Calculating forensic match probabilities for Y-STRs using a discrete Laplace distribution

Mikkel Meyer Andersen[a,*], Poul Svante Eriksen[a], Niels Morling[b]

[a] Department of Mathematical Sciences, Aalborg University, Denmark
[b] The Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark
[*] mikl@math.aau.dk

Match probability

The likelihood ratio is often formulated as

$$LR = \frac{P(E|H_p)}{P(E|H_d)}, \tag{1}$$

where

H_p: The suspect is the donor of the genetic data (prosecutor's hypothesis)

H_d: The suspect unconnected to the crime (defense attorney's hypothesis)

$P(E|H_p) = 1$ is often assumed

$P(E|H_d)$**, the 'match probability':** The probability that the suspect matches the haplotype found at the crime scene given that the suspect is unconnected to the crime (how probable is it that some random man's haplotype matches the haplotype found at the crime scene). If we knew the haplotypes of the entire population, the population frequency of the haplotype in question would be the match probability.

Estimator: A formalised way to obtain a match probability is called a match probability estimator.

Normalised allele process

Assume a **Fisher-Wright model** of evolution (constant number of N individuals and generations are discrete and non-overlapping) with a selectively neutral single-step mutation process: For individual $i = 1, 2, \dots, N$, choose the i'th individual's parent (father) at random from the previous generation (each with probability $1/N$) and inherit this parent's haplotype. For every locus at every individual in the new generation, determine if a mutation (and its direction) will happen.

Consider the normalised allele process,

$$V_g(i) := X_g(i) - X_g(N), \tag{2}$$

where $X_g(i)$ denote the allele of the i'th individual (out of N in total) in the g'th generation.

Let $Z_g(i)$ be the mutational event preceding inheritance and $q(d) := P(Z_g(1) - Z_g(2) = d)$. The distribution of the normalised allele process, quantified through the probability mass function

$$\eta_g(d) := P(V_g(i) = d), \tag{3}$$

was presented in [1] as a recurrence relation, namely

$$\eta_g = \frac{1}{N} q * \left(\sum_{i=0}^{g-2} \left[\frac{N-1}{N} \right]^i q^i \right) + \left(\frac{N-1}{N} \right)^{g-1} q^g \tag{4}$$

for $g \in \{2, 3, \dots\}$ and $\eta_1 = q$, where $*$ means the convolution and $q^i = q^{i-1} * q$ means the i'th convolution of q.

