

A Bandwidth Selection for Kernel Density Estimation of Functions of Random Variables

A. R. Mugdadi

Department of Mathematics
Southern Illinois University
Carbondale, IL 62901, U.S.A.
amugdadi@math.siu.edu

and

Ibrahim A. Ahmad

Department of Statistics and Actuarial Sciences
University of Central Florida
Orlando, FL 32816, U.S.A.
iahmad@mail.ucf.edu

Abstract

In this investigation, the problem of estimating the probability density function of a function of m independent identically distributed random variables, $g(X_1, X_2, \dots, X_m)$ is considered. The choice of the bandwidth in the kernel density estimation is very important. Several approaches are known for the choice of bandwidth in the kernel smoothing methods for the case $m = 1$ and g is the identity. In this study we will derive the bandwidth using the least square cross validation and the contrast methods. We will compare between the two methods using Monte Carlo simulation and using an example from the real life.

Keywords: Density estimation, function of random variables, bandwidth, kernel contrast.

1 INTRODUCTION

The kernel estimation method is an important tool in nonparametric density and distribution functions fitting. Suppose that a data set X_1, X_2, \dots, X_n , denotes a random sample from an unknown probability density function (*pdf*) $f(x)$, then the kernel density estimator of $f(x)$ is defined by

$$\hat{f}(x) = \frac{1}{nb_n} \sum_{i=1}^n k\left(\frac{x - X_i}{b_n}\right), \quad (1.1)$$

where k is a bounded nonnegative function satisfying $\int k(x)dx = 1$ and b_n is a sequence of positive number usually called the bandwidth.

Consider the function $g(X_1, X_2, \dots, X_m)$ that depends on $m \geq 1$ observations. The distribution function of $g(X_1, X_2, \dots, X_m)$ is defined by

$$H(t) = P(g(X_1, X_2, \dots, X_m) \leq t),$$

where $t \in R$. A function closely related to the distribution function is the density function, defined by $h(t) = H'(t)$, when it exists. The nonparametric kernel estimate of $h(t)$ is easily seen to be (cf. Frees (1994)):

$$\hat{h}(t) = \hat{h}(t, b) = \frac{1}{b \binom{n}{m}} \sum_{(n,m)} w\left(\frac{t - g(X_{i_1}, \dots, X_{i_m})}{b}\right), \quad (1.2)$$

where $b = b_n$ is the bandwidth, $1 \leq i_1 < i_2 < \dots < i_m \leq n$ is an ordered subset of $1, 2, \dots, n$, $\sum_{(n,m)}$ denotes summation over all $\binom{n}{m}$ subsets and $w(\cdot)$ is a kernel function. It is clear that if $m = 1$ and $g(x) = x$ then the estimator $\hat{h}(t)$ reduces to the estimator $\hat{f}(x)$.

The function of m identical random variables $g(X_1, X_2, \dots, X_m)$ have applications in real life such as the case where $g(X_1, X_2, \dots, X_m) = \sum_{i=1}^m X_i$ which identifies in actuarial science the total claims of, for example, individual insurances. It also identifies, in life testing, the life of a parallel system of m identical components. Other applications are indicated in Frees(1994), Ahmad and Fan (2001) and Ahmad and Mugdadi (2003 b). Also, the authors use the estimate (1.2) to derive a kernel based testing procedure for normality, cf. Ahmad and Mugdadi (2003 a).

Some asymptotic properties of the estimate (1.2) are known, for example, Frees(1994) discusses the consistency and the asymptotic normality of $\hat{h}(t)$. Precisely, he shows that under some conditions and if $b_n \rightarrow 0$ such that the bias $B_n(t) = E\hat{h}(t) - h(t) \rightarrow 0$ as $n \rightarrow \infty$, then $\hat{h}(t)$ is consistent estimate of $h(t)$. In addition Frees(1994) proves that under some conditions $n^{\frac{1}{2}}(\hat{h}(t) - h(t))$ is asymptotically normal with mean B_n and variance C . a positive constant. The above work demonstrated that the asymptotic behavior of $\hat{h}(t)$ is different from that of $\hat{f}(x)$ and utilizes new methodologies. Ahmad and Fan (2001) obtain the optimal theoretical bandwidth b in this general case. This work is expanded by Ahmad and Mugdadi (2003 b). Thus we are motivated to further the above work and discuss data-based choices of the bandwidth.

