Imperial College London

Imperial College of Science, Technology and Medicine

Department of Life Science

MSc Computational Methods in Ecology and Evolution

Using genetic frequency time-series to detect sites under selection

Author:
Chuxinyao Wang
CID:
02248307

Supervisor:
Dr. Bhavin Khatri

August 2023

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science/Research at Imperial College London Submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration

The actual data for this study came from Prof. Nastoulli of UCL Hospital (UCLH). The simulated data for this study was coded and implemented by myself. In addition, part of the code for parameter optimisation and multi-hypothesis testing analysis was written and provided by Dr. Khatri. The analysis of the data was completed by myself under the guidance of Dr. Khatri. The simulations and data analysis for this study involves large computation, so high-performance computing were required for most of the work. Regarding the availability of the code, since the code of the analysis part is still in the testing stage and the ownership belongs to Dr. Khatri Bhavin, it cannot be made public. Please contact Dr. Khatri Bhavin if necessary. For the part of simulation generating frequency data, please email the author himself.

Abstract

The Wright-Fisher(WF) model is a discrete-time method for describing the evolution of populations. Under the framework of the WF model, this study uses an approximate solution of the stochastically change of gene frequency to evaluate its feasibility of inferring the evolution of virus populations in nature. Here, we conduct simulations of single site analysis and genome wide analysis. Single site analysis is to test the ability of such method to characterise optimum parameters. And genome wide analysis is to evaluate feasibility in detecting selection sites in genes under likelihood estimation method as well as the statistical analysis of multi-hypothesis. We apply this method to two intrahost datasets of immunocompromised patients infected with SARS-CoV-2, then interpret the results in light of the simulation analysis. However, we noticed that different parameter constrained situations can affect the accuracy of detecting selection.

Contents

1 Introduction

The time series of allele frequencies is of great help in many studies in the field of population genetics and evolution (Bollback et al. 2008), such as human ancient DNA (aDNA) data (Hummel et al. 2005), virus population data(Shankarappa et al. 1999), and experimental evolutionary population data (drosophila, bacteria, phage) (Buri 1956, Woods et al. 2006, Bollback & Huelsenbeck 2007). This is because the strength of selection and the anticipated change in allele frequency over time are closely related (Bollback et al. 2008). Current methods for detecting genetic selection in evolution heavily rely on the analysis of synonymous and nonsynonymous substitutions ratio (dN/dS). Despite the established utility of the dN/dS ratio in comparing evolutionary processes among different independent populations, its application to gene sequences within a single population has shown reduced sensitivity in detecting natural selection (Kryazhimskiy & Plotkin 2008). Hence, dN/dS is not applicable for the study of segregating variation in a population. Although time-series methods for detecting selection of single site exist and are popular (Bollback et al. 2008, Malaspinas et al. 2012, Tataru et al. 2015), a genome wide selection analysis is still computationally expensive Khatri (2016). Whereas, an analytical stochastic dynamics solution would have the significant advantage of directly evaluating the likelihood function, allowing for very efficient computation of the maximum likelihood estimation for parameter optimisation (Khatri 2016). An analytical solution was developed by Khatri (2016) which can use gene frequency data changes over time can explore the impact of three genetic pressures (genetic drift, selection and mutation) on species population evolution. This study will evaluate the use of this method on estimating parameters and detecting selection in evolution at both genome-wide level and single site level.

Despite the fact that the method of estimating the selection coefficient of a single genetic locus from time-series data has been extensively used in the

study of the evolution of simple traits (Jewett et al. 2016), for the evolution of viruses, such as RNA virus, due to its high variability, high mutation rate, and low virus replication cycle characteristics (Domingo et al. 1996), their evolutionary history are more difficult to determine from direct observation of phenotypic changes. Therefore, it is considered here that starting from the genotype of the virus, the evolution of the virus can be directly obtained through the changes in the nucleotides at the locus on the sequence. Hypothesis testing for an entire gene sequence presents another problem, increasing the number of Type 1 errors (Goeman & Solari 2014). If only two entire gene sequences are compared for hypothesis testing, there might be about 500 sites occur Type I error on the gene sequence if $\alpha = 0.05$, $L = 10,000$ number of locus, which is unacceptable. Therefore, this study will also use the method of multi-hypothesis testing to verify and compare the evolution results in sequences to reduce Type I error.

In this study we simulate gene frequency change time series using the Wright-Fisher genetic model under conditions of genetic drift, selection and mutation, assuming two variant possibilities (wild-type and mutant). The three parameters of N, μ and s are estimated by Maximum Likelihood Estimation (MLE), and p-values calculated by the Likelihood Ratio Test (LRT) used as a decision statistic for multi-hypothesis testing from simulated genome wide data. Finally, the Receiver Operating Characteristic (ROC) curve (Zweig & Campbell 1993) is used to obtain Area Under the ROC Curve(AUC) to test the performance of multi-hypothesis testing. In addition, the study also used the real data of SARS-CoV-2 patients to verify the accuracy of the method, and explored the evolved loci differences in virus gene sequence in patients with immunocompromised (Choi et al. 2020, Clark et al. 2021, Markov et al. 2023) infection, specifically, locate locus under selection from virus intra-host evolution.

2 Methods

2.1 Use and implementation of mathematical methods

The study adopts the WF model and considers the effects of genetic drift, selection, and mutation in haploid RNA viruses (SARS-CoV-2 in this study) with N individuals, consider the wild type and mutant $(N-n)$ individuals (n). Genetic drift refers to the process of random changes in gene frequency due to the influence of random events in a finite population. This change is caused by random factors, and the effect is more significant in small populations. Genetic drift can be simplest described as binomial sampling in WF model, which is half the variance of the variation frequency between generations (**Eqn.1**) (Kimura 1983). Selection refers to the variation of gene frequency caused by the adaptive difference among individuals in nature, which can be favorable adaptation or unfavorable adaptation. In addition, mutations refer to random, heritable changes that occur in a gene or genome, and also has forward and backward mutations. Expressed in discrete form is **Eqn. 2**. It represents the average change in mutation frequency per generation due to selection and mutation. Where s is the selection coefficient. If it is a preference choice from the wild type to the mutant type, then $s > 0$ means positive selection, otherwise $s < 0$ means negative selection. μ_1 is the forward mutation coefficient, indicating the mutation rate from wild type to mutant type, and vice versa, the backward mutation rate from mutant type to wild type is μ_2.

$$D(x) = \frac{x(1-x)}{2N} \tag{1}$$

$$M(x) = sx(1-x) + \mu_1(1-x) + \mu_2 x \tag{2}$$

Eqn. 3 and 4 are coded in MATLAB to simulate genetic pressure. Let $n(t)$ be the number of mutant type in generation t and $x(t) = n(t)/N$ be

the corresponding frequency of mutant type. The random mating of the population leads to a count of wild type in generation $t+1$ that is binomial distributed (Eqn.6) (Wright 1931, Crow & Kimura 1970, Fisher 1999, Ewens 2004). Therefore, $n(t+1)/N$ is the frequency of the variant in generation $t+1$ (Eqn.7). This enables the implementation of random drift changes into computer simulations.

$$\frac{\delta x}{\delta t} = M(x) \tag{3}$$

$$\delta x = M(x)\delta t \tag{4}$$

$$n(t+1) \mid n(t) \sim Bin(N, x(t) + \delta x) \tag{5}$$

$$x(t+1) = \frac{n(t+1) \mid n(t) \sim Bin(N, x(t) + \delta x)}{N} \tag{6}$$

Besides, $M(x)$ (Eqn.2) can be used to simulate the impact of selection and mutation on gene frequency on computer.Yet positive selection and negative selection are categorised and used in this study because this research covered both of these types of selection. WF model was run in MATLAB in discrete time to perform independent simulations for each site (assuming that each site in the genome evolves independently). Simulate actual sampling by controlling the sampling time interval. As a result, a set of matrix f (number of time sites, length of the genome) will be obtained to store the sampling time (t), and the frequency of the mutant gene in the population corresponding to the time ($x(t)$) to evaluate the accuracy of the maximum likelihood parameter estimation .

2.2 Single site Monte-Carlo simulation with selection, mutation and genetic drift under Wright-Fisher model

Setting initial parameters, timescales and frequencies of the simulated data are based on the establishment of mutants proposed by Desai & Fisher (2007). Establishment describes a beneficial mutant needs to survive and overcome genetic drift to reach certain frequency before fixing, that frequency is established frequency. For positive selection, especially the selection advantage ($2Ns >> 1$), according to Desai & Fisher (2007), there is an average frequency ($1/2Ns$) for mutants to establish. Before this frequency is reached, mutants increases in specific rate to reach the establishment as soon as possible, it takes around $\tau_{est} = \frac{\ln 2}{s} > \frac{1}{2s}$. Once reached, it will increase at an exponential rate until it is fixed. Therefore the total mean time to fixation from zero frequency can be calculated by **Eqn. 7** (Desai & Fisher 2007).

For negative selection, there is mutation selection balance (Crow & Kimura 1970). Mutation-selection balance refers to a situation in which the rate at which new deleterious mutations are introduced into a population is balanced by the rate at which they are removed through selection. By using **Eqn. 2**, equilibrium can be found, which is $x_{eqm} = \frac{\mu}{|s|}$ when ($\mu = \mu_1 = \mu_2$), and time to reach the equilibrium is $\tau_{eqm} = \frac{1}{|s|}$.

$$\tau_{fix} = \frac{1}{2s} \times (1 + \ln(Ns)) \tag{7}$$

In order to match virus data situation (Seo et al. 2002, Sanjuán et al. 2010), $N\mu$ is set to 1 here, that is, the population size $N = 10000$, the mutation rate $\mu = 10^{-4}$.

We choose a selection coefficient $s = 0.005$ for positive selection, which

gives a mean time to fixation of $\tau \approx 261$ generations.

We choose $T = 10000$ much greater than τ for simulations under positive selection in this single-gene simulation, and with $\Delta\, t = \{10, 20, 30, 40, 50, 100, 150, 200, 300, 400, 500\}$ ($\tau \approx 1/2s = 100$ *generations*). Then, in order to explore the impact of stopping the simulation at fixation on parameter estimation, we cut off the simulation once if fixed to increase estimation accuracy.

In order to explore the influence on the selection coefficient estimation before and after the establishment, we set part of the initial frequencies below the establishment frequency and part above the establishment frequency (the establishment frequency is about $1/2Ns = 0.01$, $x_0 = \{0, 0.001, 0.01, 0.1, 0.5\}$.

For negative selection, s is set to -0.005 the initial frequency is set to the mean frequency $x_0 = \frac{\mu}{|s|} = 0.002$ in mutation-selection balance, which can be obtained from **Eqn. 1**. Same T and Δt are chosen for negative selection too. Whole single site simulation use Monte Carlo simulation to perform 500 simulations. The median value of the optimised parameters obtained by MLE is compared with the initial parameters in the range log_{10}.

2.3 Genome-wide simulation with selection, mutation and genetic drift under Wright-Fisher model

Genome wide simulations used the WF model with selection, mutation and genetic drift, including parameters population size (N), positive and negative selection coefficient ($s_{pos} > 0$, $s_{neg} < 0$), mutation rate (μ), fraction of negative selection sites in the genome (fr), initial frequency distribution ($x(0)$), total number of generations (T), sampling time interval (Δt), and genome length (L). Besides, sample by binomial distribution (**Eqn. 8**) to

simulate read-depth, and the number of reads is $ns = 100$.

$$x_{sampled}(i,k) = \frac{Bin(ns, X(i,k)}{ns} \tag{8}$$

We assume that the initial frequency is derived from a balanced distribution of allele frequencies in mutation-selection balance. When $s < 0$ and $2N|s| >> 1$ (Selective advantage) (Desai & Fisher 2007) this is approximately a gamma distribution ($shape = 2N\mu$, $scale = 2N|s|$) (Kim et al. 2017). For the initial frequency of the genome sites subject to positive selection, random sampling in the interval $[0, 1]$ is used to generate it. fr is 0.8 (in the case of $L = 10000$ genomic positions, the number of negative selection sites is 8000).

In addition, according to Kimura (1962), the value of $2N|s|$ that mutation will be under strong selection among population ($2Ns > 1$ or $2N|s| < -1$ or neutral ($0 < 2N|s| < 1$). Because of this, we assume an exponential Distribution fitness effect (DFE). DFE is a concept in genetics that describes the proportion of new mutations that are beneficial, neutral, or deleterious (Eyre-Walker & Keightley 2007). Note that for the fraction of neutral mutations in an exponential DFE is a simplification and approximation that holds under our assumptions which can be calculated by $\frac{1}{2N} = 5 \times 10^{-5}$ (Kimura 1979),so sites with $s \leq 10^{-5}$ is effectively neutral. When a fraction of sites in the genome are effectively neutral, it means that mutations occurring at those sites have minimal impact on an individual's fitness.

Based on the above, the simulation tests the impact of different possibilities of T and Δt on parameter estimation, and evaluate their performance through the ROC and AUC. For total iterative generations, there are $T = \{1000, 10000\}$. When $T = 1000$, Δt is set to $5, 10, 20, 50$, when $T = 10000$, Δt is set to $\{50, 100, 200, 500\}$.This is because under positive selection, more than generation of establishment time $\tau_{est} = \frac{\ln 2}{s} > \frac{1}{2s}$ can

better observe the fixation of mutant (Desai & Fisher 2007). So s_{pos} is set to 0.005 ($\tau_{est} = 100$ in this case). Besides, amount of sequences in this study is set to $N = 10000$, μ is 10^{-5}, satisfying $N\mu = 0.1$, and s_{neg} is set to 2×10^{-4}. The mutant gene frequencies of the whole population are stored in a matrix generated under different regimes. All matrices are executed through a set of purpose-built MATLAB programs, which contain the frequency change of each polymorphic site at time T. In order to explore the impact of sampling on the parameter estimations, simulations are also divided into two groups, one group is the result with binomial sampling as mentioned before, and the other group is the result without binomial sampling. The above simulation results are saved and used for further analysis in subsequent MLE parameter estimation and multi-hypothesis testing (see subsection 2.4 for details).

2.4 SARS-CoV-2 patient data processing and analysis

The patient data from March 2020 to February 2021 provided by Prof. Nastoulli of UCL Hospital (UCLH) are used in this study. Using the iVar tool, the data is transformed from the *bam* file format into a *tsv* file (Danecek et al. 2011). Single nucleotide variants (SNVs) and indels are called using the output of the command by iVar (Grubaugh et al. 2019). Each *tsv* file represents the gene frequency of the virus in a patient's host on a test date. By combining multiple *tsvs*, a list of SARS-CoV-2 virus nucleotide frequency records over time can be obtained. This is pre-processed to converted to '.mat' format by Dr. Khatri. And data named 'Patient 6' and 'Patient 9' are used for the analysis of this study. For each patient, 3 '.mat' files are stored. Respectively, an frequency array $f(i, j, k)$, i represents the gene frequency. The number of i should be 29903 nucleotides (Wu et al. 2020) for SARS-CoV-2; j represents the type of variation, including six types: A, T, C, G ,insertion and deletion; k represents the patient disease time, which can be used to express the time scale of frequency changes. Besides, an array

$dparray(i,k)$ records read-depth (Grubaugh et al. 2019) of each nucleotide of each patient at each recorded date. So I has 29903 rows (i) and number of dates recorded column (k).And t array indicates the number of days since positive. The above arrays are saved and used for further analysis in the MATLAB function written for subsequent MLE parameter estimation and multi-hypothesis testing (see the next subsection for details).

2.5 Parameter Optimisation based on MLE and Implementation of Multi-hypothesis Testing

The analysis of the simulated data in the above subsections and the analysis of the patient data are estimated by using MLE to obtain the optimised parameters. The MLE equation was written by Dr. Khatri and provided for use in this study. The function computes the best N, μ, and s $(Nopt, muopt, sopt)$ for each site and uses maximum likelihood to evaluate whether site-based selection is effective. For the data of whole gene, no matter simulated and real data, the above function is executed through loops to optimise the parameters of each gene site in turn, and finally gather three lists corresponding to the three parameters. The 'Simplex' optimisation method is used for likelihood estimation which developed by Nelder & Mead (1965). In the case of optimising three parameters at the same time, the function can also draw the likelihood surface for each estimate, including the 95% confidence surface. Furthermore, in order to find a better regime, i.e., what condition has the least difference for parameter estimation. This study also discusses the situation of fixing N and μ. In general, there are: 1. Optimising three parameters at the same time; 2. Fixing N to optimise μ and s, 3. Fixing μ to optimise N and s; 4. Both N and μ are fixed to optimise s. This is a classified discussion made to explore the relationship between virus evolution within a host and evolution between hosts under selection with real data.

For the analysis of multiple hypothesis testing, simulated data and real data are different. First, the simulation of a single site does not require this part of the analysis. The parameter optimisation results of whole genome simulation under different regimes are used for ROC drawing and AUC value. We plot false discoveries vs true discoveries for all sites ordered by the LRT statistic with the largest first on ROC. AUC is the area under this curve and the larger AUC the better the performance calculation through Likelihood Ratio Statistic ($d_0 = 2 \times (\ln(L_{max}) - \ln(L_{max0}))$) as a statistical decision,

where L_{max} indicates the maximum likelihood function value and L_{max0} indicates the maximum likelihood function value of the null model (s=0). A perfect ROC curve would have a vertical line from the origin to $TPR = 1$, indicating that all selected sites were detected correctly, followed by a horizontal line $FPR = 1$, indicating all neutrally evolved sites. 'perfcurve' in MATLAB can accomplish this purpose. (Fawcett 2004). Among them, the labels are classified according to whether $2Ns > 1$, that is, whether the site is neutral or under selected. The classification score is list d_0, and 1 represents the positive class. The obtained AUC values are classified according to different regimes, and line charts of AUC changes are drawn.

Multi-hypothesis testing analysis for real data uses the Benjamini-Hochberg (BH) false discovery rate controling method (Benjamini & Hochberg 1995). BH method finds all discoveries with False Discovery Rate(FDR) <0.05 to sort the $p-value$ and $q-value$ stated by Storey (2002) is roughly the FDR for each site . $q-value$ is calculated by built-in function 'mafdr' (Storey et al. 2004, Storey 2003, Storey & Tibshirani 2003). The $q-value$ is used as the score in the ROC. The label is assigned according to $p-value$. The $p-value$ is obtained from the value of the complement of the chi-square cumulative density function ('chi2cdf' in MATLAB) with the difference n between the number of degrees of freedom of full model vs null model calculated at the value in d_0 (Abramowitz & Stegun 1968). The code implementation of this part is done by Dr. Khatri.

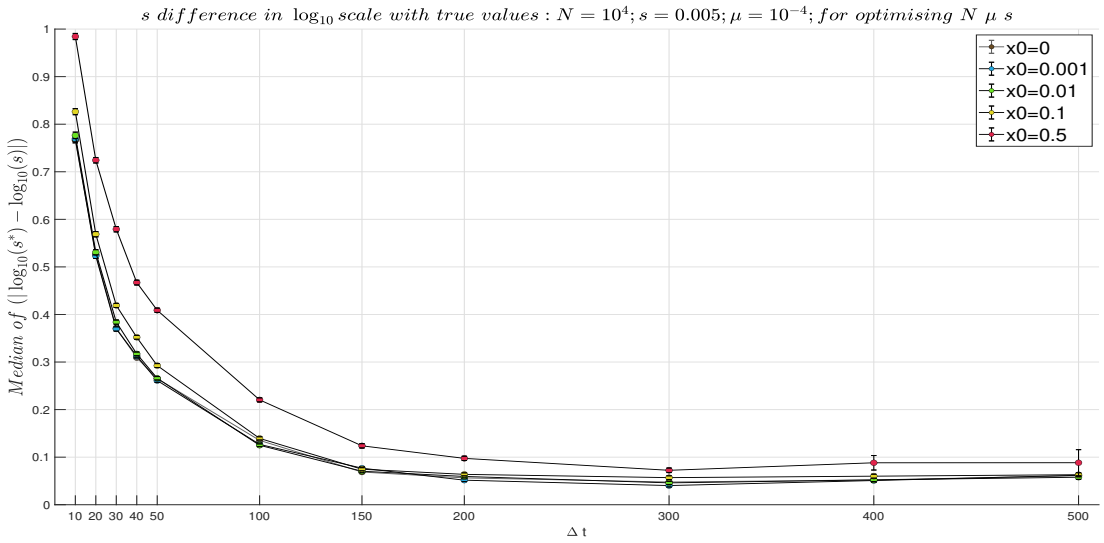
3 Results

3.1 Single site parameter estimation results

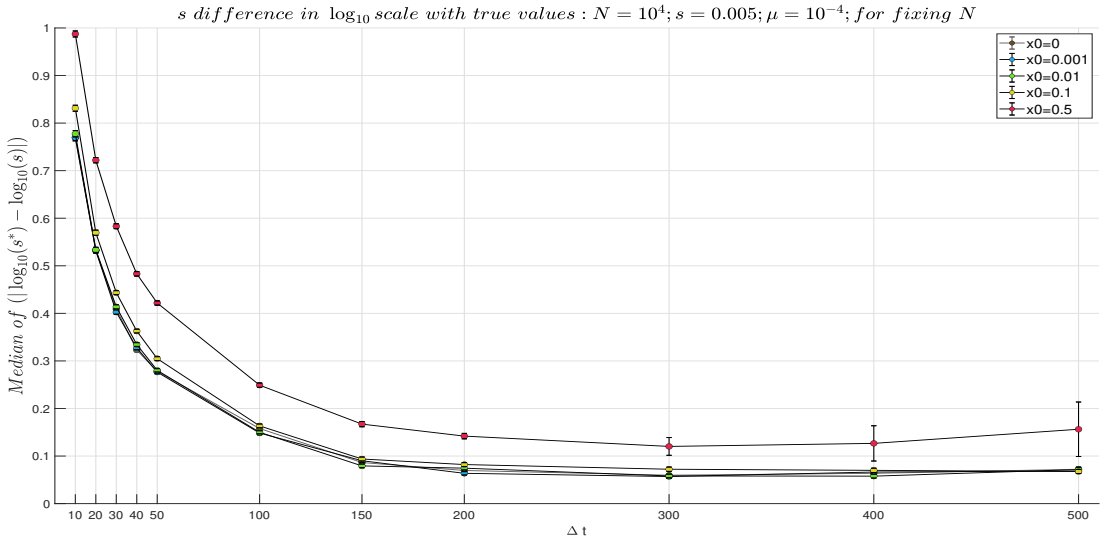
First look at the positive selection part of the single site simulation. **Figure 2-4** show the error changing of assessing parameter estimating accuracy on a log_{10} of magnitude scale between the optimised parameter value and the set parameter value under different Δt for 4 different parameter fixing conditions. For s (**Figure. 1**), the situation is simple. As Δt increases, the error between the estimated value and the initial set value gradually decreases, forming an reverse exponential downward trend. However, it is worth noting that the error at $x_0 = 0.5$ has a significant error compared with other initial frequencies. The error is about 0.5 orders of magnitude higher than other initial frequencies.

For N (**Figure. 2**), as Δt increases, the error from the initial N value gradually decreases, and the smallest error occurs in the simulation of $\Delta t = 40$ with a error of less than 0.2 log 10 scale, and then as Δt increases, the error gradually increases again. Especially for the simulation of $x_0 = 0.5$, the error at $\Delta t = 500$ has a large gap compared to other initial frequency cases. Overall, the error between optimising the three parameters at the same time is smaller than when μ is fixed.

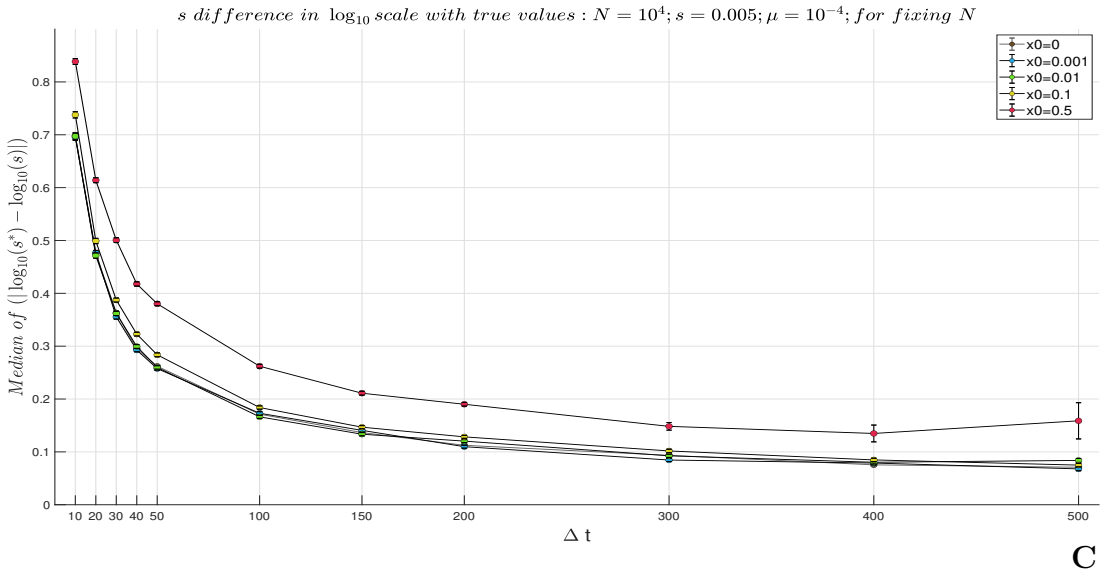
For μ (**Figure. 3**), the situation is similar to N, the smallest error occurs at $\Delta t = 50$, and the error gradually increases after $\Delta t = 50$. Unlike the error of N, the maximum error of μ is always no greater than 0.8.


$s\ difference\ in\ \log_{10} scale\ with\ true\ values: N = 10^4; s = 0.005; \mu = 10^{-4}; for\ optimising\ N\ \mu\ s$
x0=0
x0=0.001
x0=0.01
x0=0.1
x0=0.5
$Median\ of\ (|\log_{10}(s^*) - \log_{10}(s)|)$
Δ t

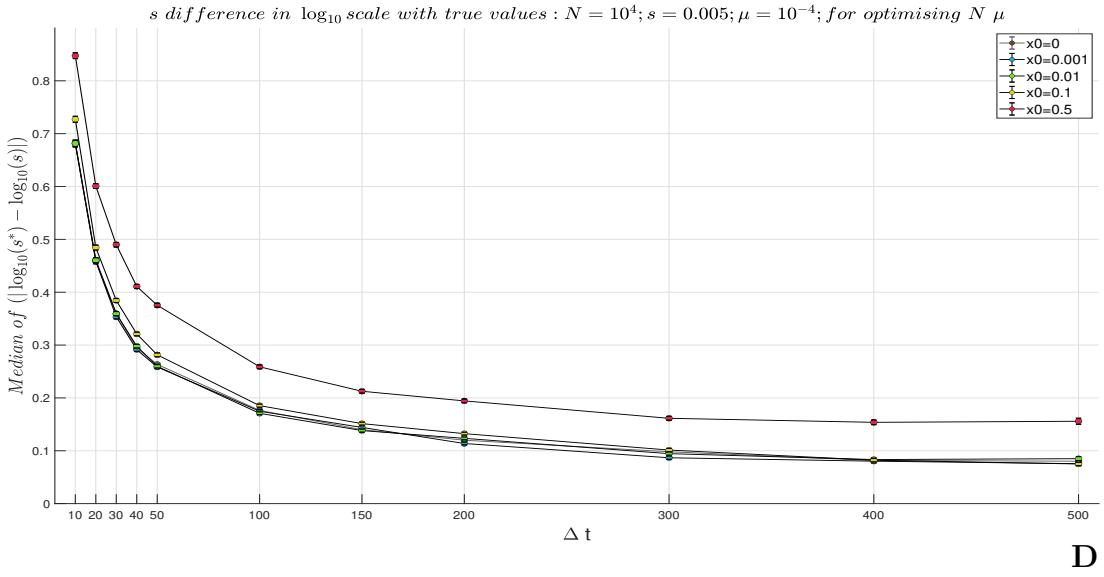
A


$s\ difference\ in\ \log_{10} scale\ with\ true\ values: N = 10^4; s = 0.005; \mu = 10^{-4}; for\ fixing\ N$
x0=0
x0=0.001
x0=0.01
x0=0.1
x0=0.5
$Median\ of\ (|\log_{10}(s^*) - \log_{10}(s)|)$
Δ t

B

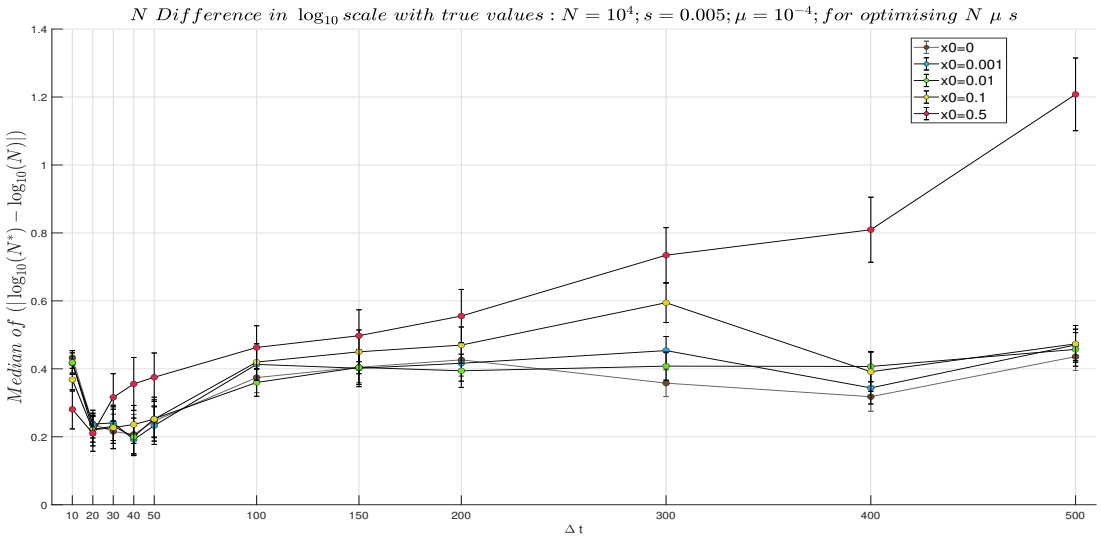


C



D

Figure 1. s **error for single site positive selection simulation.** A is result by optimising $N\mu s$, B is result by fixing N, C is result by fixing μ and D is result by fixing N μ. Different colours indicates different initial frequencies.