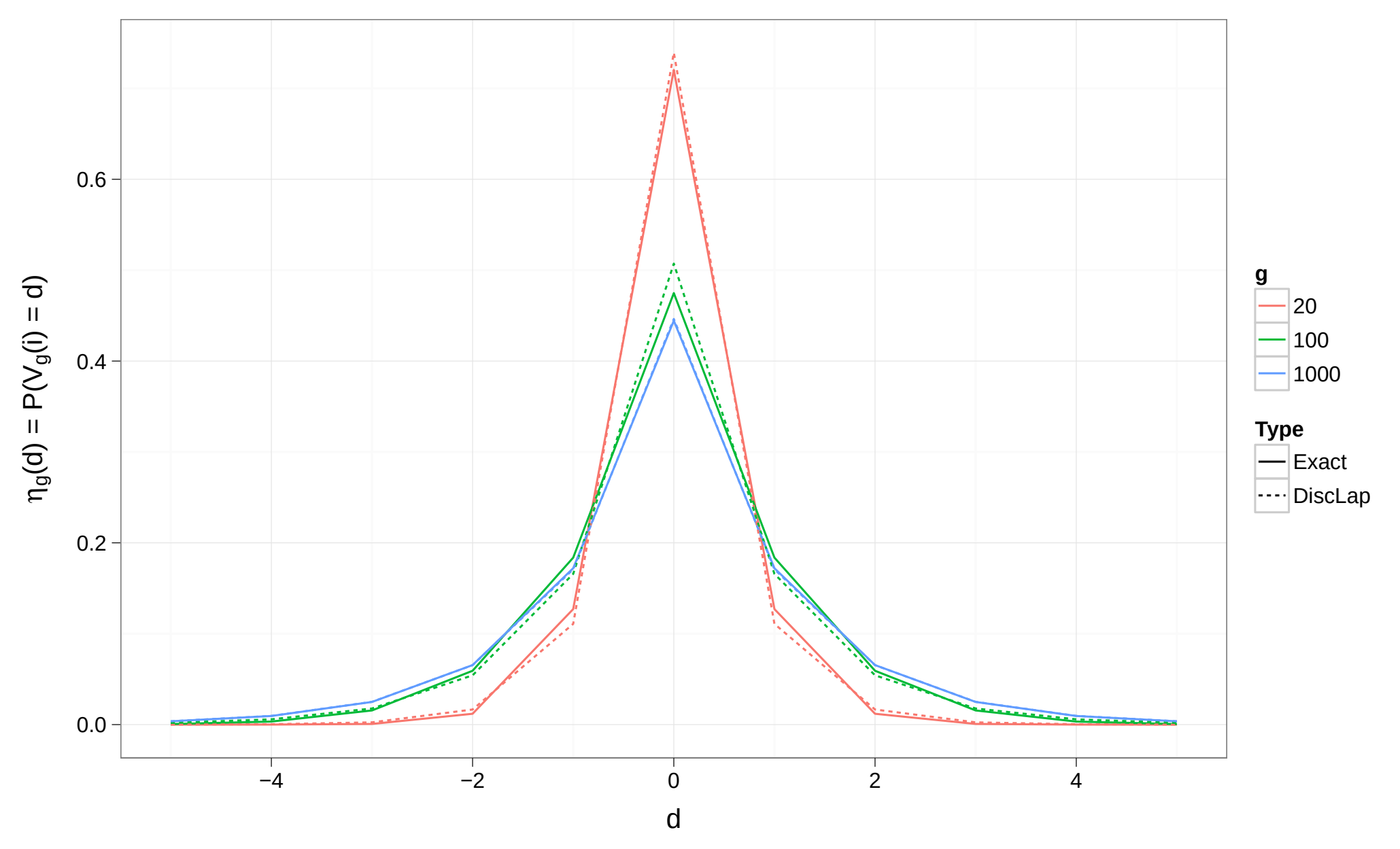
Discrete Laplace distribution

We suggest a **discrete Laplace distribution**,

$$f(d) \propto p^{|d|} \quad \Rightarrow \quad f(d) = \left(\frac{1-p}{1+p} \right) p^{|d|}, \tag{5}$$

as an approximation of $\eta_g(d)$:

For $N = 100$ individuals and a mutation rate $\mu = 0.01$



Exponential family

A reparameterisation with $\theta = \log p$ gives

$$f(d; \theta) = \exp\left(\log\left(\frac{1-e^\theta}{1+e^\theta} \right) + \theta|d| \right) = \exp\left(\theta|d| - A(\theta) \right) \tag{6}$$

with $A(\theta) = \log\left(\frac{1+e^\theta}{1-e^\theta} \right)$.

This natural exponential family is implemented in the `R` package `disclap` that also supplies it as a new generalised linear model (for example useful with the `glm` function and its cousins like the prediction function `predict`).

Statistical model for single center

Let DL(p, m) be a discrete Laplace model with dispersion parameter $0 < p < 1$ and center parameter $m \in \mathbb{Z}$ with probability mass function

$$f(d; p, m) = \left(\frac{1-p}{1+p} \right) p^{|d-m|}, \tag{7}$$

which is a non-central version of Equation (5). Hence, m can also be seen as a non-centrality parameter.

Inference of a sample $\{d_i\}_{i=1}^n$ can be made by the MLE's (maximum likelihood estimates)

$$\hat{m} = \text{median}\{d_i\}_{i=1}^n, \tag{8}$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n |d_i - \hat{m}|, \tag{9}$$

$$\hat{p} = \hat{\mu}^{-1}(\sqrt{\hat{\mu}^2 + 1} - 1). \tag{10}$$

The normalised allele process has a fixed individual as reference (local behaviour). The discrete Laplace distribution is an approximation to the distribution of the normalised allele process. We choose the reference individual non-randomly as the median of all the alleles for one-locus haplotypes. Thus, using the discrete Laplace distribution is merely a qualified guess of the global allele distribution.