Kernel smoothing provides a simple way of finding structures in data sets without knowing the probability density function. Theoretical and simulation analysis have shown that the choice of the kernel is not crucial for density estimation in the case of independent identically distributed (i.i.d.) random variables. The most important part in the kernel estimation method is to select the bandwidth. There are many situations where it is satisfactory to select the bandwidth by graphing many densities using different bandwidths, then select the most acceptable density. One of the strategies to do that by starting with small (or large) bandwidth then going up (or down) until you reach the suitable one.

Meanwhile, when we select the bandwidth, we need to consider the error in our selection. There are many methods to calculate the error such as the Mean Square Error

(*MSE*) and the Mean Integrated Square Error (*MISE*) where

$$MSE(\hat{h}(t)) = E[\hat{h}(t) - h(t)]^2,$$

and the

$$MISE(\hat{h}(t)) = E \int [\hat{h}(t) - h(t)]^2 dt$$

Also, there are many methods to select the bandwidth for the case of $\hat{f}(x)$. Most of these methods are based on minimizing the *MSE* or the *MISE*. The following methods are used to select the bandwidth for $\hat{f}(x)$:

- 1- Least Squares Cross-Validation (see Rudemo (1982) and Bowman(1984))
- 2- Biased Cross-Validation (see Scott and Terrell (1987))
- 3- Plug-in Bandwidth Selection (see Sheather and Jones (1991))
- 4- Smoothed Cross-Validation (see Hall, Marron and Park (1992))
- 5- Root-n Bandwidth Selection (see Hall, Sheather, Jones and Marron (1991))
- 6- The Contrast Method (see Ahmad and Ran(2003))

Jones, Marron and Sheather(1996) compared among the first generation methods (Rules of thumb, Least squares cross validation and biased cross validation) and the second generation methods (the plug-in approach and smoothed bootstrap) using real data examples, asymptotic analysis and simulation.

On the other hand, Ahmad and Fan (2001) derived the optimal bandwidth that minimizes the *AMISE* of $\hat{h}(t)$ by:

$$b_{opt} = \left(\frac{R(w)}{(\mu_2(w))^2 R(h^{(2)})} \right)^{\frac{1}{5}} \binom{n}{m}^{\frac{-1}{5}} \quad (1.3)$$

where, $R(f) = \int f^2(x)dx$ and $\mu_2(w) = \int x^2 w(x)dx$, but b_{opt} depends as usual on the unknown density function $h(t)$.

2 Least Square Cross Validation for $\hat{h}(t)$.

The motivation of the least square cross validation (LSCV) comes from expanding the *MISE* of $\hat{h}(t, b)$ to obtain

$$MISE(\hat{h}(t, b)) = E \int (\hat{h}(t, b))^2 dt - 2E \int (\hat{h}(t, b)h(t))dt + \int (h(t))^2 dt \quad (2.1)$$

but, $\int (h(t))^2 dt$ does not depend on b. Therefore, minimization of $MISE(\hat{h}(b))$ is equivalent to minimization of

$$MISE(\hat{h}(t, b)) - \int (h(t))^2 dt = E \int (\hat{h}(t, b))^2 dt - 2 \int (\hat{h}(t, b)h(t))dt. \quad (2.2)$$

The *LSCV* is based on abstaining an unbiased estimate of (2.2). Let $m_{(1)}$ be a fixed number equal to m and define the density estimate of $g(X_1, X_2, \dots, X_m)$ based on the $\binom{n}{m}$ cases with $X_1, X_2, \dots, X_{m_{(1)}}$ deleted by:

$$\hat{h}_{-m_{(1)}}(t, b) = \frac{1}{\left(\binom{n}{m} - 1\right)b} \sum_{(n,m), m \neq m_{(1)}} w\left(\frac{t - g(X_{i_1}, X_{i_2}, \dots, X_{i_m})}{b}\right). \quad (2.3)$$

But,

$$E \sum_{(n,m), m \neq m_{(1)}} w\left(\frac{t - g(X_{i_1}, X_{i_2}, \dots, X_{i_m})}{b}\right) = \left(\binom{n}{m} - 1\right) E\left(w\left(\frac{t - g(X_1, X_2, \dots, X_m)}{b}\right)\right), \quad (2.4)$$

and

$$E\left(w\left(\frac{t - g(X_1, X_2, \dots, X_m)}{b}\right)\right) = \int w\left(\frac{t - g(x_1, x_2, \dots, x_m)}{b}\right) dH(g(x_1, x_2, \dots, x_m)). \quad (2.5)$$