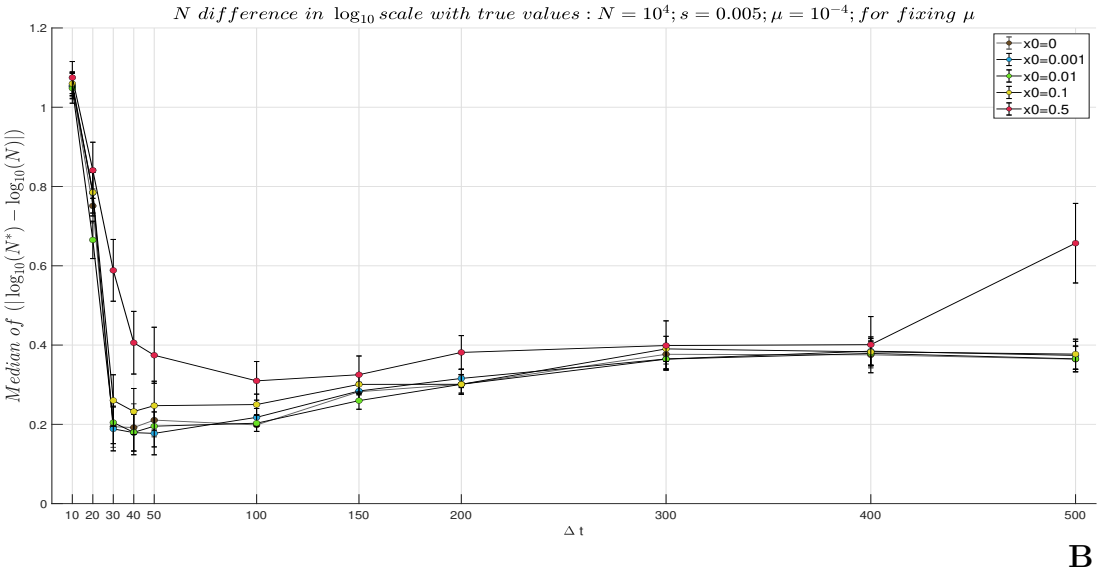
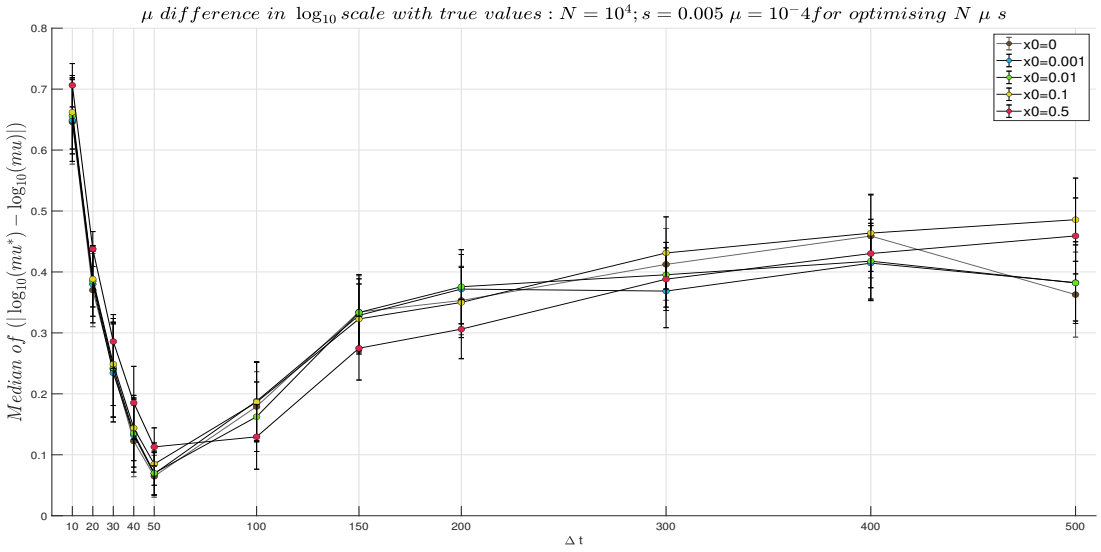
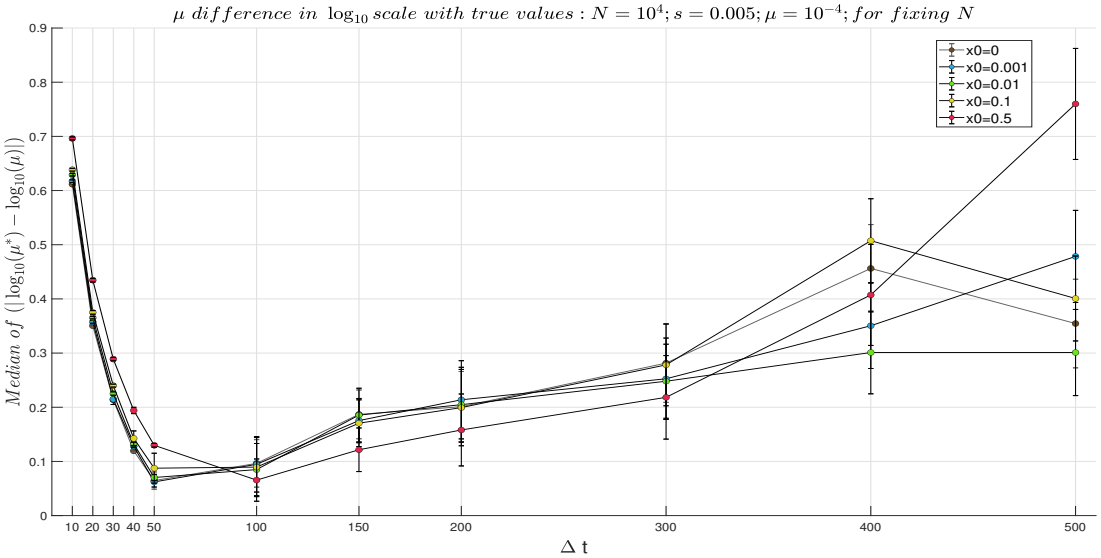


Figure 2. N **error for single site positive selection simulation.** A is result by optimising $N\mu s$, and B is result by fixing μ. Different colours indicates different initial frequencies.



A



B

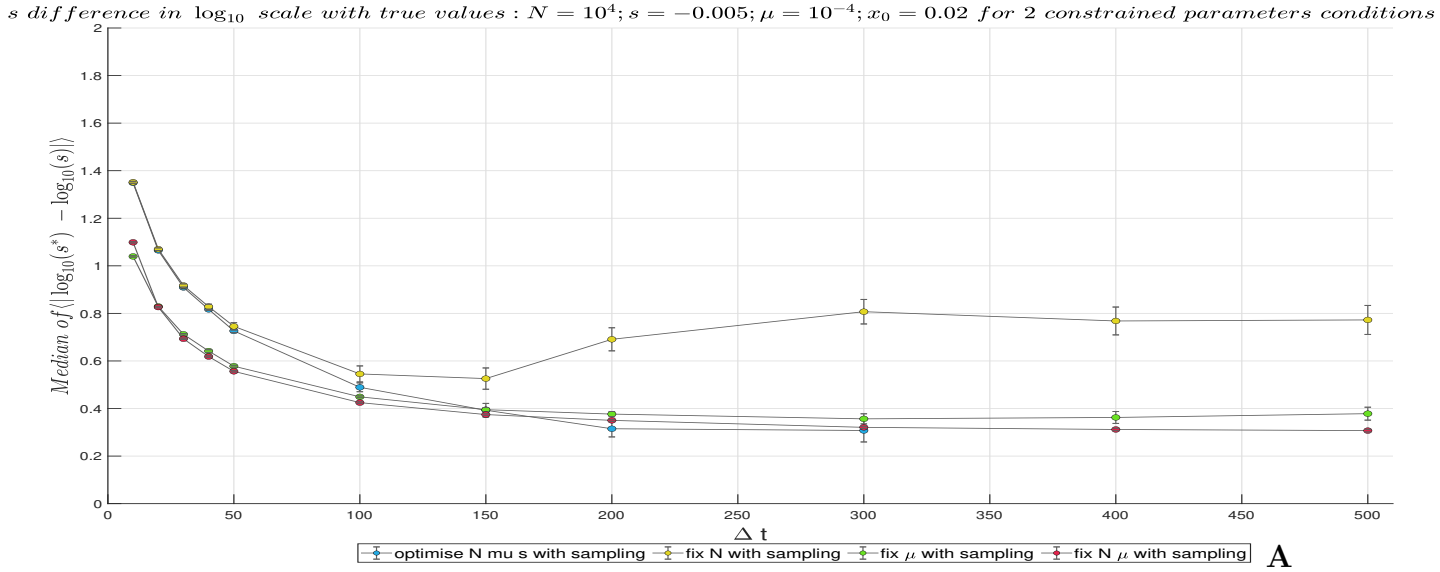
Figure 3. μ **error for single site positive selection simulation.** A is result by optimising $N\mu s$, and B is result by fixing N. Different colours indicates different initial frequencies.

For results in single site negative selection simulation (**Figure. 4**), re-

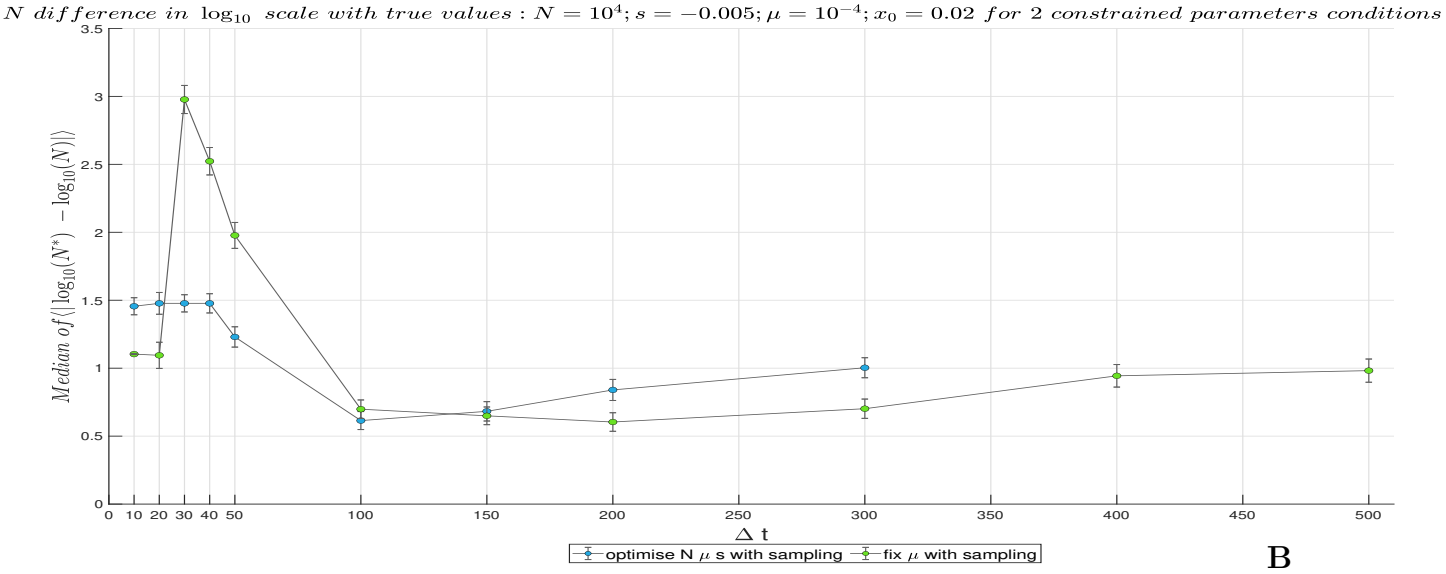
sult is roughly the same trend with positive selection in error of true s and estimated s decreases with the increase of Δt, but the error is larger than the result of positive selection. But the result of fixing N when $\Delta t = 200$ and later, the error is again gradually increased. However, N error are different from situation in positive selection. When Δt from 20 to 30, error unexpectedly increases in fixing μ condition. Then sharply decreases until $\Delta t = 100$. Besides, in optimising three parameters condition, the error of N doesn't change a lot at first 4 sampling time sites. Overall, the error in the optimization parameter N under negative selection is still greater than that under positive selection. Finally, looking at μ, the overall trend first increases and then decreases, which is similar to that of positive selection. It's just that the inflection point appears later, and there is an upward trend after $\Delta t = 100$, and it is when $\Delta t = 50$ under positive selection.

After comparing the errors of the real and estimated values of the parameters under positive selection and negative selection, it can be found that there are some differences in the estimated coefficients of different selections. For the estimation of s, the error between true value and estimated values under the positive selection is slightly lower than under the negative selection. This shows that for the estimation of s, estimating the positive selection coefficient (s<0) is more accurate than the negative selection coefficient (s<0). Under each same sampling interval, the estimation error of positive selection on s is about 0.2-0.6 order of magnitude smaller than that of negative selection. For the estimation of N, error between positive and negative selection is greater. The error range is about 0.2-1.2 orders of magnitude under positive selection, and 0.5-3 orders of magnitude under negative selection. For the estimation of μ, the error difference between positive and negative selection is not big, which is reflected in the larger sampling interval, the negative selection has a larger error in the estimation of μ than the positive selection.

It is worth noting that all error lines that optimising three parameters lack the result of $\Delta t = 400, 500$. This is because the time series analysis here cannot be completed as expected, so there are not result.



A



B

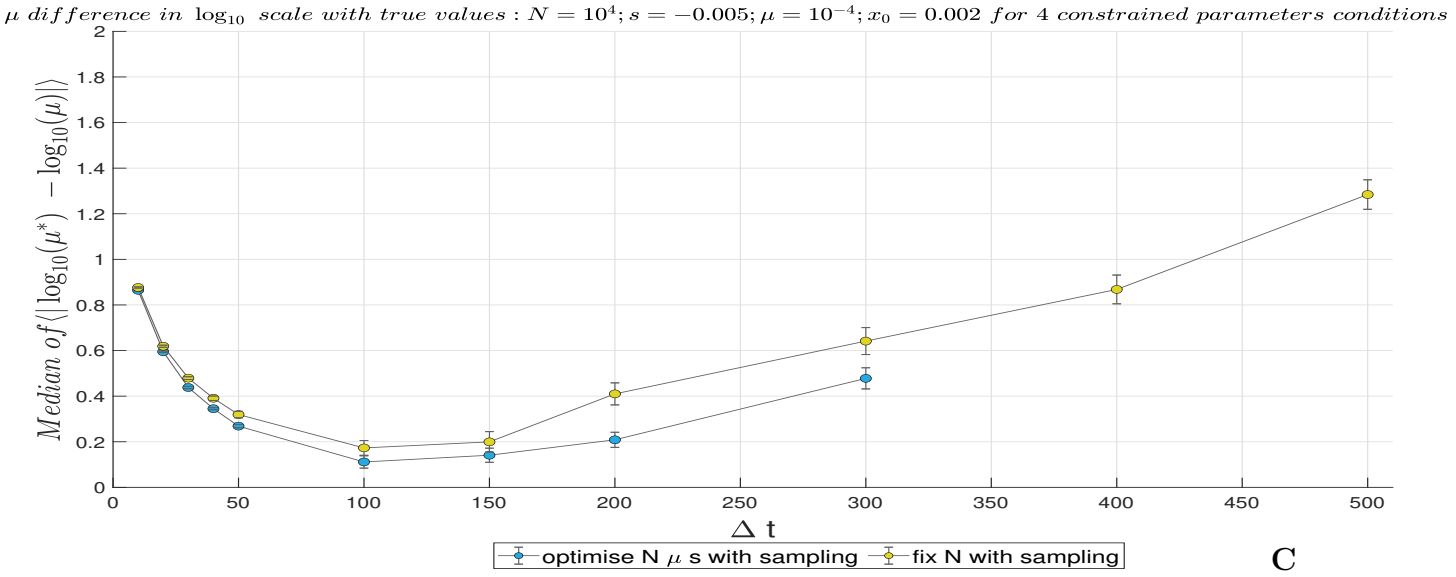


Figure 4. *s N* μ **error for single site negative selection** A is *s* error plot, B is *N* error plot and C is μ error plot. Different colours indicates different parameters constrained conditions.

3.2 Genome-wide simulation and AUC results

This AUC in this study is to test the performance of the parameter optimisation analysis method in distinguishing whether the site is under selection (ROC plots for each condition can be found in appendix). From the left plot of **Figure.5**, at $T = 1000$, the sampled performance is worse than the non-sampled results. The sampled results are also more volatile. Overall, it seems that fixing the value of μ and optimising three parameters are most stable under the condition of without sampling and can better distinguish the selection and neutral.

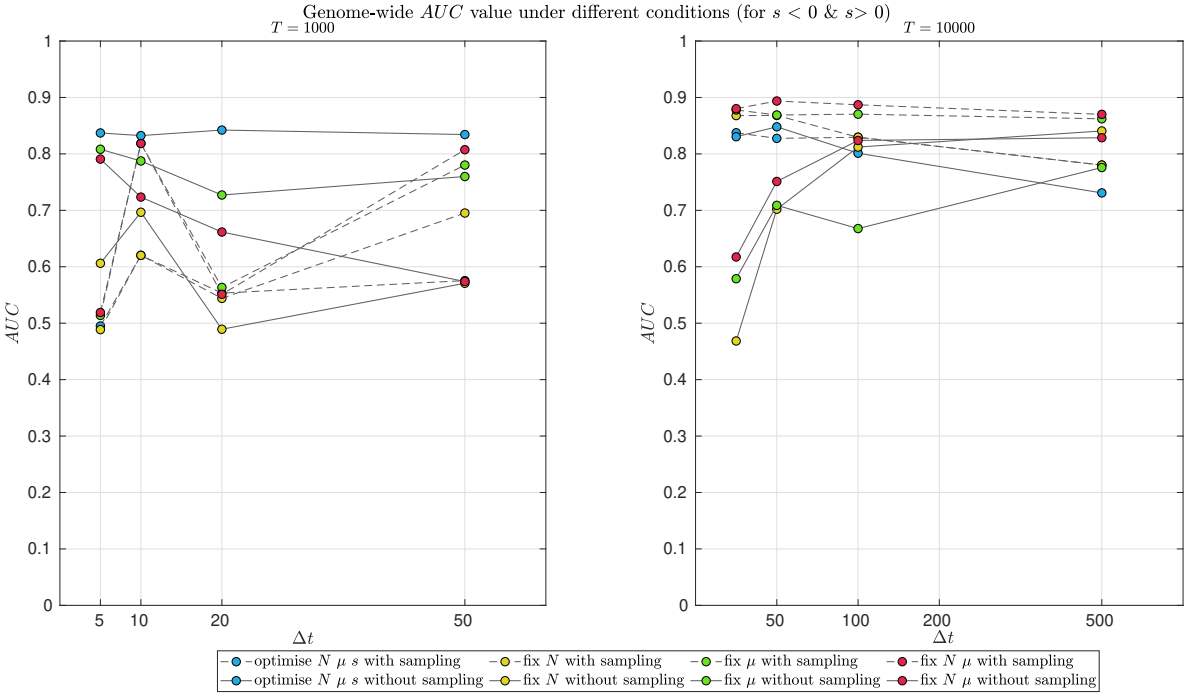


Figure 5. Genome-wide AUC value under different conditions for both positive selection and negative selection sites. This plot shows AUC value for different T (T=1000 on left, T=10000 on right). Different colours indicates whether fix N, μ parameters conditions. Dash lines are results that simulations are done by binomial sampling, whereas straight lines are not.

From the plot on the right, the result of T=10000, results with sampling are even better and stable. Besides, AUCs are mostly maintained at the level of 0.8-0.9. For the unsampled results, the results displayed by AUC at $\Delta t = 50$ and 100 indicate that they cannot distinguish the selection from the neutral situation very well. Fixing N, μ and fixing μ with sampling are optimal among all conditions when T=10000. In view of the overall situation, the sampling with fixed N μ value performed the best. Although when $\Delta t = 5, 20$, AUC are low at these conditions, but compared with other conditions at the same other sampling time, it has more higher AUC.

Obtaining the ROC curve (**Figure. 6**) of the case with the highest AUC, which is $T = 10000, \Delta t = 100 with sampling$, can see the specific situation of classifying selection and neutrality of sites. The optimal threshold selection can control the FPR less than 0.1 and the TPR slightly higher than 0.7.

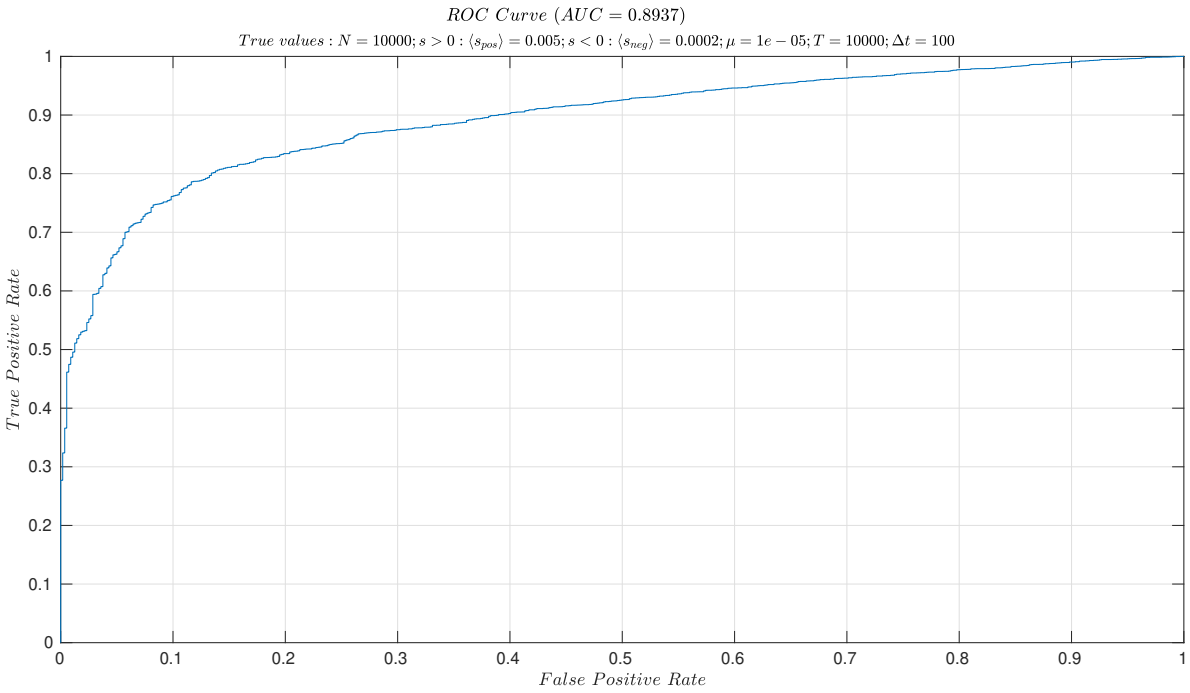


Figure 6. ROC for T=10000 dt=100. This plot shows ROC for $T = 10000, \Delta t = 100$ in fixing N μ condition. The AUC in this condition is the highest among AUC value shows in previous figure.

3.3 SARS-CoV-2 Patient analysis results

Apply analysis method proposed by Khatri (2016) based on WF model to the data of two groups of immunocompromised patients infected by SARS-CoV-2 virus, the following results can be obtained. First, **Figure. 7 (A,B)** are the histograms of the $p - values$ in different fixing parameter cases for both patient 6 and 9. The the $p - values$ increase sequentially from left to

right. The red dash lines in the figures indicate the threshold value of the adjusted $p - value$ obtained by the BH method, and all $p - values$ smaller than the dash lines are significant, and vice versa. Max count indicates the frequency of the highest bin. Ideally, p should be peaked towards 0 and tail off to a constant close to 1. From the **(Figure.7(A))**, for patient 6, it can be clearly seen that when the three sets of parameters are optimised, the proportion of the left part of the threshold is very low. When it turns to fixing N, fixing μ and fixing N μ, the results are much better then the first condition, more than 18000 locus are significant under selection. This means that detecting selection with optimising N μ s might be problematic in the case of patient 6. From the **(Figure.7(B))**, for patient 9, expect optimising N μ s has a low significant sites results, fixing μ in this case becomes worse then it in patient 6 case. Overall, regardless of the parameters constrained situations, it seems that once N is fixed the multi-hypothesis results would get better, more sites are significant.

From the perspective of the distribution of the entire $p - value$, the performance of simultaneously optimising three sets of parameters is very unsatisfactory, because it has a peak at a position where the $p - value$ is close to 1. The distributions of the other three groups are relatively good.

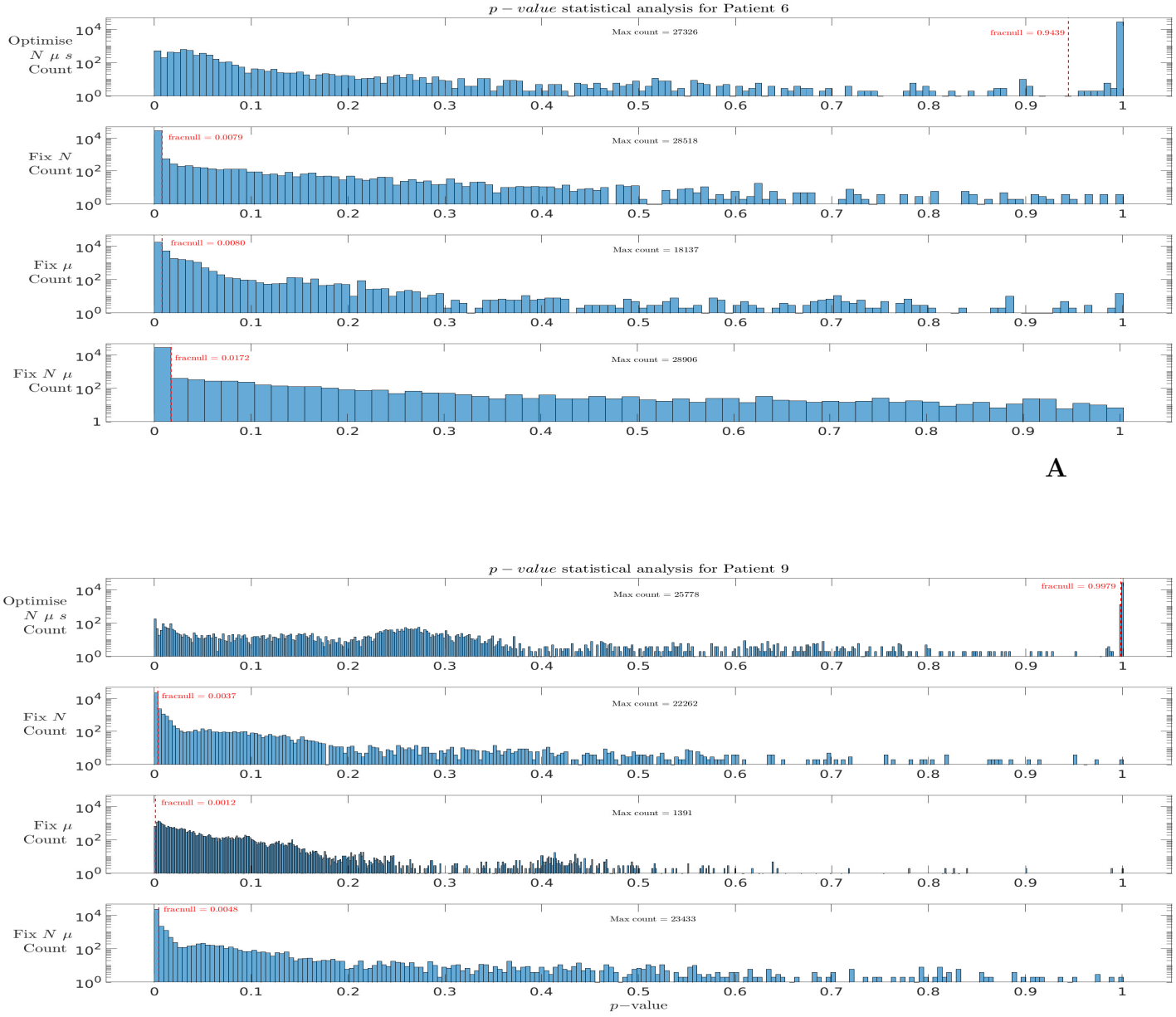
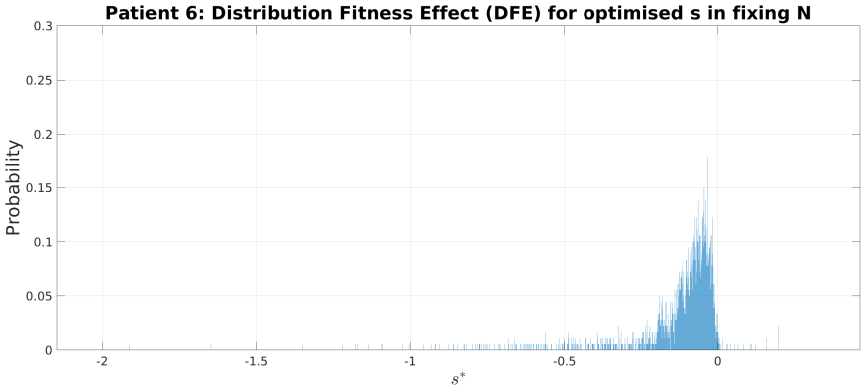


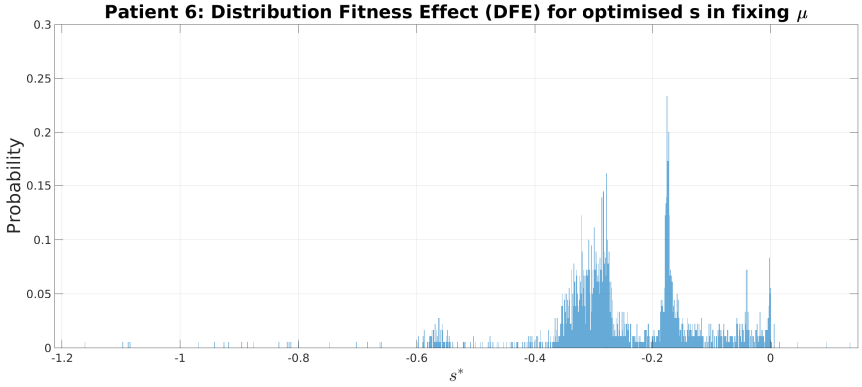
Figure 7. p-value histogram for patient 6 (A) and 9 (B) under 4 different cases. This figure is a frequency histogram of $p-value$. Four subplots represent four cases where the parameters are fixed. The red dash line indicates the threshold of the null hypothesis. If it is less than this value, it means that the p-value is significant.

