Robust estimation of mutational signatures by adaptive learning of nucleotide interaction terms in non–negative matrix factorization

Ragnhild Laursen[1], Lasse Maretty[2] and Asger Hobolth[3]

1. Department of Mathematics, Aarhus University. Email: ragnhild@math.au.dk

2. Department of Molecular Medicine, Aarhus University Hospital. Email: lasse.maretty@clin.au.dk

3. Department of Mathematics, Aarhus University. Email: asger@math.au.dk

February 18, 2022

Abstract

Somatic mutations in cancer can be viewed as a mixture distribution of several mutational signatures. The mutational signatures can be inferred using non-negative matrix factorization (NMF). Mutational signatures have previously been parametrized using either simple mono-nucleotide multiplicative models or general third-order nucleotide interaction models. We describe a flexible and novel framework for identifying biologically relevant parametrizations of mutational signatures, and in particular for estimating second-order interaction models. The estimation procedure is based on the expectation–maximization (EM) algorithm and regression in the log-linear quasi–Poisson model. We show that the second-order interaction signatures are biologically plausible, and demonstrate that they are statistically stable and sufficiently complex to fit the mutational patterns. Second-order signatures often strike the right balance between appropriately fitting the data and avoiding over-fitting. They provide a better fit to data and are biologically more plausible than first-order signatures, and the parametrization is more stable than the parameter-rich three-way interaction signatures. We illustrate our framework on two data sets of somatic mutation counts from cancer patients.

Key words: Cancer genomics, expectation-maximization (EM) algorithm, interaction terms, mutation counts, mutational signatures, Non-negative Matrix Factorization (NMF), Poisson regression.

AMS classification: Primary: 62 (Statistics), Secondary: 62F10 (Point estimation), 62F30 (Parametric inference under constraints), 62H12 (Estimation in multivariate analysis), 62P10 (Applications of statistics to biology and medical sciences), 68T05 (Learning and adaptive systems in artificial intelligence), 92B20 (Neural networks in biological studies).

	One flanking nucleotide at each side		
Signature	Factorization	Number of parameters	Key reference
Mono-nucleotide	$L + M + R$	$3 + 6 + 3 = 6 + 3 \cdot 2 = 12$	Shiraishi et al. (2015)
Di-nucleotide	$L \times M + M \times R$	$3 \cdot 6 + 6 + 3 \cdot 6 = 6 + 18 \cdot 2 = 42$	Our proposed model
Tri–nucleotide	$L \times M \times R$	$4 \cdot 6 \cdot 4 = 6 \cdot 4^2 = 96$	Alexandrov et al. (2013)

Table 1: Parametrizations of a mutational signature with one flanking nucleotide at each side and increasing complexity. We assume strand-symmetry such that only 6 mutations are relevant.

	Two flanking nucleotides at each side	
Signature	Factorization	Number of parameters
Mono-nucleotide	$L_2 + L_1 + M + R_1 + R_2$	$6 + 3 \cdot 4 = 18$
Di-nucleotide, type 1	$L_2 \times M + L_1 \times M + M \times R_1 + M \times R_2$	$6 + 3 \cdot 6 \cdot 4 = 6 + 18 \cdot 4 = 78$
Di-nucleotide, type 2	$L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$	$42 + 12 \cdot 2 = 66$
Penta-nucleotide	$L_2 \times L_1 \times M \times R_1 \times R_2$	$6 \cdot 4^4 = 1536$
Di- and mono-nucleotide	$L_2 + L_1 \times M + M \times R_1 + R_2$	$42 + 3 \cdot 2 = 48$
Tri- and mono-nucleotide	$L_2 + L_1 \times M \times R_1 + R_2$	$96 + 3 \cdot 2 = 102$

Table 2: Parametrizations of a mutational signature with two flanking nucleotides at each side. We assume strand-symmetry such that only 6 mutations are relevant. We distinguish between two types of di-nucleotide interaction models: type 1 has interaction between the flanking nucleotide and the mutation, and type 2 has interaction between the nearest neighbours.

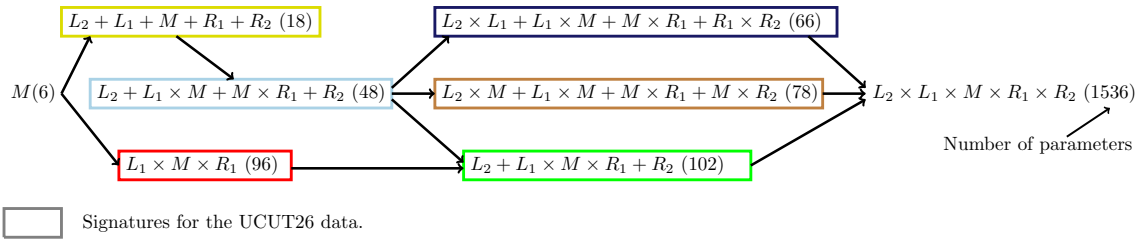

Figure 1: Factor diagram for the signatures used for the UCUT26 data set. The diagram shows the number of parameters for each signature and how the signatures are nested within each other.