Therefore,

$$E(\hat{h}_{-m_{(1)}}(t, b)) = \frac{1}{b} \int w\left(\frac{t - g(x_1, x_2, \dots, x_m)}{b}\right) dH(g(x_1, x_2, \dots, x_m)). \quad (2.6)$$

Thus, $\frac{1}{\binom{n}{m}} \sum_{(n,m_{(1)})} \hat{h}_{-m_{(1)}}(g(x_1, x_2, \dots, x_{m_{(1)}}), b)$ is an unbiased estimator of $\int \hat{h}(t, b) h(t) dt$. From the above equations, we can conclude that

$$LSCV(b) = \int (\hat{h}(t, b))^2 dt - \frac{2}{\binom{n}{m}} \sum_{(n,m_{(1)})} \hat{h}_{-m_{(1)}}(g(X_{i_1}, X_{i_2}, \dots, X_{i_{m_{(1)}}}), b) \quad (2.7)$$

is an unbiased estimator of

$$E\left(\int (\hat{h}(t, b))^2 dt - 2 \int \hat{h}(t, b) h(t) dt\right) \quad (2.8)$$

Thus, it is reasonable to choose b that minimizes $LSCV(b)$.

3 The Contrast Method for $\hat{h}(t)$.

The contrast method for the kernel density estimator $\hat{f}(x)$ proposed by Ahmad and Ran (2003). Hence we extend it to the case of the estimator $\hat{h}(t)$. The first step in the contrast method, we define the kernel density estimations $\hat{h}_j(t, b)$ based on q kernels, w_1, \dots, w_q . Thus

$$\hat{h}_j(t, b) = \hat{h}(t, b) = \frac{1}{b \binom{n}{m}} \sum_{(n,m)} w_j\left(\frac{t - g(X_{i_1}, \dots, X_{i_m})}{b}\right). \quad (3.1)$$

After choosing the contrast coefficients p_1, \dots, p_q , where $\sum_{j=1}^q p_j = 0$, select the bandwidth that minimizing the $MISE_{cont}$, where

$$MISE_{cont}(\hat{h}(t, b)) = E\left[\int \left(\sum_{j=1}^q p_j \hat{h}_j(t, b) - \sum_{j=1}^q p_j h(t)\right)^2 dx\right], \quad (3.2)$$

but $\sum_{j=1}^q p_j h(t) = h(t) \sum_{j=1}^q p_j = 0$ Therefore

$$MISE_{cont} = E\left[\int \left(\sum_{j=1}^q p_j \hat{h}_j(t, b)\right)^2 dt\right]. \quad (3.3)$$

Thus

$$ISE_{cont} = ISE(b)_{cont} = \int \left(\sum_{j=1}^k p_j \hat{h}_j(t, b) \right)^2 dt \quad (3.4)$$

is an unbiased estimator of $MISE_{cont}$. Therefore a reasonable choice for estimating b is to minimize ISE_{cont} which does not depend on the unknown density function $h(t)$. Ahmad and Mugdadi (2003) proved that the estimator based on the ISE_{cont} for $\hat{h}(t)$ is consistent. Finally, define the density estimation using the kernel contrast approach by:

$$\hat{h}(t) = \sum_j^q c_j \hat{h}_j(t, b), \quad (3.5)$$

where $\sum_j^q c_j = 1$. We can have an equal weight for the kernels by choosing q as an even integer, $p_j = -p_{2j}$, $j = 1, \dots, \frac{q}{2}$ and $c_j = \frac{1}{q}$, for $j = 1, \dots, q$.

4 Simulation Work

The *LSCV* and the contrast are data based bandwidth methods. We can use these two methods with any kernel function $w_j(x)$ such that $\int w_j(x)dx = 1$ and $w_j(x)$ is a bounded function.

In this section we will compare between the *LSCV* and the ISE_{cont} methods using Monte Carlo studies and a real data example as well.

