Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

Overlap Coefficients Based on Kullback-Leibler of Two Normal Densities: Equal Means Case

Hamza Dhaker[a,*], Papa Ngom[b], Boubakari Ibrahimou[c], Malick Mbodj[d]

[a] *Dpartement de mathmatiques et statistique,Université de Moncton, NB, Canada*
[b] *LMA,Université Cheikh Anta Diop, Dakar, Senegal*
[c] *Florida International University, Miami, FL, USA*
[d] *Bowie State University, Bowie, MD, USA*

Abstract

Overlap coefficient (OVL) represents the proportion of overlap between two probability distributions, as a measure of the similarity between them. In this paper, we define a new overlap coefficient Λ based on KullbackLeibler divergence and compare its performance to three known overlap coefficients, namely Matusia's Measure ρ, Morisita's Measure λ, Weitzman's Measure δ. We study their properties, relations between them, and give approximate expressions for the biases and the variances.

Keywords: KullbackLeibler; Overlap Coefficients; Normal Density.

1. Introduction

Measures of similarity are useful in quantifying the extent of similarity between two populations. Their complements, known as measures of dissimilarity are also commonly used by researchers. These measures are often used to make comparative inferences about two groups, such as describing the degree of inter-specific encounter or crowdedness of two species in their resources utilization, estimating the proportion of genetics deviates in segregating populations. One such similarity measure, known as the overlap coefficient (OVL) has been used for comparing fits of statistical distributions by different methods. However, due to the unknown nature of sampling distributions of these measures, decisions are often made using only point estimates.

In the literature, overlap coefficients are mostly used in ecology. Other applications include the lowest bound for the probability of failure in the stress-strength models of reliability analysis (Ichikawa [4]), an estimate of the proportion of genetic deviates in segregating populations (Federer et al. [1]), and a measures of disjunction (Sneath [11]). For more details and applications of OVL coefficients including application on income differentials, please see Mulekar and Mishra [[9, 10]], Inman and Bradley [5] and Gastwirth [3].

Let X be a random variable defined on the real line for two different populations and $f_1(x)$ and $f_2(x)$ their respective probability density functions. The overlapping coefficients are the common areas under the two functions, defined as follows:

- Matusia's Measure [7]

$$\rho = \int \sqrt{f_1(x) f_2(x)} dx$$

- Morisita's Measure [8]

$$\lambda = \frac{2 f_1(x) f_2(x)}{\int [f_1(x)]^2 dx + \int [f_2(x)]^2 dx}$$

*Corresponding author
Email address: hamza.dhaker@umoncton.ca (Hamza Dhaker)

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

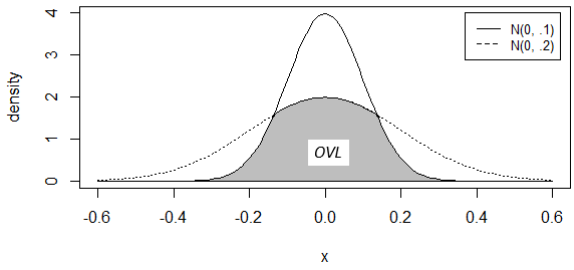

Figure 1: The overlap of two normal densities.

- Weitzman's Measure [13]

$$\Delta = \int \min\{f_1(x), f_2(x)\}dx$$

- Kullback-Leibler [6]

$$\Lambda = \frac{1}{1 + KL(f_1 \| f_2)} \tag{1}$$

with $KL(f_1\|f_2) = \int (f_1 - f_2) \log\left(\frac{f_1}{f_2}\right) dx$

Our goal in this paper is to compare the Kullback-Leibler Measure Λ's performance to Matusia's Measure ρ, Morisita's Measure λ, and Weitzman's Measure δ. We study their properties, their relations, in addition to approximating expressions for their biases and variances. The paper is organized as follow. In Section 2, we derive the expressions of the measures described above and study their properties along the lines of Mulekar and Mishra [9]. In Section 3, we provide their maximum likelihood estimators along with approximate variances and covariances. In Section 4, a simulation study is perform to evaluate and compare biases and mean square errors of OVL measures estimates. In Section 5 we give an example using a real dataset. Finally, the conclusion is presented in Section 6.