Figure 8 and **Figure 9** show the DFE of optimised s in 4 parameters constrained situations. We mentioned in the previous part that most of sites in gene are neutral selection, so most sites' s are very small, and number of s gradually decreases when s gets larger. And the distribution of s closely satisfies the distribution. In the results of **Figures 8 and 9**, it can be seen that among the results of four parameter optimisation choices, the DFE distribution under fixed N **Figures 8(b)** and fixing N μ **Figures 9(d)** is closest to such distribution from zero to negative (because in most sites s¡0, the closer selection coefficient to zero, the closer to neutral). Whereas, optimised s in optimising 3 parameters and fixing μ are some extent not in this distribution. Therefore, the results in fixing N and N μ basically satisfy the above-mentioned point of view, and the detection on the selected sites is better under fixing N and fixing N μ.

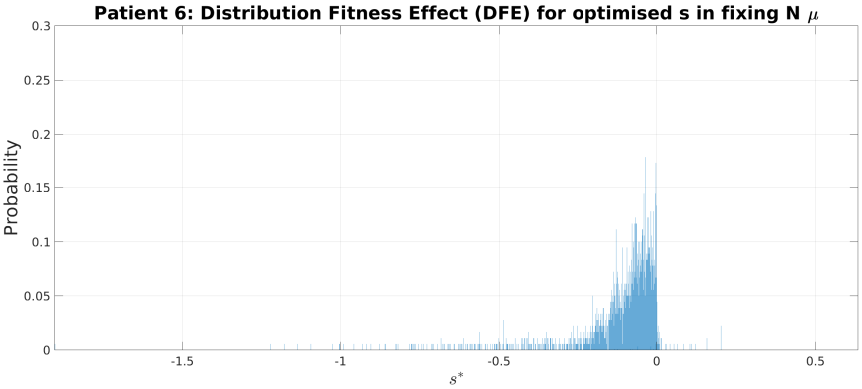
(a) DFE for optimising N μ s for patient 6



(b) DFE for fixing N for patient 6



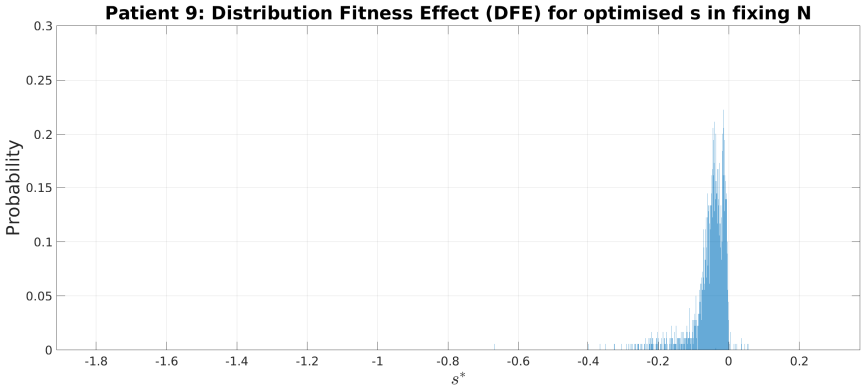
(c) DFE for fixing μ for patient 6



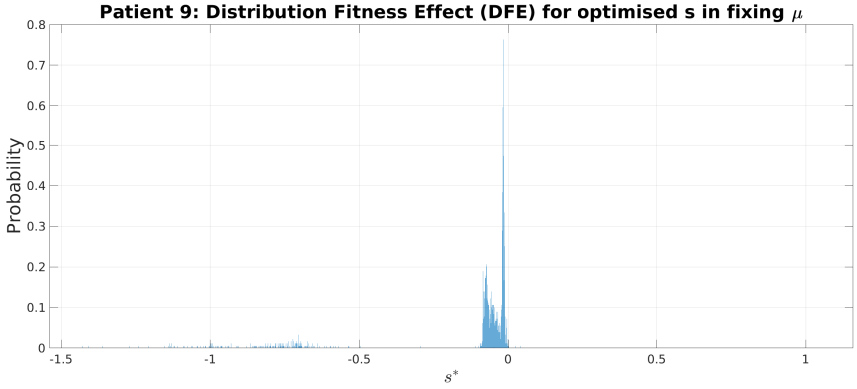
(d) DFE for fixinging N μ for patient 6

Figure 8. DFE in 4 parameters constrained situation for patient 6 This group of graphs shows the probability density distribution (pdf) of the optimised selection coefficient of patient 6, that is, the fitness effect distribution. For details, please refer to the method section.

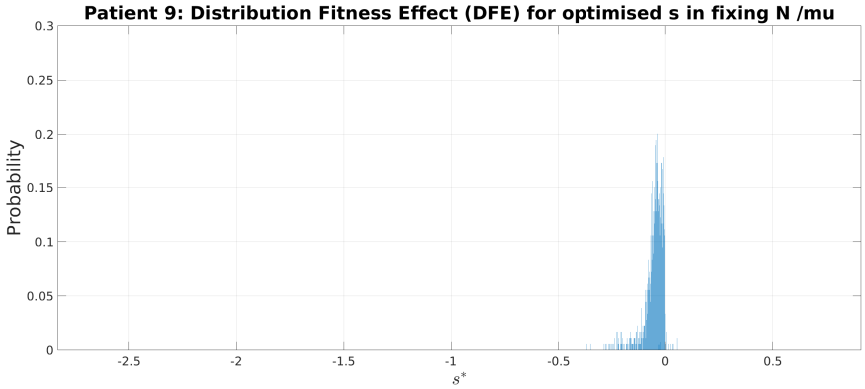
(a) DFE for optimising N μ s for patient 9



(b) DFE for fixing N for patient 9



(c) DFE for optimising μ for patient 9

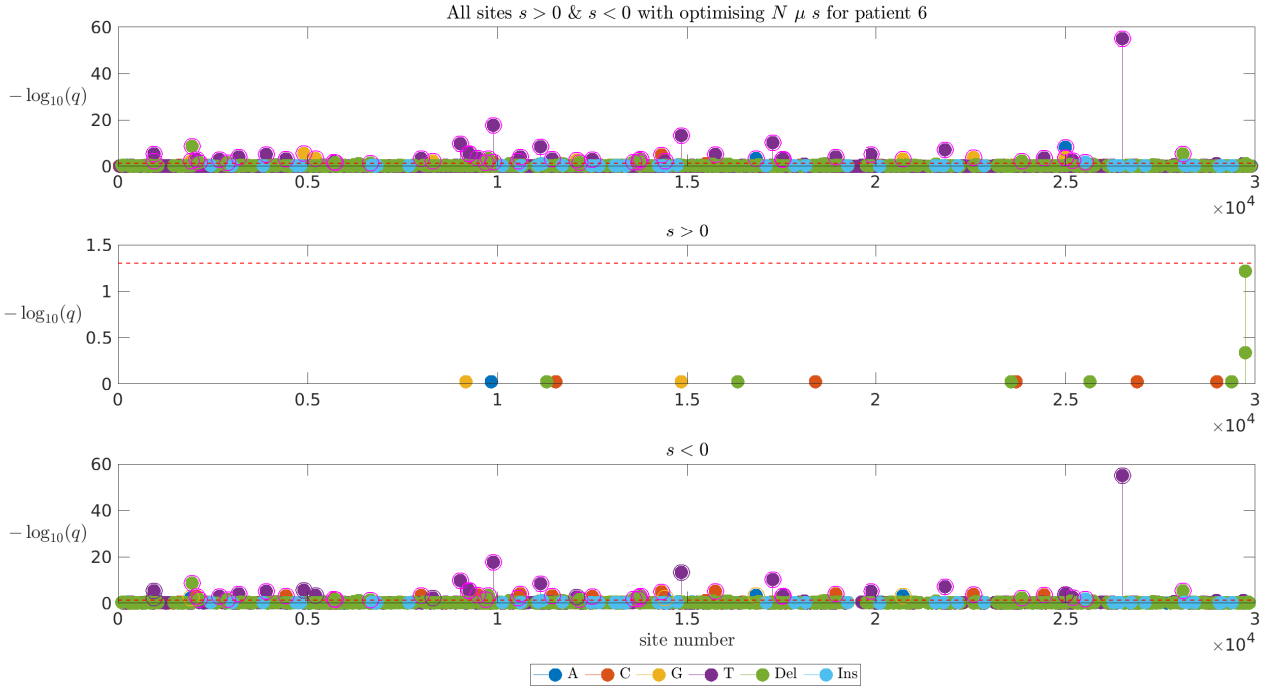


(d) DFE for optimising N μ for patient 9

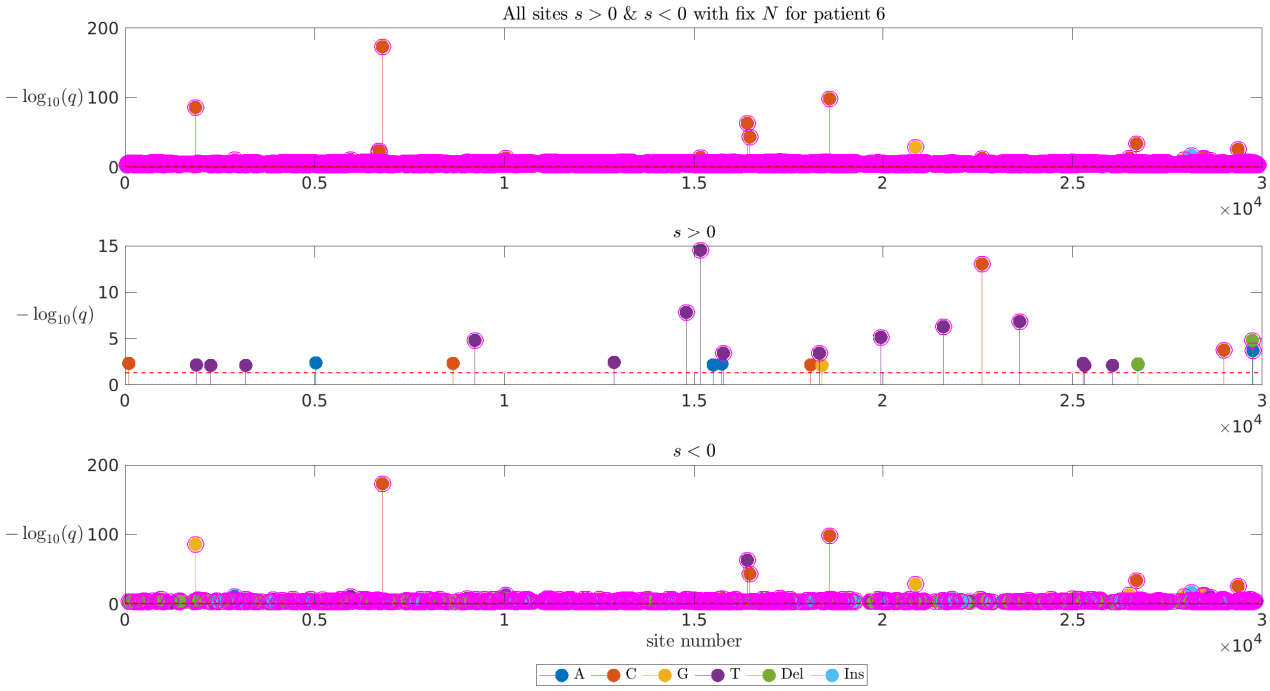
Figure 9. DFE in 4 parameters constrained situation for patient 9 This group of graphs shows the probability density distribution (pdf) of the optimised selection coefficient of patient 9, that is, the fitness effect distribution.

Figure 10 and Figure 11 show the selected sites within SARS-CoV-2 virus under 4 parameters constrained situations. According to the above situation, since the results of fixing N and fixing N μ are more reliable, we can find the corresponding selection sites lollipop plots (**Figure 10 and Figure 11**) of the two cases for verifying. It can be found that whether it is patient 6 or 9, the selected sites among positive selection sites in these two cases are also closer. Compared with the above-mentioned cases, the positive selected sites of the other two limited parameters conditions have some shortcomings.

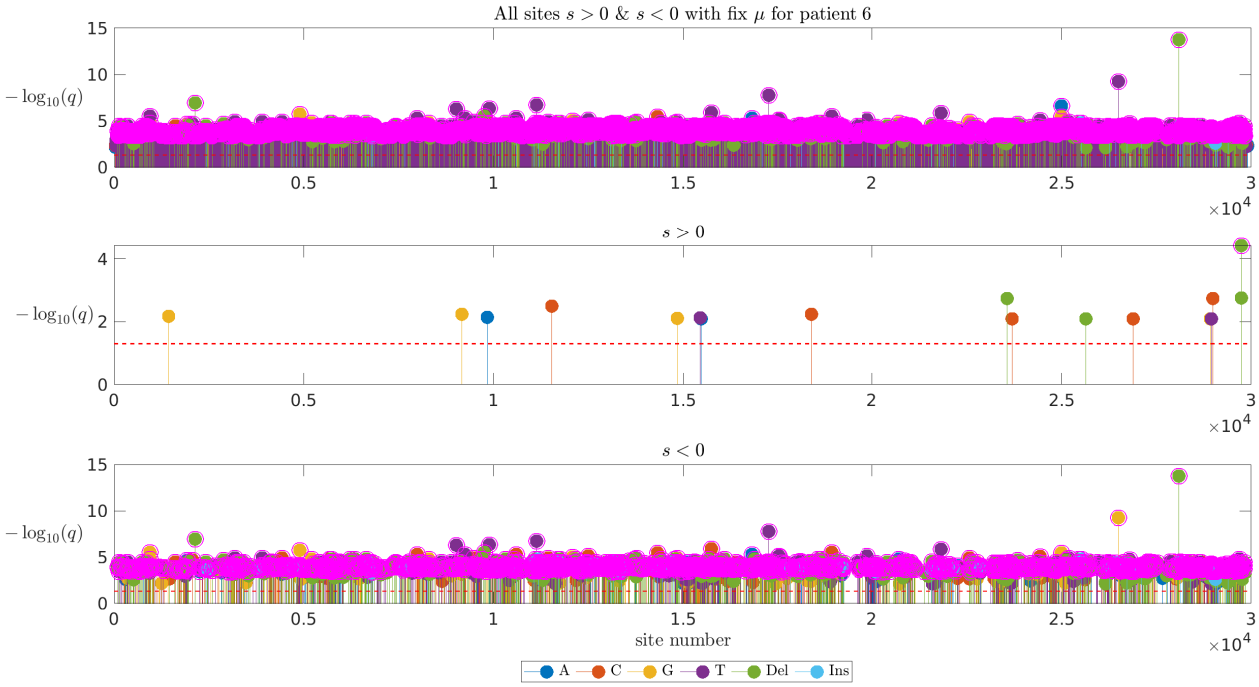
In the case of optimising the three parameters at the same time, it seems that the selected sites cannot be well picked out, and from the results of fixing μ, it seems that all sites' $-log_{10}(q)$ values are significantly higher than $-log_{10}(0.05)$, it seems that under these two constrained cases, the result is either too strict or too loose.



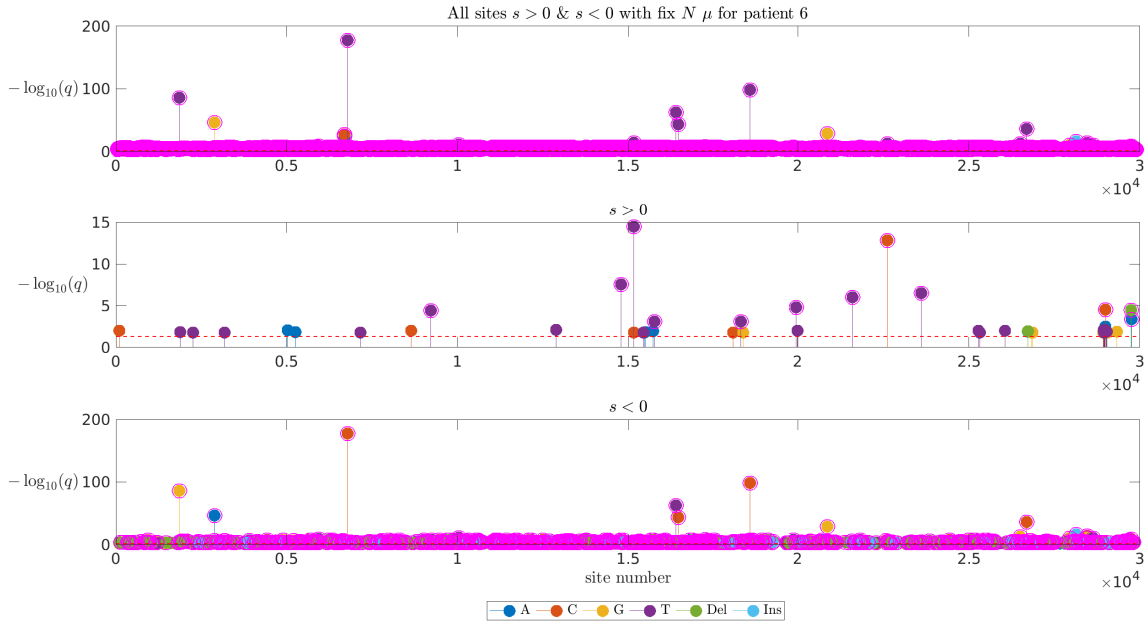
(a) Selected site for optimising N μ s in patient 6



(b) Selected site for fixing N in patient 6

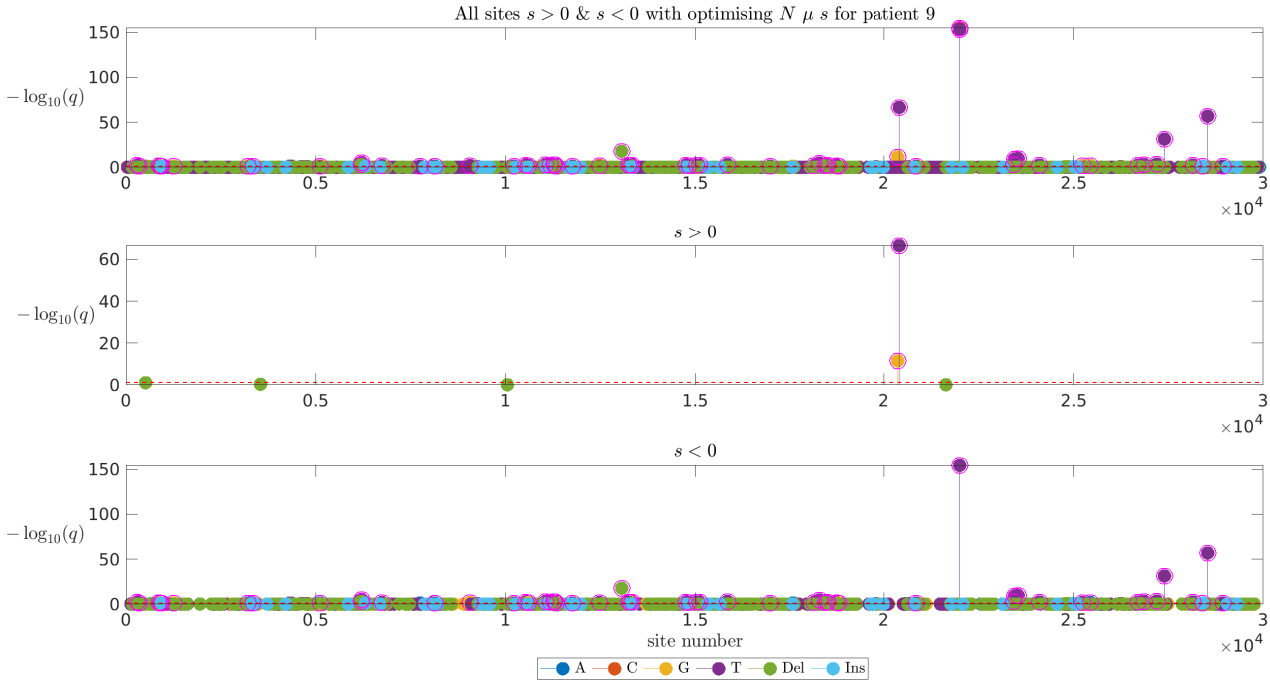


(c) Selected site for fixing μ in patient 6

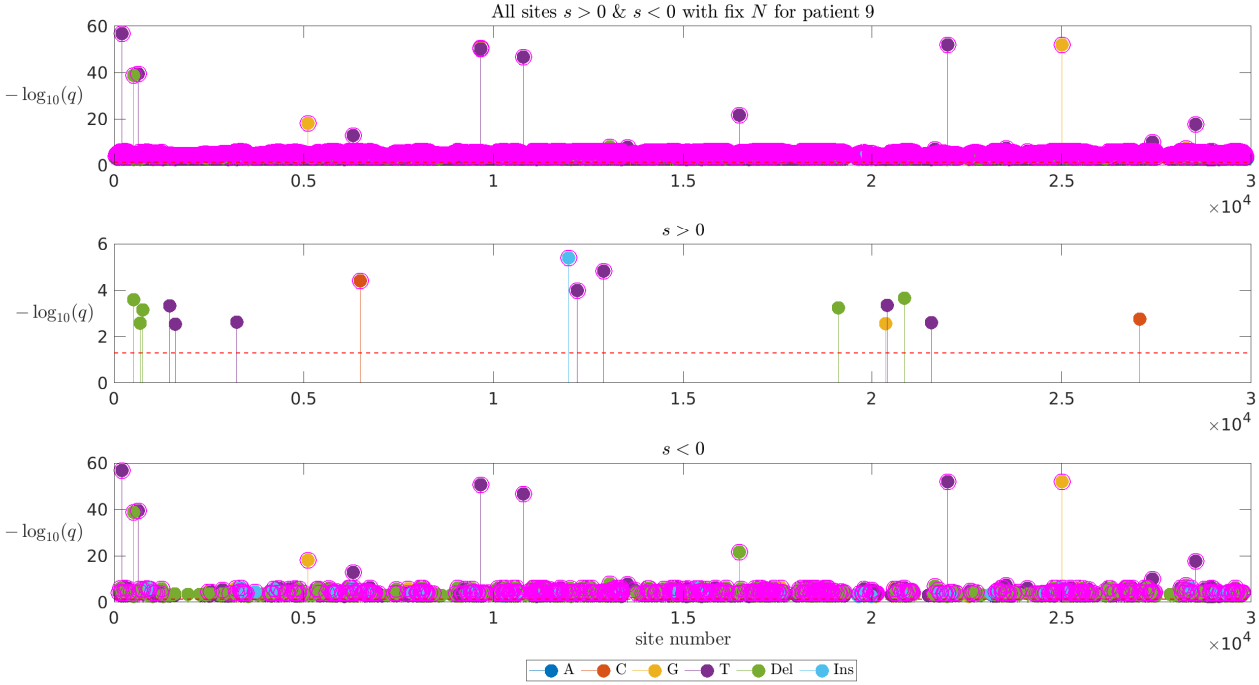


(d) Selected sites for fixing $N\mu$ in patient 6

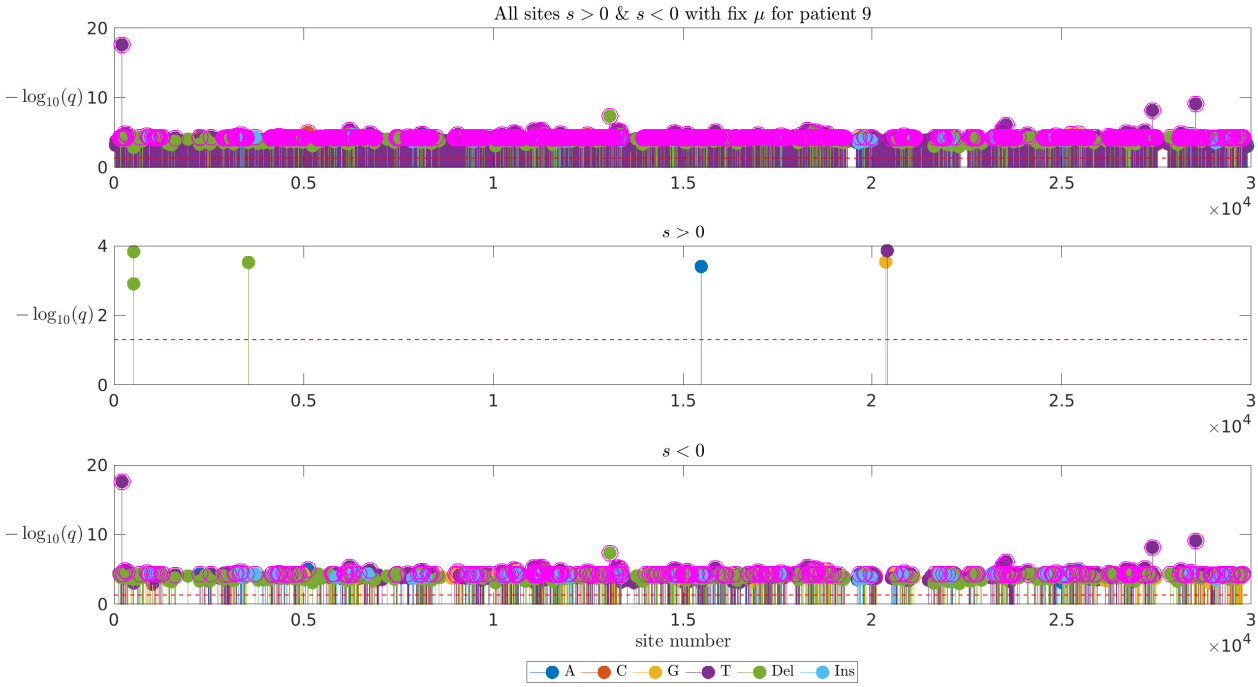
Figure 10. Selection sites lollipop plot for both positive and negative selection in patien 6 for 4 constrained parameters situations From top to bottom in each plot are the $-\log_{10}(q)$ values of all selection sites, negative selection sites and positive selection sites. The larger the value, the closer the $q - value$ of that site is to 0. All sites above the red dash line are sites greater than $-\log_{10}(\alpha = 0.05)$, indicating that they are significantly selected under the hypothesis test. All the sites with pink circles outside indicate the selected sites that have passed the screening of the BH method, and the sites screened by the BH method are also more stringent.



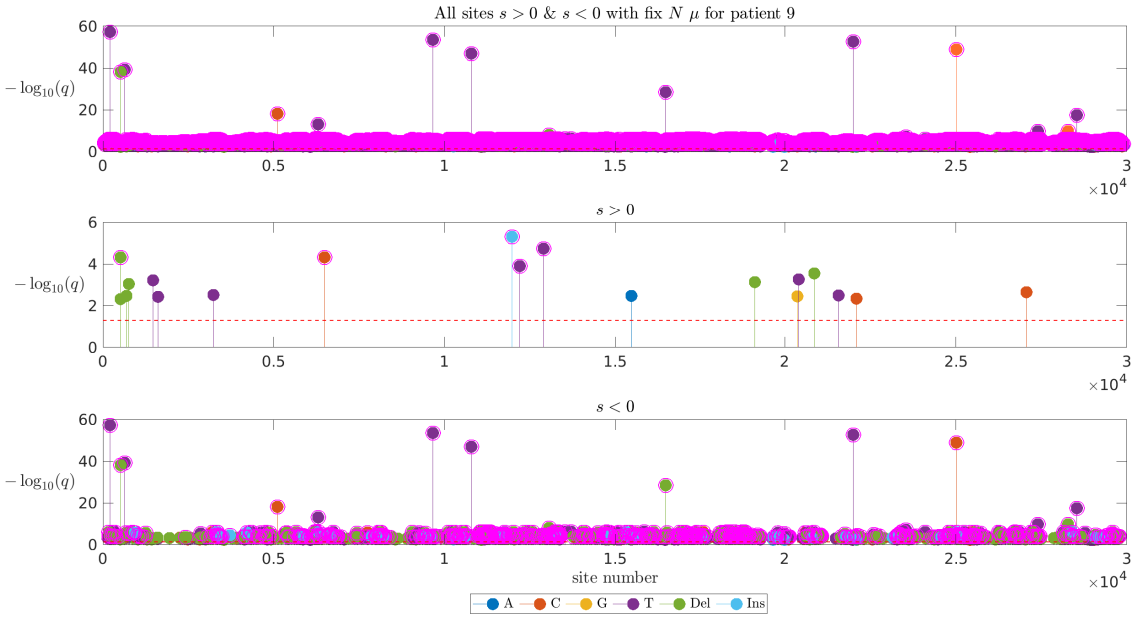
(a) Selected site for optimising $N\ \mu\ s$ in patient 9



(b) Selected site for fixing N in patient 9



(c) Selected site for fixing μ in patient 9



(d) Selected sites for fixing $N\mu$ in patient 9

Figure 11. Selection sites lollipop plot for both positive and negative selection in patien 9 for 4 constrained parameters situations From top to bottom in each plot are the $-\log_{10}(q)$ values of all selection sites, negative selection sites and positive selection sites. The larger the value, the closer the $q - value$ of that site is to 0. All sites above the red dash line are sites greater than $-\log_{10}(\alpha = 0.05)$, indicating that they are significantly selected under the hypothesis test. All the sites with pink circles outside indicate the selected sites that have passed the screening of the BH method, and the sites screened by the BH method are also more stringent.

4 Discussion

Independently detecting whether a locus is under selection through a frequency change time series of genes can be a viable approach, especially for segregating variations within a single population. This is basically based on the analysis of statistical methods to make a decision. Therefore, the accuracy of decision has become the primary criterion for evaluating the method. This brings us to first explore the accuracy of parameter estimation by using maximum likelihood estimation under four fixed parameter cases through single site simulations for both positive selection and negative selection.

There are four cases of fixing parameters are considered: Optimising N μ s, fixing N to optimise μ s, fixing μ to optimise N s, fixing $N\mu$ to optimise s. In the case of different fixed parameters, the optimal way seems to fix the population (N size and mutation rate (μ). Among the four cases, when their changing trends are mostly the same, fixing N μ can have the best result of detecting selection among 4 cases.

For the estimation of s, under positive selection mostly has lower errors than under negative selection. This shows that estimation under the positive is more accurate than the negative selection. This may be due to the difference in frequency change conditions under positive and negative selection. We mentioned in the Method that under strong positive selection ($2Ns >> 1$), there is a situation where the mutant is established (100 generations under our setting) and reaches a fixed position after an average of 261 generations, so under the same parameter setting and the same actual frequency , every simulation of positive selection shows roughly the same changing frequency trend (eventually reaches fixation); However, under negative selection, we only mention the existence of mutation and selection balance, and the change trend during the period is not yet known, so there may be different results for each simulation , which may lead to more accu-

rate estimation of the parameters by forward selection.

For the estimation of N, under positive selection, the N error has a certain Δt as the inflection point (20 for optimising the three parameters, 50 for fixing μ). The accuracy increases at low sampling intervals and decreases at high sampling intervals. And when the starting frequency is greater than the established frequency ($x^* = 1/2Ns = 0.01$), the parameter estimation error is larger than the case where the starting frequency is lower than the established frequency. This should be related to the reason that the sites that reach the establishment frequency and the sites that do not reach the establishment frequency will reach fixation at different times. As the frequency change situation is different before the establishment and after the establishment, this may lead to the above results.

