Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Statistical modeling of repertoire overlap in entire sampling spaces

Note no.	SAMBA/13/17
Authors	Lars Holden, Martin Jullum, Geir Kjetil Sandve
Date	8. mai. 2017

Authors
Lars Holden, Martin Jullum, Geir Kjetil Sandve

Norsk Regnesentral

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
office address
Gaustadalleen 23a
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Bankkonto · bank account
8200.01.48888
Org.nr. · enterprise no.
NO 952125001 VAT

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Statistical modeling of repertoire overlap in entire sampling spaces**
Authors	**Lars Holden, Martin Jullum, Geir Kjetil Sandve**

Date	8. mai. 2017
Year	2017
Publication number	SAMBA/13/17

Abstract

We analyze the distribution of T-cell clonotypes in a compartment like blood based on samples. In particular, we study how the distribution of clonotype frequencies changes between different samples. We consider this as a sampling problem and formulate the problem as a generalization of the classical statistical problem of comparing samples from an urn. Due to the low sampling size compared to the number of different clonotypes in the entire sampling space, the classical methodology that works directly with clonotype frequencies in samples is not suited. We approach this challenge by representing other properties of the sample. Our re-representation allows for easy sampling model fitting and testing under natural model conditions. Although we here focus on the application on clonotypes, the new methodology generalizes seamlessly to other applications.

Keywords	T-cell clonotypes, multinomial models, sampling with repertoire overlap
Availability	Open
Research field	Biostatistics
Number of pages	15
© Copyright	Norsk Regnesentral

Table of Content

1 Introduction

We analyze biological samples that consist of a large number of different types of cells. One example is the distribution of T-cell clonotypes in blood. In particular, we study how the distribution of clonotype frequencies changes between different samples. We consider this as a multinomial sampling problem.

The multinomial distribution is classically formulated in terms of an urn containing a number of balls with different colors, where balls are drawn with replacement. Also other standard discrete/categorical statistical models, like the Bernoulli, binomial, geometric and negative binomial distribution may be formulated as a specific sampling scheme from such an urn. Classical urn problems concern estimation of parameters of such models, and testing hypothesis based on samples from the same or different urns. See e.g. Johnson and Kotz (1977) for various applications of urn problems. We consider the comparison of clonotypes over time as an urn problem. Christophersen, et al. (2017) also describes the problem with clonotypes instead of colors and T-cells instead of balls.

What distinguishes our problem from classical urn problems is that we in principle have no knowledge about which or how many different clonotypes (i.e. colors) are in the urn. As an example, one might expect a compartment like blood or duodenum to typically contain 200-2000 unique clonotypes specific to gluten. However, in general there are in the order 10^{18} different possible clonotypes (Venturi, V. et al. 2007), of which millions or billions would likely be specific to gluten, and are a priori equally likely to be in our samples. This means that we cannot explicitly list the potential clonotypes that could be found in the sample from a specific patient. In a context of gluten-specific T-cells, the total number of T-cells in each sample will typically be in the range 5-150. Thus, the sample size is generally small compared to the number of clonotypes in the underlying compartments, such that we are only able to observe a small number of the clonotypes. Each new sample consists mainly of new clonotypes that has not been observed earlier. Instead of a standard approach for multinomial distribution, it is necessary to reduce the number of parameters and estimate parameters using a variant of M-estimation (Wilcox, R. R. 2012), i.e. by optimizing a function that is not the likelihood.

2 Notation and models

Our approach consists of a statistical model for the frequency of the different clonotypes. We want to compare a set of k different samples from a single patient, i.e. how the distribution has changed between the samples. Let $X_{.,.} = \{X_{i,j}\}_{i=1,\dots,N_T, j=1,\dots,k}$ denote all the observations from a patient where $X_{i,j}$ is the number of T-cells of clonotype i in sample j from the patient for $i = 1, \dots, N_T$ and $j = 1, \dots, k$. Also, many of the clonotypes are only observed in some of the samples, i.e. the most frequent value is $X_{i,j} = 0$. Let also $M_{j,T} = \sum_i X_{i,j}$ be the number of T-cells of sample j, i.e. its sample size, and let $N_{j,T}$ denote the number of unique clonotypes in sample j. The data may be considered as samples of k different urns; the unobserved population of T-cells/clonotypes in the blood at the time the sample was extracted. Let thus $n_{i,j}$ be the (unknown) number of T-cells of clonotype i in the blood at the time of sample j, for $i = 1, \dots, N$ and $j = 1, \dots, k$, where N denotes the total number of different clonotypes which have ever appeared in the blood for this patient. Denote also by $M_j = \sum_i n_{i,j}$ the total number of T-cells in the blood at the time of sample j.

