

**ON THE SELECTION OF THE SMOOTHING PARAMETER IN
POISSON SMOOTHING OF HISTOGRAM ESTIMATOR:
COMPUTATIONAL ASPECTS**

Yogendra P. Chaubey¹, and Pranab K. Sen²

¹ Department of Mathematics and Statistics, Concordia University, Montreal, Canada
E-mail: chaubey@alcor.concordia.ca

² Department of Statistics and Biostatistics, University of North Carolina at Chapel Hill,
Chapel Hill, NC, USA
Email: pksen@bios.unc.edu

ABSTRACT

In this paper the problem of selection of the smoothing parameter for the density estimator using the Poisson distribution [see Gowronski and Stadmüller (1980) and Chaubey and Sen (1996)] is considered. Two cross validation methods, namely *likelihood based cross validation* and *integrated squared error cross validation*, are compared through a numerical study. It is found that the choice proposed in Chaubey and Sen (1996) may not be appropriate for large samples. Instead, data adaptive choice works well for large as well as small samples. Based on this study we also claim that the smoothing parameters selected using any of the two cross validation methods are asymptotically equivalent and seem to provide the smallest Hellinger distance between the estimator and the true density.

KEYWORDS

Cross Validation; Density Estimator; Poisson Histogram Smoother; Hellinger Distance.

2000 Mathematics Subject Classification: 62G07

1 INTRODUCTION

Let $\{X_i, i \geq 1\}$ be a sequence of independent and identically distributed random variables with common distribution function $F(x)$ and the density $f(x)$ supported on \mathbb{R}^+ . Then a

smooth estimator of the density function $f(x)$ is given by

$$\tilde{f}_n(x; \lambda_n, \mathcal{D}) = \lambda_n \sum_{j=0}^N p_j(\lambda_n x) w_j(\lambda_n, \mathcal{D}), \quad (1.1)$$

where λ_n is a constant which controls the smoothness of the estimator, $w_j(\cdot, \cdot)$, $j = 1, \dots$ denote the weights depending on the data \mathcal{D} , namely $w_j(\lambda, \mathcal{D}) = F_n((j+1)/\lambda) - F_n(j/\lambda)$, F_n being the empirical distribution function (*edf*) based on the data \mathcal{D} and $N = \lambda_n \max(X_1, \dots, X_n)$ [see Gowronski and Stadmüller (1980, 1981)]. Chaubey and Sen (1996) proposed independently, a similar estimator where the weights $p_j(\lambda_n x)$ were replaced by $p_j^*(\lambda_n x) = p_j(\lambda_n x) / \sum_{i=0}^N p_i(\lambda_n x)$, so that

$$\sum_{j=0}^N p_j^*(\lambda_n x) = 1. \quad (1.2)$$

The important property of the sequence $\{\lambda_n\}_{n=1}^{\infty}$ is to be chosen such that $\lambda_n \rightarrow \infty$ and $n^{-1}\lambda_n \rightarrow 0$ as $n \rightarrow \infty$.

A convenient *stochastic choice* of λ_n was proposed by Chaubey and Sen (1996) as:

$$\lambda_{n(1)} = \frac{n}{\max(X_1, \dots, X_n)}, \quad (1.3)$$

as it satisfies the desired properties mentioned before if $E(X) < \infty$. Chaubey and Sen (1998) noticed that for the compact support this choice will not satisfy the property that $n^{-1}\lambda_n \rightarrow 0$ as $n \rightarrow \infty$. To cover this case also, they proposed the choice

$$\lambda_{n(2)} = n(\log \log r_n)^{-1} / X_{n-r_n+1:n} \quad (1.4)$$

where $X_{k:n}$ denotes the k^{th} order statistic of the random sample (X_1, \dots, X_n) and $r_n = o(\log \log n)$. This choice satisfies the properties as mentioned before. We may also recommend a deterministic choice

$$\lambda_{n(3)} = [n^{2/5}] + 1 \quad (1.5)$$

as in Gowronski and Stadmüller (1981) due to the strong convergence of the density under this condition. These recommendations are based on the asymptotic theory, however, in finite samples they may not be very satisfactory. The choice $\lambda_{n(1)}$ and $\lambda_{n(2)}$ may turn out to be very large resulting in not so smooth estimators. The choice $\lambda_{n(3)}$ may be a good guide as a desirable choice, but one would like to assert some 'optimality' to this constant. It is proposed in this article to investigate some cross validation methods for data adaptive

choice of λ_n . Note that it is explicit in Gawronski and Stadtmüller (1981) that λ_n is an integer, but this is not necessary.