Under negative selection, when μ is fixed and $\Delta t = 30$, overvaluation occurs. This is the opposite of the change under positive selection before this sampling interval. We suspect there is a reason for this. It may be due to the insufficient number of single site simulations that some occasional situations occur more frequently, resulting in a large error in the estimation of μ. It can be found that when μ is fixed, the error is reduced a lot. In the case of optimizing the three parameters, the change of N error in the first few sampling intervals is completely different from the above-mentioned fixing μ. The parameters estimate in the first few sampling intervals are very approximate, and the subsequent trend is similar to the previous case.

For the estimation of μ, the variation trend of the error of positive and negative selection is very close. In positive selection, the error gap between different initial frequencies increases with the increase of the sampling interval, which may also be related to the smaller change of the frequency change at smaller intervals, and the larger sampling interval , maybe the details of

some frequency changes are ignored, so that the parameters can find the real optimal parameter in the maximum likelihood estimation.

It should also be noted that for the estimation of parameters N and μ, when $\Delta t = 500, x_0 = 0.5$, there will be an obvious increase in the error compared with other parameter constrained cases. We suspect that this is due to the effect on the parameter estimates caused by excessively exceeding the initial frequency of the establishment frequency so that the frequency reaches fixation earlier.

Then, we explore the accuracy of finding selection sites is related to whether the population size and mutation rate are fixed. The simulated genome wide frequency change data is compared with two large datasets from SARS-CoV-2 intrahost patients. we demonstrate that it is more accurate under fixing N and fixing $N\mu$ to detect selection sites.

Our calculation and analysis of the genome-wide simulated data further show that with the increase of the sampling interval, the accuracy of detection selection is also improved, and the results are more stable when the sampling time is larger, as AUC at T=10000 $\Delta t > 100$ are kept at 0.8-0.9 which are higher and then in T=1000 and lower sampling interval (bouncing between 0.5-0.8).

There are four cases of fixing parameters are considered: Optimising N μ s, fixing N to optimise μ s, fixing μ to optimise N s, fixing $N\mu$ to optimise s. In the case of different fixed parameters, the optimal way seems to fix the population (N size and mutation rate (μ). Among the four cases, when their changing trends are mostly the same, fixing N μ can have the best result of detecting selection among 4 cases.

However, both positive and negative selection are simulated during genome wide simulation, but they does not analysed separately. The two selection cases should be separated to examine the effect of the four cases of parameter constraining on different type of selection sites detection. Also, separate positive selection and negative selection can better align with real data, then compare the accuracy under each selection scenario.

Such result can also be reflected in the results of patient data. In the two groups of patient data, more significant sites were brought about by fixing N μ. However, different from the simulated situation, the time interval for sampling and analysis in reality is often not regular. The average total generation of SARS-CoV-2 virus in the two patient data in this study are 181 and 350 generations according to gamma (2generations per day) of SARS-COV-2 (Markov et al. 2023), respectively, and the average sampling interval was 16 generations and 18 generations. If the two sets of data are put in the analysis of the AUC results analysed by the simulated data, patient 6 is close to sampling interval of 20 generations in the simulated data, and patient 9 is closest to the sampling interval of 50 generations. The optimal constrained parameter conditions for them is fixing N μ. But in fact, we didn't set suitable intervals that fits to patient data. Perhaps for the real sampling interval of patients, there will be different fixed parameter selections that can make the results of screening selection sites more accurate.

In general, whether to fix N is the key to the impact on the accuracy of detecting selection. From the results of comprehensive simulation results and actual data, it seems that the results brought about by a only fixed μ are not much improved compared with optimising N μ $s.$, which can be clearly reflected in the patient data analysis whichever the $p - value$ distribution, DFE or Sites under selection results. By comparing the performance of fixing N μ and fixing μ results, we can prove that N fixed is the key factor for

improving the accuracy of the finding selection sites.

Moreover, whether the frequency can finally be fixed is not necessarily for the diagnosed patient, it is an unknown situation, we cannot know how the frequency will change after receiving treatment, so the simulation here does not consider intercepting different frequencies to see the impact on parameter estimation. There is also no consideration of the impact of intervention of treatment on changes in virus frequency and population size, which can be considered in the simulation to more closely match the reality.

In what actually happens, there may be linkage (Lewontin & Kojima 1960), the genetic link between certain loci, between loci in the genome. This could lead to mutations spreading across the genome differently than expected from independent allelic models. Therefore, when performing genome-level simulations, factors such as linkage effects may need to be considered, and even more complex models may need to be used (Slatkin 2008). These actual factors have not been well considered and involved in this simulation. It is just that the loci that have received selection are screened out through multiple hypothesis testing analysis independently.

5 Conclusion

This study analysed the time series of gene frequency changes, and explored the effect and accuracy of the four fixed parameters conditions on the detection site under selection. First, through single site analysis, the influence of four fixed parameters on the accuracy of parameter estimation was evaluated, and the results were obtained that fixing N and μ have more accurate parameters estimation at low sampling intervals, and the accuracy of s estimation increases with sampling intervals increasing. In addition, the study also simulated the genome wide data to conduct multi-hypothesis testing to detect the selection sites. Then the gene frequency change data of SARS-CoV-2 in immunocompromised patients were used to compare with simulated data. It was found that fixing parameters $N\mu$ has a better outcomes than fixing N, fixing μ and optimising N μ s at same time.

6 Data and code availability

Code for likelihood estimation and multi-hypothesis part and data can't be released. For the generation code of frequency simulation data based on WF model, contact the author by email to obtain permission. (https://github.com/chuxinyaowang/MSc_project.git)

7 Reference list

References

Abramowitz, M. & Stegun, I. A. (1968), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Vol. 55, US Government printing office.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

Bollback, J. P. & Huelsenbeck, J. P. (2007), 'Clonal interference is alleviated by high mutation rates in large populations', *Molecular biology and evolution* **24**(6), 1397–1406.

Bollback, J. P., York, T. L. & Nielsen, R. (2008), 'Estimation of 2 n es from temporal allele frequency data', *Genetics* **179**(1), 497–502.

Buri, P. (1956), 'Gene frequency in small populations of mutant drosophila', *Evolution* pp. 367–402.

Choi, B., Choudhary, M. C., Regan, J., Sparks, J. A., Padera, R. F., Qiu, X., Solomon, I. H., Kuo, H.-H., Boucau, J., Bowman, K. et al. (2020), 'Persistence and evolution of sars-cov-2 in an immunocompromised host', *New England Journal of Medicine* **383**(23), 2291–2293.

Clark, S. A., Clark, L. E., Pan, J., Coscia, A., McKay, L. G., Shankar, S., Johnson, R. I., Brusic, V., Choudhary, M. C., Regan, J. et al. (2021), 'Sars-cov-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms', *Cell* **184**(10), 2605–2617.

Crow, J. & Kimura, M. (1970), 'An introduction to population genetics theory. new york: Harper & row'.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T. et al. (2011), 'The variant call format and vcftools', *Bioinformatics* **27**(15), 2156–2158.

Desai, M. M. & Fisher, D. S. (2007), 'Beneficial mutation–selection balance and the effect of linkage on positive selection', *Genetics* **176**(3), 1759–1798.

Domingo, E., Escarmís, C., Sevilla, N., Moya, A., Elena, S. F., Quer, J., Novella, I. S. & Holland, J. J. (1996), 'Basic concepts in rna virus evolution', *The FASEB Journal* **10**(8), 859–864.

Ewens, W. J. (2004), *Mathematical population genetics: theoretical introduction*, Vol. 27, Springer.

Eyre-Walker, A. & Keightley, P. D. (2007), 'The distribution of fitness effects of new mutations', *Nature Reviews Genetics* **8**(8), 610–618.

Fawcett, T. (2004), 'Roc graphs: Notes and practical considerations for researchers', *Machine learning* **31**(1), 1–38.

Fisher, R. A. (1999), *The genetical theory of natural selection: a complete variorum edition*, Oxford University Press.

Goeman, J. J. & Solari, A. (2014), 'Multiple hypothesis testing in genomics', *Statistics in medicine* **33**(11), 1946–1978.

Grubaugh, N. D., Gangavarapu, K., Quick, J., Matteson, N. L., De Jesus, J. G., Main, B. J., Tan, A. L., Paul, L. M., Brackney, D. E., Grewal, S. et al. (2019), 'An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using primalseq and ivar', *Genome biology* **20**(1), 1–19.

Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B. & Oppermann, M. (2005), 'Detection of the ccr5-δ32 hiv resistance gene in bronze age skeletons', *Genes & Immunity* **6**(4), 371–374.

Jewett, E. M., Steinrücken, M. & Song, Y. S. (2016), 'The effects of population size histories on estimates of selection coefficients from time-series genetic data', *Molecular biology and evolution* **33**(11), 3002–3027.

Khatri, B. S. (2016), 'Quantifying evolutionary dynamics from variant-frequency time series', *Scientific reports* **6**(1), 32497.

Kim, B. Y., Huber, C. D. & Lohmueller, K. E. (2017), 'Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples', *Genetics* **206**(1), 345–361.

Kimura, M. (1962), 'ON THE PROBABILITY OF FIXATION OF MUTANT GENES IN A POPULATION', *Genetics* **47**(6), 713–719.

Kimura, M. (1979), 'Model of effectively neutral mutations in which selective constraint is incorporated', *Proceedings of the National Academy of Sciences* **76**(7), 3440–3444.

Kimura, M. (1983), *The neutral theory of molecular evolution*, Cambridge University Press.

Kryazhimskiy, S. & Plotkin, J. B. (2008), 'The population genetics of dn/ds', *PLoS genetics* **4**(12), e1000304.

Lewontin, R. & Kojima, K.-i. (1960), 'The evolutionary dynamics of complex polymorphisms', *Evolution* pp. 458–472.

Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. (2012), 'Estimating allele age and selection coefficient from time-serial data', *Genetics* **192**(2), 599–607.

Markov, P. V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N. I. & Katzourakis, A. (2023), 'The evolution of sars-cov-2', *Nature Reviews Microbiology* **21**(6), 361–379.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *The computer journal* **7**(4), 308–313.

Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. (2010), 'Viral mutation rates', *Journal of virology* **84**(19), 9733–9748.

Seo, T.-K., Thorne, J. L., Hasegawa, M. & Kishino, H. (2002), 'Estimation of effective population size of hiv-1 within a host: a pseudomaximum-likelihood approach', *Genetics* **160**(4), 1283–1293.

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X. et al. (1999), 'Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection', *Journal of virology* **73**(12), 10489–10502.

Slatkin, M. (2008), 'Linkage disequilibrium—understanding the evolutionary past and mapping the medical future', *Nature Reviews Genetics* **9**(6), 477–485.

Storey, J. D. (2002), 'A direct approach to false discovery rates', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **64**(3), 479–498.

Storey, J. D. (2003), 'The positive false discovery rate: a bayesian interpretation and the q-value', *The annals of statistics* **31**(6), 2013–2035.

Storey, J. D., Taylor, J. E. & Siegmund, D. (2004), 'Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **66**(1), 187–205.

Storey, J. D. & Tibshirani, R. (2003), 'Statistical significance for genomewide studies', *Proceedings of the National Academy of Sciences* **100**(16), 9440–9445.

Tataru, P., Bataillon, T. & Hobolth, A. (2015), 'Inference under a wright-fisher model using an accurate beta approximation', *Genetics* **201**(3), 1133–1141.

Woods, R., Schneider, D., Winkworth, C. L., Riley, M. A. & Lenski, R. E. (2006), 'Tests of parallel molecular evolution in a long-term experiment with escherichia coli', *Proceedings of the National Academy of Sciences* **103**(24), 9107–9112.

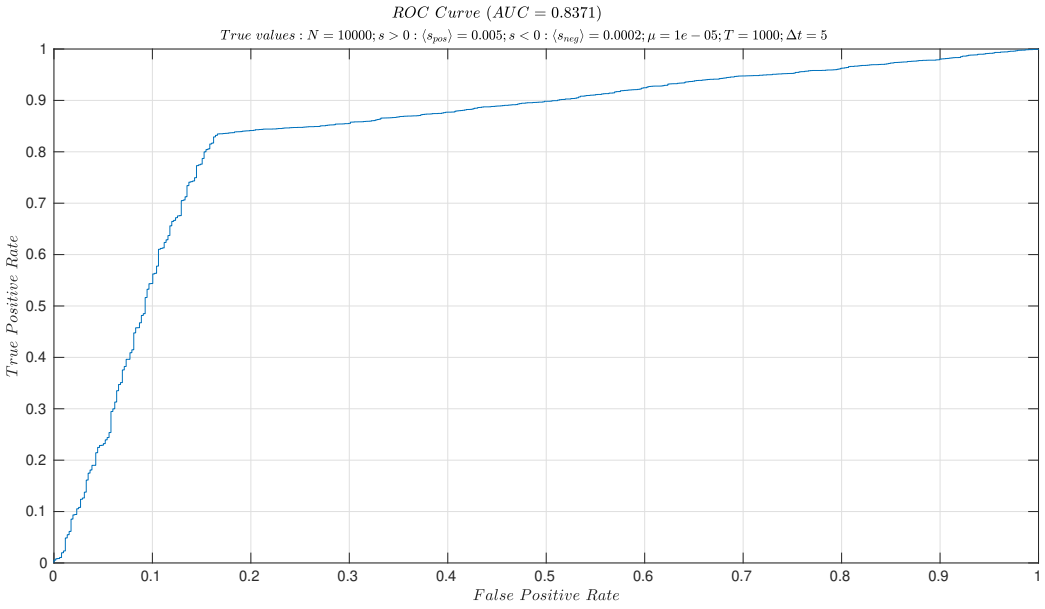
Wright, S. (1931), 'Evolution in mendelian populations', *Genetics* **16**(2), 97.

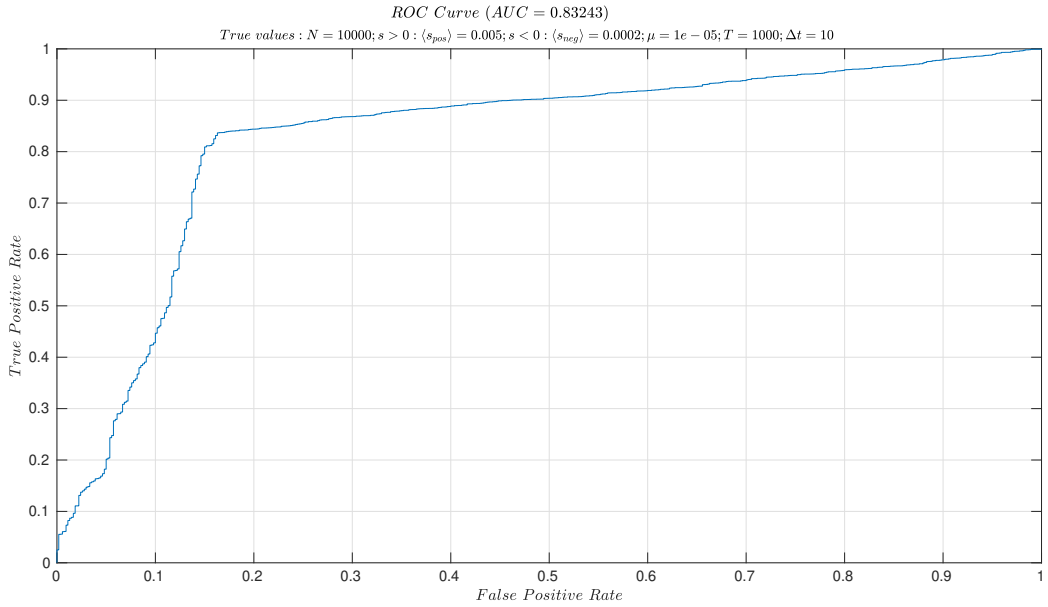
Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y. et al. (2020), 'A new coronavirus associated with human respiratory disease in china', *Nature* **579**(7798), 265–269.

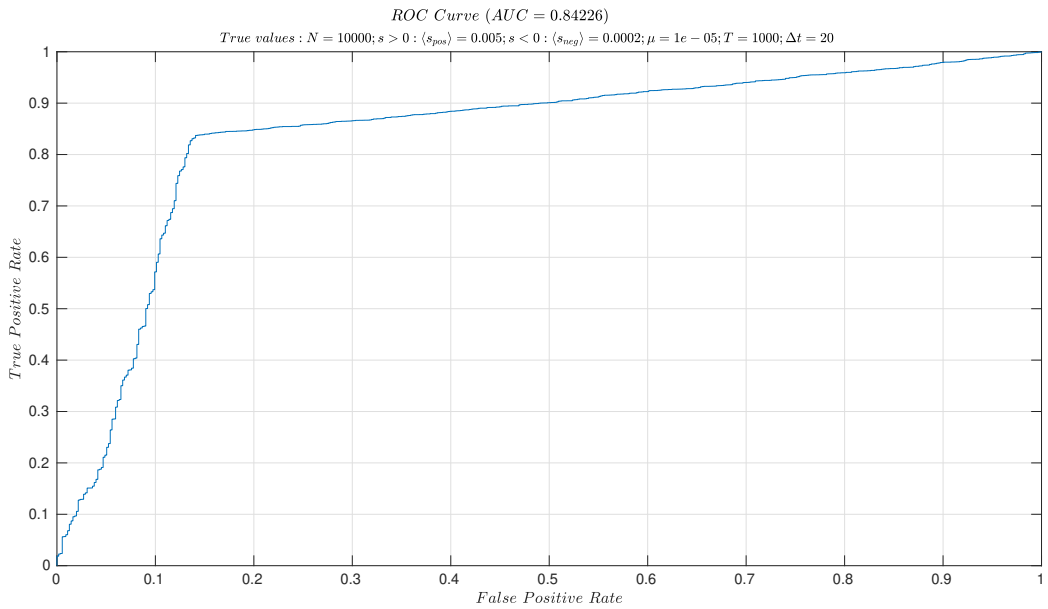
Zweig, M. H. & Campbell, G. (1993), 'Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine', *Clinical chemistry* **39**(4), 561–577.

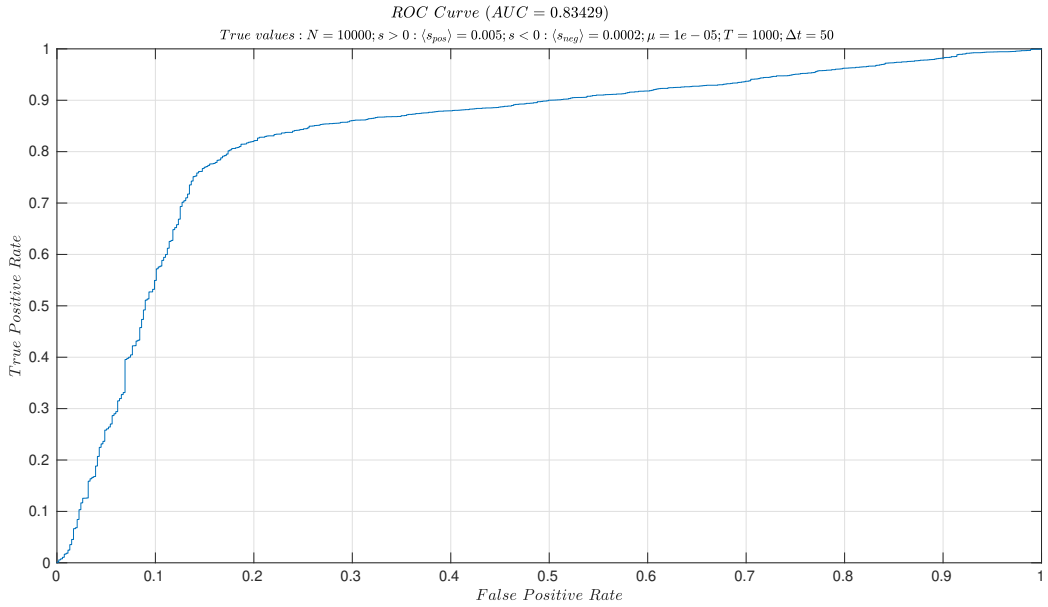
8 Appendix

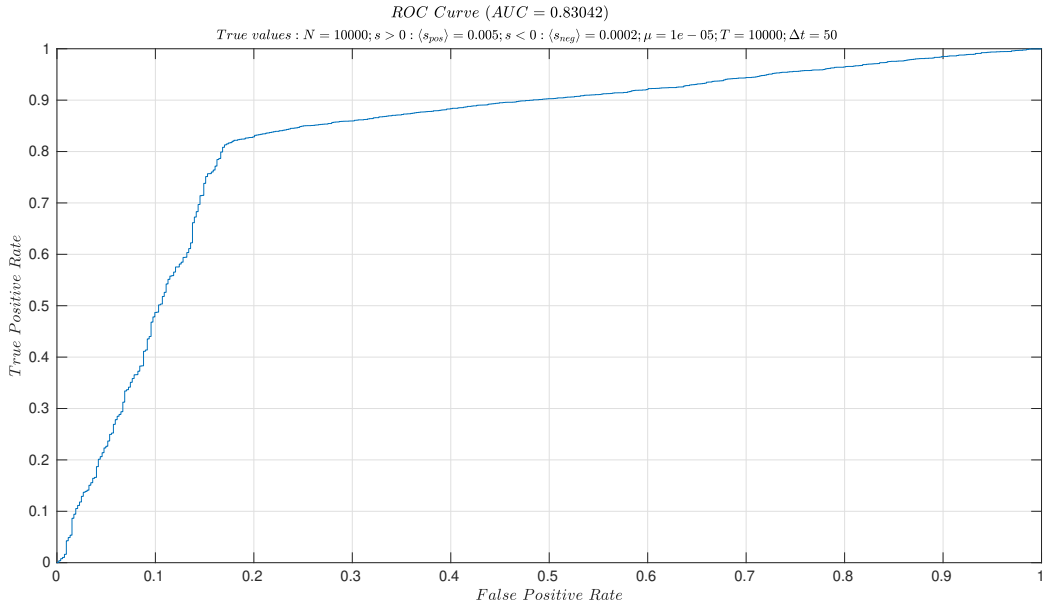
8.1 ROC plots for optimising N μ s with and without sampling in genome-wide simulation