4.1 Monte Carlo Studies

During our simulation study we will use the following three kernels

- 1- The normal kernel with mean 0 and variance c .
- 2- The Epanechnikov kernel
- 3- The Biweight kernel

Also, we will simulate from the following populations:

- I- The normal distribution with mean 0 and variance 1.
- II- The exponential distribution with mean 1.
- III- The Cauchy distribution with pdf

$$f(x) = \frac{1}{(\pi)(1+x^2)}$$

To evaluate the performance and to compare between the *LSCV* and the ISE_{cont} methods we simulate random samples of sizes 25 and 50 from the standard normal distribution and also from the exponential distribution with mean 1. In Figures 1 through 4 we compare between the exact densities and the estimated densities of $g(X_1, X_2) = X_1 + X_2$, where X_1 and X_2 are random variables having standard normal distribution (Figures 1 and 2) and having exponential distribution with mean 1 (Figures 3 and 4). From these Figures we conclude that both density estimates describe the data precisely, but the one based on the contrast method is smoother.

Table 1: Using two normal kernels

Population I	ISE_{cont}		LSCV	
n	Var.	Bias	Var.	Bias
25	0.0090	0.1339	0.0474	0.1986
30	0.0092	0.0879	0.0390	0.2299
35	0.0095	0.0979	0.0396	0.1699
40	0.0092	0.0779	0.0589	0.1519
45	0.0054	0.0593	0.0640	0.6666
50	0.0084	0.0876	0.0354	0.2173

Table 2: Using Epanechnikov and Biweight kernels

Population I	ISE_{cont}		LSCV	
n	Var.	Bias	Var.	Bias
25	0.0728	-1.322	0.0217	-2.384
30	0.0729	-1.294	0.0735	-2.323
35	0.0192	-1.155	0.0867	-2.277
40	0.0248	-1.130	0.0549	-2.362
45	0.0519	-1.202	0.0521	-2.148
50	0.0219	-1.223	0.0870	-2.400

Tables 1 through 6 give us, in a more precise fashion, the differences between the ISE_{cont} and the $LSCV$ methods.

We compare the variance of \hat{b} , when we select \hat{b} minimizing the $ISE(b)_{con}$, defined in (3.4) with the one by minimizing the $LSCV(b)$ defined in (2.7). Also, we compare the biases, where the biases are calculated by comparison with (theoretically) "optimal" bandwidth choices of (1.3). We simulated n values from each of the above pdf's based on a set with 30 \hat{b} values. To calculate \hat{b} using the ISE_{cont} and the $LSCV$ we used the following combinations:

1. w_1 and w_2 are two kernels distributed $N(0,1)$ and $N(0,4)$ respectively,
2. w_1 and w_2 are two kernels distributed Epanechnikov and Biweight, respectively,
3. w is a kernel from $N(0,1)$, and
4. w is a kernel from Epanechnikov distribution.

Also, in the contrast method we used $p_1 = -p_2$ and $c_1 = c_2 = \frac{1}{2}$.

Tables 1-6 provide us with more information about the differences between the contrast and the least square cross validation methods to derive the bandwidth. We can conclude from the Tables the following:

- 1- The variance of \hat{b} based on the contrast method is less than the variance based on the $LSCV$ method when we simulate from normal and when we use two normal kernels (Table 1).
- 2- In most of the cases the variance of \hat{b} based on the ISE_{cont} method is less than the variance based on the $LSCV$ method when we simulate from normal distribution and

Table 3: Population II: using two normal kernels

Population II	ISE_{cont}		LSCV	
n	Var.	Bias	Var.	Bias
25	0.0728	-1.322	0.0217	-2.384
30	0.0729	-1.294	0.0735	-2.323
35	0.0192	-1.155	0.0867	-2.277
40	0.0248	-1.130	0.0549	-2.362
45	0.0519	-1.202	0.0521	-2.148
50	0.0219	-1.223	0.0870	-2.400

Table 4: Population II: using Epanechnikov and Biweight kernels

Population II	ISE_{cont}		LSCV	
n	Var.	Bias	Var.	Bias
25	0.0787	0.3149	0.0342	-0.658
30	0.0433	0.3916	0.0015	-0.719
35	0.0573	0.3649	0.0018	-0.739
40	0.0203	0.4796	0.0091	-0.734
45	0.0165	0.4669	0.0075	-0.764
50	0.0282	0.4029	0.0068	-0.770