Statistical model for mixtures of populations

- r loci instead of just one (mutations across loci are assumed to happen independently)
- c subpopulations centered at $y_j = (y_{j1}, y_{j2}, \dots, y_{jr})$ for $j = 1, 2, \dots, c$
- n observed haplotypes, $x_i = (x_{i1}, x_{i2}, \dots, x_{ir})$ for $i = 1, 2, \dots, n$
- $d_{ijk} := |x_{ik} - y_{jk}|$ is the distance at the k'th locus between the i'th haplotype x_i and the j'th center y_j
- z_i is the unobserved, latent variable identifying the subpopulation from which the i'th haplotype originated (such that $z_i = j$ when the i'th haplotype originated from the j'th subpopulation)
- $v_{ij} := P(z_i = j \mid x_i)$, such that $v_{i+} = \sum_{j=1}^c v_{ij} = 1$
- $\tau_j := P(z_i = j)$ is the a priori probability for originating from the j'th subpopulation yielding the constraint $\sum_j \tau_j = 1$
- τ_j can be estimated by $\hat{\tau}_j = \hat{v}_{+j}/n = \sum_{i=1}^n \hat{v}_{ij}/n$, where $\hat{v}_{ij}$ is an estimate of v_{ij}.

We only observe $\{x_i\}_{i=1}^n$. A haplotype stems from only one subpopulation, but which one is unknown: Use EM algorithm to estimate this latent variable of which subpopulations the haplotype belongs to.

- One parameter, α_j, per subpopulation j corresponding to the age of the center (how long time the center has been present in the population)
- One parameter, β_k, per locus (related to mutation rate)
- Assume additive effects using the linear predictor $\log p_{jk} = \theta_{jk} = \alpha_j + \beta_k$ for $j = 1, 2, \dots, c$ and $k = 1, 2, \dots, r$

R packages

`disclap`: Discrete Laplace exponential family for models such as a generalized linear model: `http://cran.r-project.org/package=disclap`

`disclapmix`: Inference in a mixture of Discrete Laplace distributions using the EM algorithm: `http://cran.r-project.org/package=disclapmix`

Example of estimation using `R`:

```
library(disclapmix)
data(simpop)
db <- simpop[rep(1:nrow(simpop), simpop$n), 1:7]
res <- disclapmix(db, centers = 1:5,
   use.parallel = TRUE, verbose = 0)
summary(res$best.fit)
disclap.estimates <- predict(res$best.fit,
   newdata = simpop[, 1:7])
```

Simulation study

We simulated 12 different population types by taking all possible combinations of

- Loci: $r = 7$
- Mutation rate: $\mu = 0.01; 0.003$ or 0.001
- Generations: $g = 500$ or $1{,}000$
- Initial population size: $k = 10{,}000$ or $50{,}000$.

We assumed a population growth, α, such that the expected populaition size, $\alpha^g k$, after g generations was 20,000,000. The populations were simulated using the `R` package `fwsim`. For each combination of the parameters, 5 realisations of the population were simulated. For each of these populations, 50 datasets of size 500; 1,000; and 5,000 were drawn. In total $12 \cdot 5 \cdot 3 \cdot 50$ = 9,000 datasets were sampled and used as a basis for comparison.

For all singletons (haplotypes observed only once) in the dataset, the discrete Laplace distribution approach was compared to the naive $1/n$ estimator and to Brenner's $(1-\kappa)/n$ estimator [2], where α is the number of singletons in the dataset and $\kappa = \alpha/n$ as inspired by [3].

As performance measures, the observed bias and the Kullback-Leibler divergence (the distance between two probability distributions that can interpreted as a prediction error) were calculated. If a haplotype has population frequency p and is estimated to $\hat{p}$, then the Kullback-Leibler divergence is $D_{KL}(\hat{p}; p) = \hat{p} \log\left(\frac{\hat{p}}{p}\right) + (1-\hat{p}) \log\left(\frac{1-\hat{p}}{1-p}\right)$. For a haplotype dataset $H = \{h_i\}_{i=1}^n$ with singletons $\{h_i\}_{i \in S}$ and population frequencies $\{p_i\}_{i \in S}$ estimated as $\{P_{E(H)}(h_i)\}_{i \in S}$ by an estimator E, the bias is

$$B_{H,S}(E) = \frac{1}{|S|} \sum_{i \in S} (P_{E(H)}(h_i) - p_i). \tag{11}$$

and the distribution of Kullback-Leibler divergences for singletons $\{h_i\}_{i \in S}$ is

$$D_{H,S}(E) = \{D_{KL}(P_{E(H)}(h_i); p_i)\}_{i \in S}. \tag{12}$$

References

[1] A Caliebe, A Jochens, M Krawczak, U Rösler. A Markov chain description of the stepwise mutation model: Local and global behaviour of the allele process. Journal of Theoretical Biology 266 (2010) 336-342.