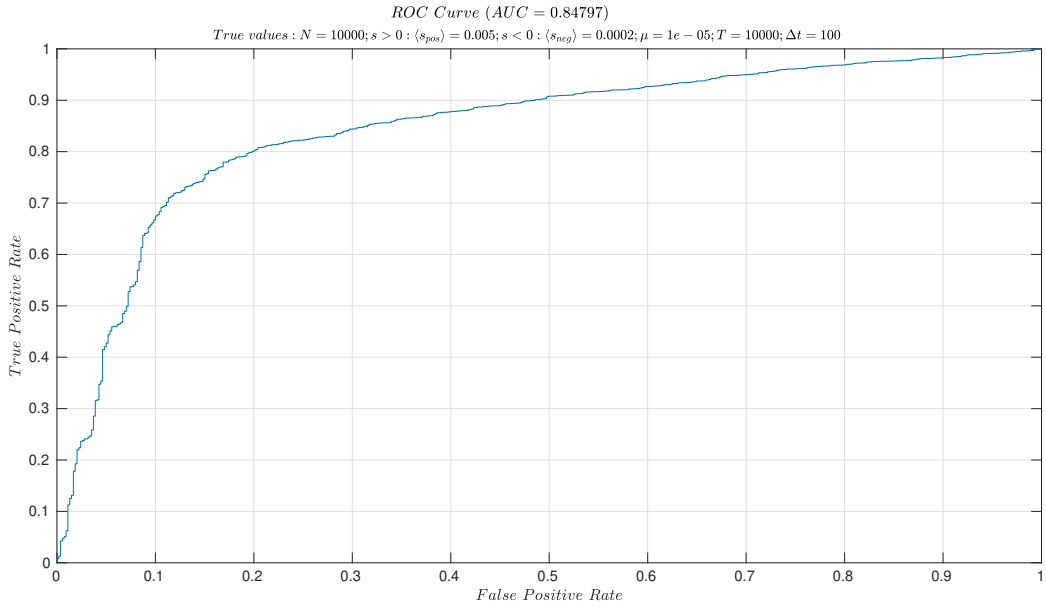


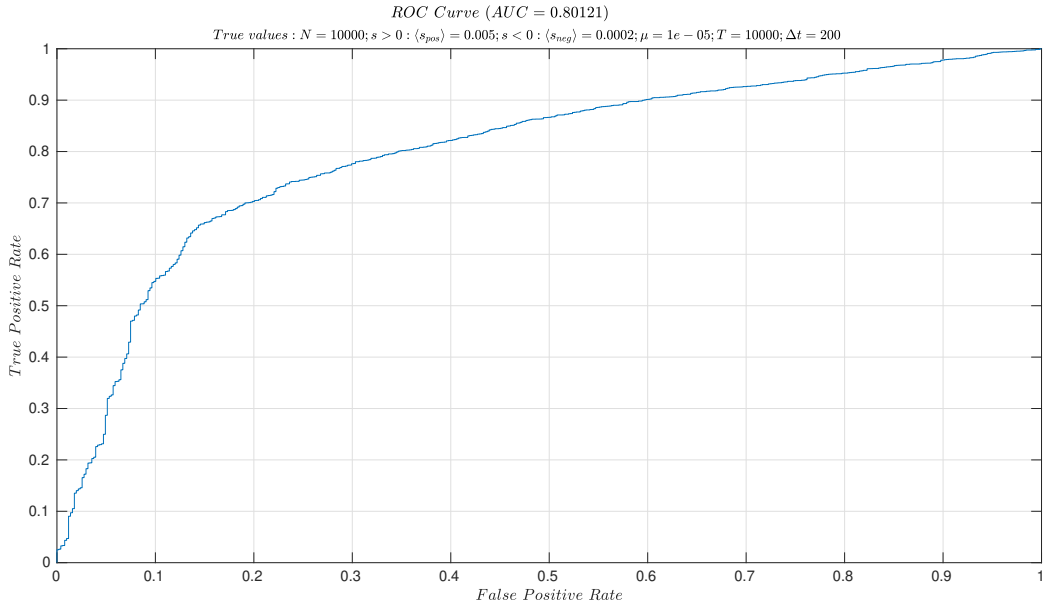

ROC Curve (AUC = 0.83243)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

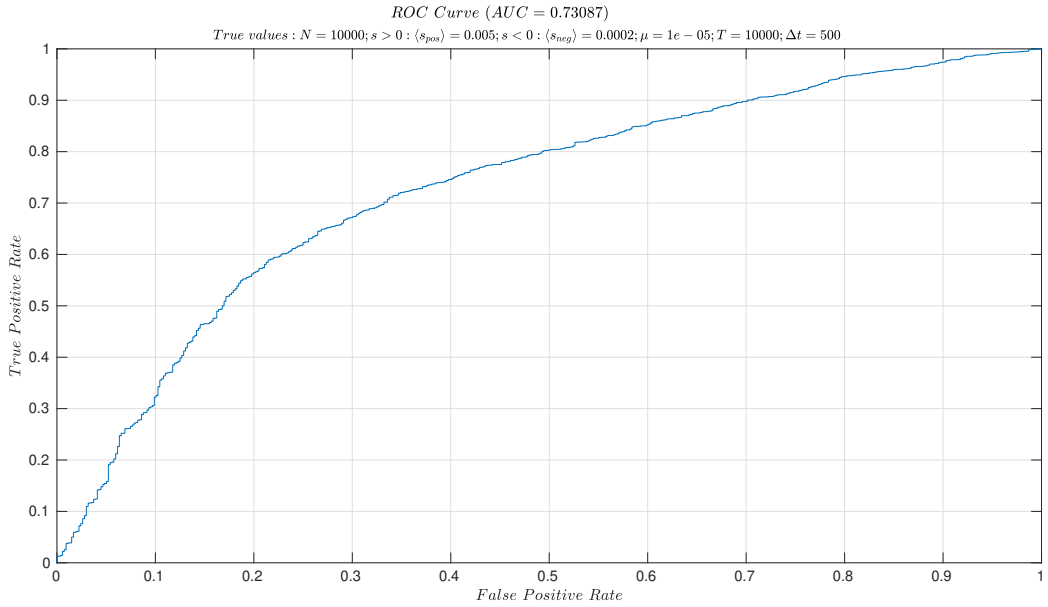

ROC Curve (AUC = 0.84226)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

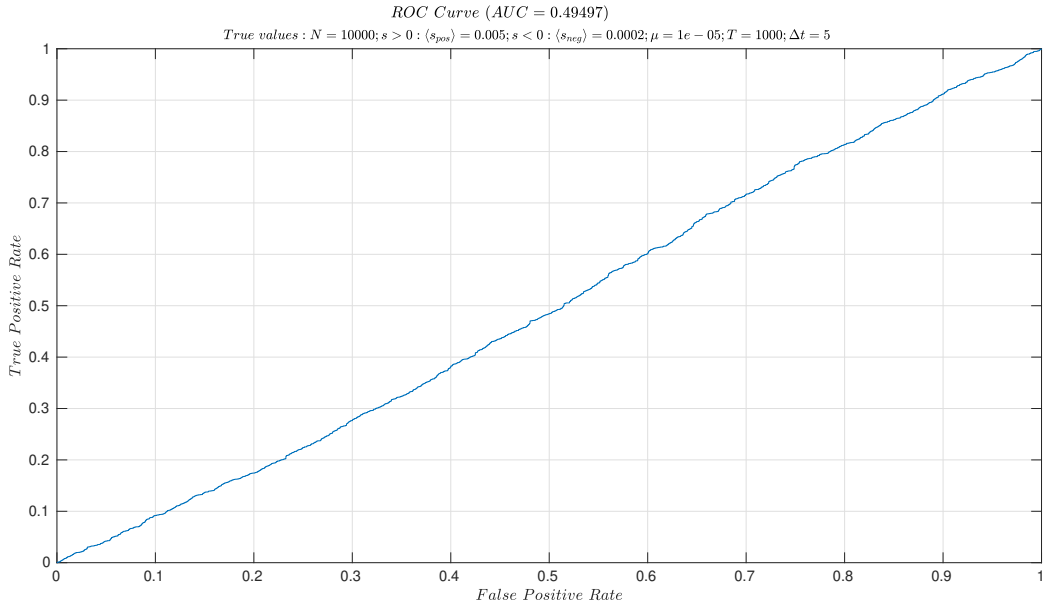

ROC Curve (AUC = 0.83429)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

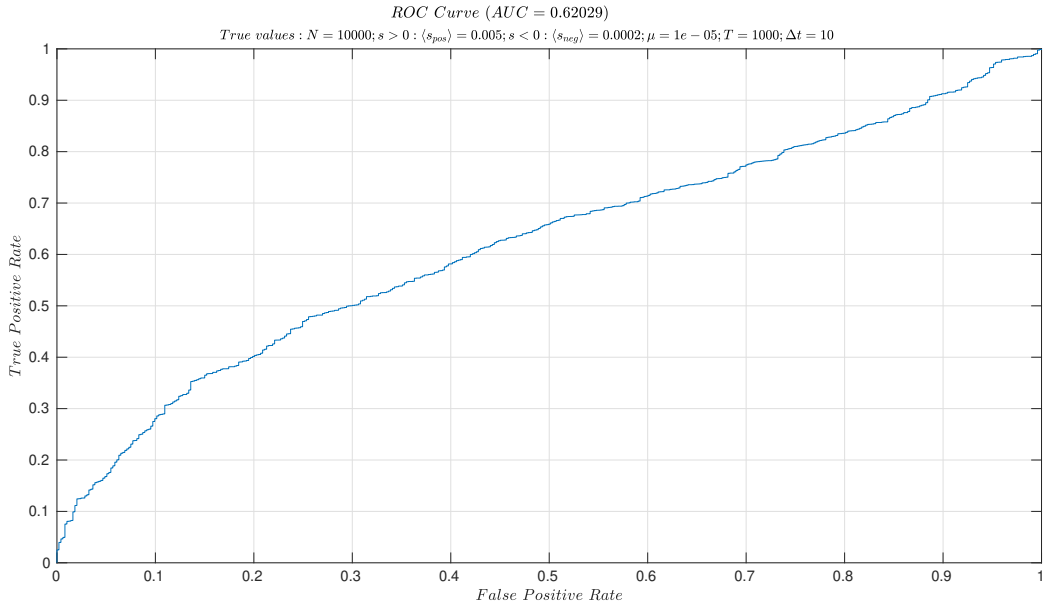

ROC Curve (AUC = 0.83042)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

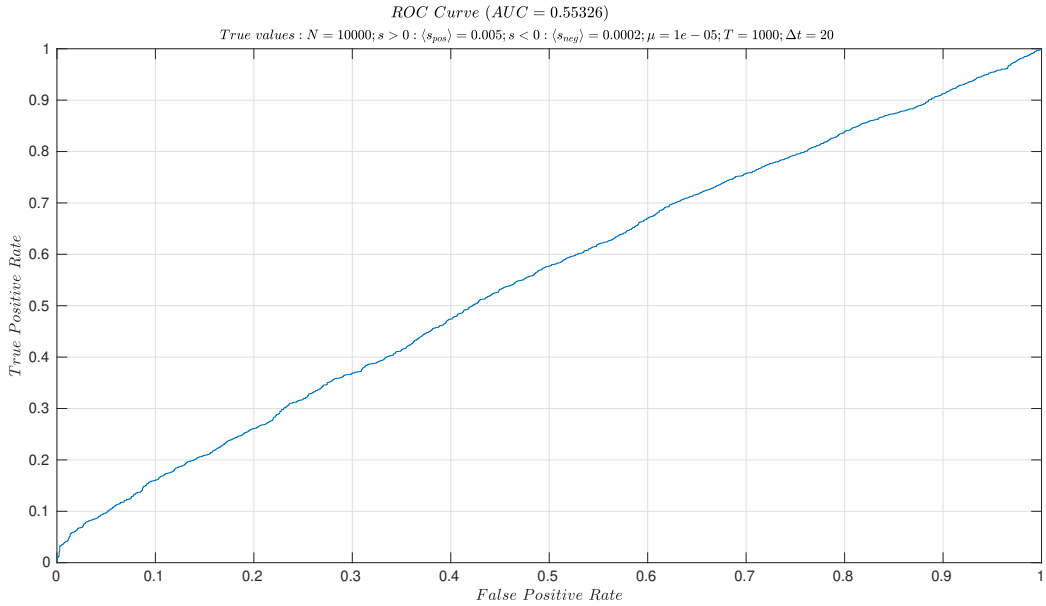

ROC Curve (AUC = 0.84797)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

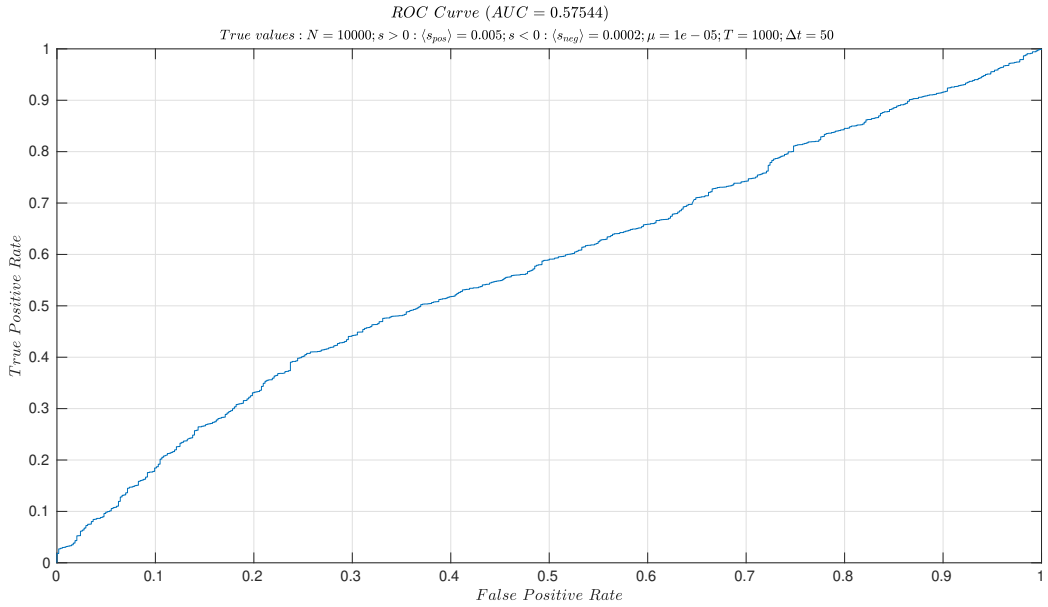

ROC Curve (AUC = 0.80121)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

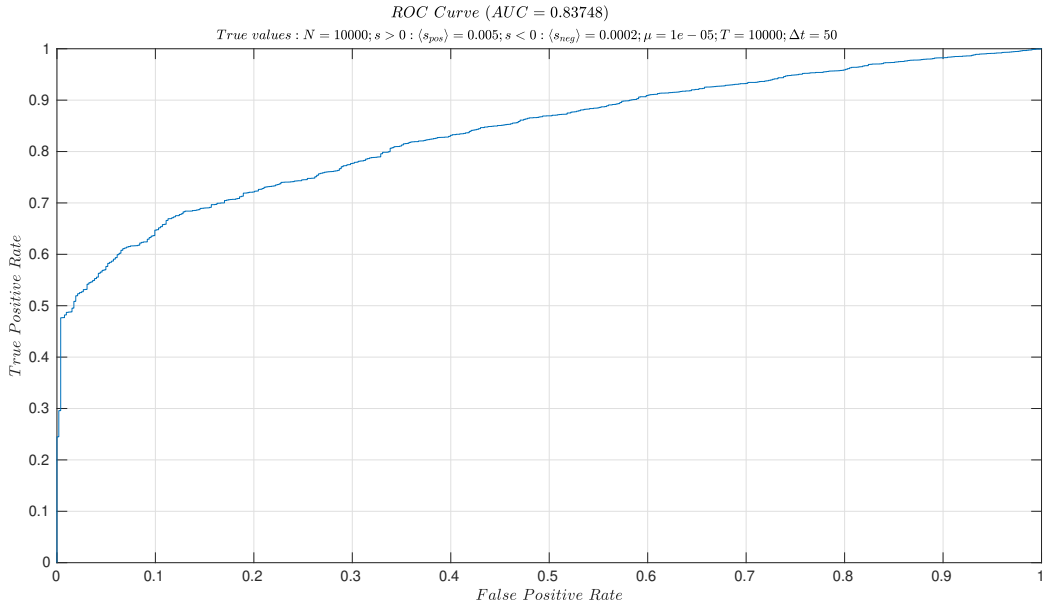

ROC Curve (AUC = 0.73087)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 500
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

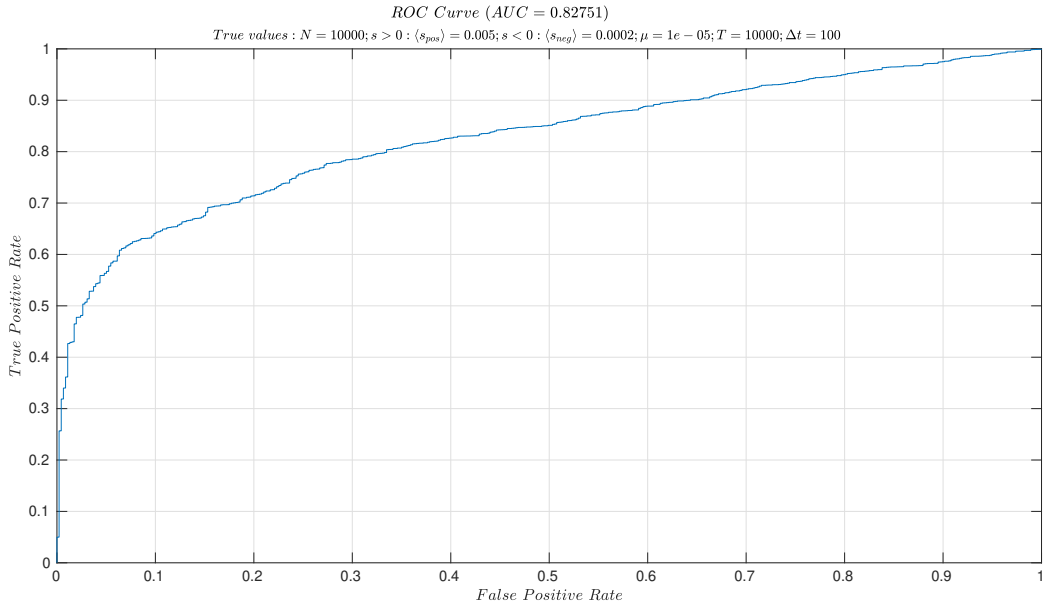

ROC Curve (AUC = 0.49497)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

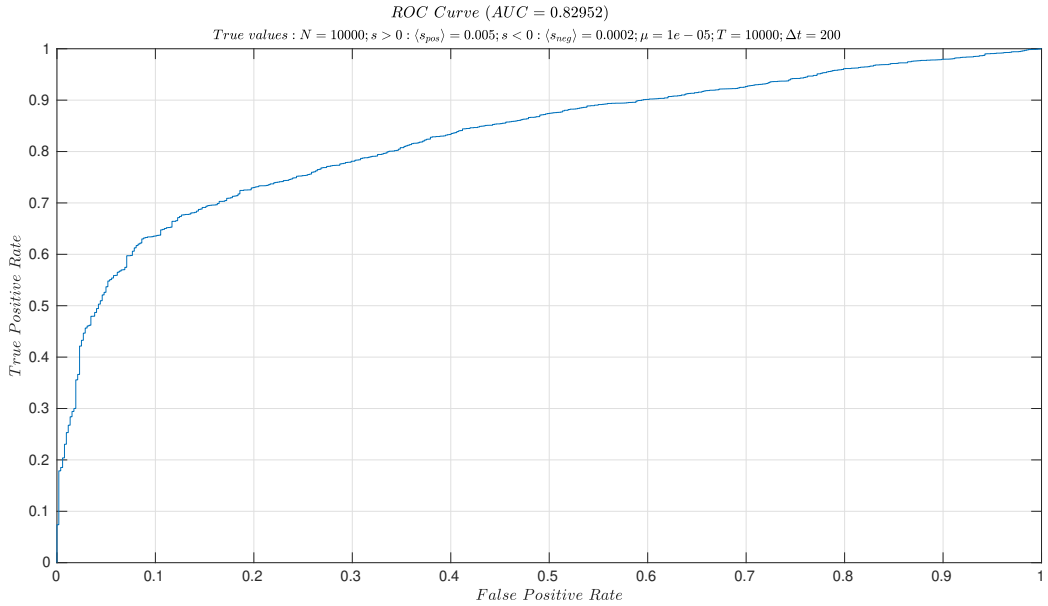

ROC Curve (AUC = 0.62029)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

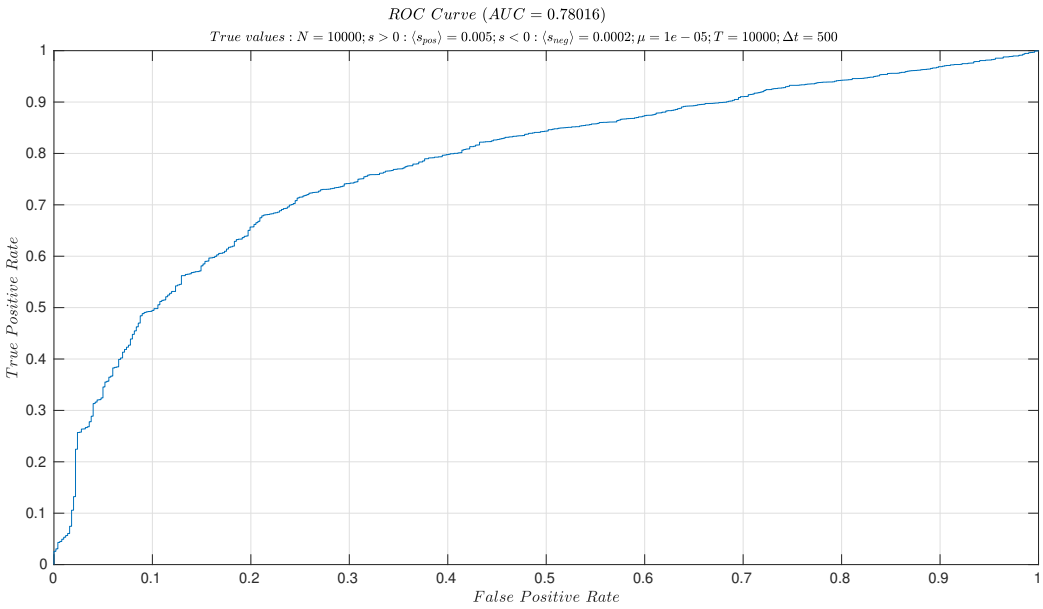

ROC Curve (AUC = 0.55326)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


ROC Curve (AUC = 0.57544)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

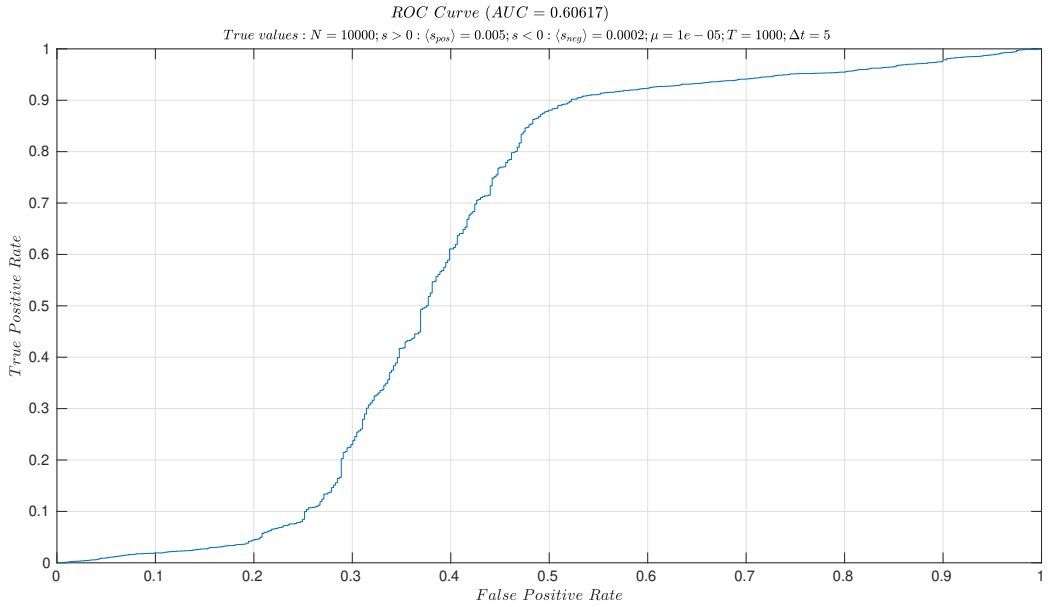

ROC Curve (AUC = 0.83748)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

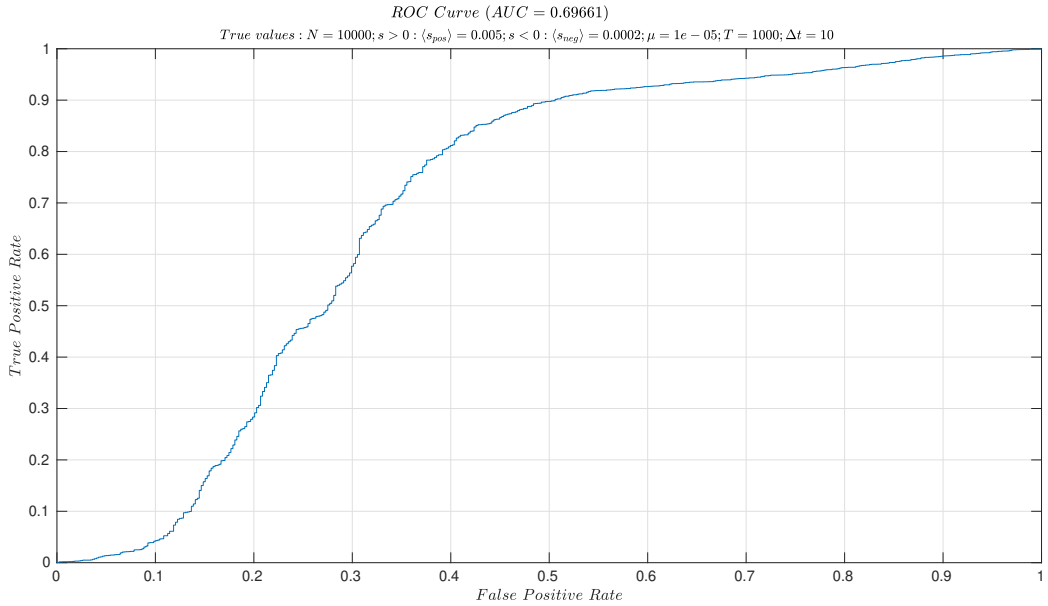

ROC Curve (AUC = 0.82751)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

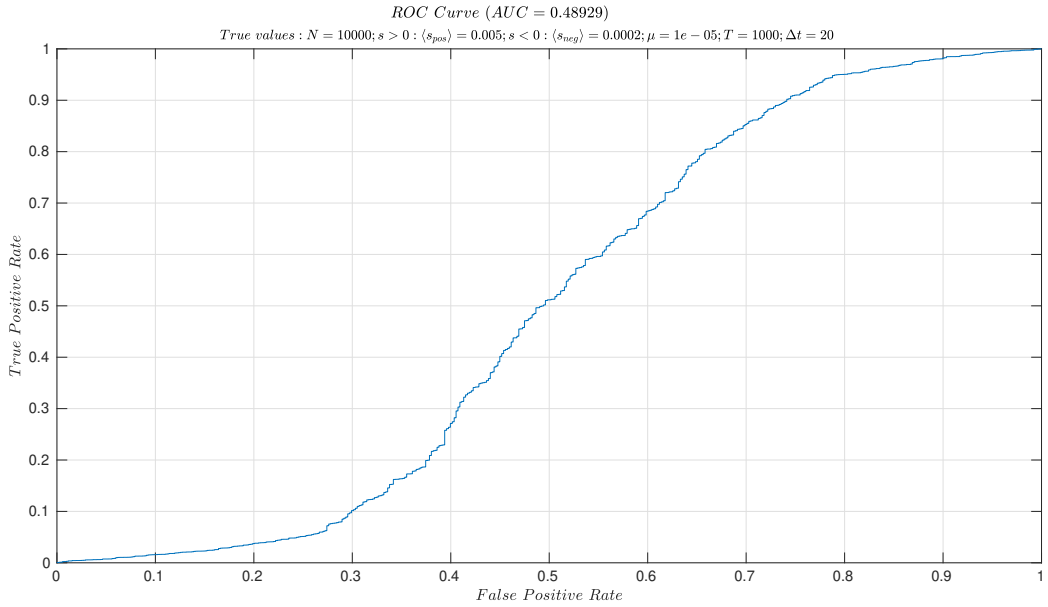

ROC Curve (AUC = 0.82952)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

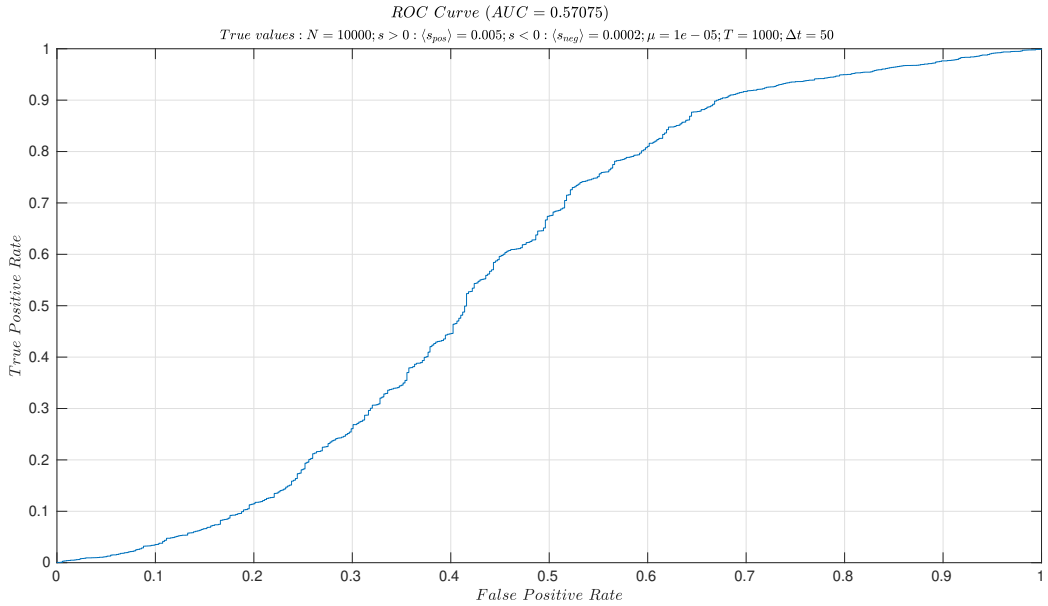


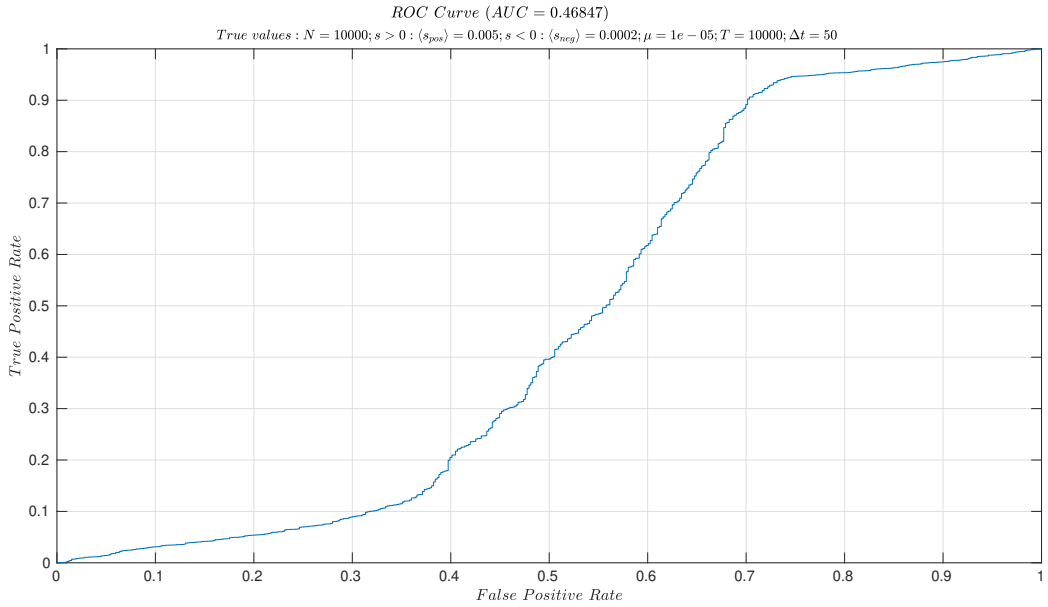
8.2 ROC plots for fixing N with and without sampling in genome-wide simulation