2. Properties of Different Overlap Measures

Let $f_1(x)$ and $f_2(x)$ represent the two populations normal densities with common expectation parameter μ and variances σ_i^2 $(i = 1, 2)$ respectively. We define $C = \sigma_1/\sigma_2$ $(C \geq 0)$ as the ratio of standard deviations. Under the equal means condition, the four similarity measures of interest are given by:

$$\rho = \sqrt{\frac{2C}{1 + C^2}} \tag{2}$$

$$\lambda = \frac{2\sqrt{2}}{\sqrt{1 + C^2}} \left(\frac{C}{1 + C}\right) \tag{3}$$

$$\Delta = \begin{cases} 1 - 2\Phi(b) + 2\Phi(Cb) & if \quad 0 < C < 1 \\ 1 + 2\Phi(b) - 2\Phi(Cb) & if \quad C \geq 1 \end{cases} \tag{4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

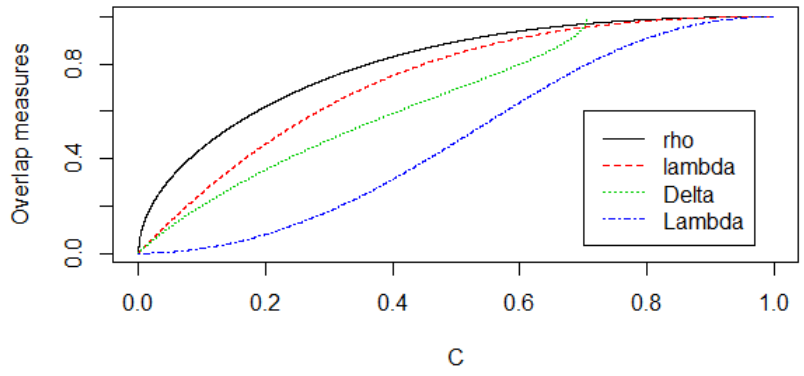

Figure 2: Measures of similarity as functions of C.

$$\Lambda = \frac{2C^2}{C^4 + 1} \tag{5}$$

where $b = \sqrt{-\ln C^2/(1 - C^2)}$ and $\Phi(.)$, the cumulative distribution function of a standard normal deviate. Figure 1 shows overlap of $N(0, 0.1)$ and $N(0, 0.2)$.

It is important to note that all these similarity measures are independent of the population mean μ and depend only on the population variances. In other words, all these similarity measures are completely determined by the ratio of the population variances. The numerical value of similarity measures is zero if and only if $C = 0$ or $C = \infty$. Also, these measures take a numerical value of 1 if and only if $C = 1$. All four measures possess properties of reciprocity, invariance, and piecewise monotonicity.

- Symmetry with respect to C: All the measures are "symmetric" in C around 1, i.e., $\rho(C) = \rho(1/C)$, $\lambda(C) = \lambda(1/C)$, $\Delta(C) = \Delta(1/C)$, and $\Lambda(C) = \Lambda(1/C)$. As in the case of the exponential, this "symmetry" is not equivalent to the mirror-image symmetry. It is a multiplicative equivalent of the "even functions," where typically a function is considered an even function in the additive sense.

- Monotonicity property: All these measures are monotonically increasing with respect to C for $0 \leq C \leq 1$, and monotonically decreasing in C for $C > 1$.

For ρ, λ, Δ and Λ defined in equations 2-5, we have

(i) $\lambda \leq \rho$ and $\Delta \leq \rho$.

(ii) $\rho \leq \Lambda$

For all $C > 0$, equality holding if $C = 1$.

