Boostrap Approach to Find the Dp Divergence

Given:

N = Length of the data set
D = Number of dimensions of the data set
$\boldsymbol{n_i}$ = The various Monte Carlo sample sizes
B = Number of bootstrap iterations (Number of Monte Carlo iterations at each sample size)

1) I choose the smallest and largest Monte Carlo sample size $\boldsymbol{n_i}$, and I use approximately 100 Monte Carlo sample sizes.

   Usually, the lowest sample size $\boldsymbol{n_1}$, is set equal to **D**, the dimension of the data set. The largest sample size $\boldsymbol{n_{max}}$, is set to **N/2**, half the length of the data set.

   The maximum is conservatively set to **N/2**, because this is the largest sample size for which it is possible to choose 2 completely different samples from the data during the Monte Carlo iterations.

I have found that for some data sets it is possible to use a maximum value of $\boldsymbol{n_i}$ of larger than **N/2** and maintain asymptotic convergence, but for other data sets using values much larger than **N/2** destroys the convergence.

2) I define the sample size intervals for $\boldsymbol{n_i}$. I do this somewhat heuristically. I use a power law curve to find an asymptotic $\boldsymbol{D_p}$ value. Therefore, it is necessary to have more "resolution" for small sample sizes, where the curve is quickly changing, than at larger sample sizes, where the curve almost constant (as you can see in the two plots I have).

   Therefore for small values of sample size, $\boldsymbol{n_i}$, the sample sizes $\boldsymbol{n_i}$ have smaller intervals in between them, but as sample size becomes larger, so does the interval between the samples.

   For example if my $\boldsymbol{n_{max} = 1000}$ and $\boldsymbol{n_i = 10}$, I might define the following sampling interval in Matlab:

   sample_sizes = [10:5:100 100:10:500 500:20:1000];

3) I define the number of Monte Carlo boostrap iterations, **B**.

   The only limitation to note here is that we must consider $\boldsymbol{n_{max}}$, the largest sample size. Given that **N** is the length of the data set, and $\boldsymbol{n_{max}}$ is the largest Monte Carlo sample size:

If **N** choose $\boldsymbol{n_{max}}$ is less than **B**,$\binom{\mathbf{N}}{n_{max}} < \boldsymbol{B}$, we should note that some of the Monte Carlo iterations of size $\boldsymbol{n_{max}}$ will contain the exact same sub sample of the data set.

4) I perform **B** Monte Carlo iterations at each specified sample size $\boldsymbol{n_i}$, computing the $\boldsymbol{D_p}$ value **B** times for every sample size. Then the mean of the $\boldsymbol{D_p}$ value for those B trials is computed at every $\boldsymbol{n_i}$.

5) The following power law curve:

$$\boldsymbol{L_n(k) = an^b + c}$$

is used to find the asymptotic value of $\boldsymbol{D_p}$ as a function of the sample size (given in this paper: http://www.ams.jhu.edu/~priebe/.FILES/BootstrapIntervalEstimator_forBayesOptimalError_SSP_Final-1.pdf):

This curve provides the estimate $\boldsymbol{L_n(k)}$ as a function of the sample size '**n**.' The model assumes that the exponent 'b' is negative. So as n becomes arbitrarily large, the first term drops, and the estimate $\boldsymbol{L_n(k)}$ asymptotically converges to the constant '**c**.'

I use the Matlab "fit" function to easily calculate the values of **a**, **b**, and **c** given in the expression for $\boldsymbol{L_n(k)}$ with the following syntax:

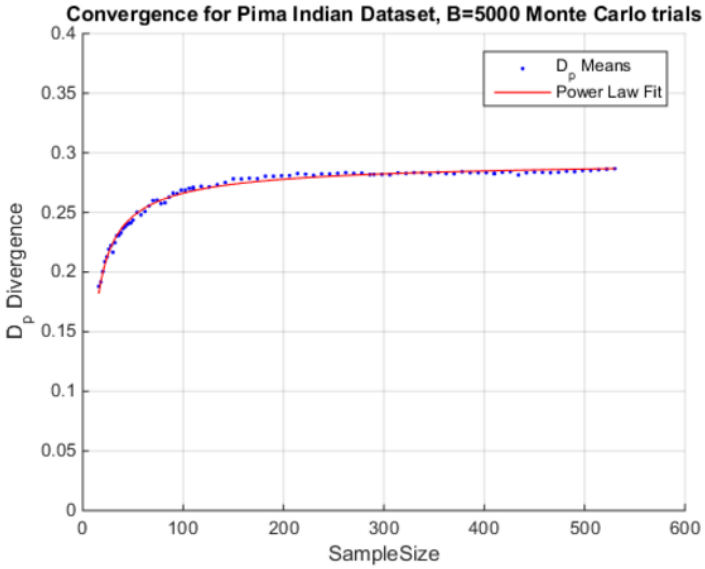
```
power_law_fit=fit(sampleSizes,Dp_divergence_means,'power2')
```

The constant '**c**' can be obtained with:

```
data1coef1=coeffvalues(power_law_fit);
asymptotic_convergence_value=data1coef1(3);
```

Results:

This is the result for the method evaluated on the Pima Indians data set:



The asymptotic value is $\boldsymbol{D_p = 0.295}$, corresponding to a lower bound on the Bayes error rate of **22.83%**, which is consistent with values I've found in the literature.

Example 2:

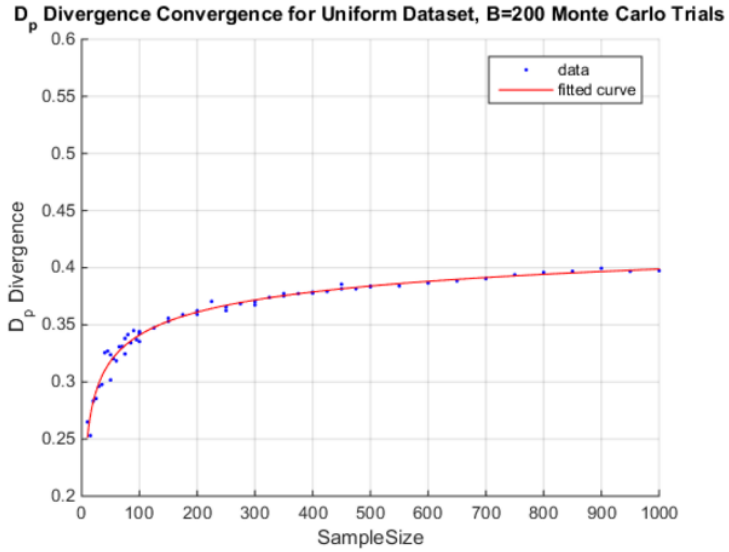
Given the following 8-dimensional uniformly distributed data set, it is possible to analytically calculate that:

$$\boldsymbol{D_p = 0.5}$$

Table 1: Uniform Dataset for Bootstrap Analysis of D_p

D_0								
μ_0	0	0	0	0	0	0	0	0
σ_0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$
D_1								
μ_1	$\frac{1}{2}$	0	0	0	0	0	0	0
σ_1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$

Here is the result of the simulation:



The asymptotic value I obtained from the power law fit was as follows:

c = 0.5044, (0.4478, 0.5611) 95% confidence interval

Which is consistent with the the analytical calculation of $\boldsymbol{D_p = 0.5}$