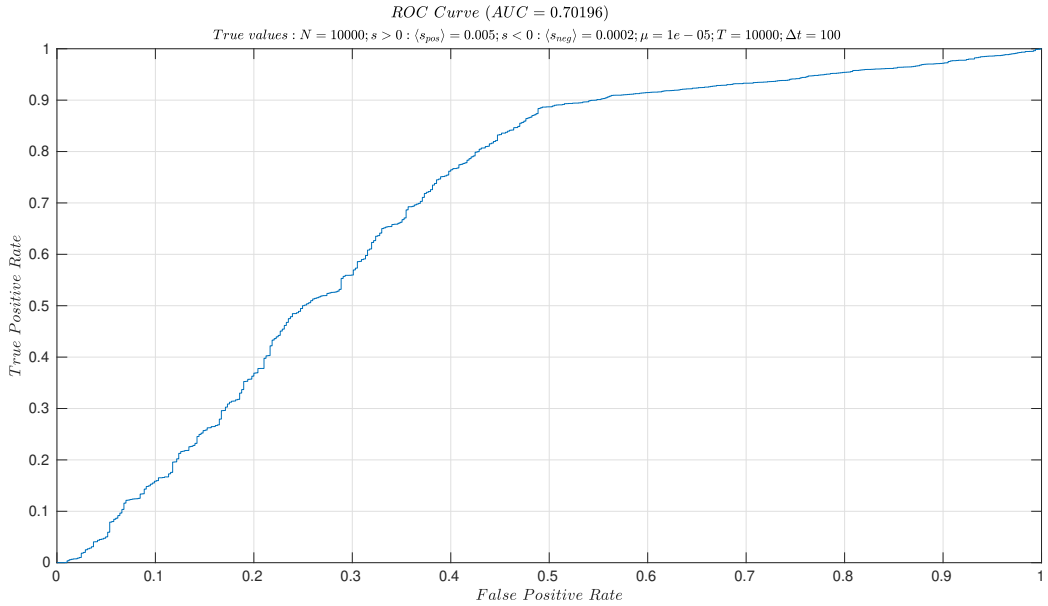

ROC Curve (AUC = 0.60617)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

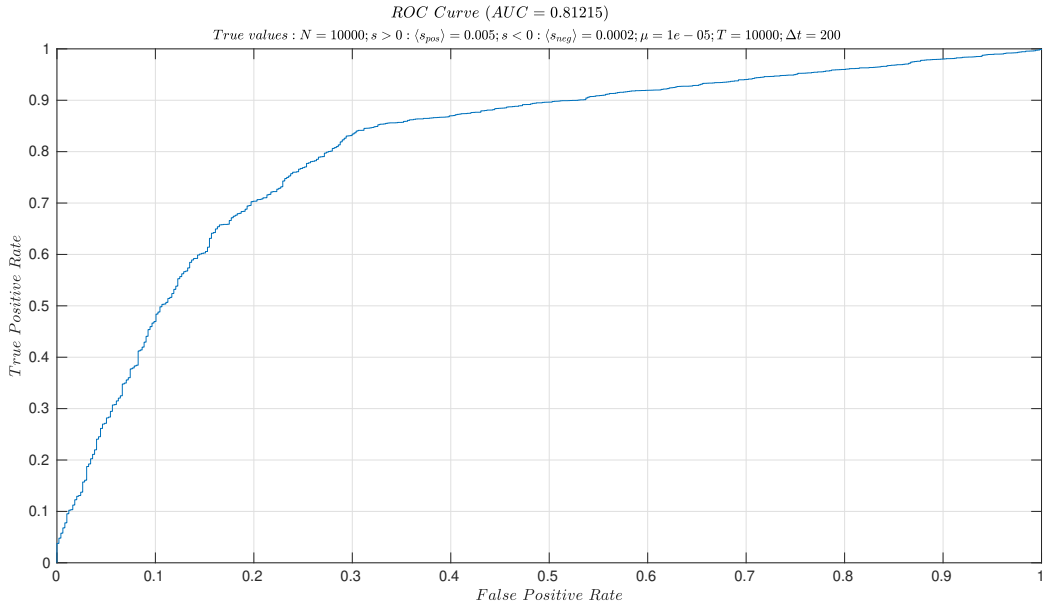

ROC Curve (AUC = 0.69661)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

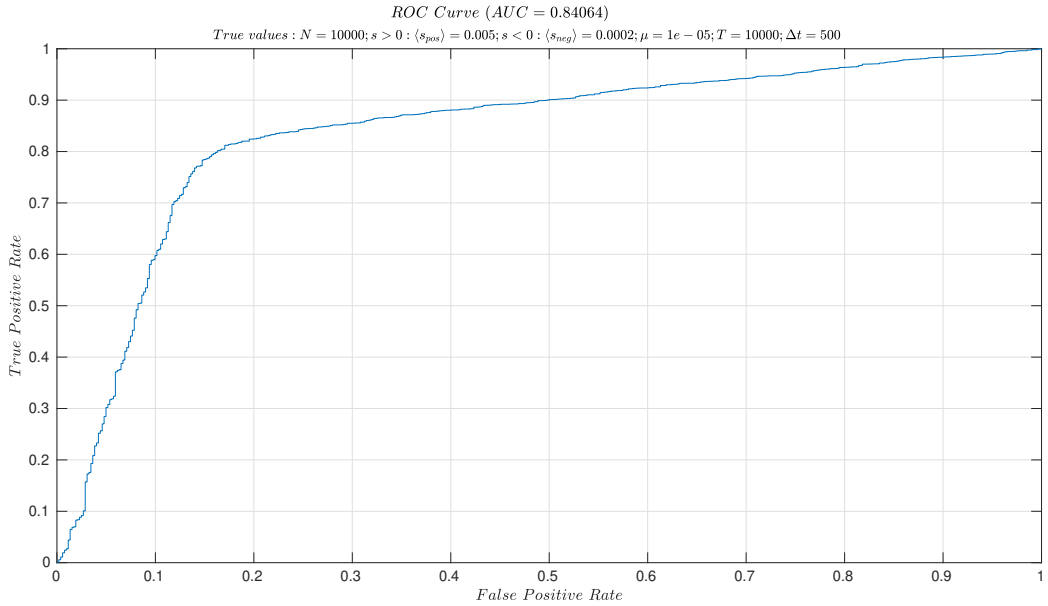

ROC Curve (AUC = 0.48929)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

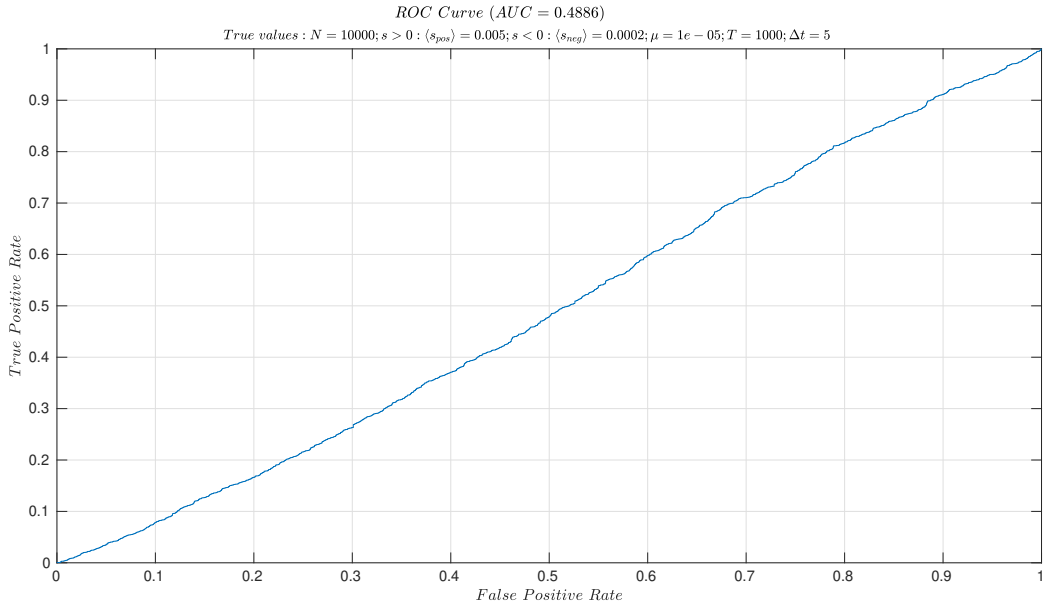

ROC Curve (AUC = 0.57075)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

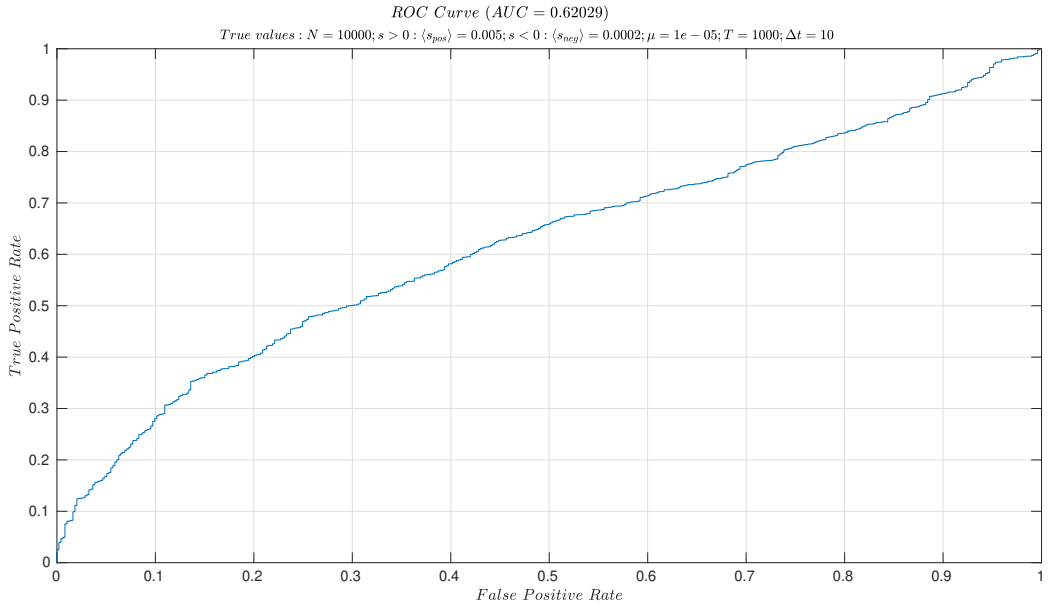

ROC Curve (AUC = 0.46847)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

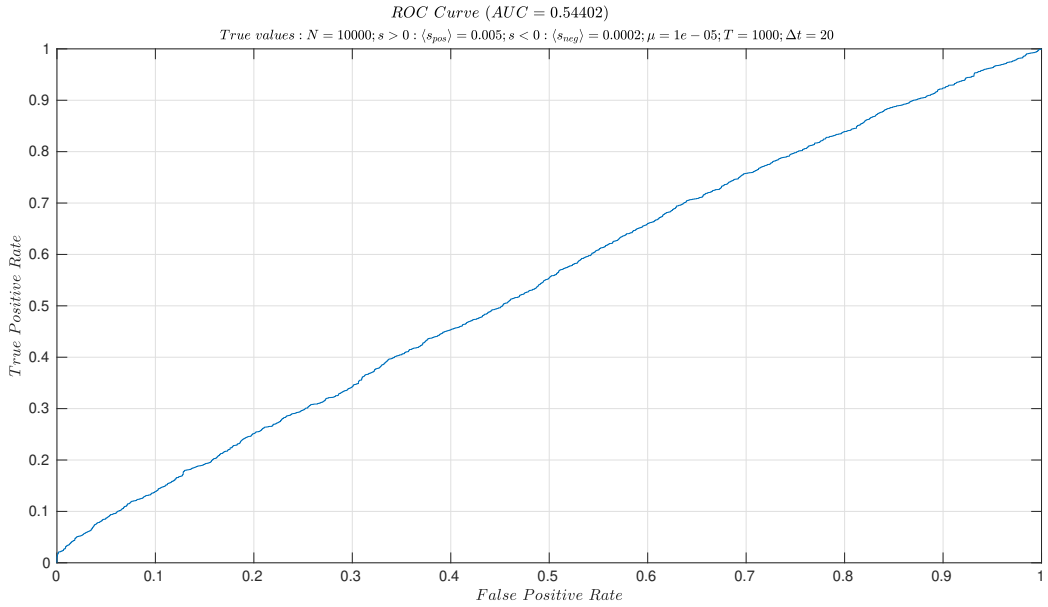

ROC Curve (AUC = 0.70196)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

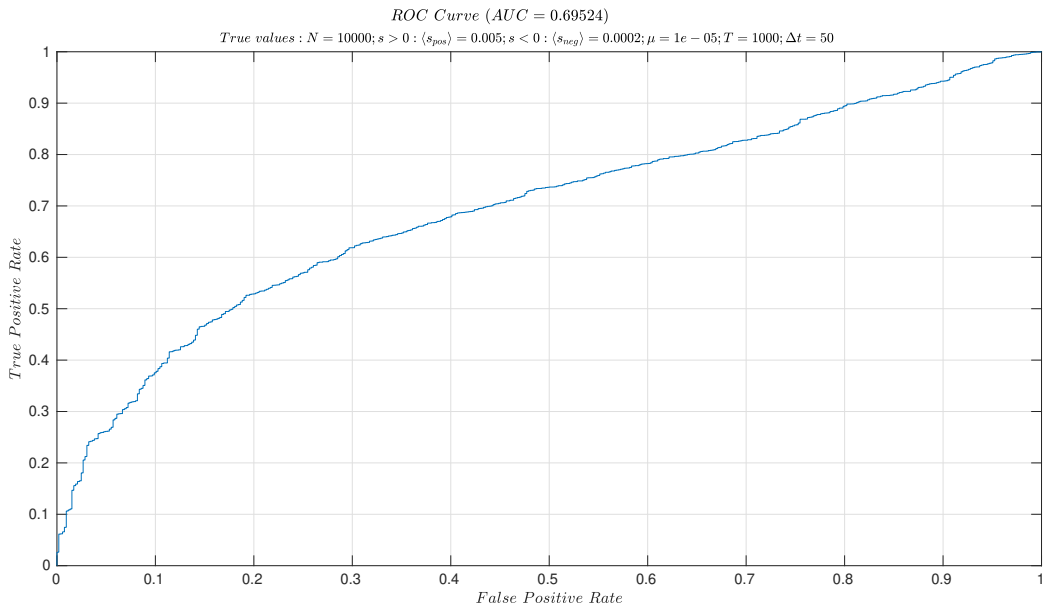

ROC Curve (AUC = 0.81215)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

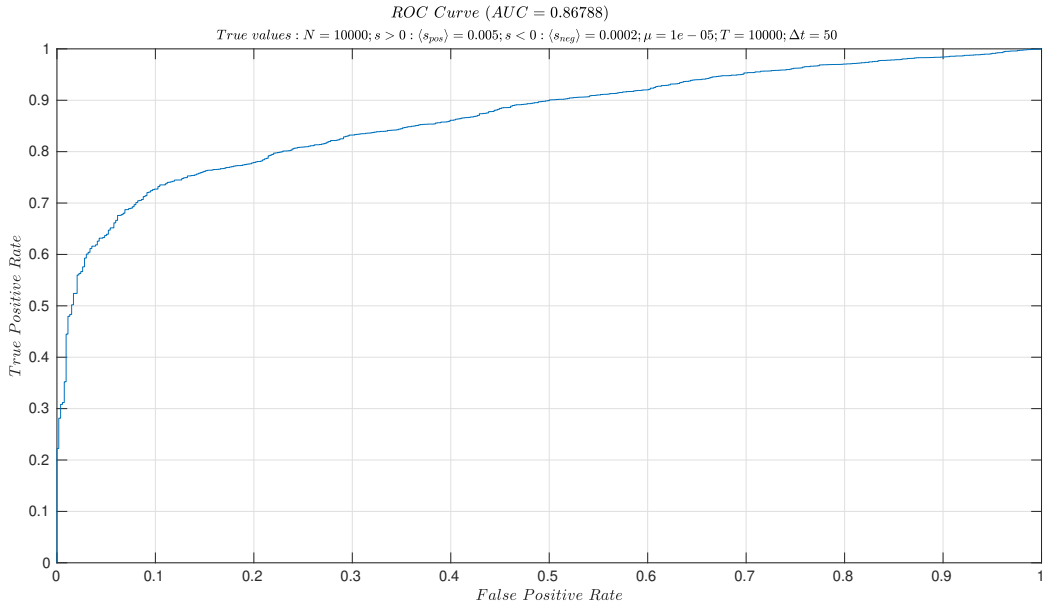

ROC Curve (AUC = 0.84064)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 500
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

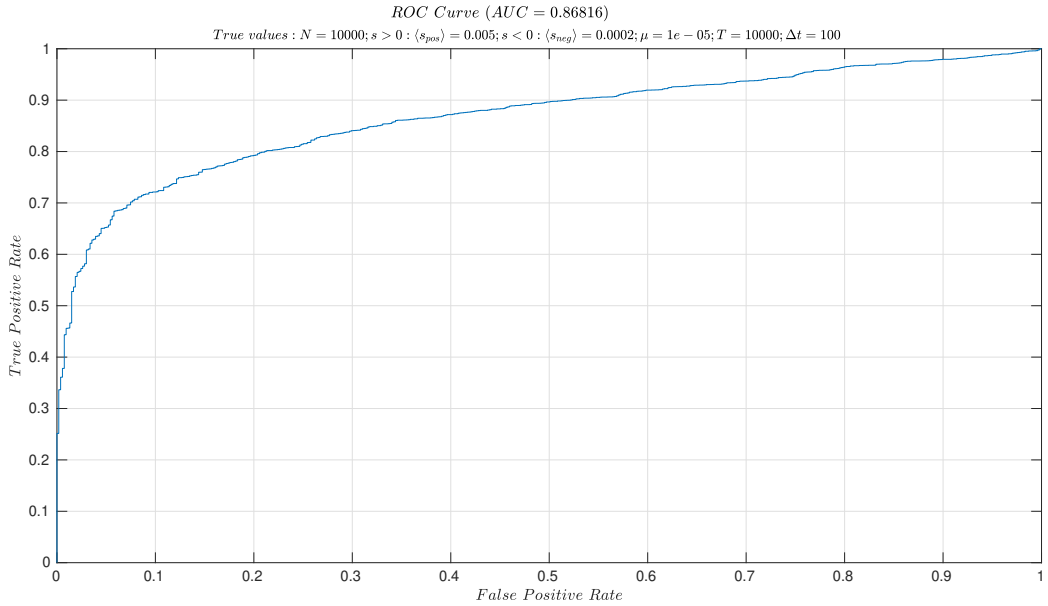

ROC Curve (AUC = 0.4886)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

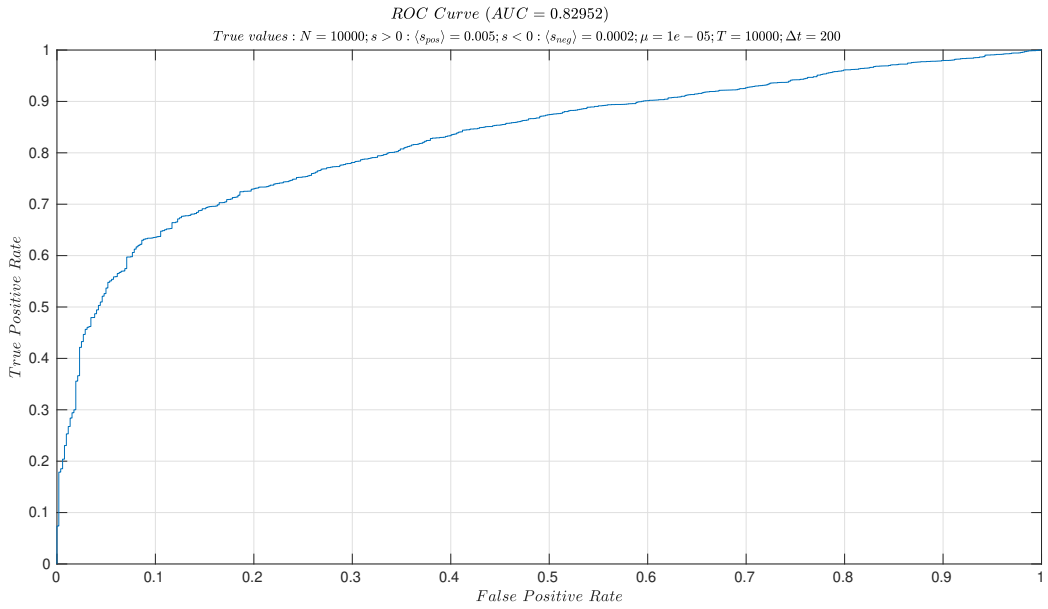

ROC Curve (AUC = 0.62029)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

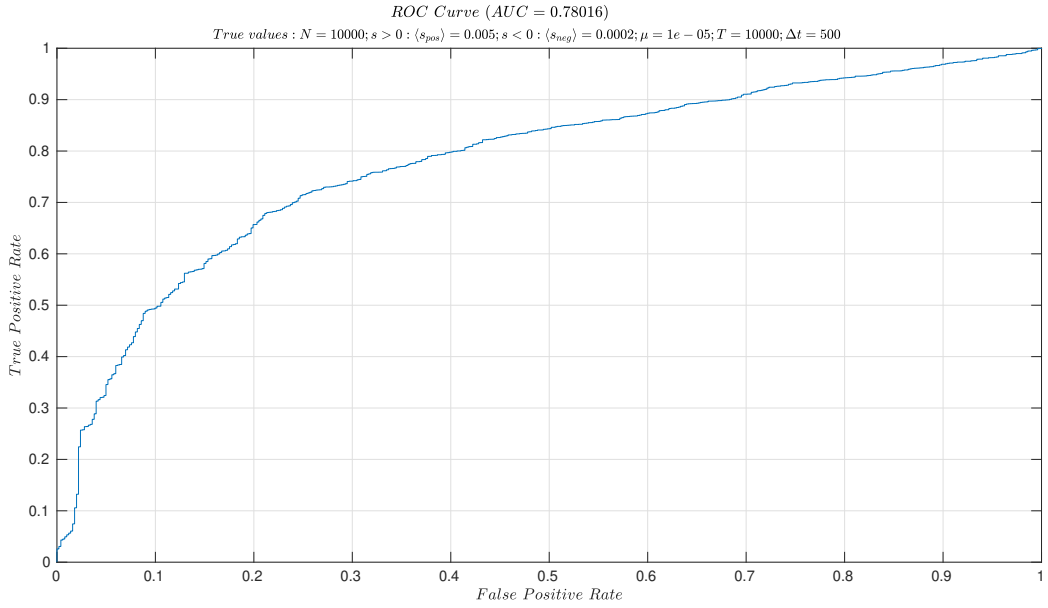

ROC Curve (AUC = 0.54402)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


ROC Curve (AUC = 0.69524)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

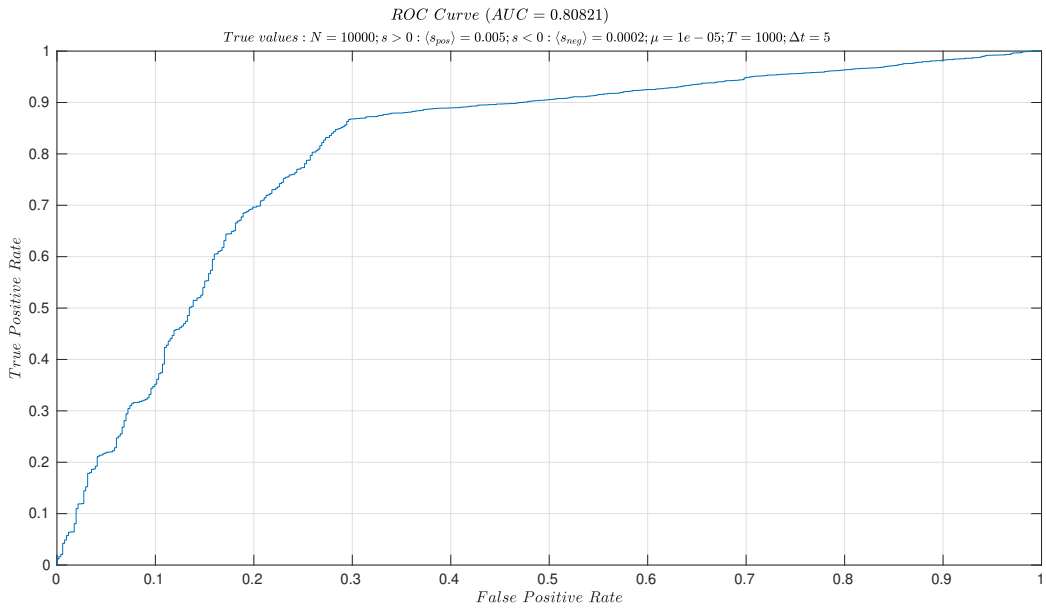

ROC Curve (AUC = 0.86788)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

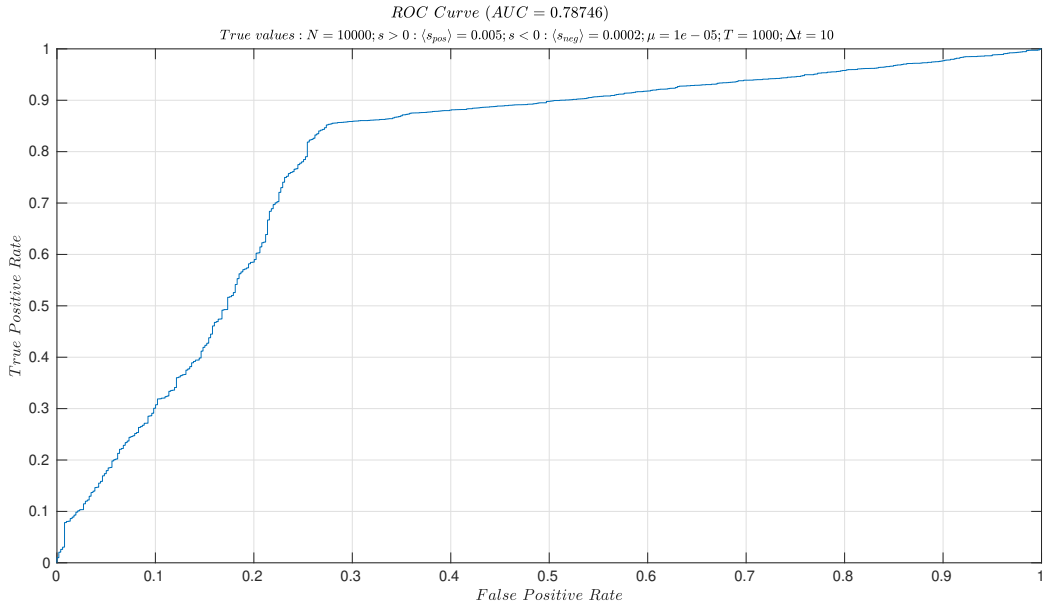

ROC Curve (AUC = 0.86816)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

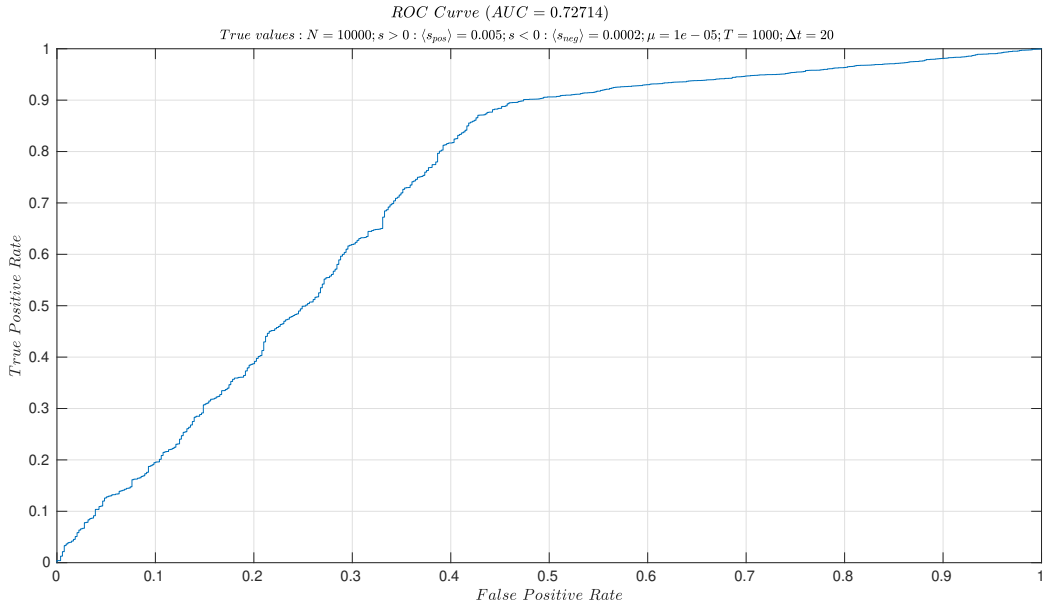

ROC Curve (AUC = 0.82952)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

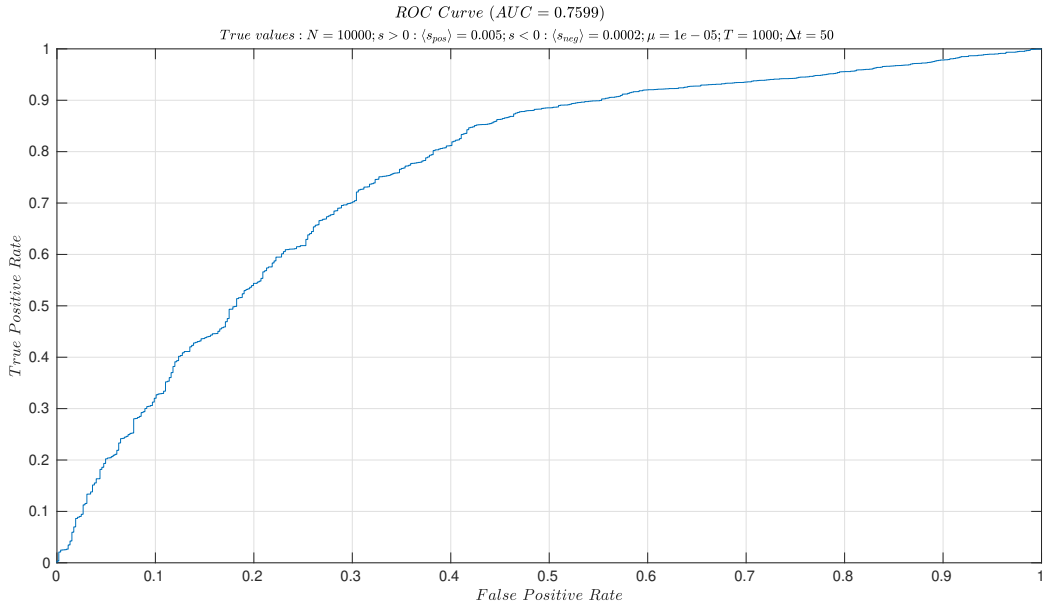

ROC Curve (AUC = 0.78016)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 500
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

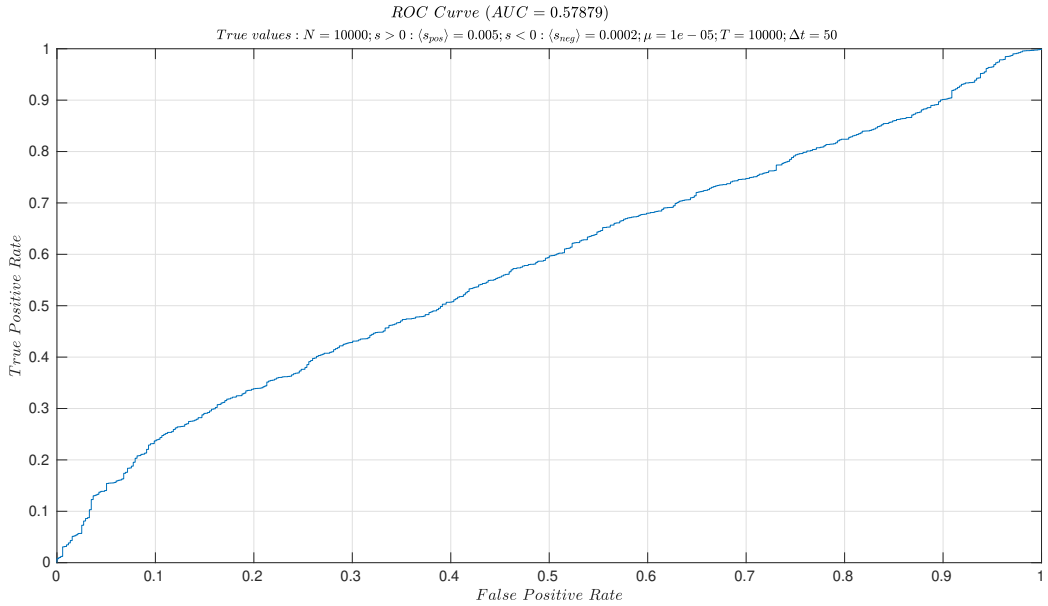
8.3 ROC plots for fixing μ with and without sampling in genome-wide simulation