In correspondence with the standard urn sampling case, we assume that for each sample the $M_{j,T}$ T-cells are sampled randomly and independent with replacement from the M_j T-cells in the blood with a probability $M_{j,T}/M_j$. Whether we assume that the sampling is performed with or without replacement is of minor importance here since $M_{j,T} \ll M_j$. Typical sizes for the above quantities are $N_{j,T}$: $5-50$ (unique colors in sample j), N_T: $15-100$, (unique colors in all the samples from a patient), $M_{j,T}$: $5-150$ (number of balls in sample j) and M_j: $30\,000 - 500\,000$ (number of balls in the urn when extracting sample j). It is difficult to quantify the size of N, but a rough guess is that it is in the range 200-2 000. That is, we have $N_T \ll N$.

Our interest concerns the frequencies of the clonotypes in the blood, $p_{i,j} = n_{i,j}/M_j$, for the clonotypes $i = 1, \dots, N$ and for samples $j = 1, \dots, k$, for each patient, rather than the counts $n_{i,j}$ themselves. It is easier to estimate the frequencies than the nominator and denominator of the fraction. The variables $n_{i,j}$ for all values of i may increase at the same time as a biological reaction and $p_{i,j}$ seam to be more stable quantities. We are interested in the distribution of the frequencies and assume the ordered frequencies follow the parametric form $p_{(i),j} = c_j/(a_j - 1 + i)$ for a constant c_j set such that $\sum_i p_{(i),j} = 1$. Here, we use the ordered variables, i.e. the frequencies satisfy $p_{(i),j} \geq p_{(i+1),j}$ for all i, j. We have tested other functional forms of $p_{(i),j}$ and found out that the above function fits the data best. This model describes the distribution of frequencies of clonotypes with one parameter, a_j. Here there are some highly frequent clonotypes and many clonotypes with low frequency.

Our analysis of the data shows that it is not possible to estimate N, the number of clonotypes in the blood. For most samples we get an equally good fit with the data for any value of N between the observed N_T and more than 1000. There exists parameters such that $max_i |p_{(i),j,N=200} - p_{(i),k,N=1000}| < 0.01$, making it impossible to differentiate between samples from $p_{(i),j,N=200}$ and $p_{(i),k,N=1000}$. The reason for this is that we are not able to differentiate between many clonotypes with low frequencies and even more clonotypes with even smaller frequencies. Most of these clonotypes will not be observed, and if observed, they will only be observed once. Therefore, we need to assume a total number of different clonotypes that have ever appeared in the blood for this patient, say $N = 500$, and interpret the say 450 least frequent clonotypes as representatives of the low frequent clonotypes that we cannot expect to observe (and if observed, this will most likely be only once). It is not possible to classify this in one group since we are not able to distinguish between a T-cell from a medium frequent clonotype with one observation (among the 50 most frequent) and a very rare clonotype (not among the 50 most frequent). A different value of N, would change the estimated c_j and a_j, but only have marginal effect on the most frequent $p_{i,j}$. This implies that the frequency distribution can be represented by one parameter a_j for each sample. In the analyses we have conducted, we have also explored alternative values of N to assess robustness of any interpretations.

Our data indicate that $0.01 < a_j < 100$. With only one parameter determining the frequency distribution, there is a functional relationship between the a_j parameter and other diversity measures like Shannon entropy and D50. D50 is the fraction of the clonotypes that are necessary in order to include 50% of the T-cells. This may be measured both in the population and in a sample, see Figure 1.