0.1 Analysis of BRCA21 data

Recall that the BRCA21 count matrix has $T = 96$ mutation types and $N = 21$ patients. The number of observations is thus $n_{\text{obs}} = T \cdot N = 96 \cdot 21 = 2016$. We follow Alexandrov et al. (2013) and fix the number of signatures at $K = 4$.

Firstly, consider the three models where all four signatures are either mono-nucleotide, di-nucleotide or tri-nucleotide. The number of parameters n_{prm}, penalization term for number of parameters $n_{\text{prm}} \log n_{\text{obs}}$, GKL and BIC for each of the three models are summarized in Table 3. The GKL difference between the mono-nucleotide and di-nucleotide model is very large which means that the fit to the data is poor for the mono-nucleotide model.

Model for signatures	Number of parameters n_{prm}	Model complexity $n_{\text{prm}} \log n_{\text{obs}}$	Fit to data GKL	Model selection BIC
Sole mono-nucleotide	$4 \cdot 12 = 48$	365	4647	9659
Sole di-nucleotide	$4 \cdot 42 = 168$	1278	2120	5518
Sole tri-nucleotide	$4 \cdot 96 = 384$	2922	1479	5881
Mixture	$12 + 42 + 42 + 96 = 192$	1461	1870	5200

Table 3: Summary statistics for the three basic models for the BRCA21 data where all signatures are the same, and a flexible mixture model where signatures can vary. The mixture model consists of one mono-nucleotide interaction signature, two di-nucleotide interaction signatures, and one tri-nucleotide interaction signature, and has the smallest BIC. We have $K = 4$ and the number of observations is $n_{\text{obs}} = T \cdot N = 2016$.