Table 5: Using two normal kernels

Population III	ISE_{cont}		LSCV	
n	Var.	Bias	Var.	Bias
25	0.0120	1.2465	0.0001	-0.933
30	0.0119	1.2105	0.0002	-0.923
35	0.0152	1.1919	0.0010	-0.941
40	0.0070	1.1665	0.0009	-0.891
45	0.0115	1.1912	0.0001	-0.874
50	0.0092	1.1645	0.0001	-0.871

Table 6: Using Epanechnikov and Biweight kernels

Population III	ISE_{cont}		LSCV	
	Var.	Bias	Var.	Bias
25	0.0001	0.0196	0.0000	-0.388
30	0.0007	0.0089	0.0001	-0.389
35	0.0010	-0.004	0.0001	-0.388
40	0.0281	-0.041	0.0001	-0.387
45	0.0037	-0.031	0.0002	-0.368
50	0.0554	-0.215	0.0031	-0.372

we use Epanechnikov and Biweight kernels (Table 2). Also, when we simulate from exponential with two normal kernels (Table 3).

3- The variance of \hat{b} based on the $LSCV$ method is less than the variance based on the ISE_{cont} when we simulate from Cauchy distribution (Table 5 and 6) and when we simulate from exponential and we use Epanechnikov and Biweight kernels for the contrast methods.

4- The absolute value of the bias of \hat{b} based on the ISE_{cont} method is less than the one based on based on the $LSCV$ method when we simulate from normal and when we simulate from exponential (Tables 1,2,3 and 4) and when we simulate from Cauchy when we use Epanechnikov and Biweight kernels using the ISE_{cont} method (Table 6).

5- The absolute value of the bias of \hat{b} based on the $LSCV$ method is less than the one using the contrast with two normal kernels.

6- The mean square error using the ISE_{cont} is less than the one using the $LSCV$ when we simulate from normal and exponential distributions (Tables 1,2,3 and 4), however, the mean square error using the $LSCV$ is less than the one using the ISE_{cont} when we simulate from Cauchy (Tables 5 and 6). Also, it is clear that in all the cases the mean square error is decreasing as the sample size is increasing.

Based on these simulation the two methods are provide us different bandwidths as well appropriate choices of the bandwidth for $\hat{h}(t)$.

4.2 Real Data Example

An important measure of the performance of any bandwidth selection method is how well it performs in practice. There are many applications for $\hat{h}(t)$, cf. Frees (1994) and Ahmad and Fan (2001). During our study we will choose the bandwidth using the contrast method and the least square cross validation method, and we want to check whether our conclusion in the real life example about the smoothness of the density function is consistent with the conclusion in the simulations examples.

Let X_1, X_2, \dots, X_n be a random sample of insurance claims, a particular line of business. Table 2.7 represents the claims in thousands of dollars received by one of the Auto Insurance companies in the United States in September 2001.

From the stand point of the insurer, of interest is the distribution of the sum of claims X_1, X_2, \dots, X_m , interpret m to be the expected number of claims in a specified

Table 7: The Insurance Data

0.981	2.040	8.734	1.340	6.578	1.134	2.670
0.503	6.089	2.345	1.110	1.005	2.457	1.774

financial period, for example a day or month. In this example, we discuss the case $g(x_1, \dots, x_m) = x_1 + \dots + x_m$, when $m = 2$. Using the standard normal kernel in the LSCV the bandwidth $b = 0.29$. In the contrast method the choice of the bandwidth doesn't depend on the constants $c_j, j = 1, \dots, q$, thus in Figures 6 and 7 the bandwidth is the same which is $b = 0.42$, while in Figure 8 the two bandwidths are $b = 0.42$ and $b = 0.24$. In the contrast method, we used the normal kernel with mean 0 and with variances 1,4,9, and 16 when the contrast coefficients are p_1, p_2, p_3 and p_4 , respectively.

From Figures 5 through 8, we conclude the following:

- 1- The LSCV do not provide a smooth density estimation, which is consistent with the case for $m = 1$, see Jones, Marron and Sheather (1996) (Figure 5).
- 2- When the contrast coefficients are 1/2 and 1/2 we have almost an identical density estimations with the case of 1/3 and 2/3, but they are close to case 1/10 and 9/10, but not identical (Figures 6 and 7).
- 4- The density estimations based on two or four kernels when the kernels have the same weight is almost identical (Figure 8).
- 5- The bandwidth when we use four contrast coefficients is smaller than the one using the LSCV, but because we used the four kernels in density estimation, the contrast method provides us with a smoother bandwidth.