We will investigate two choices of cross validation methods, one is based on the *likelihood* and the other is based on *mean integrated squared error* (MISE). The latter method is a popular one in the literature on kernel smoothing and one could hope this to be a preferred method here also. However, we have found through extensive simulations that likelihood based cross validation is numerically more convenient and almost equivalent to MISE cross validation. Section 2 describes these methods in detail and Section 3 presents the results of extensive simulations. Comments on computational aspects are also detailed there. Section 4 gives conclusions of the study.

2 Likelihood and Integrated Squared Error Cross Validation

2.1 Likelihood Based Cross Validation

Kullback-Liebler divergence between the estimated density \tilde{f}_n and the true density f is given by

$$KL(\lambda_n) = \mathbb{E} \int \log \frac{f(x)}{\tilde{f}_n(x)} dF(x). \quad (2.1)$$

In practice the optimum cross-validation method estimates such divergence from the data for a given smoothing parameter and chooses one which gives the smallest estimated divergence. Bowman (1981) shows that this procedure is equivalent to the minimization of the negative likelihood:

$$CV_{KL}(\lambda_n) = -\log \prod_{i=1}^n \tilde{f}_n(x; \mathcal{D}_i) = -\sum_{i=1}^n \log(\tilde{f}_n(x; \mathcal{D}_i)), \quad (2.2)$$

where \mathcal{D}_i denotes data with X_i removed from \mathcal{D} . The solution of the above minimization problem will be denoted by λ_{nKL} . When the whole sample is used in constructing the smooth density estimator, we will simply denote the density by $\tilde{f}_n(x)$.

2.2 Integrated Square Error Cross Validation

According to this criterion we determine λ_n that minimizes the criterion related to the *mean integrated squared error*,

$$MISE(\lambda_n) = \mathbb{E} \int (\tilde{f}_n(x) - f(x))^2 dx. \quad (2.3)$$

Estimating this from the data and minimizing it is equivalent to the minimization of [see Silverman (1986)]:

$$CV_{ISE}(\lambda_n) = \int \tilde{f}_n^2(x; \lambda_n, \mathcal{D}) dx - 2 \frac{1}{n} \sum_{i=1}^n \tilde{f}_{n-1}(X_i; \lambda_n, \mathcal{D}_i). \quad (2.4)$$

The first term can be explicitly obtained and the result is shown below:

$$\int_0^\infty \tilde{f}_n(x)^2 = \frac{\lambda_n}{2} \sum_{j=0}^N \sum_{k=0}^N \frac{(j+k)!}{j!k!} \left(\frac{1}{2}\right)^{j+k} w_j(\lambda_n, \mathcal{D}) w_k(\lambda_n, \mathcal{D}). \quad (2.5)$$

The solution to the above minimization problem is denoted by λ_{nISE} .

2.3 Hellinger Distance

The Hellinger distance between two densities f and g is given by

$$H(f, g) = \int \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx. \quad (2.6)$$

This measure is appropriate as a benchmark to establish the closeness of the estimated density to the true density in finite samples, since this is a bounded measure unlike the measures in the above subsections. One notices that since

$$\int \sqrt{f(x)g(x)} dx \leq (1/2) \left(\int f(x) dx + \int g(x) dx \right) = 1/2, \quad (2.7)$$

$$0 \leq H(f, g) \leq 2.$$

A value closer to 0 signifies a closer resemblance of f and g . We will use $H(\tilde{f}_n, f)$ to compare the values of different choices of λ_n for different simulated samples from f in the next section. The optimum value here will be denoted by λ_{nH} . We conjecture that as $n \rightarrow \infty$ different choices of λ_n are equivalent.

3 Simulation Studies

3.1 Some Comments on Computations

Here we simulate samples from some standard distributions, such as the exponential, the Lognormal, the gamma, the Weibull and mixture of exponentials for sample sizes $n = 10(10)50, 100, 1000$. For each sample we obtain the optimum choice of λ_n by KL and MISE cross-validation methods. To judge the closeness of the estimated density with the

true density we have listed the Hellinger distances $H(\tilde{f}_n, f)$ for each choice of λ_n . The optimum solution may be obtained using any optimizing subroutines, however, care must be taken because the function $CV(\lambda_n)$, in general is not a very smooth function due to the discrete nature of the weights $w_j(\lambda_n, \mathcal{D})$.

We have found that the best way to optimise $CV(\lambda_n)$ is to use the search method. Towards this the suitable strategy is to search over a wider area near a possible local minima. In this investigation we have extensively used R- package of statistical analysis [see Ihaka and Gentleman (1996)]. The R- subroutine `nlm` may be used to provide a local minima and then the R- subroutine `optimise` may be used to search for the minima over an interval around the local minima.