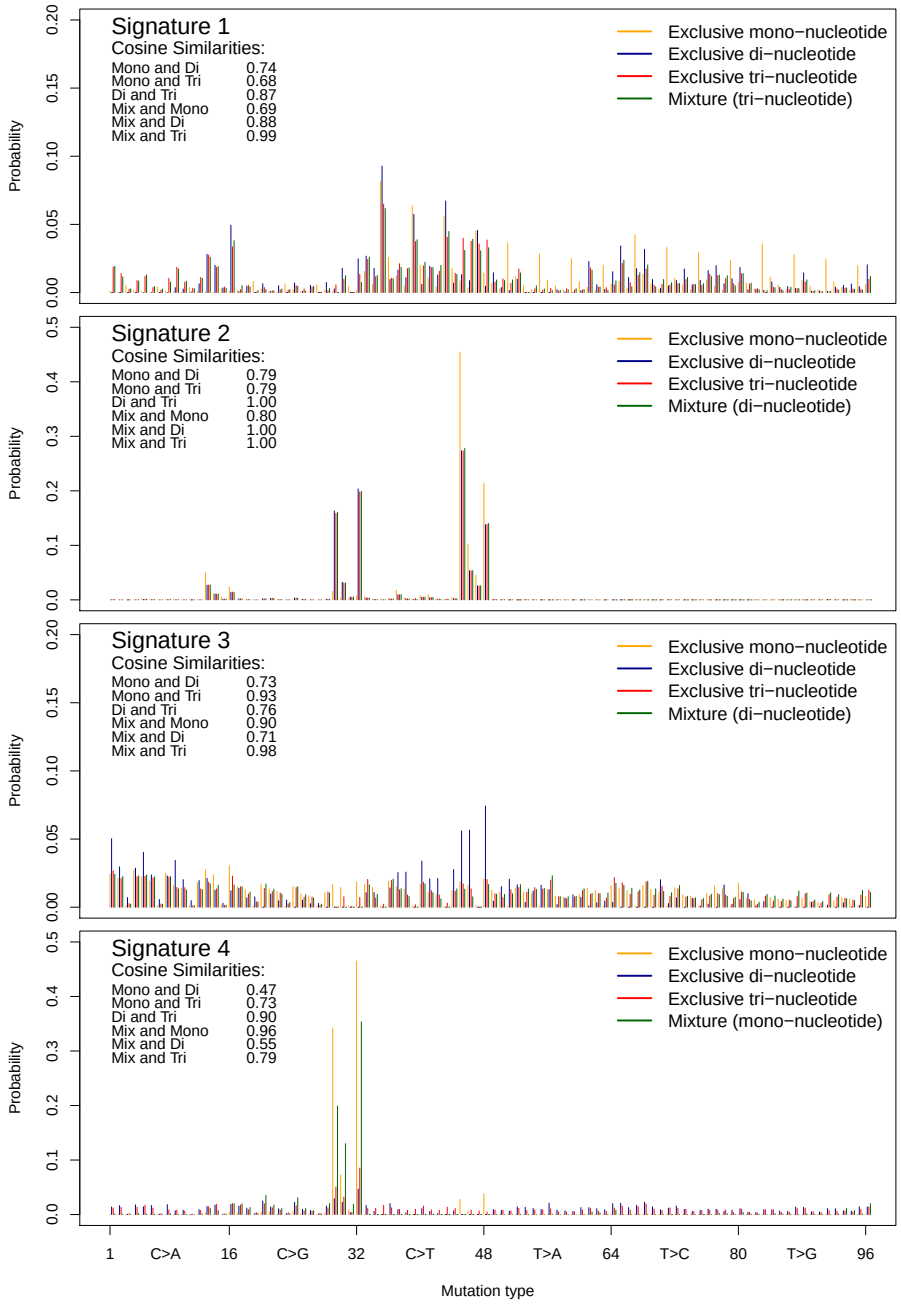

Figure 2: Comparison of signatures for four models for the BRCA21 data. The four models are three models where all four signatures are mono-nucleotide, di-nucleotide or tri-nucleotide, and one mixture model where one signature is tri-nucleotide, two are di-nucleotide and one is mon-nucleotide. The latter model has the smallest BIC value.

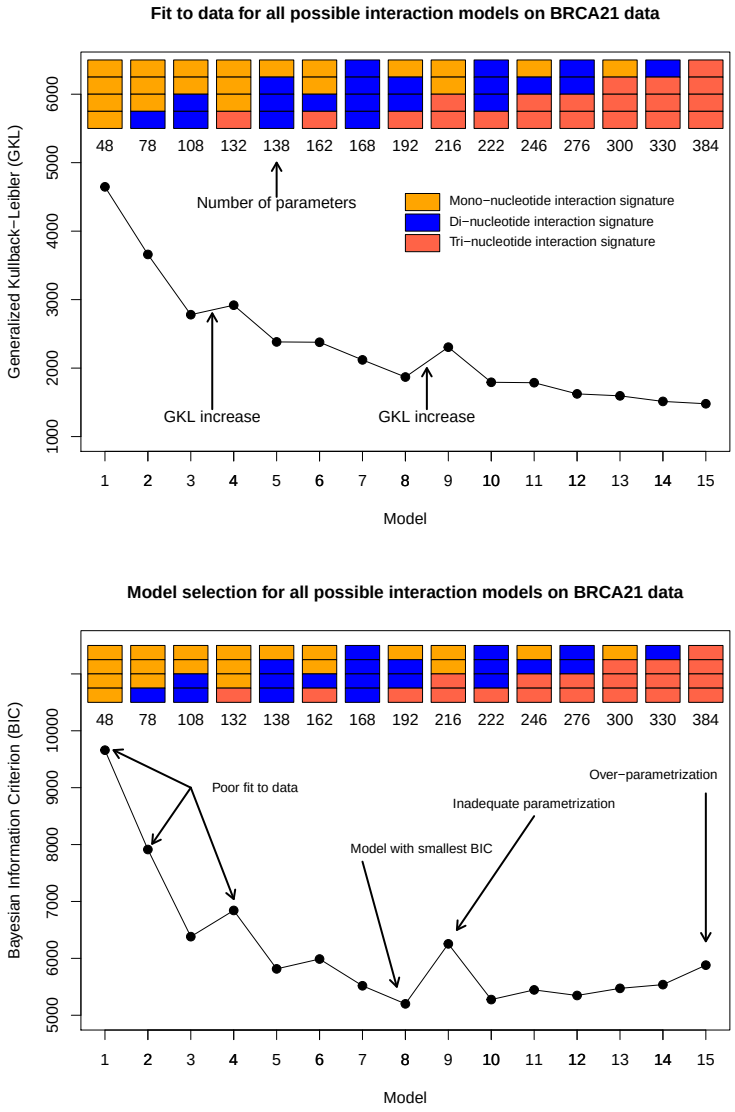

Figure 3: The Generalized Kullback-Leibler (GKL) and Bayesian Information Criteria (BIC) for all 15 models with 4 signatures for the BRCA21 data set. The models are ordered according to the total number of parameters for the 4 signatures; e.g. $4 \cdot 12 = 48$ for the sole mono-nucleotide model and $4 \cdot 96 = 384$ for the sole tri-nucleotide model.