From Figures 1 through 9, we conclude that the choice of the bandwidth based on the real data is consistent with the one based on simulations which is the ISE_{cont} method gives us a smoother density estimation.

References

- [1] Ahmad, I. A. and Fan, Y., (2001), Optimal bandwidth for kernel density estimators of functions of Observations. *Statistics and Probability Letter*, **51** 3, 245-251
- [2] Ahmad, I. A. and Mugdadi A. R., (2003a), Testing Normality using kernel method, *Journal of Nonparametric Statistics*, **15**, 273-288
- [3] Ahmad, I. A. and Mugdadi A. R., (2003b), Analysis of kernel density estimation of functions of random variables. *Journal of Nonparametric Statistics*, to appear
- [4] Ahmad, I. A. and Ran, I. S., (2003), Selection of smoothing parameters via kernel contrasts, *Journal of Nonparametric Statistics*, to appear.
- [5] Bowman, A. W., (1984), An alternative method of cross-validation for the smoothing density estimates. *Biometrika*, **71**, 353-360.

- [6] Frees, E., (1994), Estimating densities of functions of observations. *Journal of American Statistical Association*, **89**, 17-525.
- [7] Hall, P., Marron, J. S. and Park, B. U., (1992), Smoothed cross validation. *Probability Theory Related Field*, **90**, 149-173.
- [8] Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S., (1991), On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, **78**, 263-269.
- [9] Jones, M. C., Marron, J. S. and Sheather, S. J., (1996), A brief survey of bandwidth selection for density estimation. *Journal of American Statistics Association*, **91**, 401-407
- [10] Rudemo, M., (1982), Empirical Choice of histograms and kernel density estimation. *Scand. Journal of Statistics*, **9**, 65-78.
- [11] Scott, D. W. and Terrel, G. R., (1987), Biased and unbiased cross validation in density estimation. *Journal of American Statistics Association*, **82**, 1131-46.
- [12] Sheather, S. J. and Jones, M. C., (1991), A reliable data-based bandwidth selection method for kernel density estimation. *Journal of Royal Statistics Society*, **B 53**, 683-690.

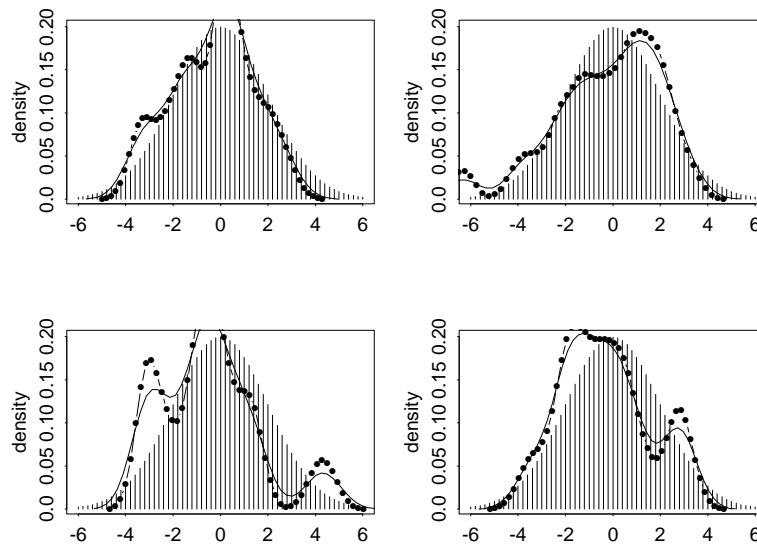


Figure 1: Simulation from normal distribution with $n = 25$, columns: exact distribution, dots: using LSCV, solid: using ISE_{cont}