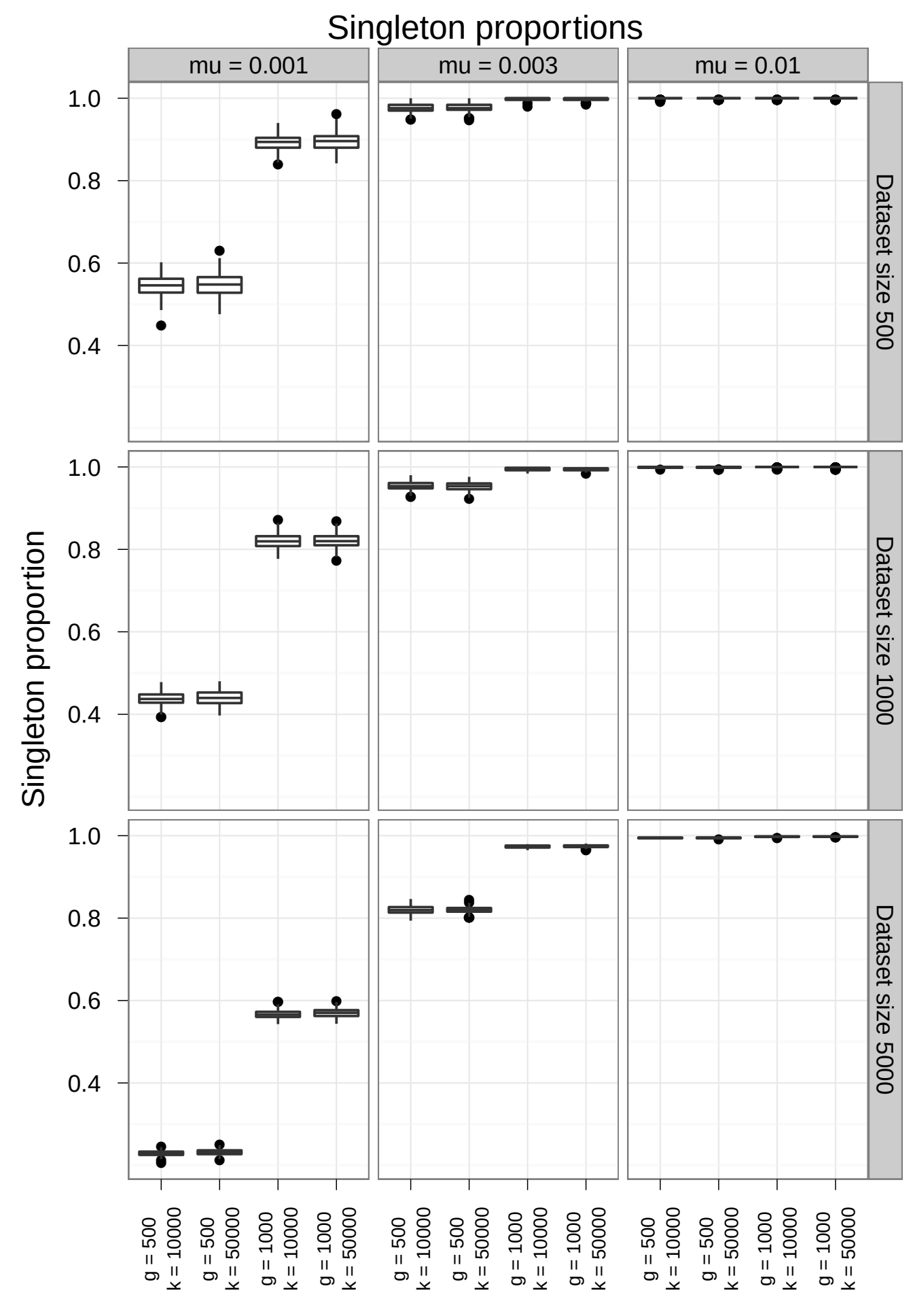
[2] C Brenner. Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. Forensic Sci. Int. Genet. 4 (2010) 281-291.