Proof. (i) the proof of the first part follows from Lemma 2 of Mulekar and Mishra [9].
(ii) For the second part, we define $h(C)$ by

$$h(C) = \Lambda/\rho = \frac{\sqrt{2}\sqrt{C^3 + C^5}}{1 + C^4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

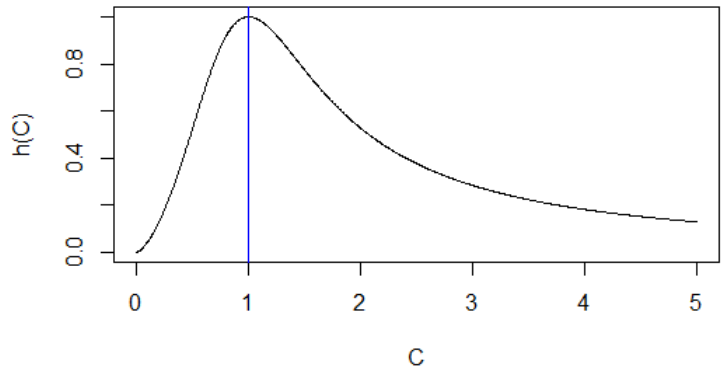

Figure 3: $h(C) = \Lambda/\rho$ as functions of C.

and the derivative with respect to C

$$h^{'}(C) = \frac{3C(1-C)(1+C)(C^2+\frac{4-\sqrt{7}}{3})(C^2+\frac{4+\sqrt{7}}{3})}{\sqrt{2}\sqrt{C+C^3}(1+C^4)^2}$$

for $C > 0$ $h^{'}(C) = 0 \Longrightarrow C = 1$. Then $h(C)$ is an increasing function of C for $0 < C < 1$ and a decreasing function of C for $C > 1$ (See Figure 3). Also the sup $h(C) = 1$ is attained at $C = 1$. Thus, $0 < h(C) < 1$ for all $(0 < C < 1)$, which gives the desired result.

3. Estimation of OVL Measures

As in Folks and Chhikara [2], parallel results to those of the two inverse Gaussian populations can be established for the normal populations with common mean. This is based on the following results (Lemma 2 below) from Mulekar and Mishra [9].
Suppose $(X_{ij}; j = 1, \cdots, n_i; i = 1, 2)$ denote independent observations from two independent normal populations.
Let $f_i(x)$ denote the Gaussian density with expectation μ and standard deviation σ_i, $i = 1, 2$. Let $C = \sigma_1/\sigma_2$ be the ratio of standard deviations as above. An unbiased estimate of σ_i^2 is given by

$$\sigma_i^2 = S_i^2 = \frac{1}{n-1}\sum_{i}^{n_i}(X_i - \overline{{}_iX})$$

Lemma 1.

$$\mathbb{E}(\widehat{C^2}) = \gamma_1 C^2 \quad Var(\widehat{C^2}) = \gamma_1 C^4$$

where γ_1 and γ_2 are constants in $\mathbb{R}^+$, then γ_1 and γ_2 can be determined as functions of n_1 and n_2 only as,

$$\gamma_1 = \frac{n_2 - 1}{n_2 - 3}, \qquad \gamma_2 = \frac{(n_2-1)^2(n_1+1)}{(n_1-1)(n_2-3)(n_2-5)} - \gamma_1^2$$

provided $n_1 > 1$, and $n_2 > 5$.

Proof. For the Proof, please see Mulekar and Mishra [9]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

OVL	$Bias(\widehat{OVL})$	$Var(\widehat{OVL})$
ρ	$\frac{(\gamma_1-1)\rho}{4}\frac{(1-C^2)}{1+C^2}$	$H_{\widehat{\rho}}$
λ	$\frac{\lambda(\gamma_1-1)}{2}\left[\frac{(1-C)(1+C+C^2)}{(1+C)(1+C^2)}\right]$	$H_{\widehat{\lambda}}$
Δ	$(\gamma_1-1)\left\{(C\phi(Cb)-\phi(b))\left(\frac{C^2b^2-1}{b(1-C^2)}\right)+Cb\phi(Cb)\right\}I_C$	$H_{\widehat{\Delta}}$
Λ	$(\gamma_1-1)\Lambda\frac{1-C^4}{1+C^4}$	$H_{\widehat{\Lambda}}$

