684–90. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btn526. URL: http://www.ncbi.nlm.nih.gov/pubmed/18852176.

[4] Hyunsoo Kim and Haesun Park. “Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis.” In: *Bioinformatics (Oxford, England)* 23.12 (2007), pp. 1495–502. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btm134. URL: http://www.ncbi.nlm.nih.gov/pubmed/17483501.

[5] D D Lee and HS Seung. “Algorithms for non-negative matrix factorization”. In: *Advances in neural information processing systems* (2001). URL: http://scholar.google.com/scholar?q=intitle:Algorithms+for+non-negative+matrix+factorization\#0.

```
# The cell type is used to label rows and columns
consensusmap(res.multirun, annCol = esGolub, tracks = NA)
```

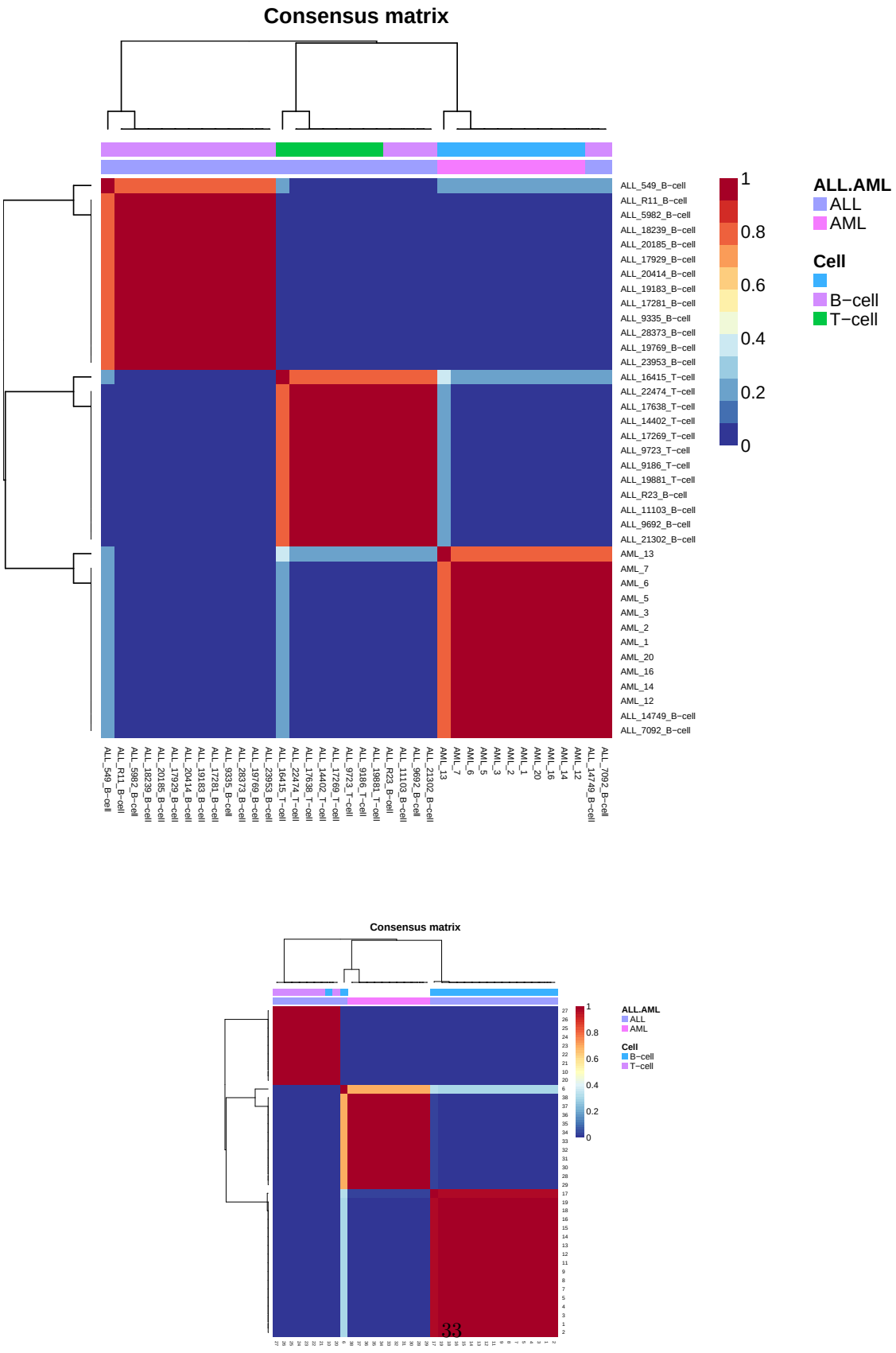

Figure 6: Heatmap of consensus matrices from 10 runs on the limited dataset (left) and from 100 runs on the complete Golub dataset (right).

```
plot(res.multi.method)
```