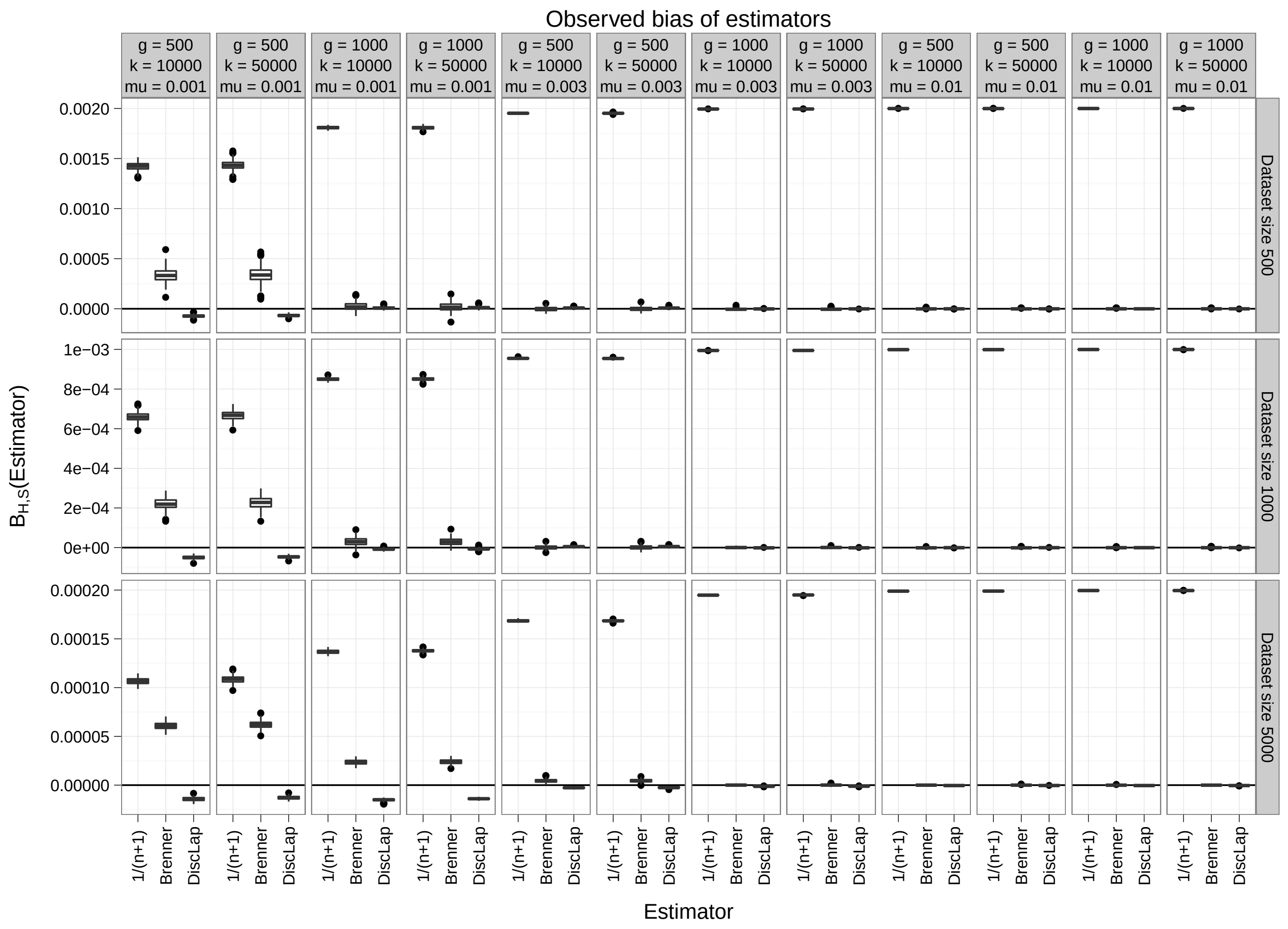
[3] H Robbins. Estimating the total probability of the unobserved outcomes of an experiment. The Annals of Mathematical Statistics 39 (1968) 256-257.