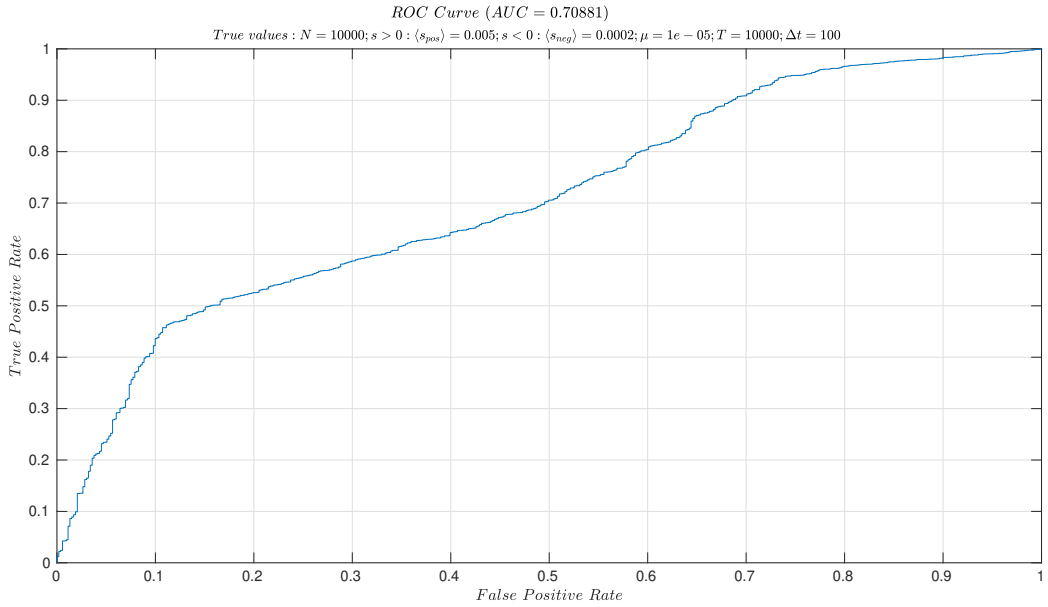


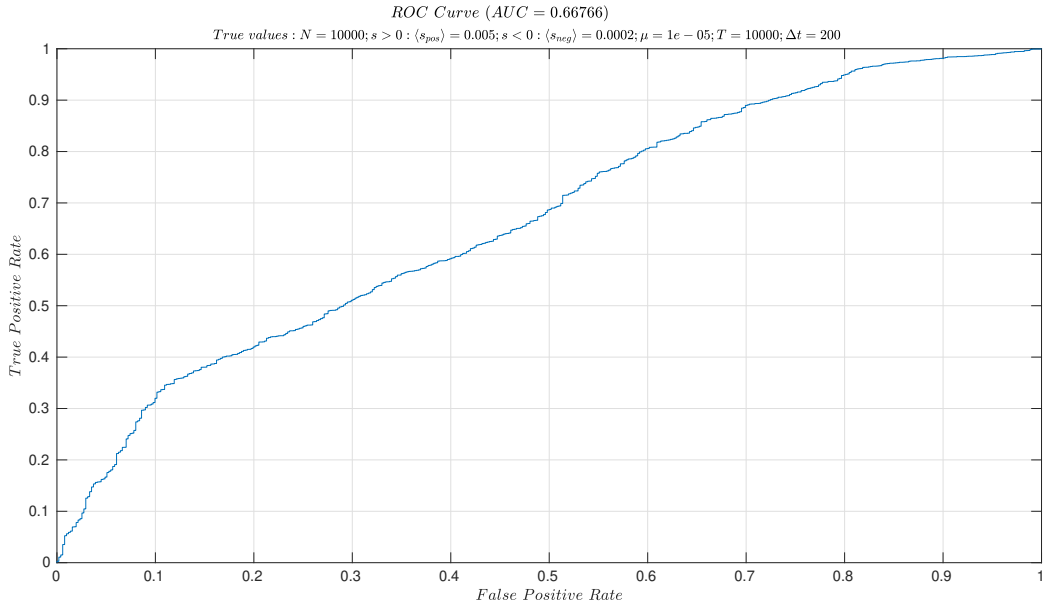

ROC Curve (AUC = 0.78746)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

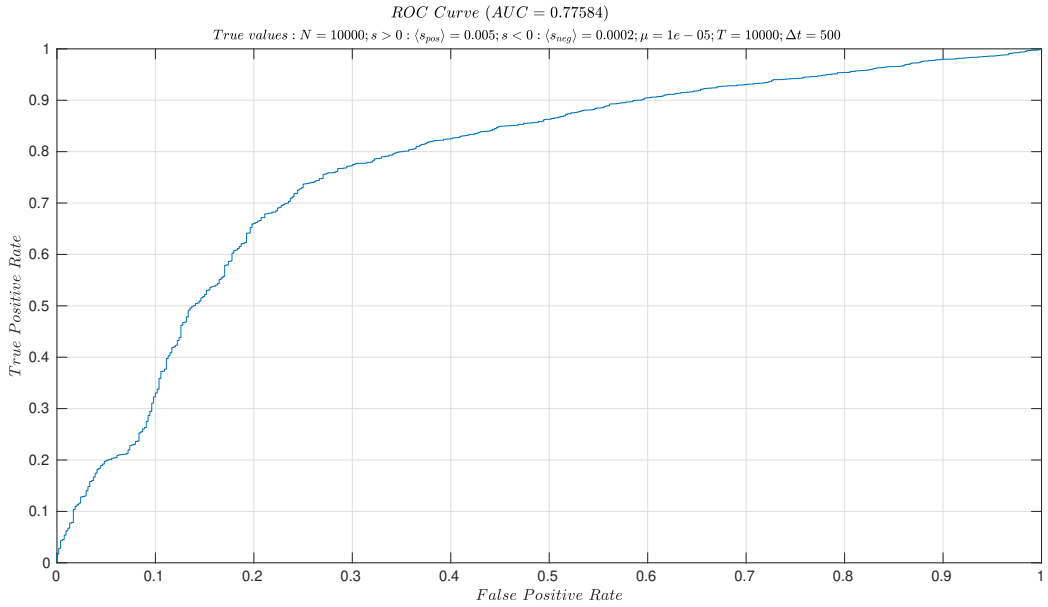

ROC Curve (AUC = 0.72714)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

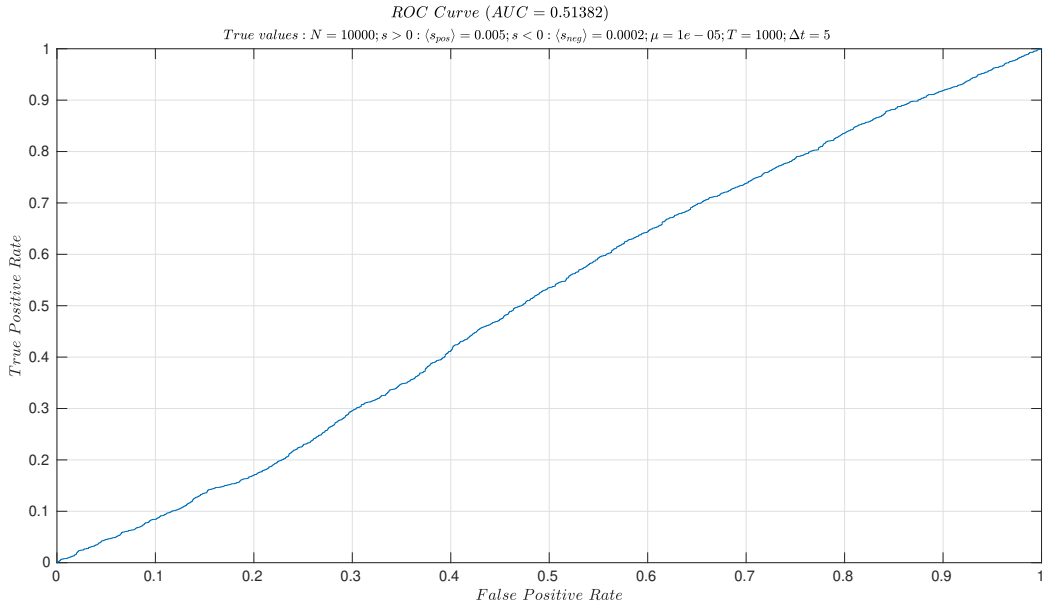

ROC Curve (AUC = 0.7599)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

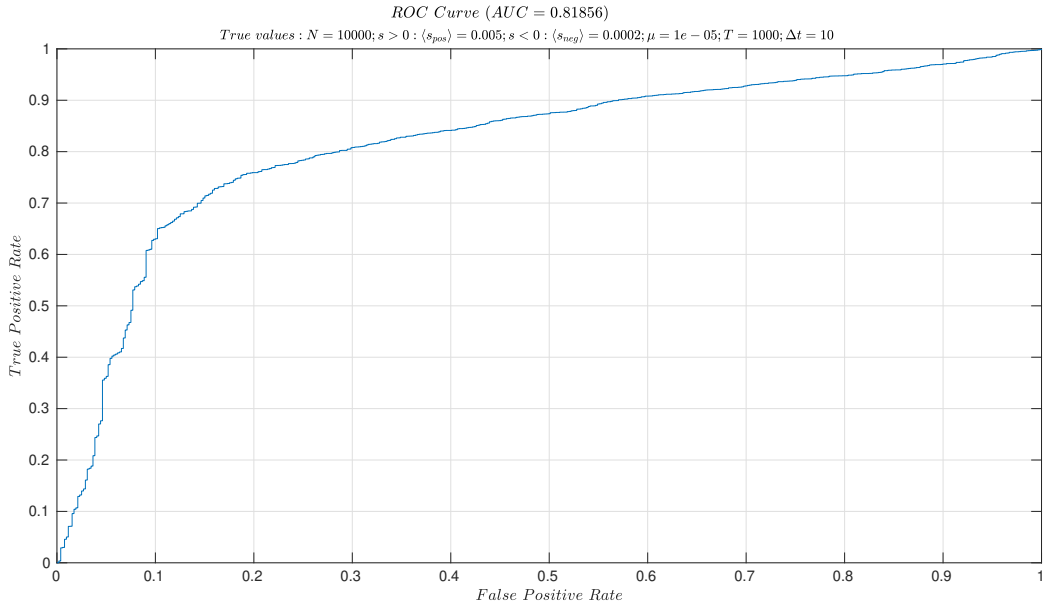

ROC Curve (AUC = 0.57879)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

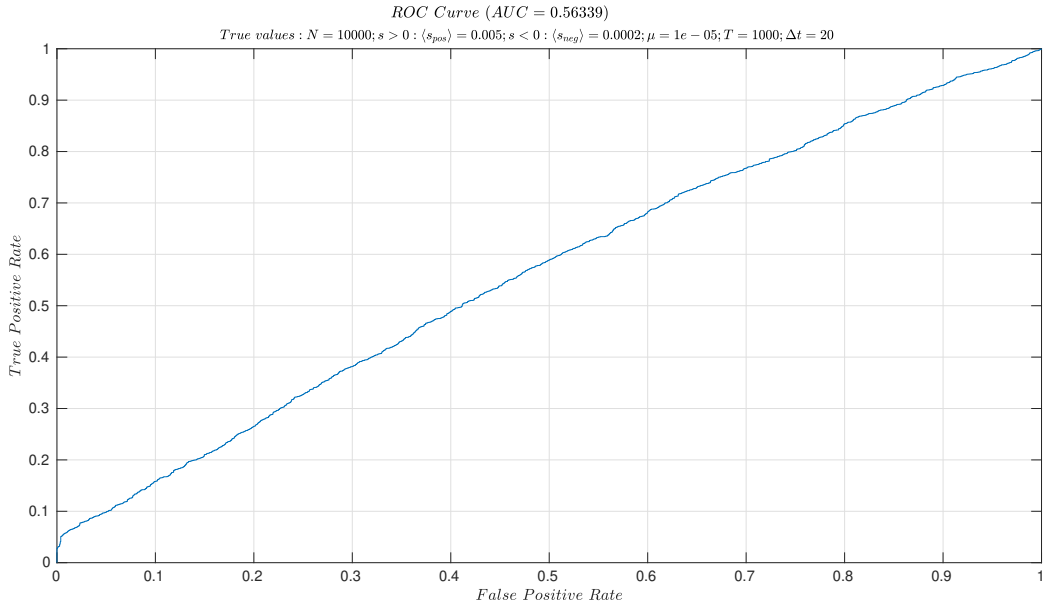

ROC Curve (AUC = 0.70881)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

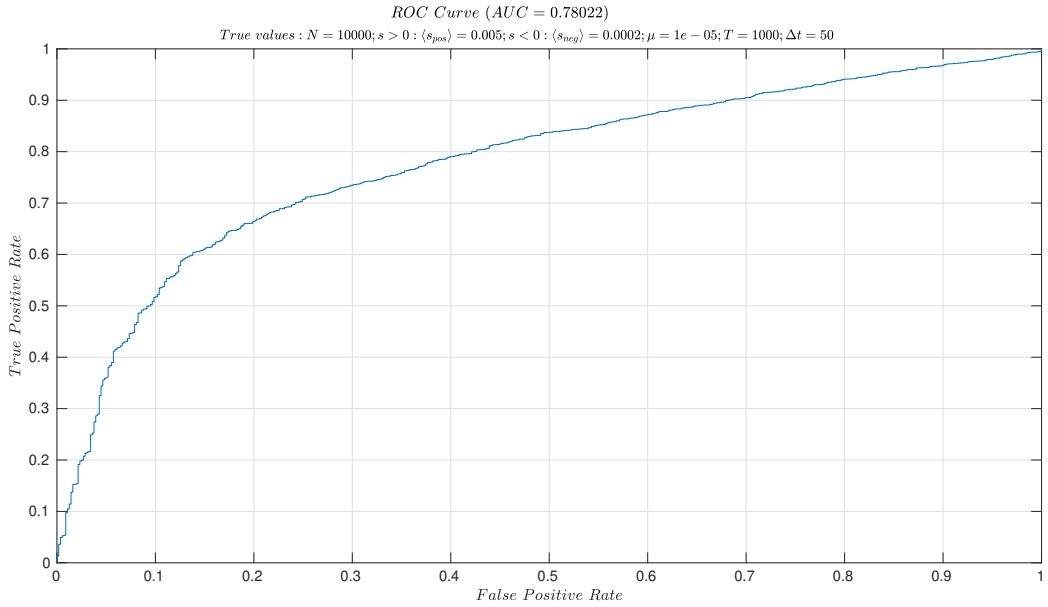

ROC Curve (AUC = 0.66766)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

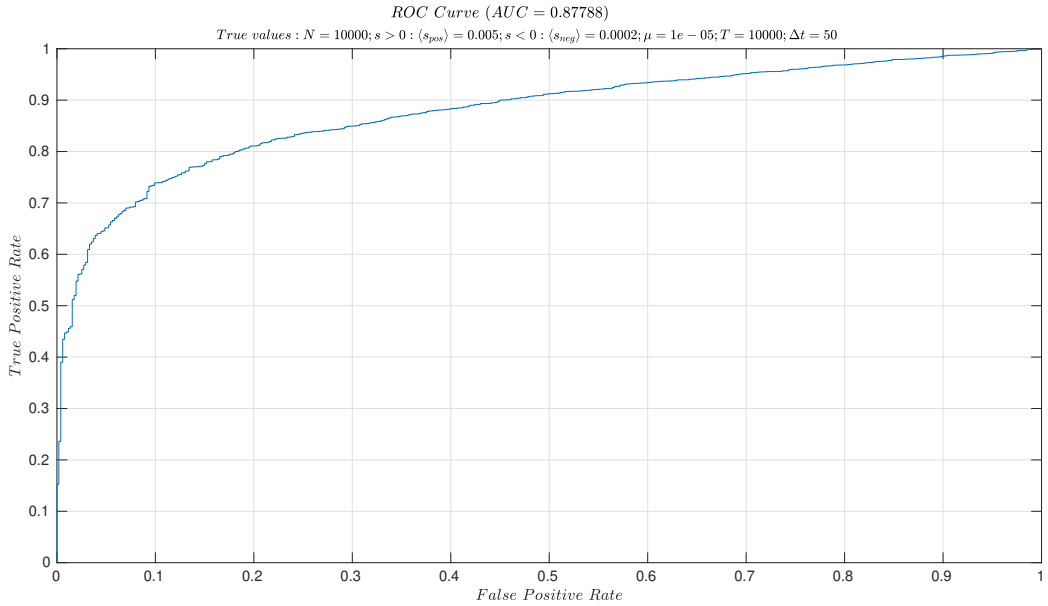

ROC Curve (AUC = 0.77584)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 500
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

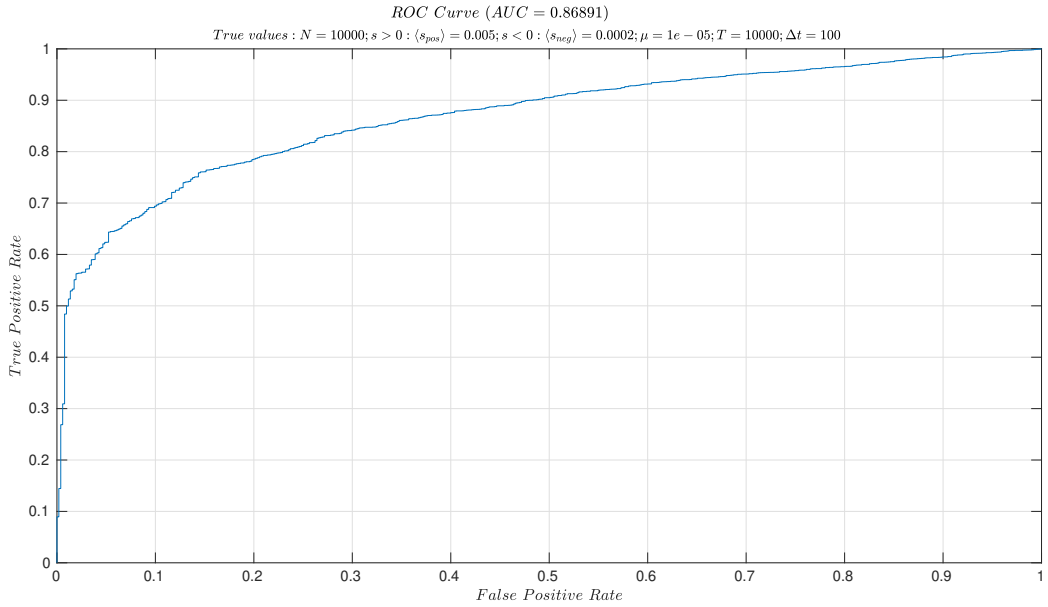

ROC Curve (AUC = 0.51382)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

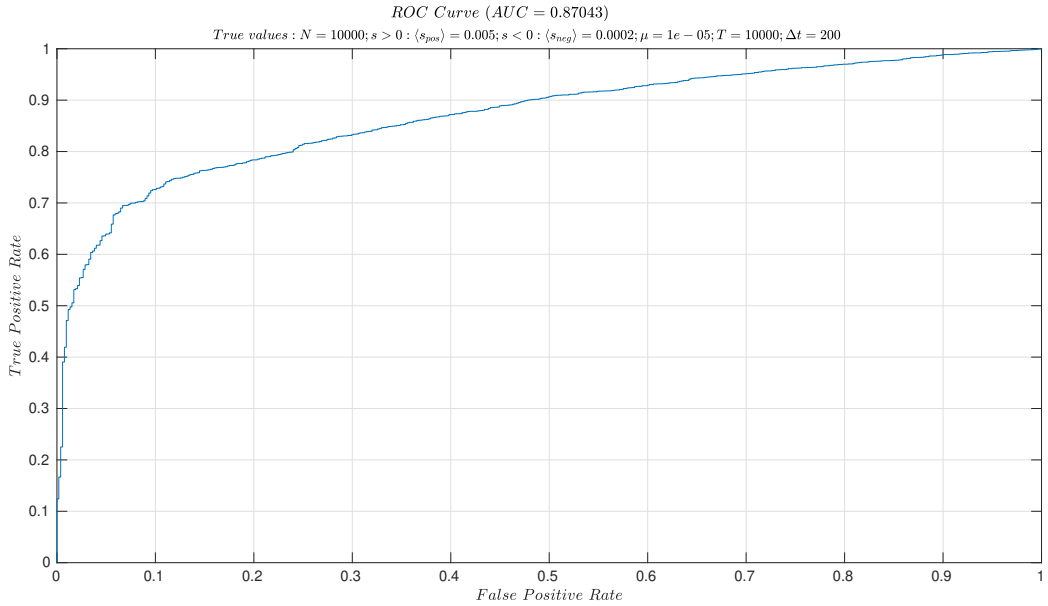

ROC Curve (AUC = 0.81856)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

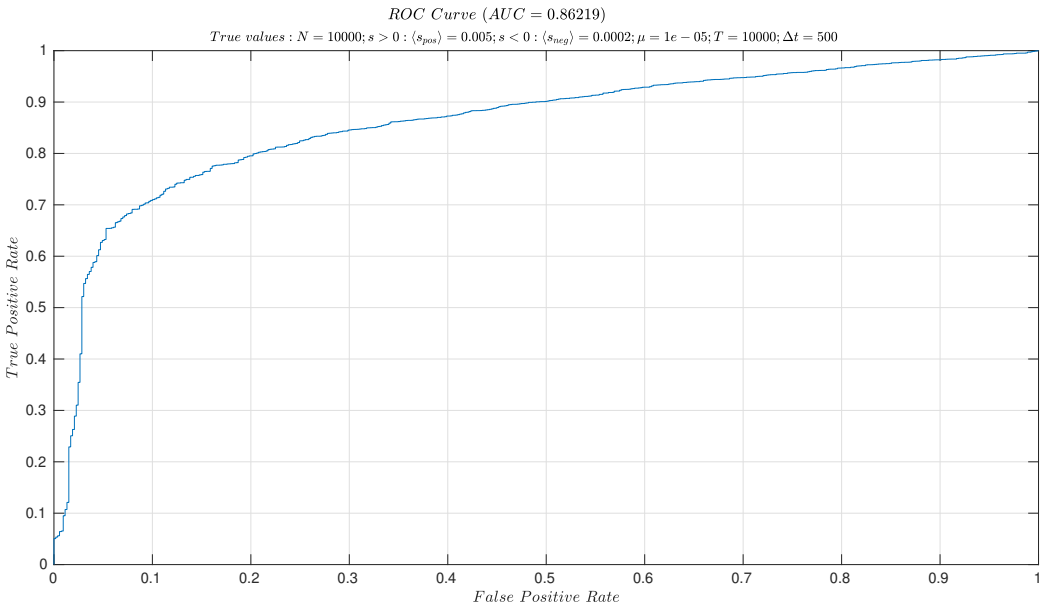

ROC Curve (AUC = 0.56339)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


ROC Curve (AUC = 0.78022)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

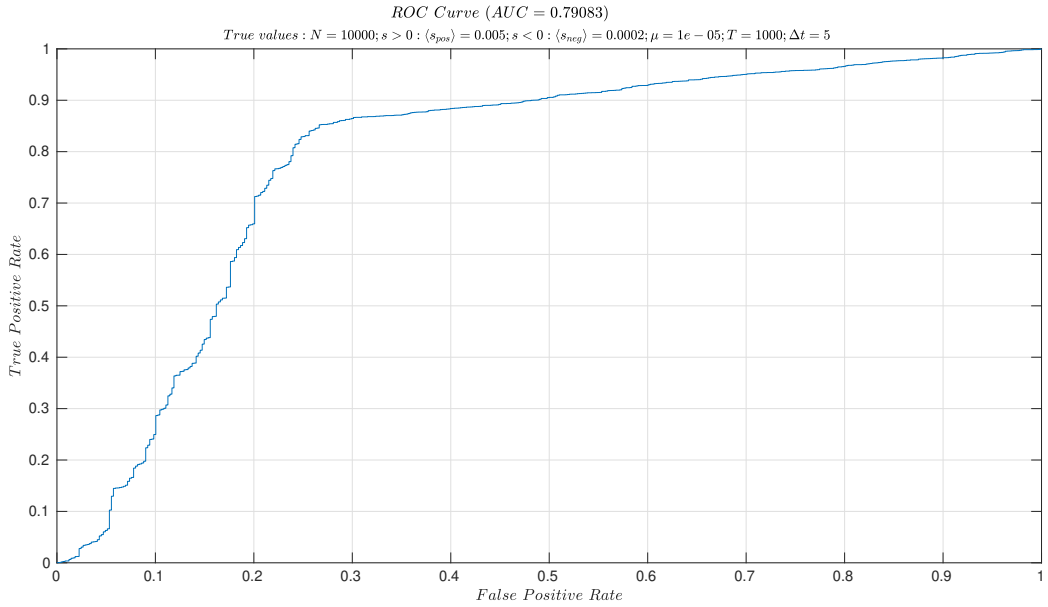

ROC Curve (AUC = 0.87788)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

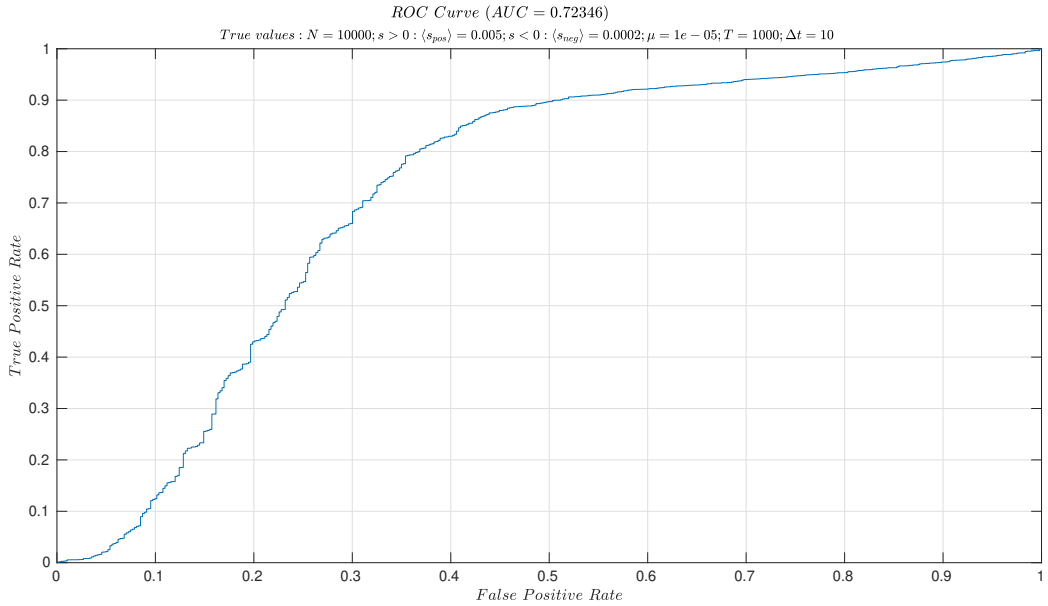

ROC Curve (AUC = 0.86891)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

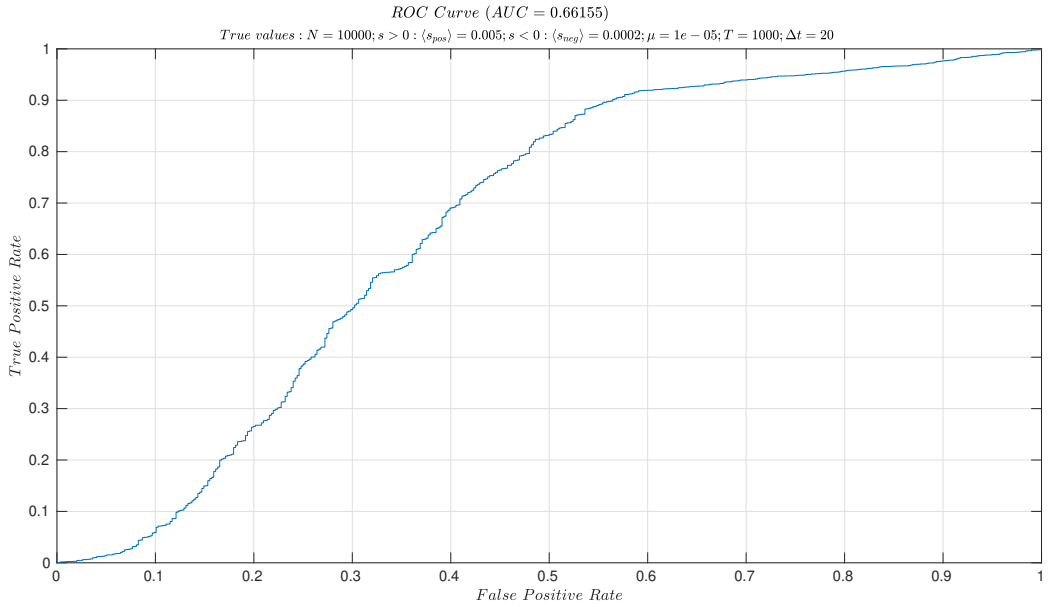

ROC Curve (AUC = 0.87043)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

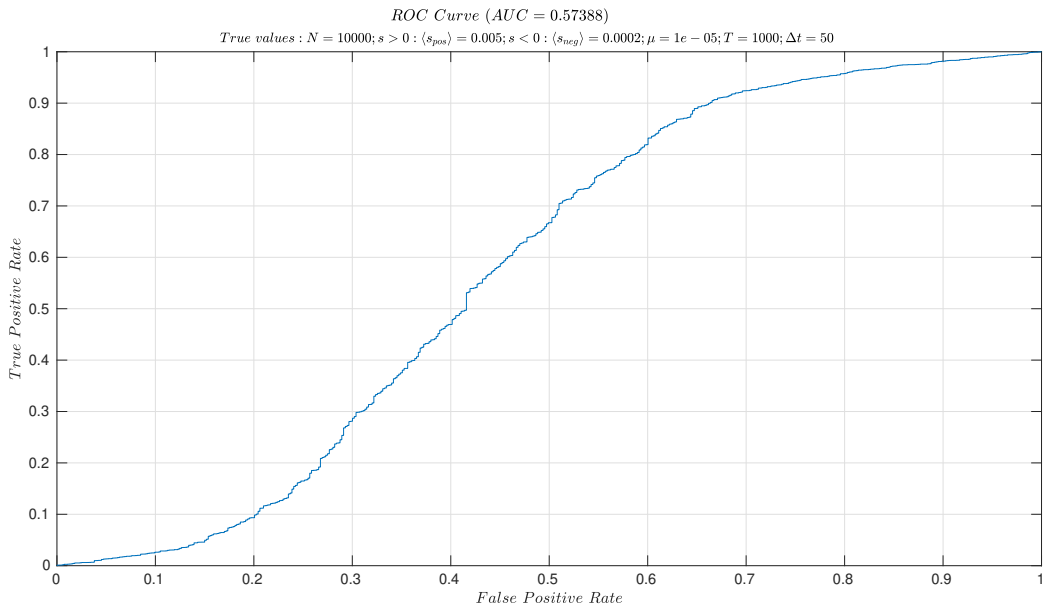


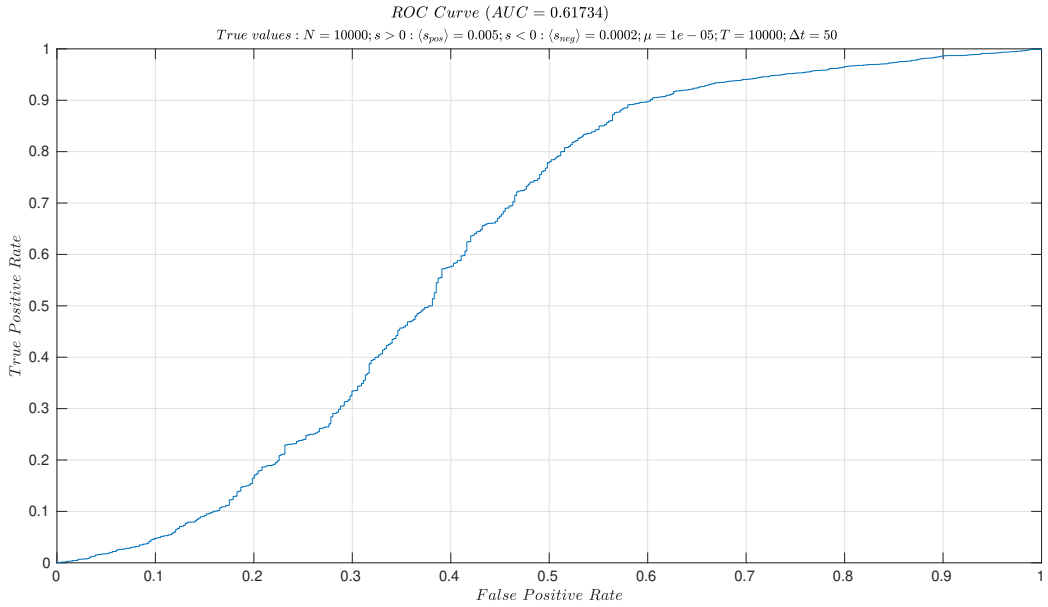
8.4 ROC plots for fixing N μ with and without sampling in genome-wide simulation