Theorem 1. *Let $\widehat{\rho}$, $\widehat{\lambda}$, $\widehat{\Delta}$ and $\widehat{\Lambda}$ be the estimates of ρ, λ, Δ and Λ respectively, by substituting $\widehat{C^2}$ for C^2, then for $n_1 > 1$ and $n_2 > 5$ we have the approximate expressions for bias and variance of $\widehat{\rho}$, $\widehat{\lambda}$, $\widehat{\Delta}$ and $\widehat{\Lambda}$ given above.*

where $H_{\widehat{OVL}} = \gamma_2(\gamma_1-1)Bias^2(\widehat{OVL})$, $\phi(.)$ is the density function of standard normal variate and

$$I_C = \begin{cases} 1 & if \quad 0 < C < 1 \\ \\ -1 & if \quad C \geq 1 \end{cases}$$

Proof. Let $g(\theta)$ a one parameter function of θ and let $\widehat{\theta}$ be an almost sure consistent estimate of θ. Then the mean and variance of $g(\widehat{\theta})$ may be obtained from the linear Taylor approximation around θ.

$$g(\widehat{\theta}) = g(\theta) + (\widehat{\theta}-\theta)g^{'}(\theta) \tag{6}$$

For example, letting $\theta = C^2$, the estimator of $\widehat{\Lambda}$:

$$\widehat{\Lambda} = g(\widehat{\theta}), \quad g(\theta) = \frac{2\theta}{1+\theta^2}.$$

Since, in this case,

$$g^{'}(\theta) = \frac{2(1-\theta^2)}{(1+\theta^2)^2}$$

from (6)

$$\begin{aligned} Bias(\widehat{\Lambda}) = \mathbb{E}(\widehat{\Lambda}) - \Lambda = \mathbb{E}(\widehat{\theta}-\theta)g^{'}(\theta) &= (\gamma_1-1)\Lambda\frac{1-\theta^2}{1+\theta^2} \\ &= (\gamma_1-1)\Lambda\frac{1-C}{1+C} \end{aligned} \tag{7}$$

Similar arguments can be used for the other overlap coefficients.

The MLEs for the two-parameter Normal distribution are asymptotically efficient and they are asymptotically normally distributed (see, Mulekar and Mishra [9]). However, the OVL measures are functions of the Normal distribution parameters. Therefore, by using the Delta-method, the OVL measures estimators are asymptotically normally distributed. Thus, the $100(1-\alpha)\%$ approximate confidence intervals are given by

$$\widehat{OVL} \pm Z_{1-\alpha/2}\sqrt{Var(\widehat{OVL})}$$

where $Z_{1-\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution.
For large samples these confidence intervals work fairly well. However, for small sample sizes more refined versions of the above confidence intervals can be obtained by

$$\left\{\widehat{OVL} - Bias(\widehat{OVL}) - Z_{1-\alpha/2}\sqrt{Var(\widehat{OVL})}, \widehat{OVL} - Bias(\widehat{OVL}) + Z_{1-\alpha/2}\sqrt{Var(\widehat{OVL})}\right\}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

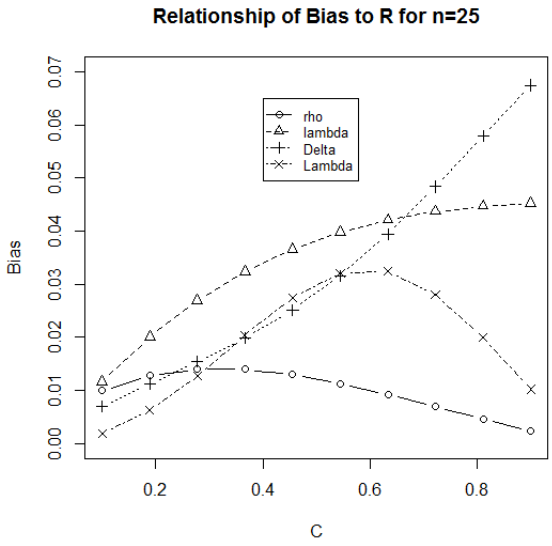

Figure 4: The bias estimates of overlap coefficients by C.