This strategy is described through the following example. Figure 1 below gives histogram of a random sample (x_1, \dots, x_n) from Lognormal(0,1) distribution with $\max(x_1, \dots, x_n) = 9.102234$. The `nlm` subroutine gives 3.287119 and 6.554863 for the minima for KL and ISE cross validation respectively. The code returned from the optimizing routine is 3 for the KL criterion and 2 for the ISE criteria. These codes imply that the result is “probably the solution” and therefore we must confirm these values. The program is based on a gradient method and when the gradient tolerance is reached (or other criteria is reached) the solution is reported. Therefore, many times local minima are reported. For the error thresholds here, the default values were used. However, changing the gradient tolerance to lower level 1e-16 did not produce different results. To confirm the solution, therefore we decide to plot the criterion function in the range [1, 20] (see Figure 2.) These plots suggest that the correct minima for the first criterion is in the interval [2, 5] and for the second criterion it is inside the interval [2, 8]. Hence, the reported values from the routine `nlm` seem to give correct solutions. So we give the starting value 2, instead of 1. Now the reported solution is 2.002482 in both the cases. This looks reasonable in the first case but not in the second case.

Therefore, we must examine the function $CV(\lambda)$ in the neighborhood of the solution. Now we use the routine `optimise` which allows to input an interval for the solution. An interval of [2, 5] gives the solution 2.70999 for the KL criterion and that in the interval [2, 8] gives a solution of 4.294595. The solution in the first case is not far from that produced by the `nlm` routine but that in the second case is quite different. Giving a wider interval of [1, 20], the solutions are respectively given by 2.629449 and 5.215846 and seem reasonable by looking at the plots in Figure 2. This procedure picks up the minima as 13.94625 $H(\tilde{f}_n, f)$ where as `nlm` routine gives a local minima of 6.551021 for a starting value of 1. The value of $n/\max(x_1, x_2, \dots, x_n)$ for this sample is given by 8.191776. The Hellinger dis-

tances of the estimated densities using Chaubey-Sen, KL and MISE choices for λ_n with the true lognormal density are respectively given by 0.02909614, 0.03221081 and 0.02638340 which are close to true distance of 0.02316986 if we knew the true density.

To see the effect of different choices on the shape of the density, we plot these in Figure 3. It may be concluded that as long as the value of λ_n is in the close neighborhood of the global minima, the estimated density is not very different from the optimum choice. The four densities corresponding to above values of λ_n as plotted in Figure 3 do not show much difference in them, qualitatively. In this particular data the plot obtained using the MISE criteria may be preferred though, over the others as it comes closer to the one obtained under the true minimum Hellinger distance and it is not as rough. This is the strategy followed for the simulation in the next section.