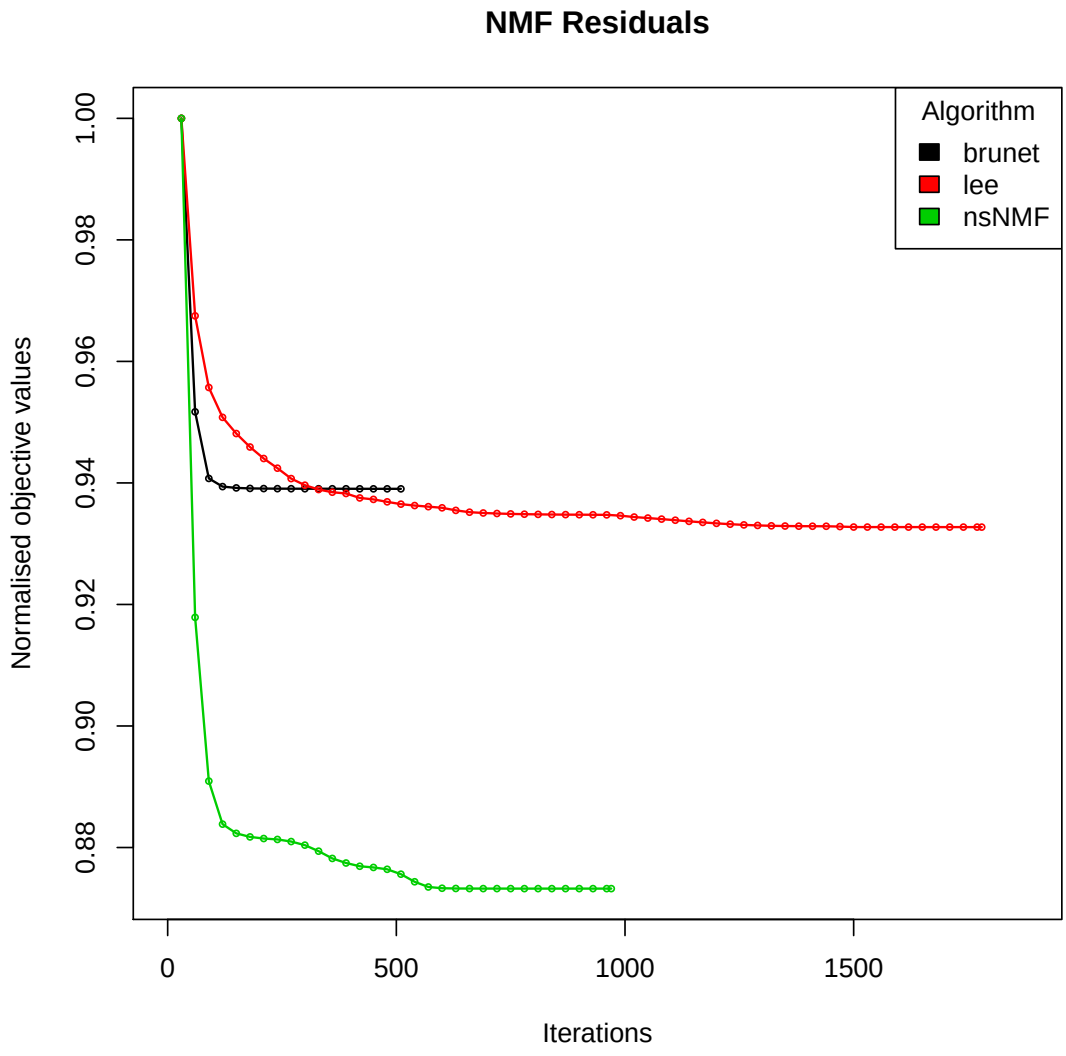

Figure 7: Error tracks comparing methods `'brunet', 'lee', 'nsNMF'`