4. Simulation Study

We performed a numerical study to examine the behavior of overlap coefficients and for comparing the approximate formula for biases and mean square errors derived in the previous section for different OVL measures. Samples of sizes $n = 10, 25, 50, 100, 200$, and 500 each were generated. From each pair of generated samples, the similarity measures ρ, λ, Δ and Λ were estimated and the amount of biases and the standard deviations of the estimates were determined. The mean squared error (MSE) and bias values for $C = 0.05, 0.25, 0.5, 0.75, 0.95$ are reported in Table 1. Tables 1 indicates that the bias of proposed OVL estimators is negligible and decreases as the sample size n increases. As expected, both bias and MSE decrease steadily as the sample size increases.

The bias estimates for $n = 25$ are plotted in Figure 3. Only one plot of bias values is presented because a similar pattern is observed for other sample sizes. For $C < 0.6$ the bias estimates of the measures $\widehat{\lambda}$, $\widehat{\Delta}$ and $\widehat{\Lambda}$ behave more similarly, but for the bias of $\widehat{\rho}$ shows a different pattern. For $C > 0.6$, the bias estimates of the measures $\widehat{\lambda}$ and $\widehat{\Delta}$ are still growing, but for that of $\widehat{\rho}$ and $\widehat{\Lambda}$ are decreasing and tends towards 0.

The standard deviation estimates for the overlap coefficients with sample size 25 are plotted in Figure 4. Again, only one figure for standard deviations is presented because similar pattern is observed for the other sample sizes. The standard deviation estimates for all four coefficients show the same behavior as the bias estimates of the measures.

The estimate of MSE are plotted in Figure 5 for all four overlap coefficients. For $C < 0.3$, the MSE estimates for the overlap coefficients have almost the same values. the MSE estimates of the measures $\widehat{\lambda}$ and $\widehat{\Delta}$ are still growing, but for that of $\widehat{\rho}$ is decreasing and tends towards 0, and for $\widehat{\Lambda}$ has a peak at $C = 0.6$ and declining steadily thereafter as C increases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

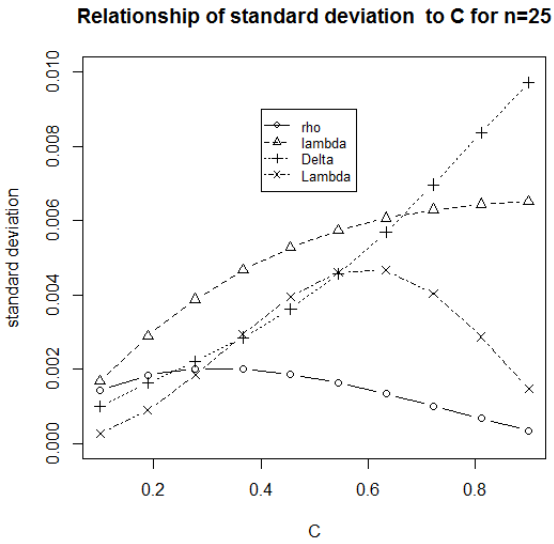

Figure 5: The standard deviation estimates of overlap coefficients by C.

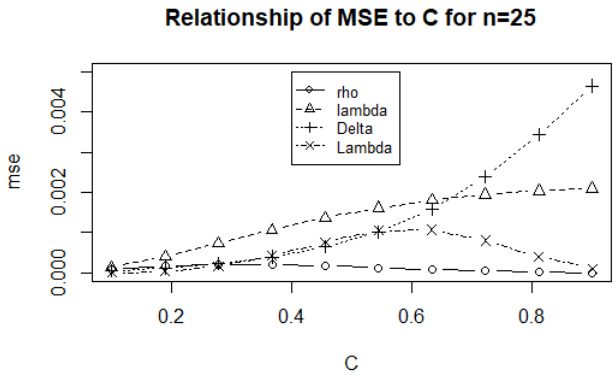

Figure 6: The MSE estimates for overlap coefficients by C.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