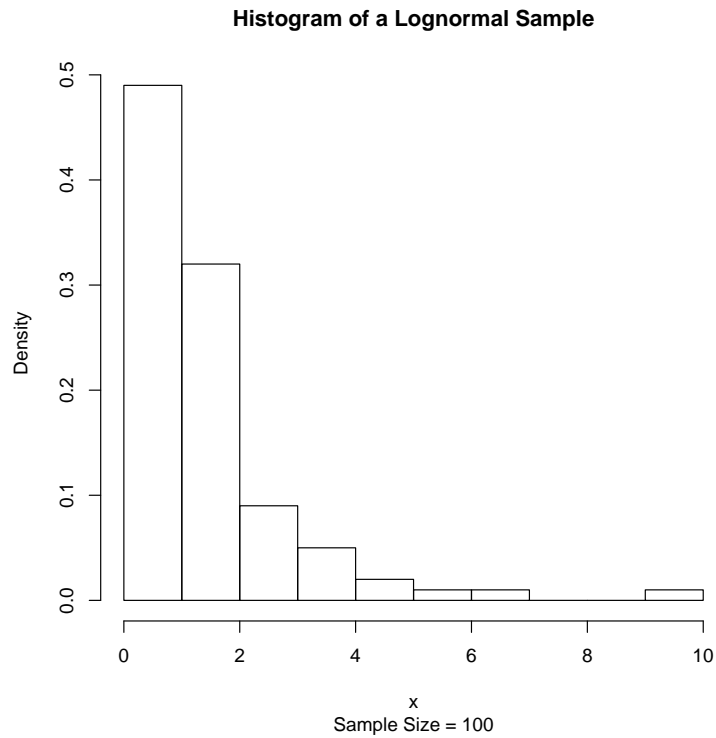


Figure 1: Histogram of a Lognormal Sample, Sample Size = 100

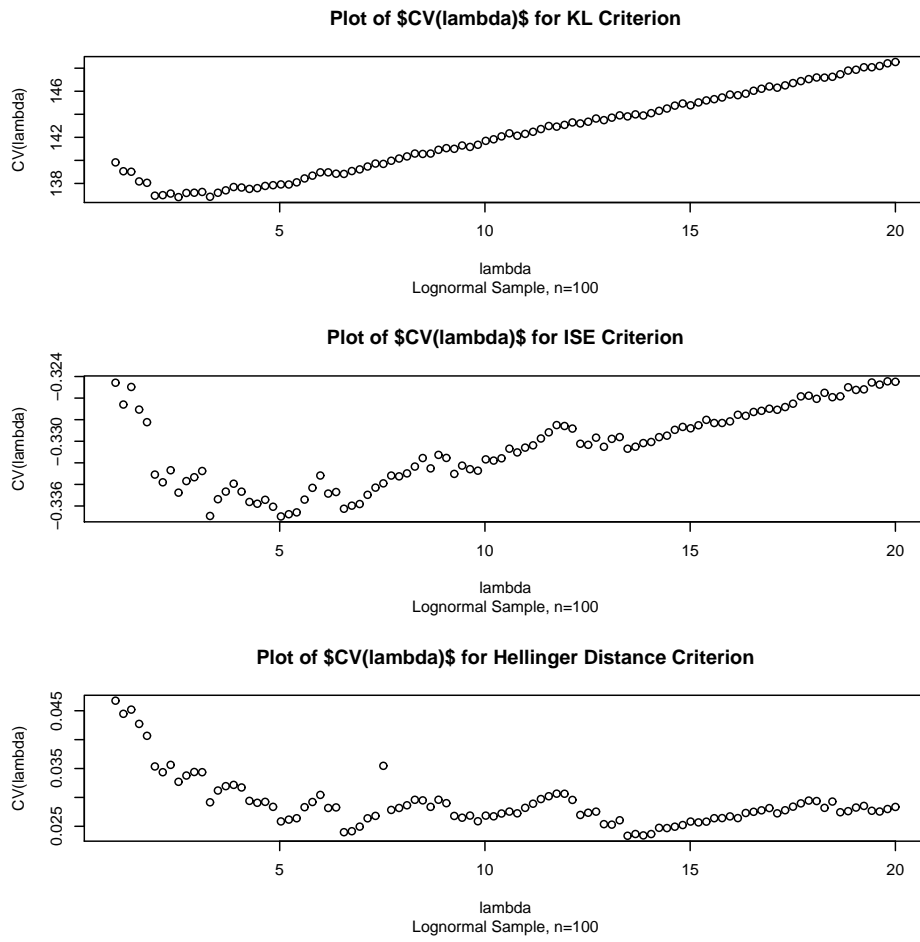


Figure 2: $CV(\lambda)$ Plots for a Lognormal Sample, Sample Size = 100

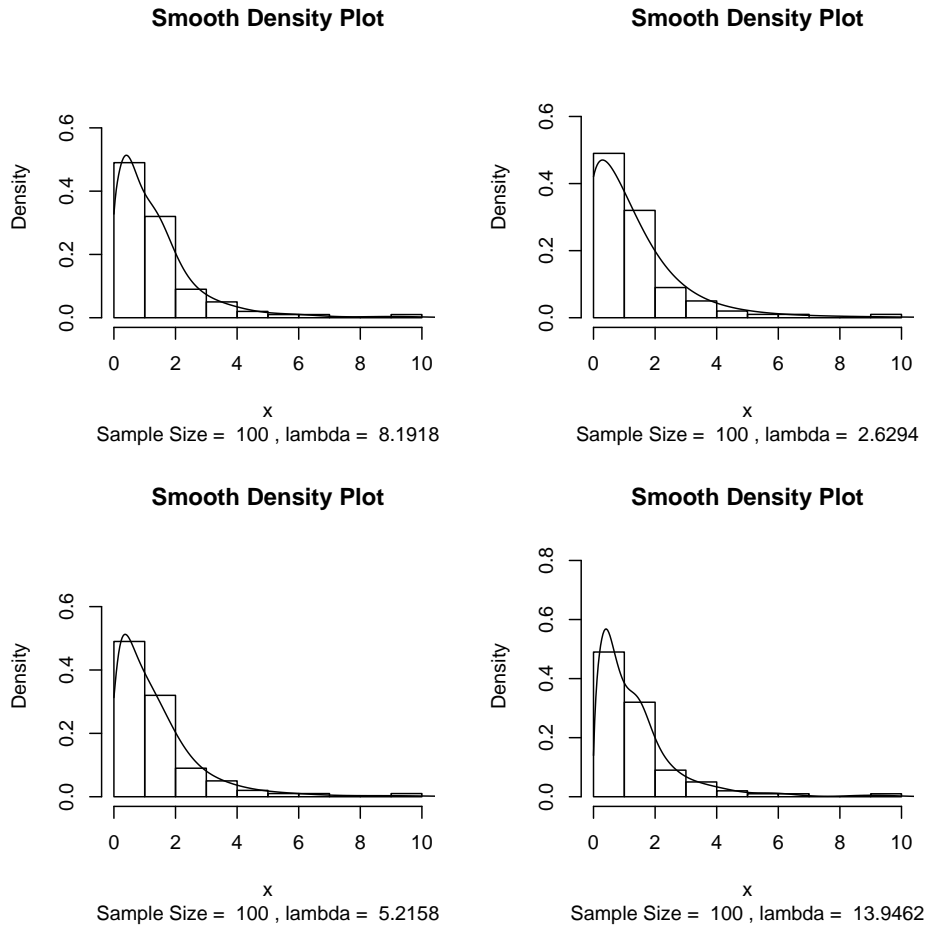


Figure 3: Smooth Density Plot for a Lognormal Sample, Sample Size = 100

3.2 Simulation for Some Standard Distributions

We have simulated from the following densities:

- (1). Exponential Distribution

$$f(x) = \exp(-x)I(x > 0)$$

- (2). Lognormal Distribution

$$f(x) = \frac{1}{x\sqrt{2\pi}\exp\{-\frac{1}{2}(\log x)^2\}}I(x > 0)$$

- (3). Gamma(α) Distribution

$$f(x) = \frac{1}{\Gamma(\alpha)} \exp(-x)x^{\alpha-1}I(x > 0)$$

- (4). Weibull(α) Distribution

$$f(x) = \alpha x^{\alpha-1} \exp(-x^\alpha)I(x > 0)$$

- (5). Mixtures of two Exponential Distributions

$$f(x) = [\Pi \frac{1}{\theta_1} \exp(-x/\theta_1) + (1 - \Pi) \frac{1}{\theta_2} \exp(-x/\theta_2)]I(x > 0),$$

where we choose $\theta_1 \geq \theta_2$ and $\Pi \neq 0.5$. In the simulations, we have fixed $\theta_2 = 1$.