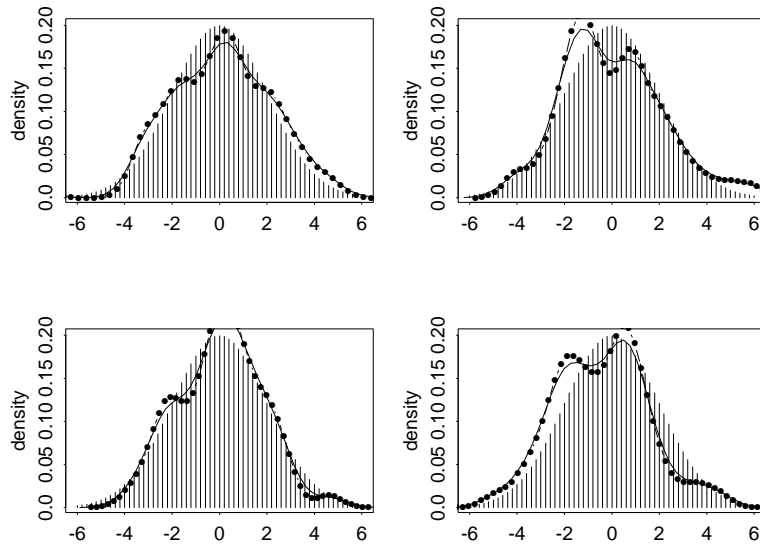


Figure 2: Simulation from normal distributing with $n = 50$,
columns: exact distribution, dots: using LSCV, solid: using ISE_{cont}

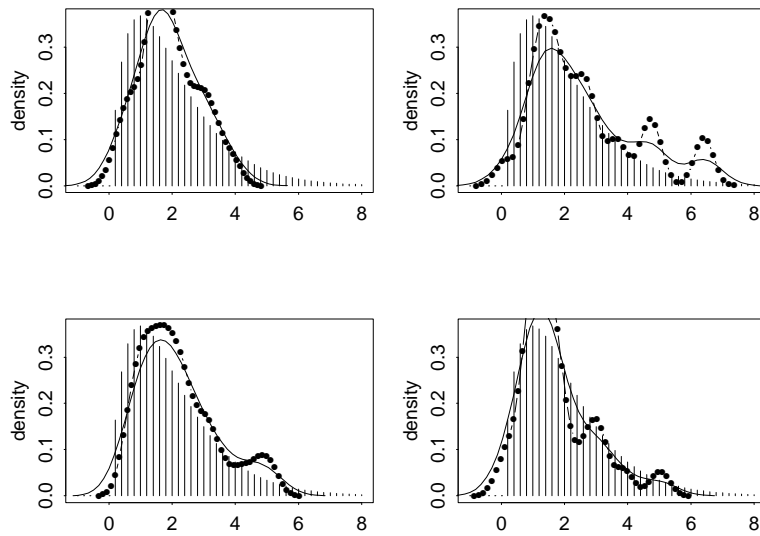


Figure 3: Simulation from exponential distribution with $n = 25$,
columns: exact distribution, dots: using LSCV, solid: using ISE_{cont}

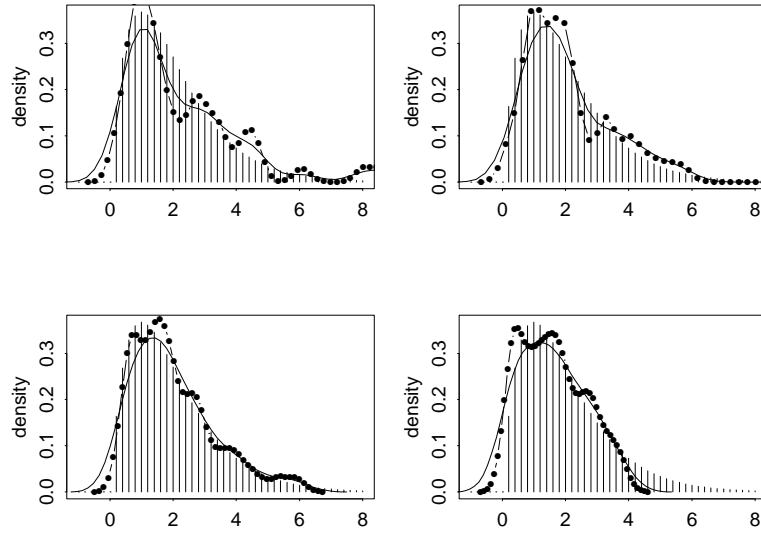


Figure 4: Simulation from exponential distribution with $n = 25$, columns: exact distribution, dots: using LSCV, solid: using ISE_{cont}