We are mainly interested in the change of frequencies in the compartments between the time of the different samples of the same patient. We expect there to be some changes in the ordering, i.e. we may have $p_{i,j} > p_{i+1,j}$ and $p_{i,k} < p_{i+1,k}$ for some value of i. We are mainly interested in pairwise comparison of order and compare this with time between samples, simultaneous sampling of different compartments of the body etc. Let the subscript i denote the ordering of sample j, i.e. $p_{(i),j} = p_{i,j}$. We model the sorting of the sample k as the ordering by $g_{j,k}(i)$ i.e. $p_{(g_{j,k}(i)),k} = p_{i,k}$. We define $g_{j,k}(i)$ as the sorting of $V_{i,j,k} = i + b_{j,k}Q_{i,j,k}$ where $b_{j,k} \geq 0$ are the parameters in the model and $Q_{i,j,k}$ are independently distributed from the distribution $Q_{i,j,k} \sim \mathrm{N}(0,1)$. For $b_{j,k} = 0,$ the ordering is the same. For $b_{j,k} > 1500$ and $N = 500$ we consider the ordering of the two samples as independent or as close to independent that we are not able to identify a possible dependence. Our model of two samples is described with the parameters $\theta = \{a_j, a_k, b_{j,k}\}$. The definition may be generalized to more samples by selecting one of the samples and make comparisons between the selected sample and the other samples. We have chosen this model for describing the change of ordering between the samples since it is easy to simulate from the model.

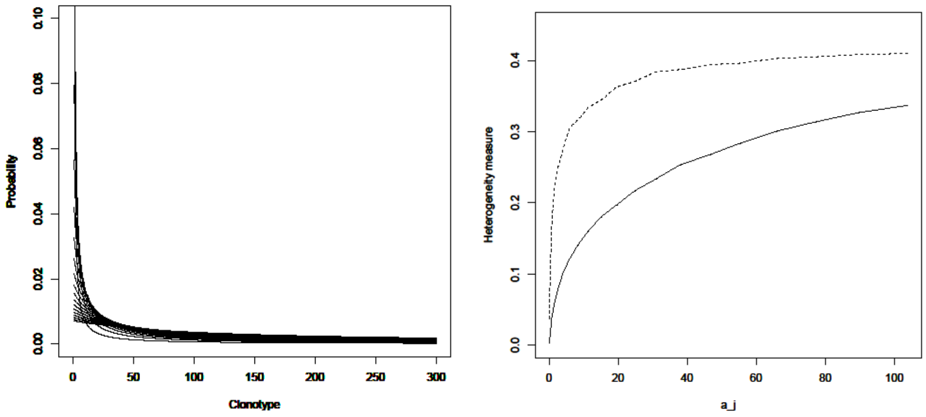

Figure 1. The left panel shows 20 different $p_{i,j}$ curves for $0.1 < a_j < 100$. We have $p_{(1),j} = 0.61$ for $a_j = 0.1$. The right panel shows the D50 heterogeneity measures for the population (curve) and for a sample with 100 T-cells (dashed). N=300 in the figure.

It is possible to measure the difference in ordering by the L_1norm

$$d_{j,k} = \sum_i |p_{i,j} - p_{i,k}| \geq \sum_i |p_{(i),j} - p_{(i),k}| = d_{j,k,min}$$

We have $0 \leq d_{j,k} \leq 2$. If the ordering of the two samples are the same at the time of the two samplings, we obtain the minimum L_1norm difference, $d_{j,k,min}$. Then we have equality in the above formula, otherwise we have an inequality. Since most of the frequencies are very small, independence of the frequencies for sample j and k implies that $d_{j,k}$ is close to 2, except when both a_j and a_k are large, giving $p_{i,j}$ almost the same for all values of i. We have introduced the L_1norm since it is easier to interpret $d_{j,k}$ than the variables $b_{j,k}$.

We define $d_{j,k,ind}$ as the value of $d_{j,k}$ when the order of clonotypes is independent between the two samples. Further we calculate a ratio $r = (d_{j,k} - d_{j,k,\min})/(d_{j,k,ind} - d_{j,k,min})$ as a measure for how close the order is. The ratio is equal to 0 for the same order and equal to 1 if the order is independent. If $b_{j,k} = 1000$, for $N = 500$, then $r \approx 0.9$ which in practice is close to independent. In order to get a better intuition on the variable $d_{j,k}$, we have made Table 1.

	a_j	0.1	1	10	100
a_k	$p_{(1),j}, p_{(2),j}$	0.60, 0.055	0.15, 0.074	0.025, 0.023	0.00557, 0.00551
0.1	$d_{j,k,min}, d_{j,k,ind}$	0, 1.76	0.91, 1.65	1.23, 1.60	1.42, 1.56
1	$d_{j,k,min}, d_{j,k,ind}$	0.91, 1.65	0, 1.37	0.50, 1.24	0.89, 1.15
10	$d_{j,k,min}, d_{j,k,ind}$	1.23, 1.60	0.50, 1.24	0, 1.0	0.47, 0.86
100	$d_{j,k,min}, d_{j,k,ind}$	1.42, 1.56	0.89, 1.15	0.47, 0.86	0, 0.53

Table 1 The second row shows the two highest probabilities $p_{i,j}$. The four lowest rows show $d_{j,k,min}$, and $d_{j,k,ind}$. Here we assume N=500.