Note that we have generally not incorporated the scales in these distributions, because of the following invariance property. Denote by $\tilde{f}_{nX}(x, \lambda_n)$ as the smooth density based on X data using parameter λ_n . Suppose that X goes through a scale transformation $Y = X/c$ where c is a positive constant. Then it can be easily seen that

$$\begin{aligned} \tilde{f}_{nY}(y; ?) &= c\tilde{f}_{nX}(cy; \lambda_n) \\ &= c\lambda_n \sum_{j=0}^N p_j(\lambda_n cy) w_j(\lambda_n, \mathcal{D}) \\ &= c\lambda_n \sum_{j=0}^N p_j(\lambda_n cy) \left[G_n \left(\frac{j+1}{c\lambda_n} \right) - G_n \left(\frac{j}{c\lambda_n} \right) \right] \\ &= \tilde{f}_{nY}(y; \lambda_n^*), \end{aligned}$$

where $\lambda_n^* = c\lambda_n$, here G_n denotes the *edf* of the transformed data $X_1/c, \dots, X_n/c$.

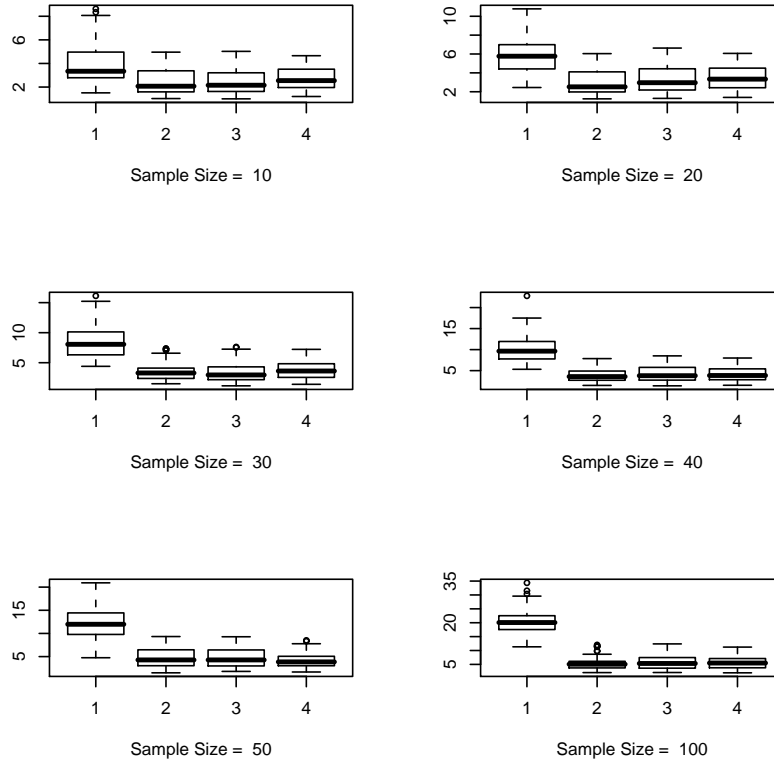


Figure 4: Box Plot for λ_n for 100 Exponential Samples; 1: Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validation, 4: Optimum Hellinger Distance