0.2 Analysis of BRCA119 data

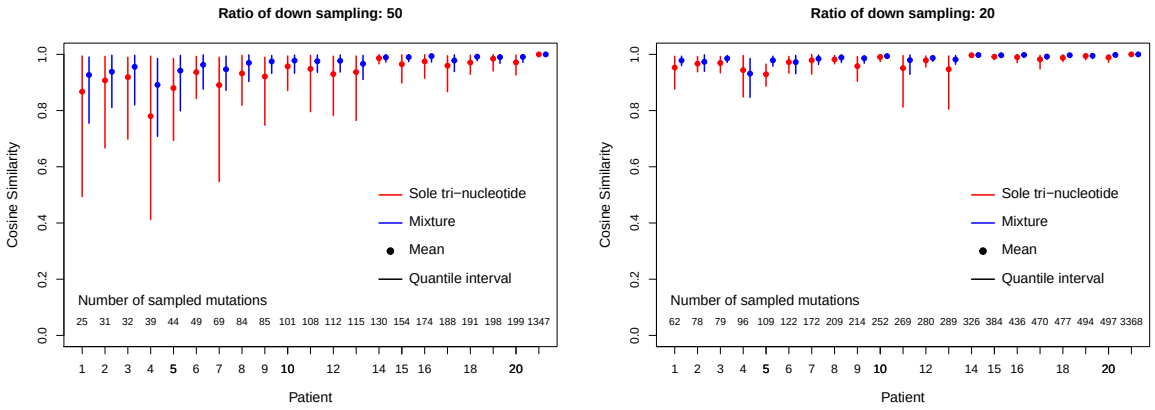

Figure 4: Exposure recovery after down–sampling the number of mutations. The exposures of each signature is better recovered using the di-nucleotide model compared to the tri-nucleotide model. For each patient the mean cosine similarities between the down-sampled exposure vectors and the full data exposure vector are closer to one for the di-nucleotide model, and the quantile intervals are narrower.

0.3 Analysis of UCUT26 data

Model for the two signatures	Number of parameters n_{prm}	Model complexity $n_{\text{prm}} \log n_{\text{obs}}$	Fit to data GKL	Model selection $\triangle$BIC
$L_1 \times M \times R_1$	$2 \cdot 96 = 192$	1645	10182	2861
$L_2 + L_1 + M + R_1 + R_2$	$2 \cdot 18 = 36$	308	10422	2005
$L_2 + L_1 \times M \times R_1 + R_2$	$2 \cdot 102 = 204$	1748	9438	1477
$L_2 + L_1 \times M + M \times R_1 + R_2$	$2 \cdot 48 = 96$	823	9788	1252
$L_2 \times M + L_1 \times M + M \times R_1 + M \times R_2$	$2 \cdot 78 = 156$	1336	9528	1246
$L_2 \times L_1 + L_1 \times M + M \times R_1 + R_1 \times R_2$	$2 \cdot 66 = 132$	1130	9008	0

Table 4: Summary statistics for the six basic models for the UCUT26 data where both signatures have the same parametrization. The models are ordered according to their BIC value. The number of signatures is $K = 2$ and the number of observations is $n_{\text{obs}} = 5260$.

References

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259.

Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS genetics*, 11(12):e1005657.