3 Estimation

From the data $X_{.,.}$, we observe the empirical frequencies $p_{i,j,D} = X_{i,j}/\sum_k X_{k,j}$. The probability for one of the samples, the observation $X_{.,j}$, is the sum of the multinomial distribution for all possible permutations H,

$$P(X_{.,j}) = \frac{1}{\#H} \sum_{H} \frac{M_{j,T}!}{X_{1,j}! \dots X_{N_T,j}!} (p_{1,j})^{X_{1,j}} \dots (p_{1,j})^{X_{N,j}}$$

The number of permutations, $\#H$, is very large. It is feasible to find the maximum likelihood estimate of a_j since there is only one parameter and the most likely ordering of the high frequencies is close to the empirical ordering. However, the maximum likelihood estimate should be very close to the value $\hat{a_j}$ that minimize

$$T(a_j) = E\{ \sum_{i} \left| p_{(i),j,D} - p_{(i),j,S(a_j)} \right| \}$$

where $p_{(i),j,S(a_j)}$ are the frequencies from sampling using the distribution with the value a_j. The use of (i) in the subscript indicates that we use the ordered frequencies, i.e. each difference in the sum is not necessarily representing the same clonotype, i.e. we take the difference between the i'th most frequent clonotype in the data, $p_{(i),j,D}$, and the i'th most frequent clonotype from the sampling, $p_{(i),j,S(a_j)}$. The $S(a_j)$ sampling has the same number of T-cells as the original D sampling from the data. The estimate $\hat{a_j}$ is easily found by simulating from the multinomial distribution. We also find the uncertainty in the estimate by assuming $\widehat{a_j}$ is the correct value, simulating using this value, and then find $\widehat{a_j}$ for each simulated data set.

Similarly, it is possible to estimate $\widehat{b_{j,k}}$ for the change in order of each pairwise combination by minimizing

$$W(b_{j,k}) = (E\{\ \textstyle\sum_i \left|p_{i,j,S(\widehat{a_j})} - p_{i,k,S(\widehat{a_k})}\right|\ \}\ \text{-}\textstyle\sum_i \left|p_{i,j,D} - p_{i,k,D}\right|\ )^2$$

Here we minimize the simulated L_1 difference of the two samples with the observed difference of the empirical frequencies. We prefer to estimate $a_j,\ a_k$ and $b_{j,k}$ separately in order to avoid different estimates for a_j for each pairwise combination and since this will only have minimum influence of the estimate. It is not feasible to use maximum likelihood estimate for $b_{j,k}$ based on the multinomial distribution since the different ordering of the samples increases the number of permutations considerably. Based on the estimated $\widehat{a_j}$, $\widehat{a_k}$ and $\widehat{b_{j,k}}$ we may calculate both

$$d_{j,k} = \sum_i |p_{i,j} - p_{i,k}|$$

and

$$d_{j,k,min} = \sum_i |p_{(i),j} - p_{(i),k}|$$

describing the difference between the two samples using the parameter value $\widehat{b_{j,k}}$.

We may estimate the uncertainty for these parameter estimates by assuming that the estimated values $a_j, a_k, b_{j,k}$ are correct, and then simulate $p_{i,j,S(a_j)}, p_{i,k,S(a_k)}$ based on these parameters and find the uncertainty in these estimates using the approach described above.

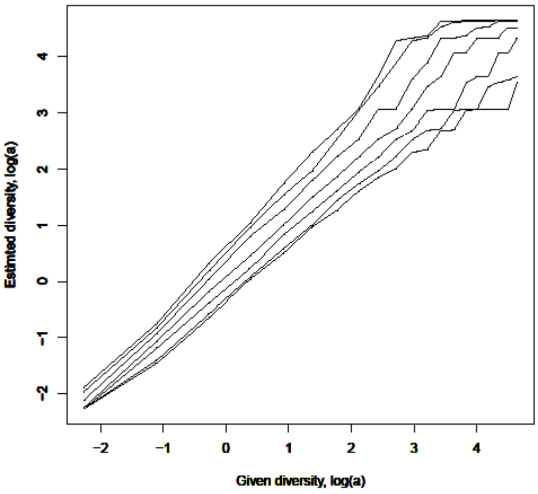

Figure 2. Estimated diversity, a in a log-log plot. The estimated quantile based on 300 simulations and 20 different a values.