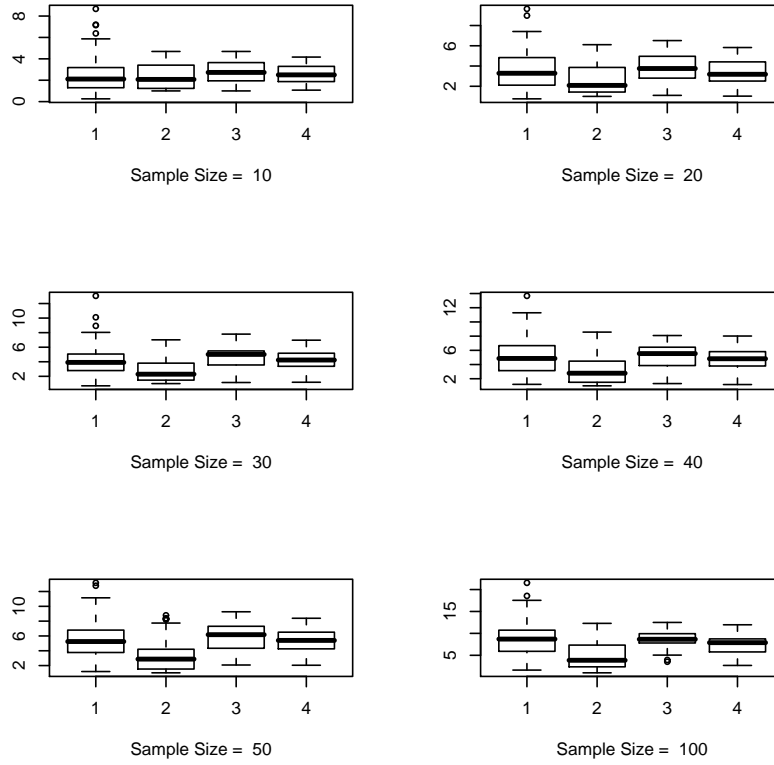


Figure 5: Box Plot for λ_n for 100 Lognormal Samples; 1: Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validation, 4: Optimum Hellinger Distance

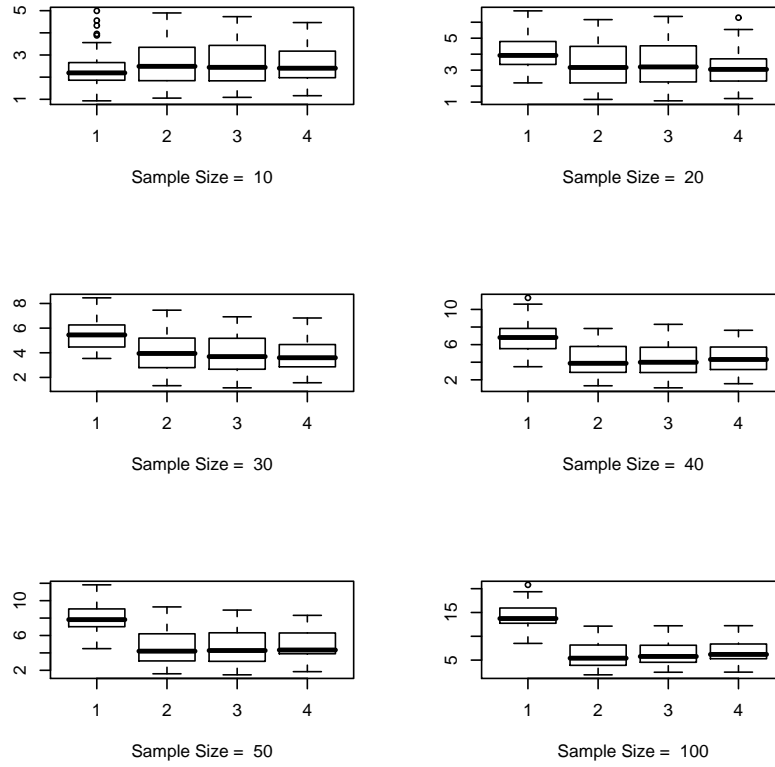


Figure 6: Box Plot for λ_n for 100 Gamma Samples, $\alpha = 2$; 1: Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validation, 4: Optimum Hellinger Distance