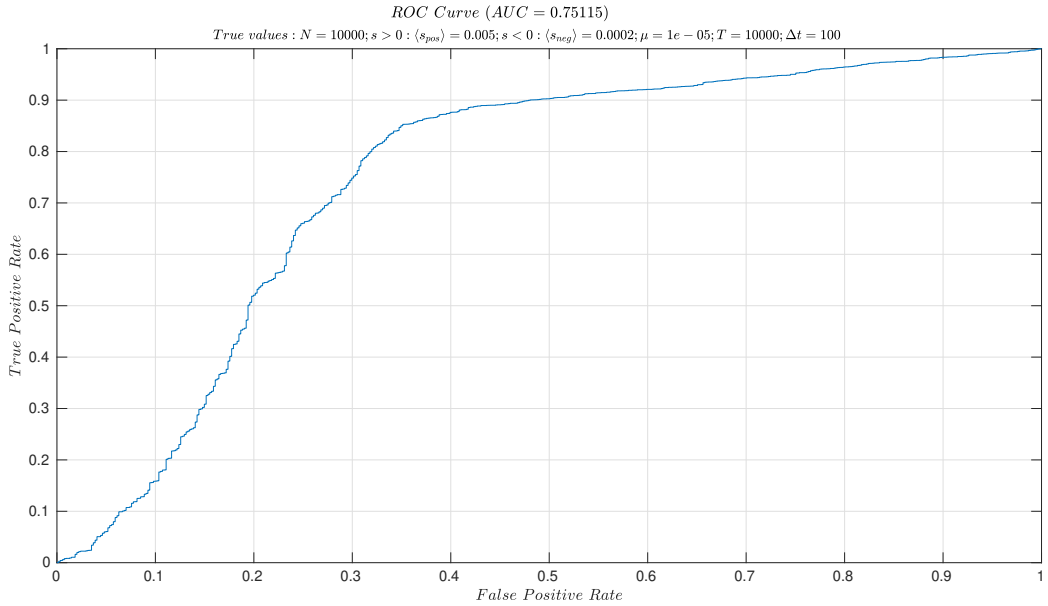

ROC Curve (AUC = 0.79083)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

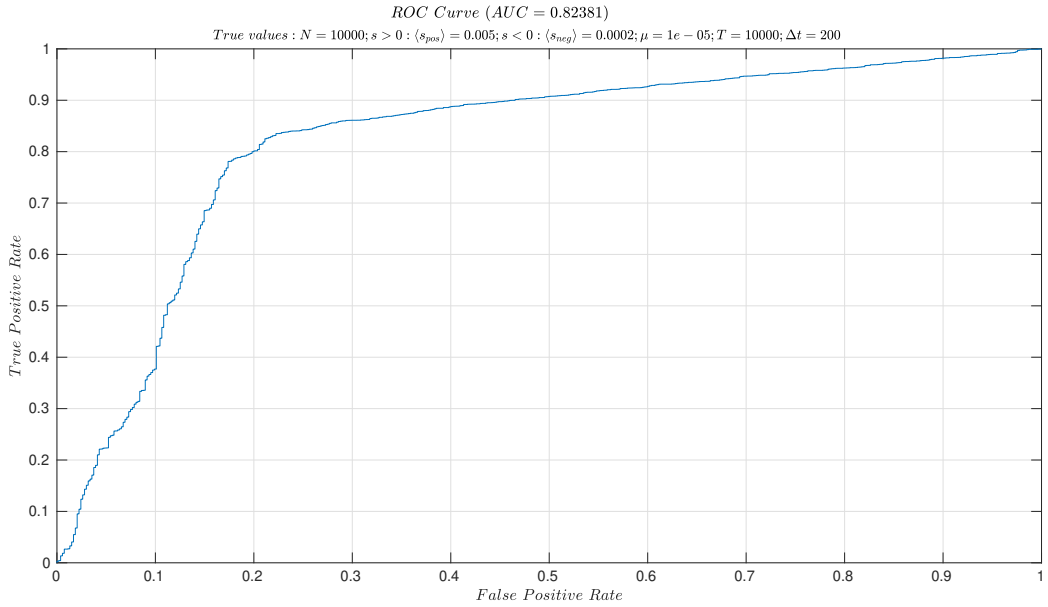

ROC Curve (AUC = 0.72346)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

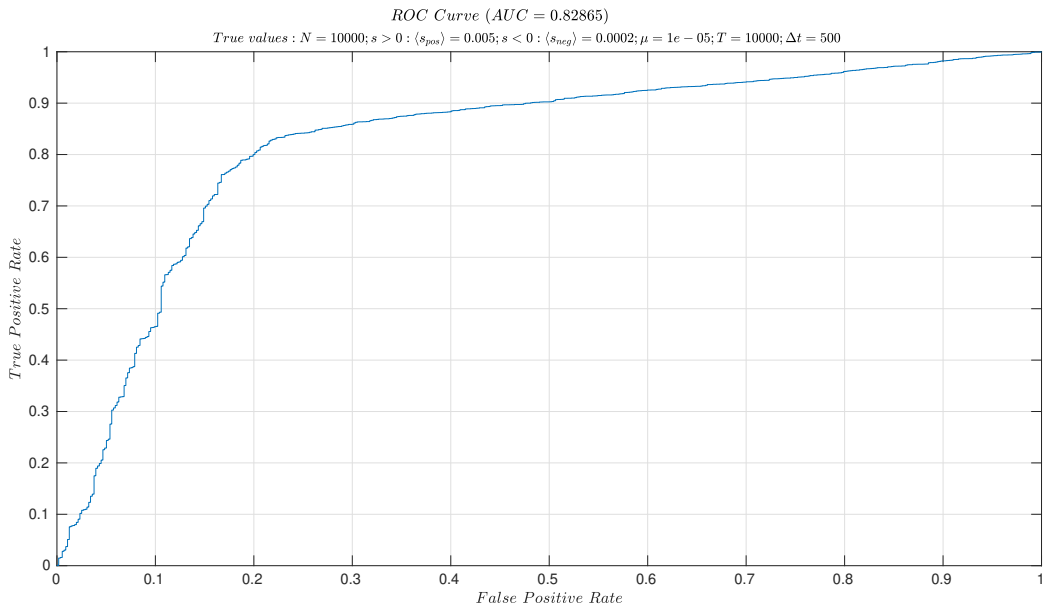

ROC Curve (AUC = 0.66155)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

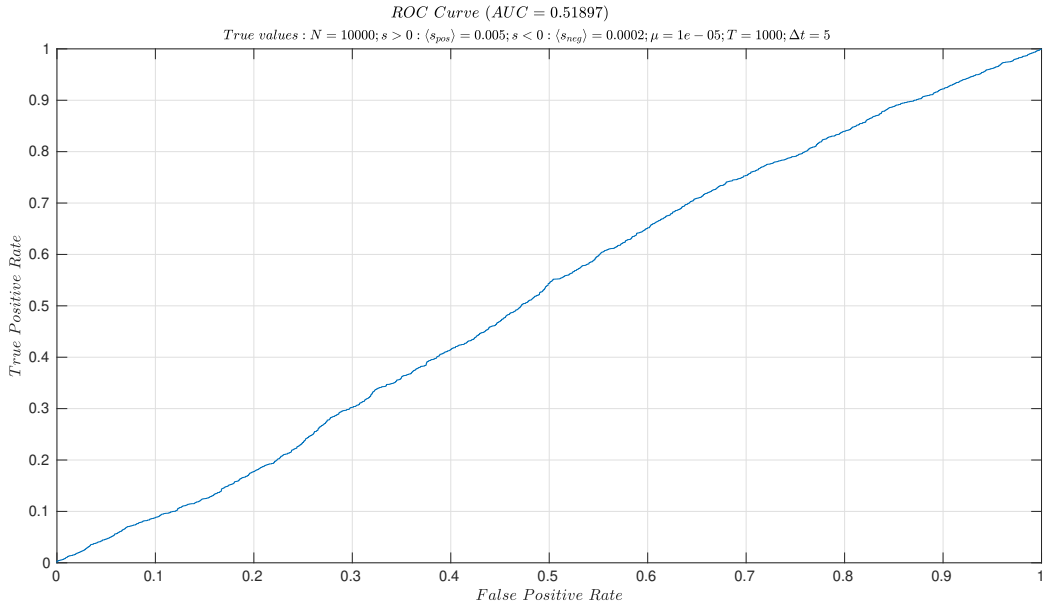

ROC Curve (AUC = 0.57388)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

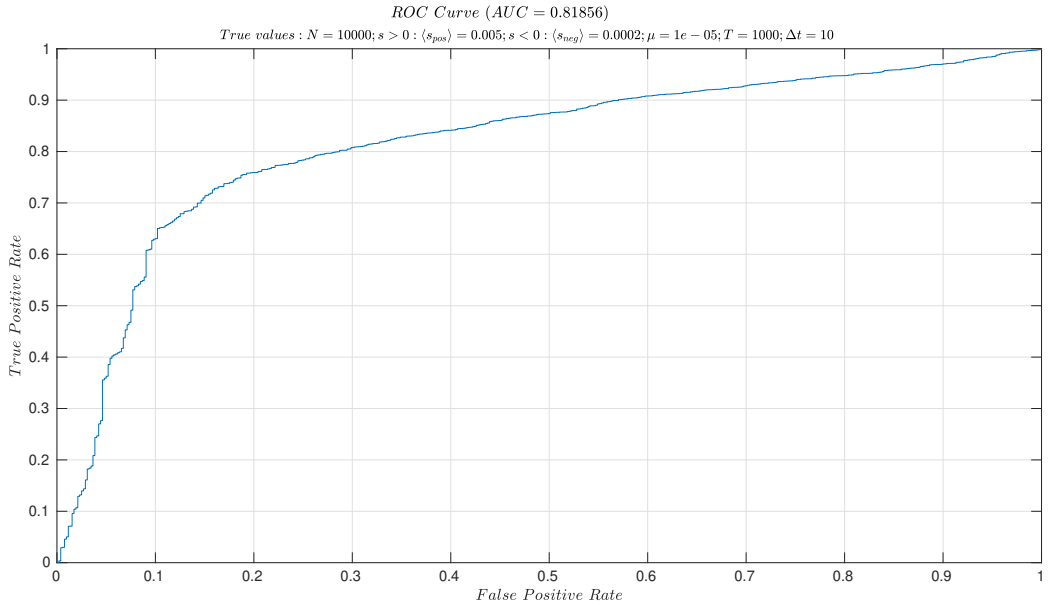

ROC Curve (AUC = 0.61734)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

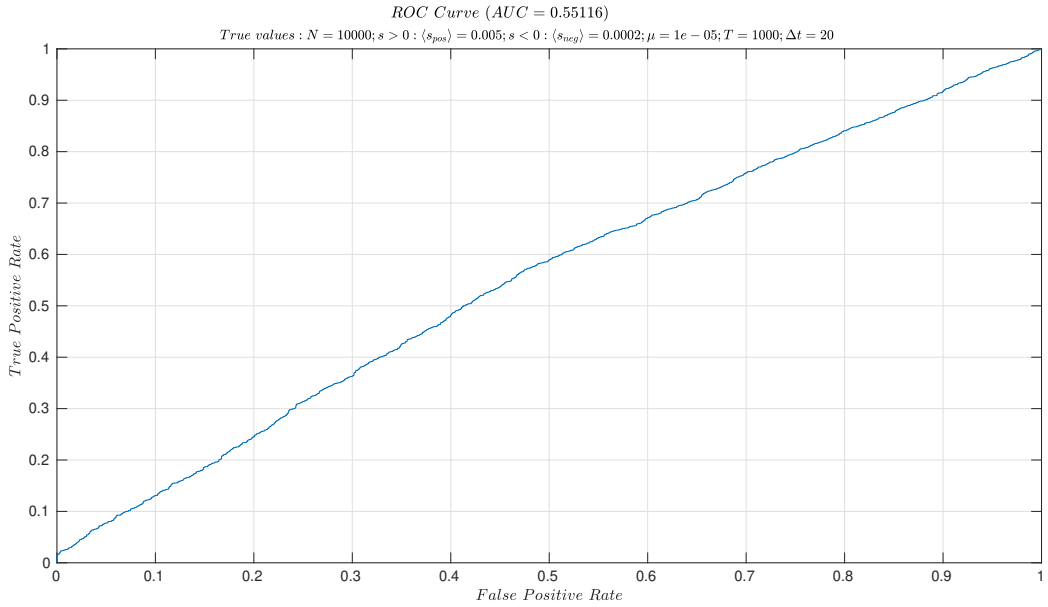

ROC Curve (AUC = 0.75115)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

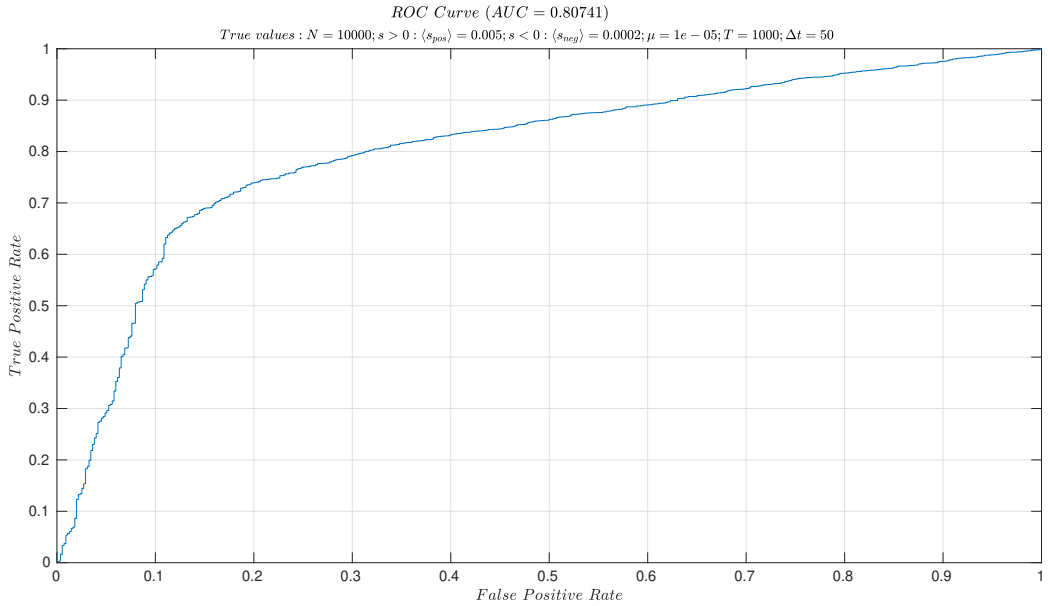

ROC Curve (AUC = 0.82381)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

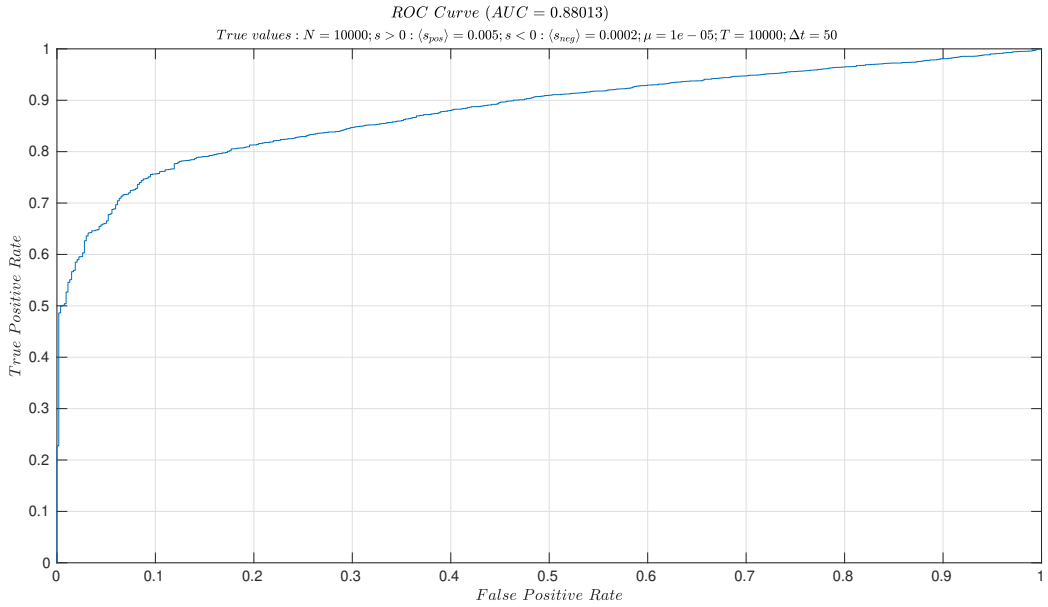

ROC Curve (AUC = 0.82865)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 500
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

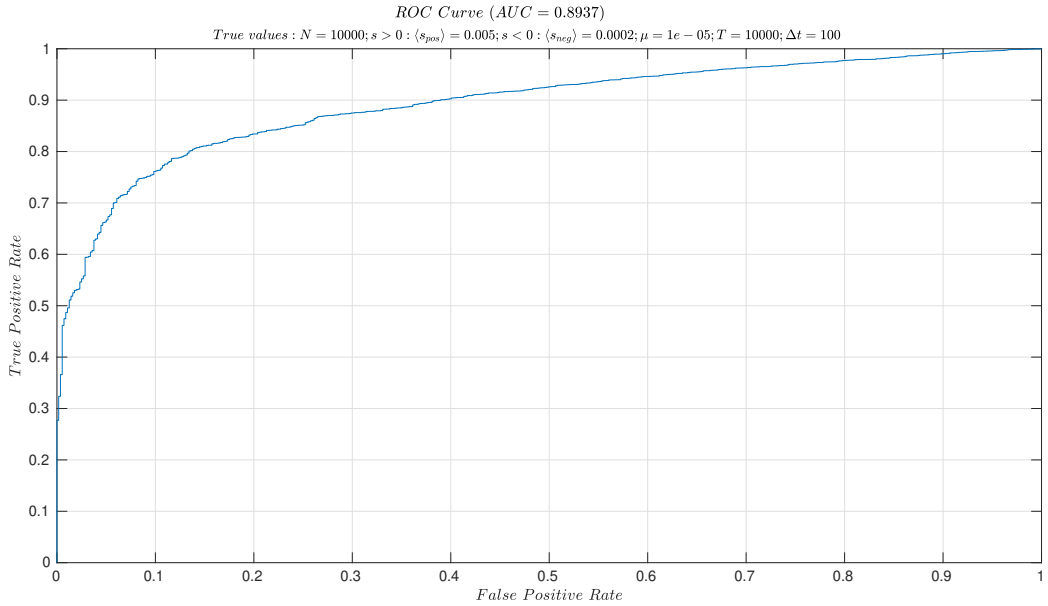

ROC Curve (AUC = 0.51897)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 5
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

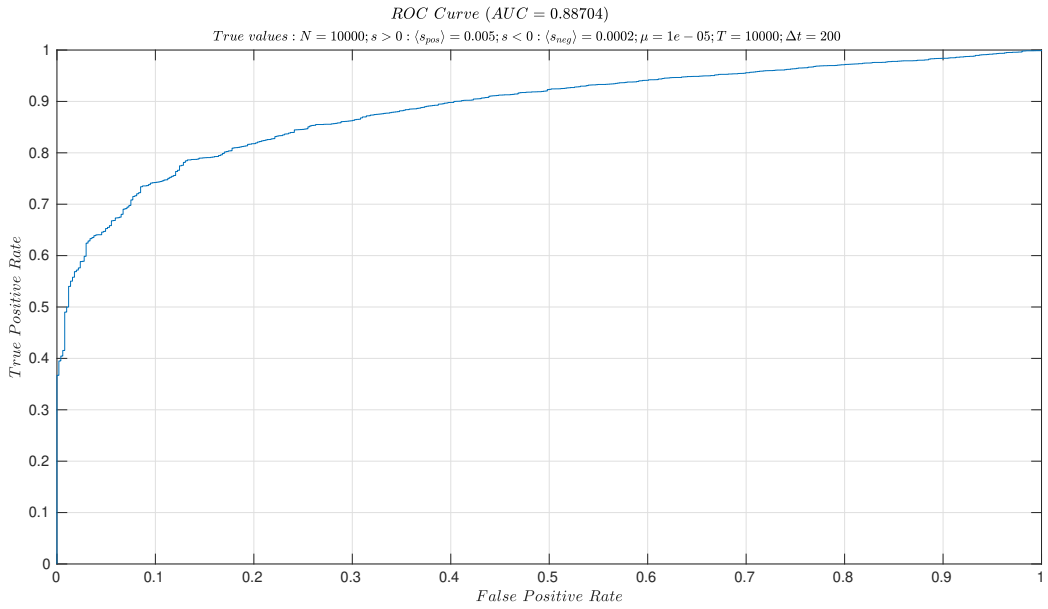

ROC Curve (AUC = 0.81856)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 10
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

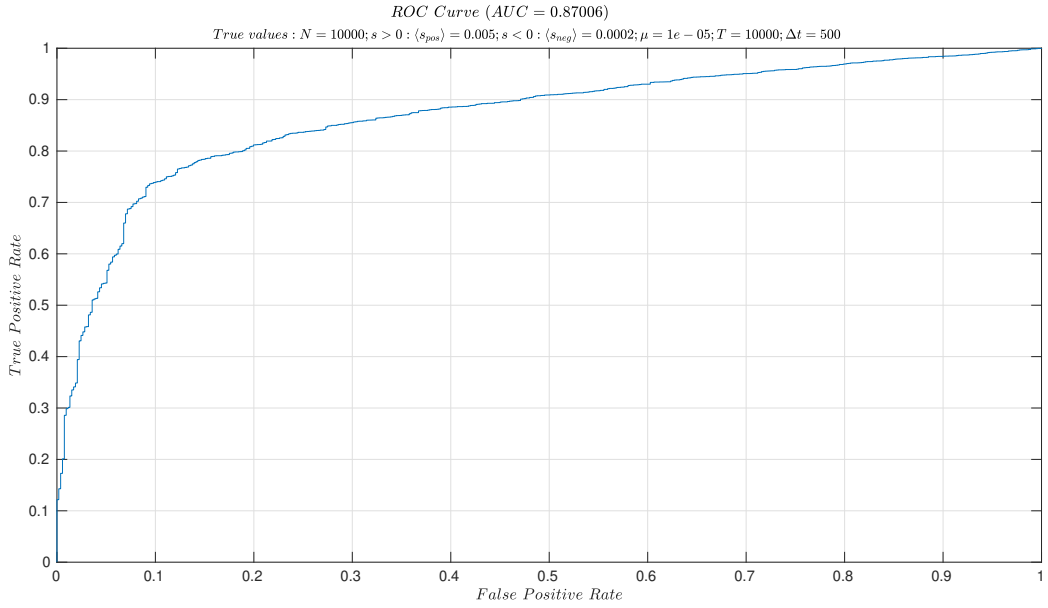

ROC Curve (AUC = 0.55116)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 20
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


ROC Curve (AUC = 0.80741)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 1000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


ROC Curve (AUC = 0.88013)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 50
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

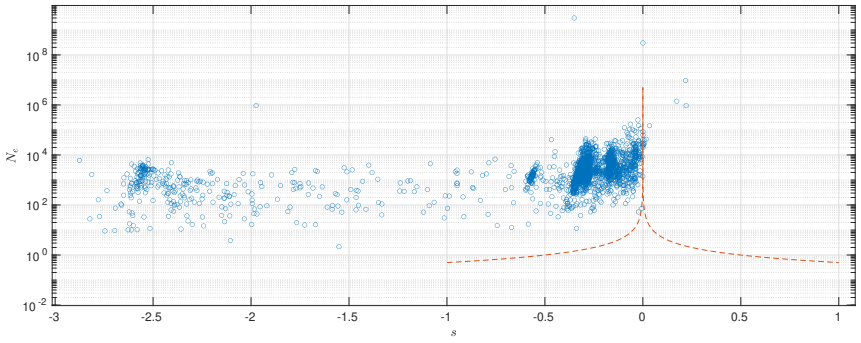

ROC Curve (AUC = 0.8937)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 100
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

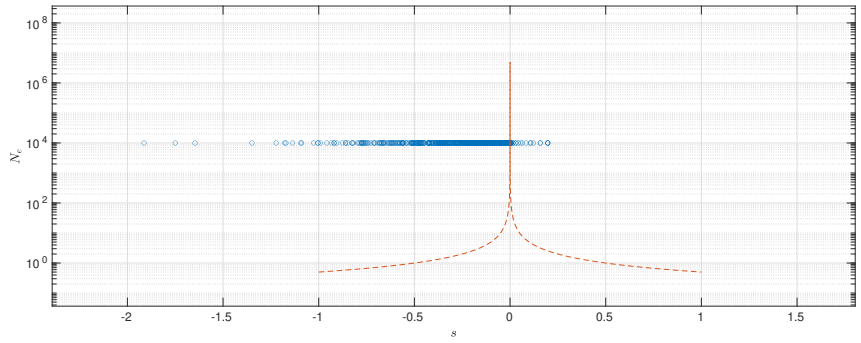

ROC Curve (AUC = 0.88704)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 200
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

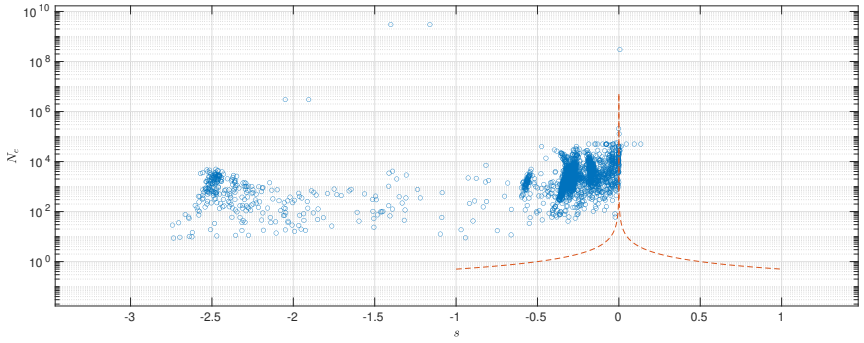

ROC Curve (AUC = 0.87006)
True values : N = 10000; s > 0 : ⟨s_pos⟩ = 0.005; s < 0 : ⟨s_neg⟩ = 0.0002; μ = 1e − 05; T = 10000; Δt = 500
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

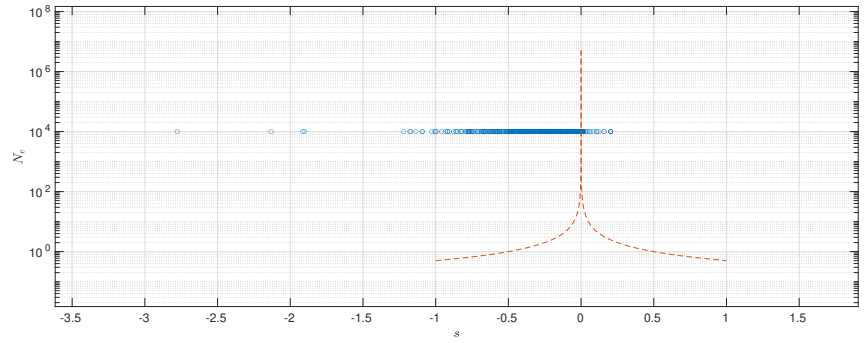
8.5 Scatter plots for N vs. μ for patient 6 and 9 under 4 parameters constrained conditions

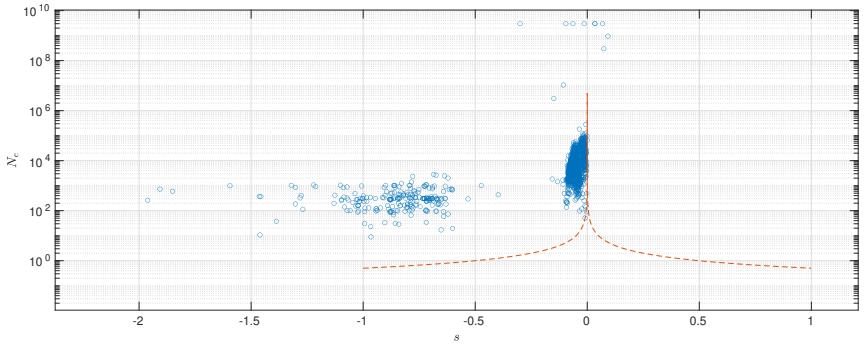
From top to bottom are:
Optimising N μ s,
Fixing N,
Fixing μ,
Fixing N μ.
First four figures are from patient 6, then patient 9

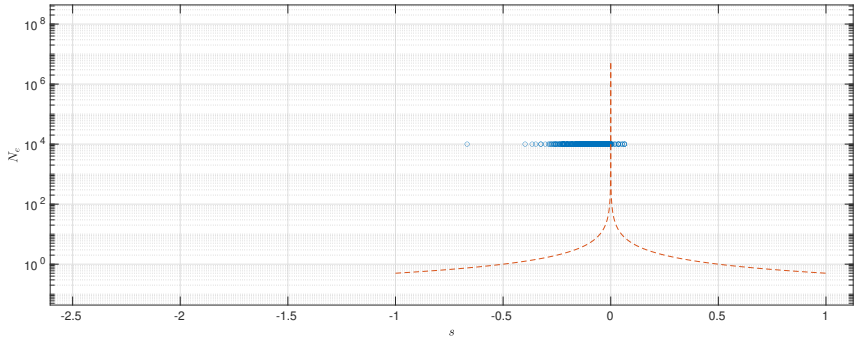


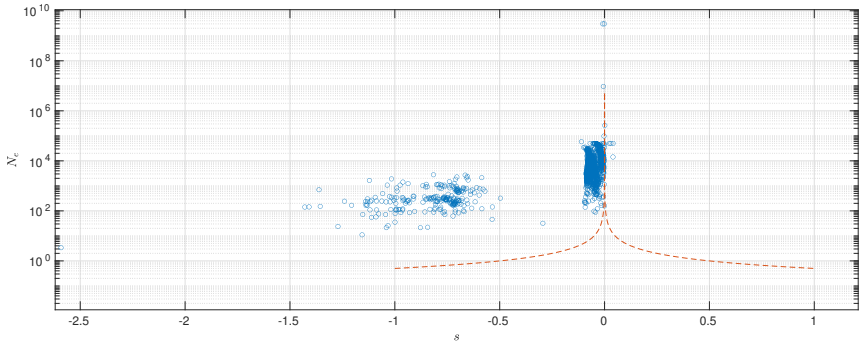


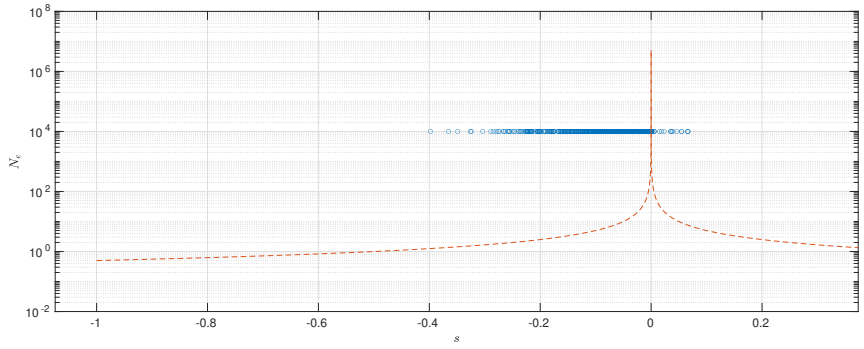

10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
10^{0}
N_e
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
s


10^{8}
10^{6}
10^{4}
10^{2}
10^{0}
N_e
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
s


10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
10^{0}
N_e
-2
-1.5
-1
-0.5
0
0.5
1
s


10^8
10^6
10^4
10^2
10^0
N_e
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
s


10^{10}
10^8
10^6
10^4
10^2
10^0
N_e
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
s


10^8
10^6
10^4
10^2
10^0
10^{-2}
N_e
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
s