In the following we give some examples on the estimation from the model. In order to fully understand the model, it is necessary with a more thorough analysis. In all the examples we assume the sample consist of 100 T-cells. In all the figures we show the seven curves for the 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, and 0.95 quantiles for the estimated value. In Figure 2 we simulate only one sample with a fixed diversity value in the interval $0.1 < a_1 < 100$ and estimate the diversity value based on the sample. In Figure 3 and 4 we simulate two samples of 100 T-cells and vary the change of order, i.e. the $b_{j,k}$ value in the interval $0.1 < b_{j,k} < 10.000$. We only search for optimal a_j and $b_{j,k}$ values in the same interval implying that the quantiles may be truncated in the ends. Note the plots for a_j and $b_{j,k}$ are log-log since the values vary many orders of magnitude. In Figure 3 we show estimated $b_{j,k}$ and r values and in Figure 4 we show estimated $d_{j,k}$ and $d_{j,k,emp}$ values. Note from the figures that we are able to estimate a_j quite accurately from a sample with 100 T-cells. Naturally, the uncertainty in the $b_{j,k}$ is much larger. From Figure 4 we note that there is a large difference in the $d_{j,k}$ and $d_{j,k,emp}$ values and these values depend strongly on the $b_{j,k}$ value.

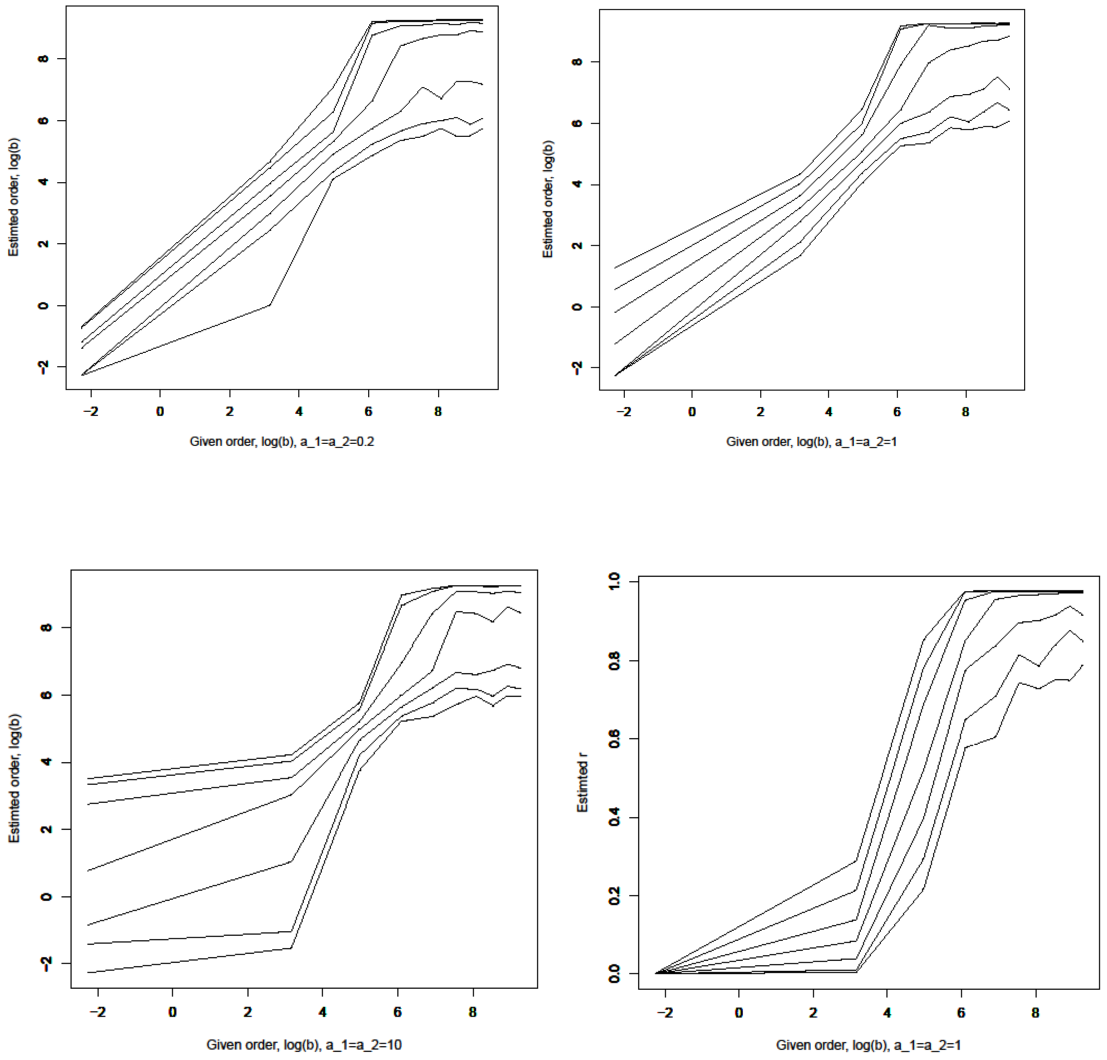