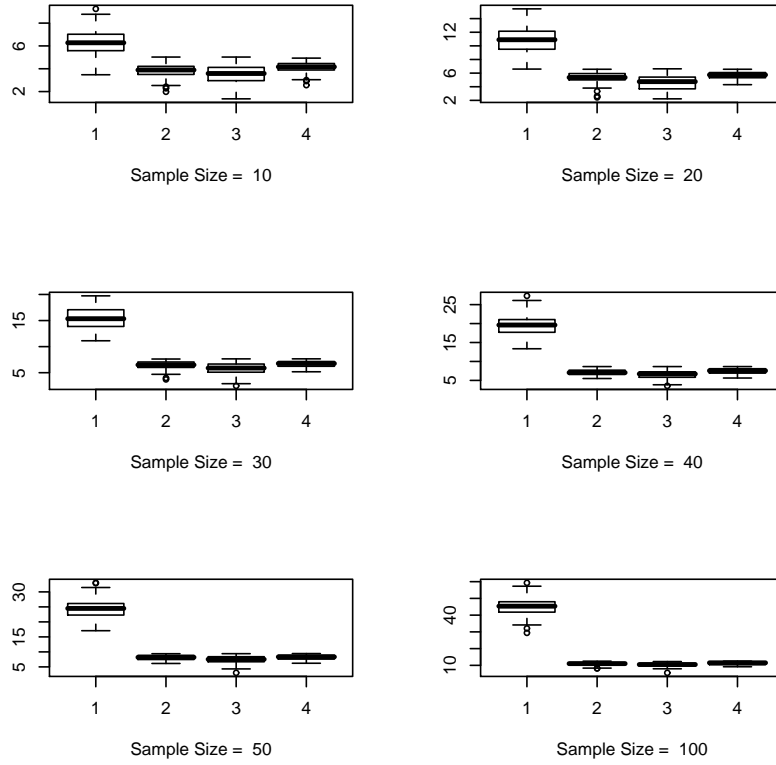


Figure 7: Box Plot for λ_n for 100 Weibull Samples, $\alpha = 2$; 1: Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validation, 4: Optimum Hellinger Distance

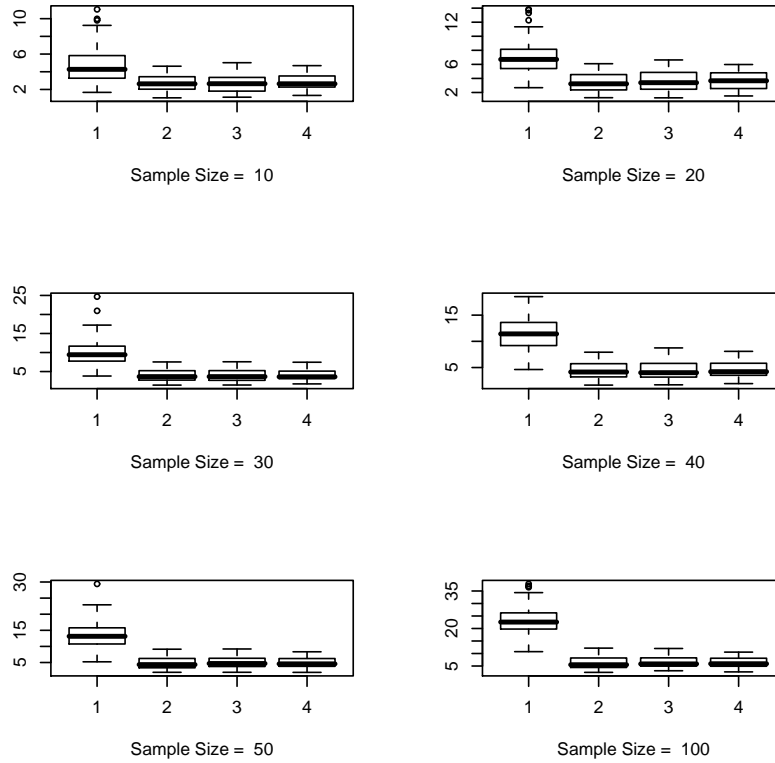


Figure 8: Box Plot for λ_n for 100 Exponential Mixture Samples, $\theta_1 = 2$, $\theta_2 = 1$, $\Pi = .4$; 1: Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validation, 4: Optimum Hellinger Distance