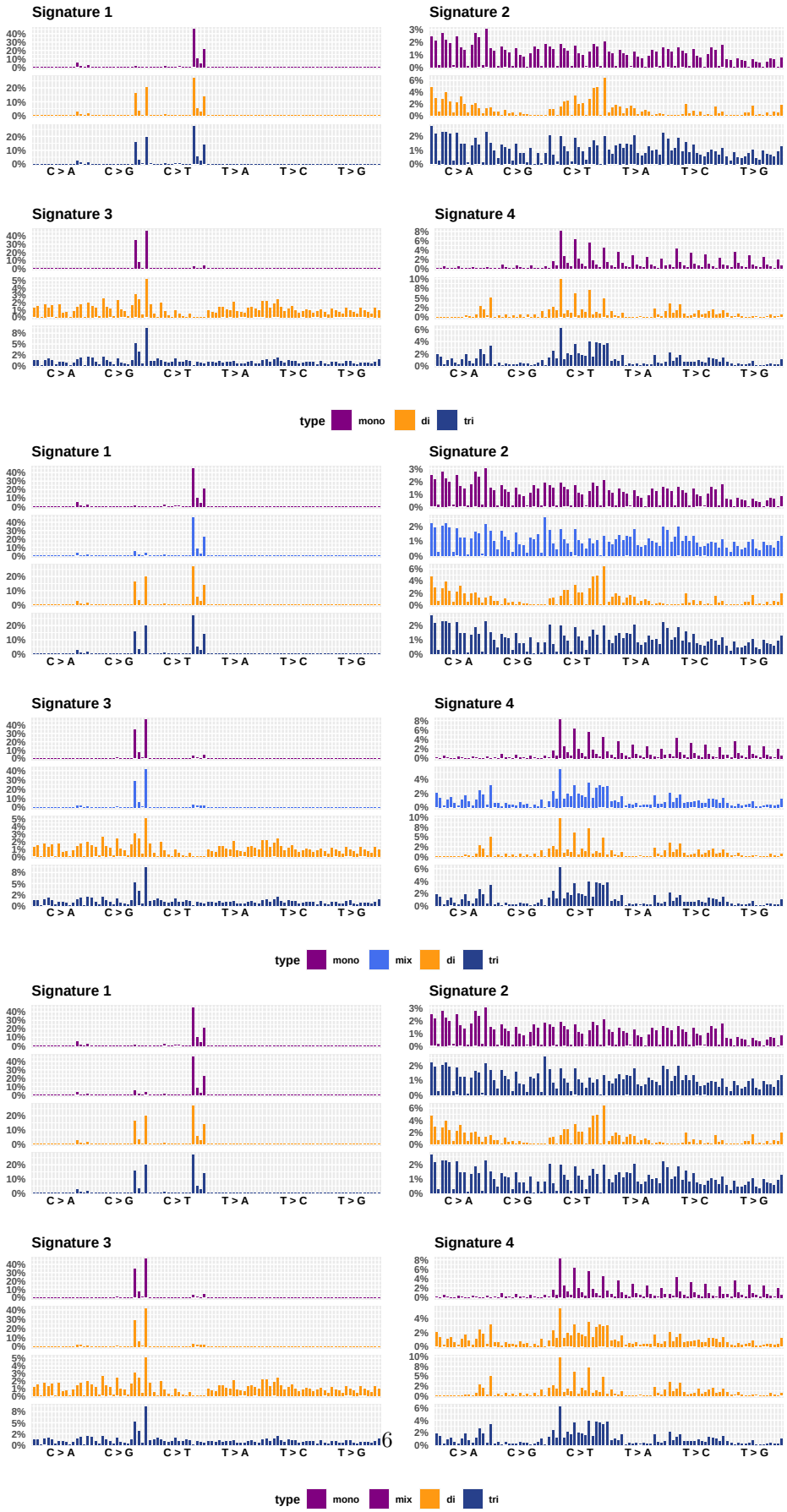

Figure 5: Caption

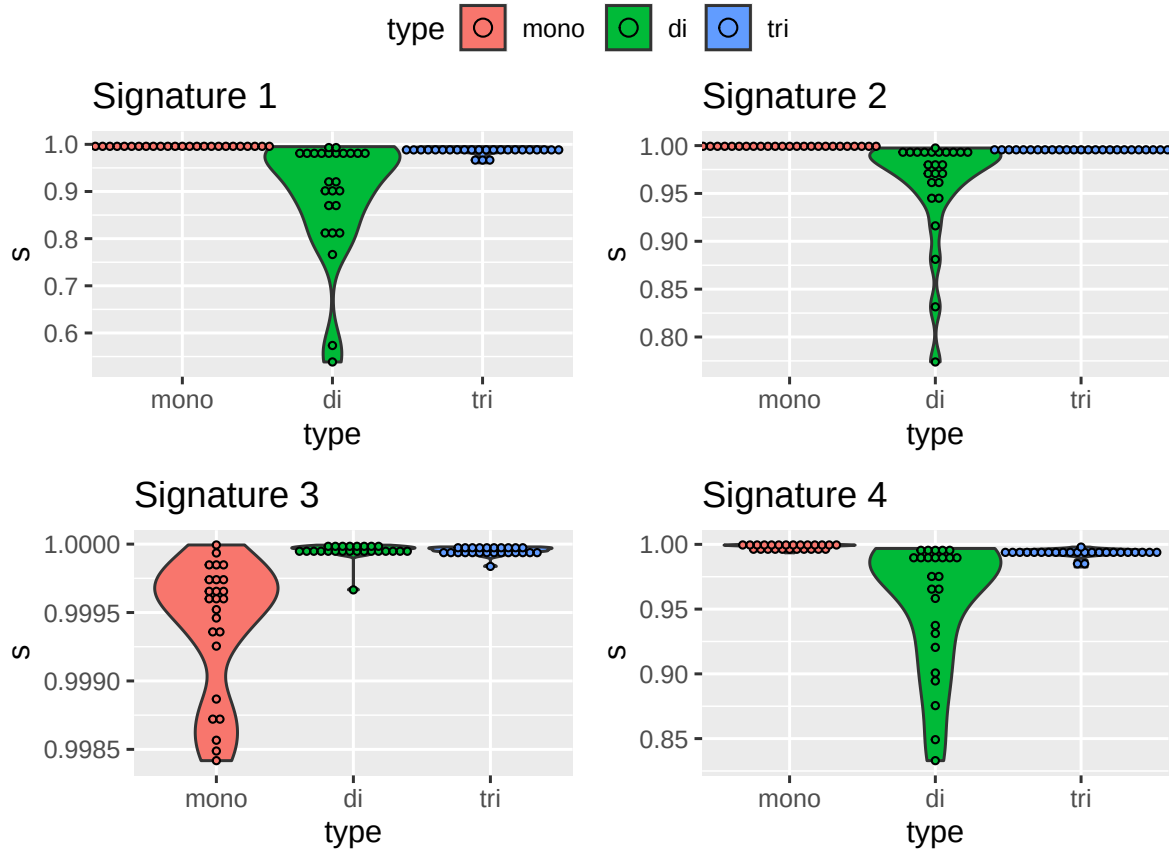

Figure 6: Caption

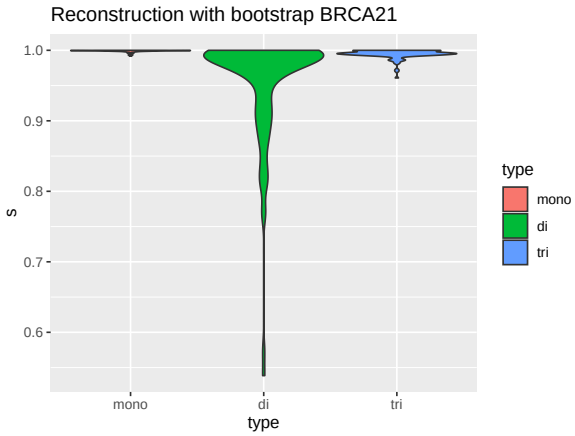

Figure 7: Caption

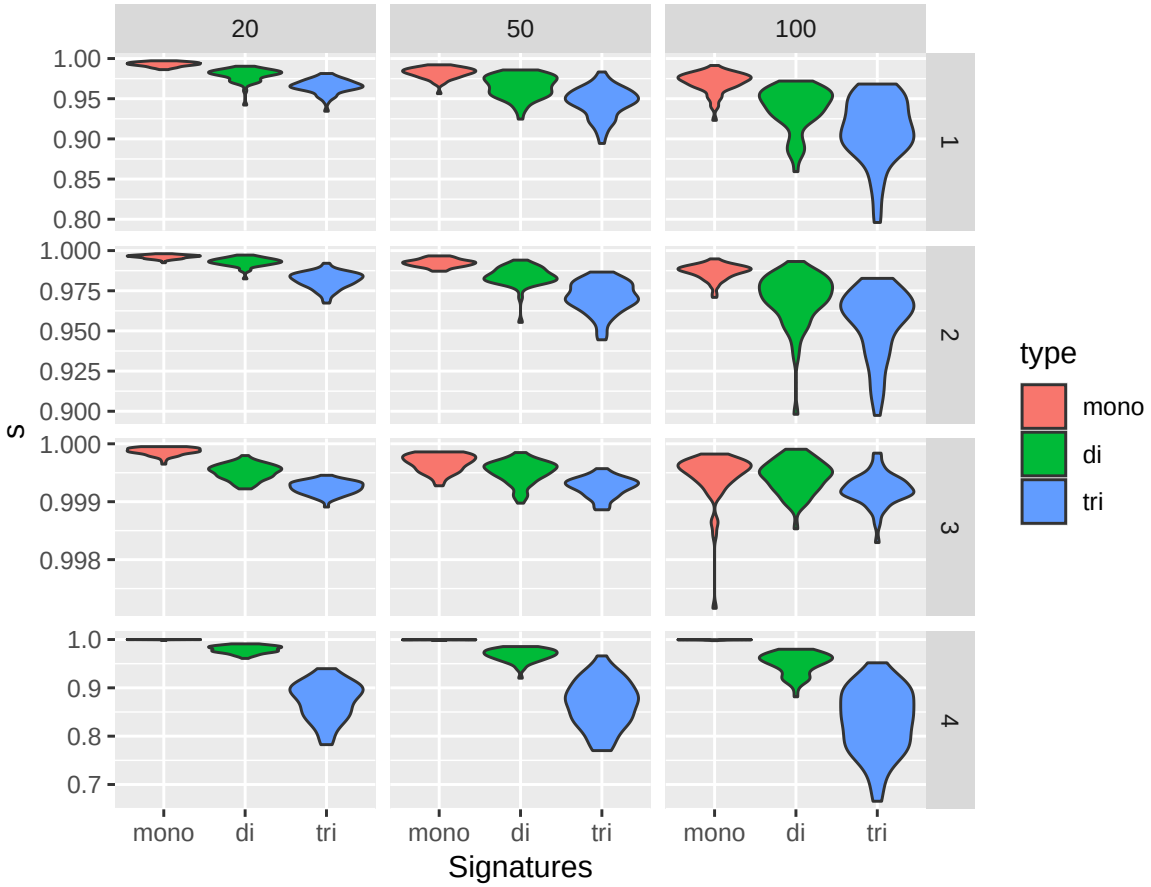