Figure 3. Estimated change in order variable $b_{j,k}$ in first three panels $a_j = a_k = 0.2$ in upper left panel, $a_j = a_k = 1$ in upper right panel and $a_j = a_k = 10$ in lower left panel

and r in the lower right panel with $a_j = a_k = 1$. The estimation is based on 100 simulations and 10 different $b_{j,k}$ values.

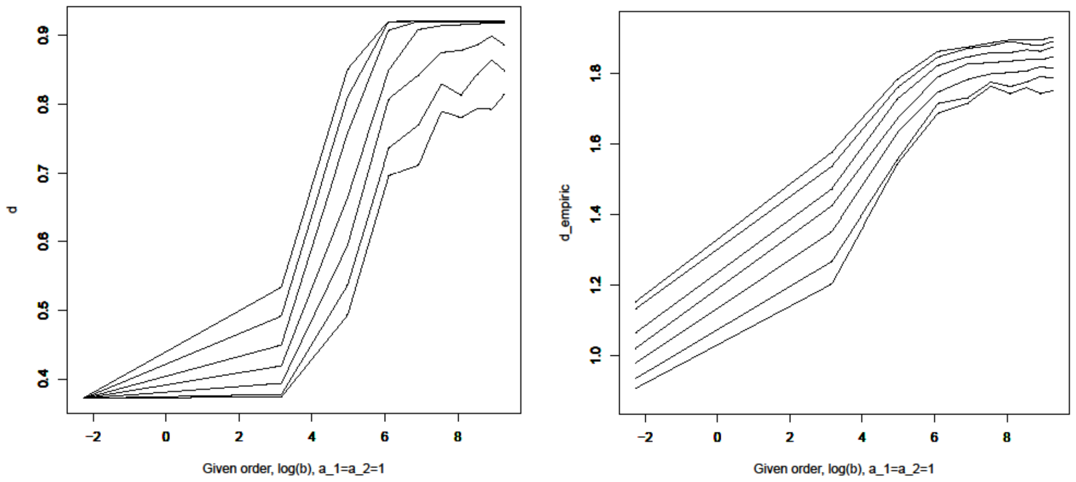

Figure 4. The estimated $d_{j,k}$ in left panel and $d_{j,k,emp}$ in right panel b varies in the interval correspond to $0.1 < b_{j,k} < 10.000$ and with $a_j = a_k = 1$. The estimation is based on 100 simulations and 10 different $b_{j,k}$ values.

References

Christophersen, A., Risnes L.F., Neumann R.S., Holden L, Sandve G.-K., Qiao S.W., Lundin K.E. and Sollid L.M., Disease-driving CD4+ T-cell clonotypes persist for decades in blood and gut tissue in celiac disease (in preparation)

Johnson, Norman L., and Kotz, Samuel (1977); Urn Models and Their Application: An Approach to Modern Discrete Probability Theory, Wiley ISBN 0-471-44630-0

Lythe, G., Callard, R., Hoare, R., Molina-Paris, C. How many TCR clonotypes does a body maintain? J. of Theoretical Biology, 389 (2016) 214-224.

Venturi, V., Kedzierska, K., Turner S.J., Doherty, P.C., Davenport, M.P., Methods for comparing the diversity of samples of the T cell receptor repertoire. Journal of Immunological Methods 321 (2007) 182-195.

Wilcox, R. R. (2012). Introduction to Robust Estimation and Hypothesis Testing, 3rd Ed. San Diego, CA: Academic Press.