5. Example

As application of the new method, consider the dataset discussed by Federer et al. [1]. The authors estimated the superior genetic deviates in segregating populations of sugar beets. The population is assumed to be normally distributed with mean μ (as in Mulekar and Mishra [9]). Using a sample of size 320, the estimated total and environmental deviations are $\widehat{\sigma}_t^2 = 0.911$ and $\widehat{\sigma}_e^2 = 0.509$, respectively (total variance= genetic variance + environmental variance). The estimate of the ratio $\widehat{C}$ is given as $\widehat{C} = \widehat{\sigma}_e/\widehat{\sigma}_t = 0.7475$. Applying Theorem 1, the estimates of bias, variance, and the 95%, confidence intervals for the OVLs are obtained and presented in Table 2.
From Table 2, all four confidence intervals does not include the value 1 the population distributions for the two groups should not be considered to be identical. However, the large lower bounds for the OVLs (For the four similarity coefficients) indicates substantial similarity between the two distributions.

6. Conclusion

In this paper we considered four measures of overlap, namely Matusia's measure ρ, Morisita's measure λ, Weitzman's measure Δ and Kullback-Leibler Λ. We used these measures in the case two Normal distributions having the same expectations and different standard deviations. The overall conclusion is that the biases and MSE of each of the OVL measures are close to zero and approximations are adequate for samples of size as small as 50. The values of the OVL measures are very similar, the coefficient based on Kullback-Leibler is always one of the best for having small values of Bias and MSE. It is clear, in general, that the approximations to bias and MSE presented here may require extremely large samples for example $n > 50$.

References

[1] Federer, W.T., L.R. Powers and M.G. Payne, (1963). Studies on statistical procedures applied to chemical genetic data from sugar beets. Technical Bulletin, Agricultural Experimentation Station, Colorado State University 77.
[2] Folks, J. L., Chhikara, R. S. (1978). The Inverse Gaussian distribution and its statistical applicationA review. J. Roy. Statist. Soc. Ser. B 40:263–289.
[3] Gastwirth, J.L., (1975). Statistical measures of earnings differentials, American Statistician, 29, 32–35.
[4] Ichikawa, M., (1993). A meaning of the overlapped area under probability density curves of stress and strength. Reliab. Eng. System Safety 41, 203–204.
[5] Inman H.F. and Bradley, E.L., (1989) .The Overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Comm. Statist. Theory Methods 18, 3851–3874.
[6] Kullback, S. and Leibler, R. A. (1951). On information and sufciency. Annals of Mathematical Statistics 22, 79–86. 1, 11.
[7] Matusita, K. (1955). Decision rules based on the distance for problem of fir, two samples, and Estimation. Ann. Math. Statist. 26 , 631–640.
[8] Morisita, M. (1959) Measuring interspecific association and similarity between communities. Memoirs of the faculty of Kyushu University. Series E. Biology 3, 36–80.
[9] Mulekar, M. S., and Mishra, S. N. (1994). Overlap Coefficient of two normal densities: equal means case. J. Japan Statist. Soc., 24, 169–180.
[10] Mulekar, M.S. and Mishra, S.N., (2000). Confidence interval estimation of overlap: equal means case. Comp. Statist. Data Analysis 34,121–137.
[11] Sneath, P.H.A., (1977). A method for testing the distinctness of clusters: a test of the disjunction of two clusters in Euclidean space as measured by their overlap. Math. Geol. 9, 123–143.
[12] Weibull, W. (1939). A statistical theory of the strength of materials. Ing. Vetenskaps Akad. Handl. 151, 1–45.
[13] Weitzman, M. S. (1970). Measures of overlap of income distributions of white and Negro families in the United States. Technical paper No. 51 22, Departement of Commerce, Bureau of Census, Washington, D. C.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