Figure 8: Caption

Figure 9: Caption

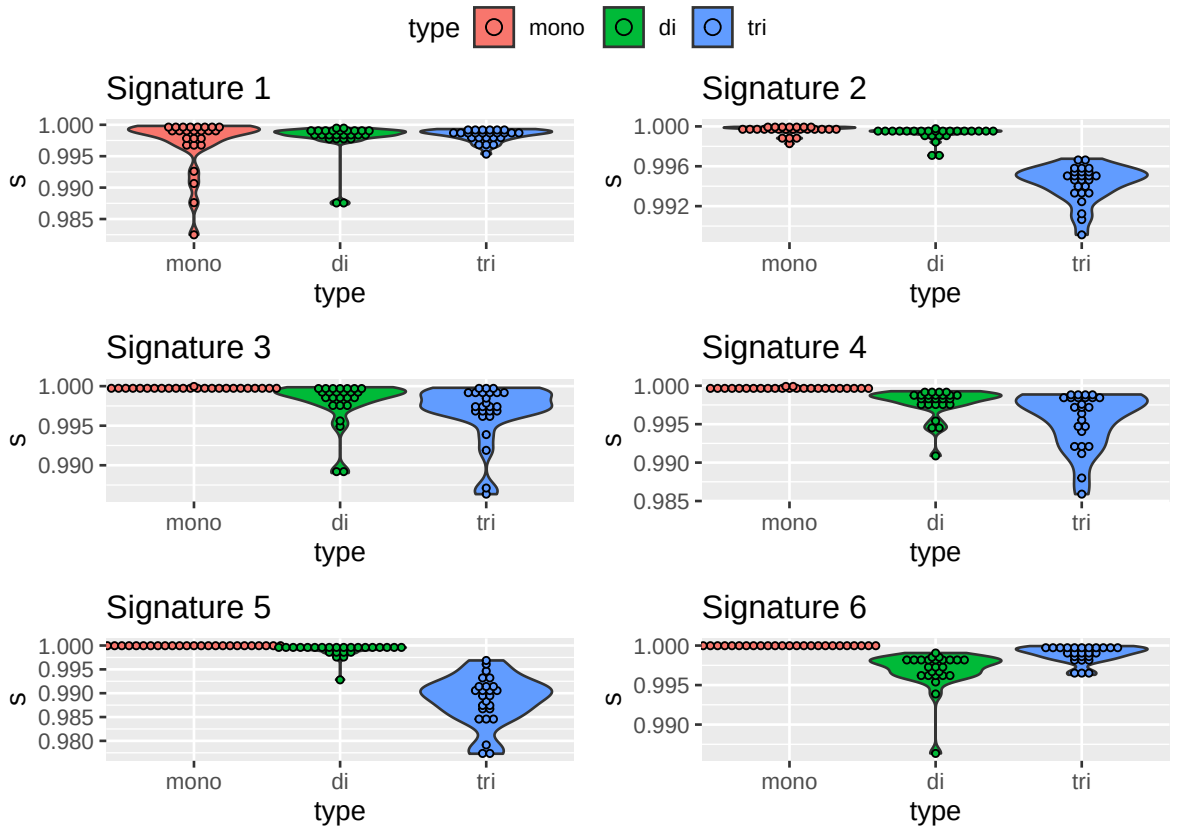

Figure 10: Caption

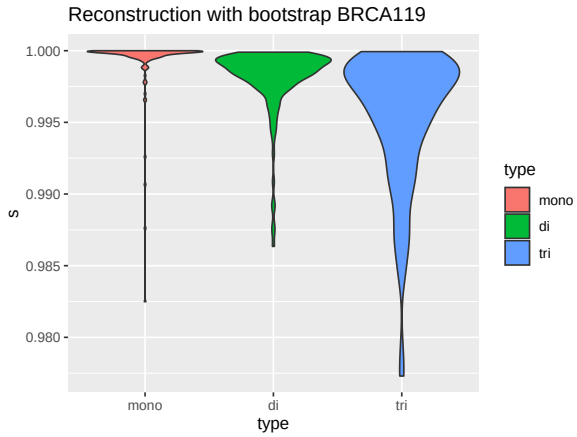

Figure 11: Caption

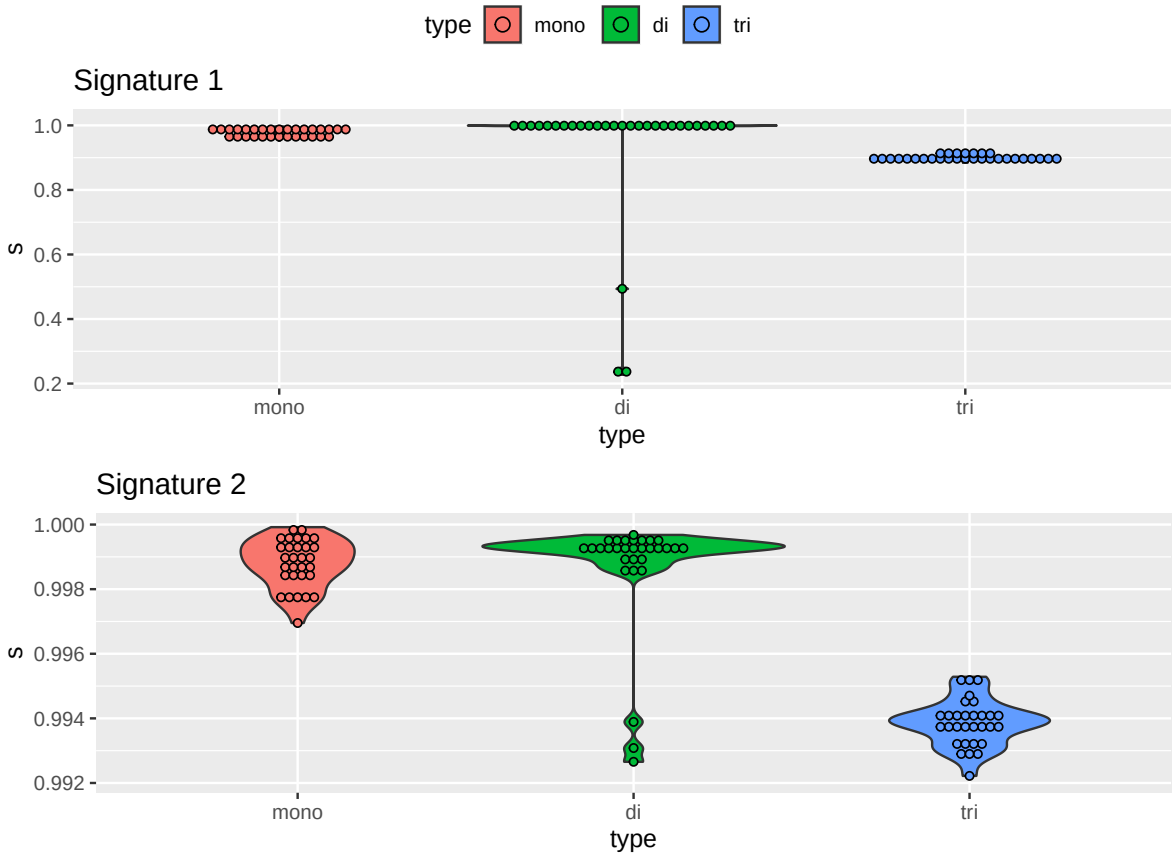