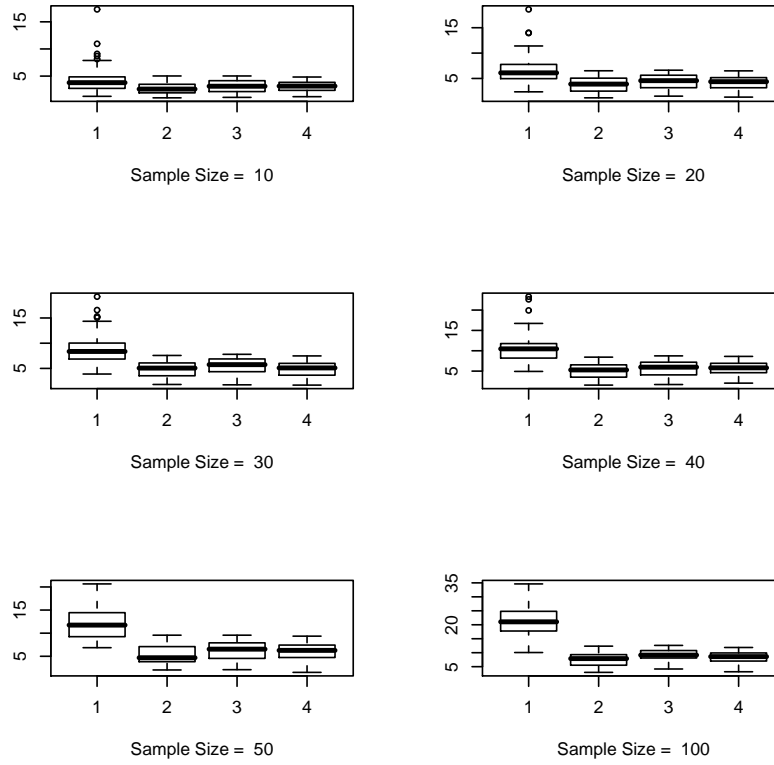


Figure 9: Box Plot for λ_n for 100 Exponential Mixture Samples, $\theta_1 = 10$, $\theta_2 = 1$, $\Pi = .2$;
 1: Chaubey-Sen Choice, 2: KL Cross-Validation, 3: ISE Cross-Validation, 4: Optimum Hellinger Distance

4 Conclusions

Denote by λ_{1O} the value which minimizes the Kullback-Liebler divergence

$$KL(\lambda_n) = \mathbb{E} \int \log \frac{f(x)}{\tilde{f}_n(x)} dF(x),$$

λ_{2O} the minimizer of

$$MISE(\lambda_n) = \mathbb{E} \int (\tilde{f}_n(x) - f(x))^2 dx$$

and λ_{3O} the minimizer of the expected Hellinger distance,

$$h(\lambda_n) = \mathbb{E} \int (\sqrt{\tilde{f}_n(x)} - \sqrt{f(x)})^2 dx.$$

We have suppressed the index n , in λ_{iO} to differentiate it from the stochastic data dependent choice λ_{in} .

It is seen that

1. Chaubey-Sen choice usually produces large values of the smoothing parameters, especially, for large samples. Because of the invariance property of the estimator, choice of the scale of the data does not affect the optimum value.
2. Chaubey-Sen choice is much more variable even in the cases on an average it is close to the true optimum.
3. The two cross-validation criteria generally produce similar results, especially for larger samples and they converge to the true optimum under the known density.
4. We conjecture that suppose λ_{iO} denotes the true value of λ_n which minimizes criterion i , $i = 1, 2, 3$, and λ_{in} is the minima based on the data, then

$$(i) \lim_{n \rightarrow \infty} \frac{\lambda_{in}}{\lambda_{iO}} = 1 \text{ a.s.}$$

$$(ii) \lim_{n \rightarrow \infty} \frac{\lambda_{1O}}{\lambda_{HO}} = \lim_{n \rightarrow \infty} \frac{\lambda_{2O}}{\lambda_{HO}} = 1 \text{ a.s.,}$$

where λ_{HO} is the true minimizer of the expected Hellinger distance between \tilde{f}_n and f .

Acknowledgements

This research was partially completed while Y.P. Chaubey was on a sabbatical leave at the Department of Biostatistics, University of North Carolina at Chapel Hill. Research facilities provided here are gratefully acknowledged. Partial support for this research through a Discovery Grant from Natural Sciences and Engineering Research Council of Canada (NSERC) to Y. P. Chaubey is also gratefully acknowledged.

REFERENCES

- [1] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.
- [2] Chaubey, Y. P., and Sen, P. K. (1996). On smooth estimation of survival and density functions. *Statist. Decisions* **14** 1–22.
- [3] Chaubey, Y. P., Sen, P. K. (1998). On smooth estimation of hazard and cumulative hazard functions. In *Frontiers of Probability and Statistics*, S.P. Mukherjee *et al.* (eds.) Narosa: New Delhi; 91–99.
- [4] Gowronski, W. and Stadmüller, U. (1980). On density estimation by means of Poisson's distribution. *Scand. J. of Statist.*, **7** 90-94.
- [5] Gowronski, W. and Stadmüller, U. (1981). Smoothing of histograms by means of lattice- and continuous distributions. *Metrika*, **28** 155-164.
- [6] Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.
- [7] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London.