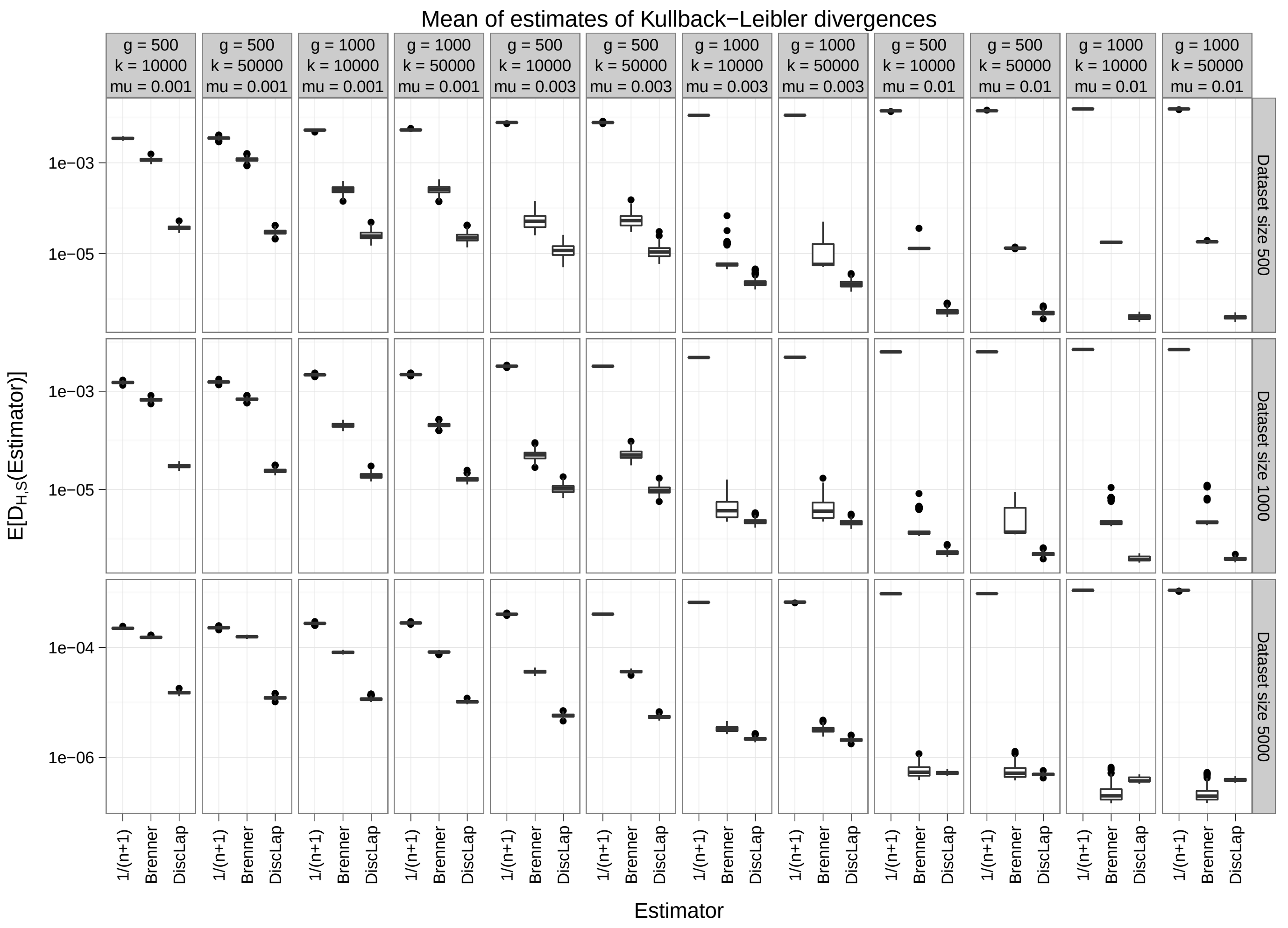
Download this poster:

`http://bit.ly/Py974f`

Results of simulation study







g is the number of generations, k is the number of individuals in the initial population, and μ is the mutation rate per locus per generation.