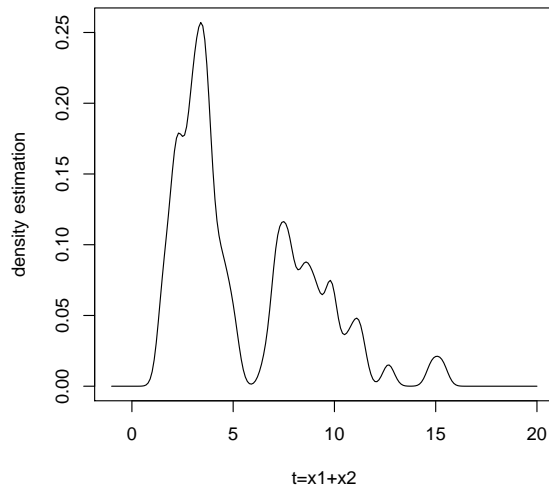


Figure 5: Density estimation for the insurance data using LSCV and

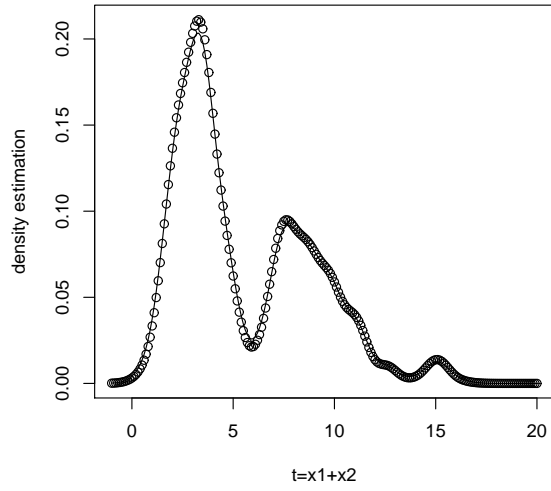


Figure 6: Density estimation for the insurance data using ISE_{cont}

***: $p_1 = 1, p_2 = -1, c_1 = c_2 = 1/2$

—: $p_1 = 1, p_2 = -1, c_1 = 1/3, c_2 = 2/3$

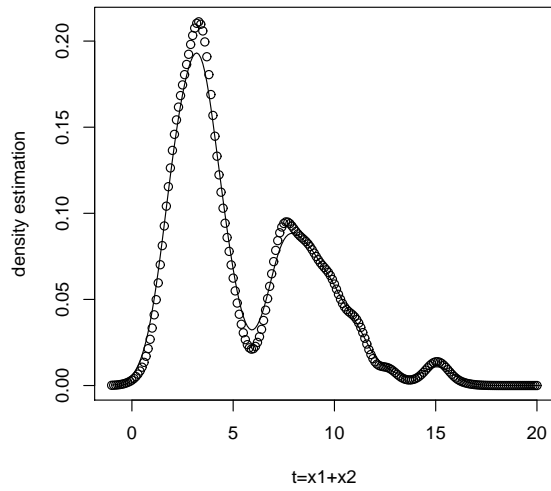


Figure 7: Density estimation for the insurance data using ISE_{cont}

***: $p_1 = 1, p_2 = -1, c_1 = c_2 = 1/2$

—: $p_1 = 1, p_2 = -1, c_1 = 1/10, c_2 = 9/10$

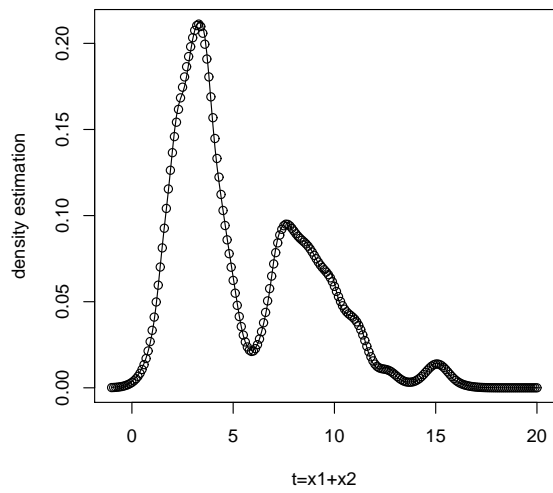


Figure 8: Density estimation for the insurance data using ISE_{cont}

***: $p_1 = 1, p_2 = -1, c_1 = c_2 = 1/2$

—: $p_1 = p_3 = 1, p_2 = p_4 = -1, c_1 = c_2 = c_3 = c_4 = 1/4$