Figure 12: UCUT bootstrap

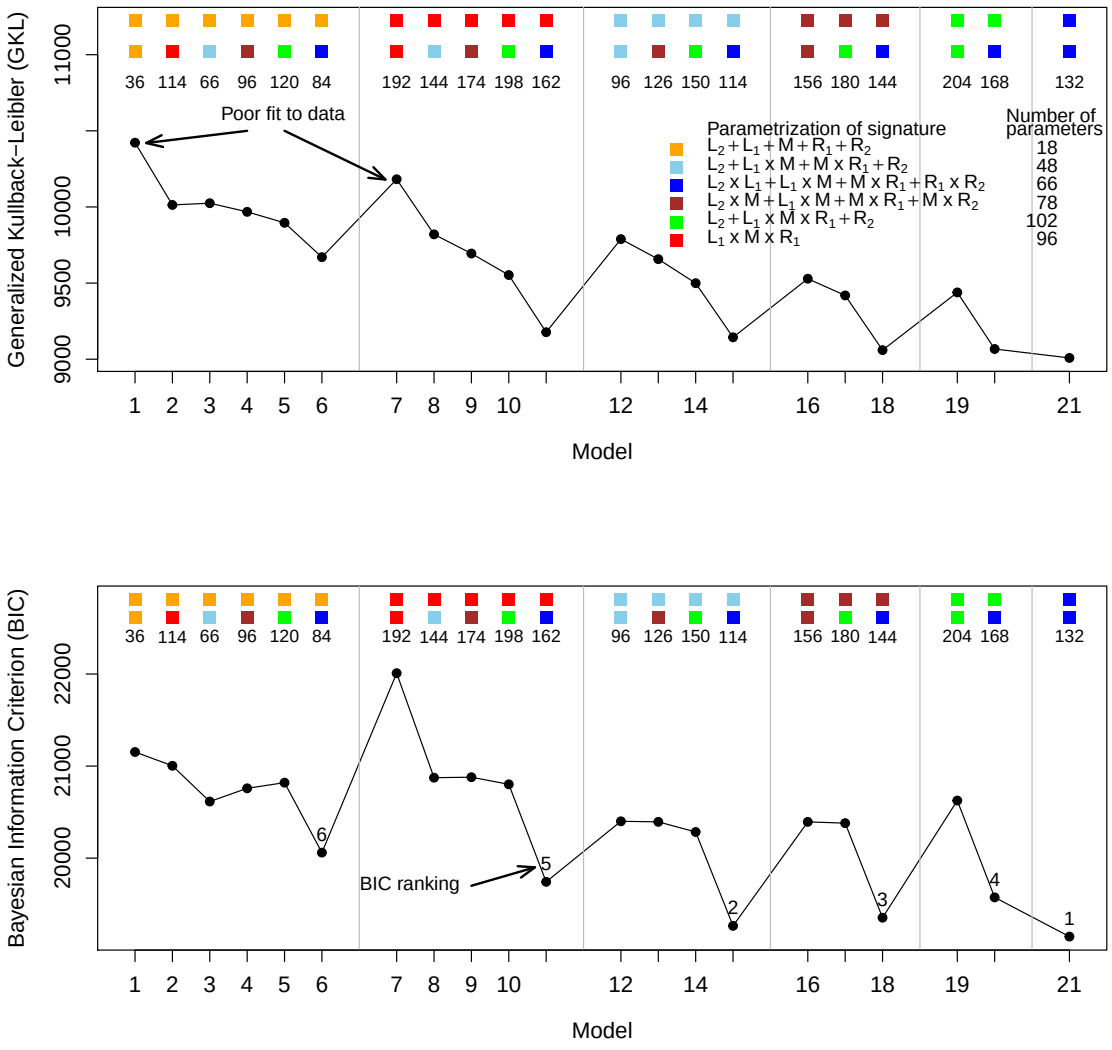

Figure 13: The Generalized Kullback-Leibler and Bayesian Information Criteria for all 15 models with two signatures for the UCUT26 data set. The models are ordered according to the total number of parameters for the two signatures; e.g. $2 \cdot 18 = 36$ for the sole mono-nucleotide model and $2 \cdot 48 = 96$ for the tri-nucleotide model.