Table 1: Bias and MSE of Estimates of OVLs

	$\widehat{\rho}$		$\widehat{\lambda}$		$\widehat{\Delta}$		$\widehat{\Lambda}$	
n	$Bias$	MSE	$Bias$	MSE	$Bias$	MSE	$Bias$	MSE
c=0.05	$\rho = 0.316$		$\lambda = 0.134$		$\Delta = 0.112$		$\Lambda = 0.005$	
10	0.022	0.008	0.018	0.005	0.014	0.003	0.001	0.000*
25	0.007	0.001	0.006	0.0009	0.004	0.0006	0.0004	0.000*
50	0.003	0.0006	0.003	0.0004	0.002	0.0002	0.0002	0.000*
100	0.0016	0.0003	0.0013	0.0002	0.001	0.0001	0.0001	0.000*
200	0.0008	0.0001	0.0006	0.000*	0.0005	0.000*	0.000*	0.000*
500	0.0003	0.000*	0.0002	0.000*	0.0002	0.000*	0.000*	0.000*
c=0.25	$\rho = 0.686$		$\lambda = 0.549$		$\Delta = 0.418$		$\Lambda = 0.124$	
10	0.043	0.029	0.058	0.052	0.045	0.031	0.035	0.02
25	0.014	0.005	0.0185	0.01	0.0142	0.006	0.011	0.004
50	0.006	0.0022	0.009	0.004	0.007	0.002	0.005	0.001
100	0.0031	0.001	0.0042	0.0018	0.0032	0.001	0.0025	0.0007
200	0.0015	0.0005	0.0021	0.0009	0.0016	0.0005	0.0012	0.0003
500	0.0006	0.0002	0.0008	0.0003	0.0006	0.0002	0.0005	0.0001
c=0.5	$\rho = 0.894$		$\lambda = 0.843$		$\Delta = 0.677$		$\Lambda = 0.470$	
10	0.038	0.023	0.056	0.049	0.061	0.058	0.12	0.217
25	0.012	0.0042	0.018	0.0091	0.019	0.011	0.038	0.041
50	0.0057	0.0017	0.0084	0.0037	0.0092	0.0045	0.0177	0.017
100	0.0028	0.0008	0.004	0.0017	0.0044	0.0020	0.008	0.0076
200	0.0014	0.0004	0.002	0.0008	0.0022	0.001	0.0042	0.004
500	0.0005	0.0001	0.0008	0.0003	0.0009	0.0004	0.0017	0.001
c=0.75	$\rho = 0.988$		$\lambda = 0.970$		$\Delta = 0.862$		$\Lambda = 0.855$	
10	0.019	0.006	0.029	0.013	0.068	0.07	0.127	0.025
25	0.006	0.001	0.009	0.002	0.021	0.013	0.04	0.046
50	0.0029	0.0004	0.0044	0.001	0.01	0.005	0.019	0.019
100	0.0014	0.0002	0.0021	0.0005	005	0.002	0.009	0.009
200	0.0007	0.000*	0.001	0.0002	0.002	0.001	0.0045	0.001
500	0.0003	0.000*	0.0004	0.000*	0.0009	0.0005	0.002	0.001
c=0.95	$\rho = 0.999$		$\lambda = 0.999$		$\Delta = 0.975$		$\Lambda = 0.995$	
10	0.004	0.0002	0.005	0.0005	0.069	0.07	0.03	0.01
25	0.0012	0.000*	0.002	0.000*	0.022	0.01	0.009	0.002
50	0.0005	0.000*	0.0008	0.000*	0.01	0.006	0.004	0.0009
100	0.0003	0.000*	0.0004	0.000*	005	0.002	0.002	0.0004
200	0.0001	0.000*	0.0002	0.000*	0.002	0.001	0.001	0.0002
500	0.000*	0.000*	0.000*	0.000*	0.0009	0.0005	0.0004	0.000*

*$|value| < 0.00001$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 November 2018 doi:10.20944/preprints201811.0197.v1

Table 2: Results based on the real data of Federer et al. (1963)

	ρ	λ	Δ	Λ
$\widehat{OVL}$	0.9793	0.9691	0.8701	0.8516
$Bias(\widehat{OVL})$	0.00044	0.00065	0.00216	0.00281
$Var(\widehat{OVL})$	0.00006	0.00014	0.0015	0.0025
95% confidence	(0.963, 0.994)	(0.945, 0.991)	(0.806, 0.911)	(0.772, 0.930)