Report on the Accuracy of Ker_LSCV_OUT.R in reproducing Ker_LSCV_OUT.m -Dimensionality/Independence -

Duccio Gamannossi degl'Innocenti

25/01/2016

Overview

In this report it is assessed the accuracy of the R function `Ker_LSCV_OUT.R` in reproducing the MATLAB function `Ker_LSCV_OUT.m` proposed in:

"Optimal bandwidth selection for conditional efficiency measures: a data-driven approach." European Journal of Operational Research 201.2 (2010): 633-640.
by Luiza Badin, Cinzia Daraio and Leopold Simar.

The analysis is focused on the reliablity of the `R` script in the two case where the Outputs are affected or are independent from Environmental Variables. In particular, the investigation is performed using simulated data produced according to the two DGPs presented in the above-mentioned paper. In the First DGP (the Unidimensional one from now on), the vector of Output is independent from a unidimensional vector of Environmental Variables. Conversely, in the second DGP (the Multidimensional one) Outputs are affected by one of the elements of a bi-dimensional vector of Environmental Variables (while the other Environmental Variable still plays no role).

Procedure

The assessment procedure is realized in three steps embodied in the `R` scripts:

1. `validation.R`
2. `validation.m`
3. `validation_assessment.Rmd`

First Step

The script takes as input two parameters: `n`, the number of observations in each sample, and `r`, the number of samples/datasets to be generated. In the present report, `n=100` and `r=100`. Three outputs are generated in this step:

- r `MATLAB` datasets (`validation_r.mat` files) whose data is produced according to the DGPs (both Unidimensional and Multidimensional) described in the simulation paragraph of the above-mentioned paper
- A comma separated values `seeds.csv` , where seeds used in the generation of every dataset are stored to ensure reproducibility

- For every generated dataset, two csv files are produced: `CV_ban_Uni_R_r.csv` and `CV_ban_Mul_R_r.csv`. These files store the Least Squares Cross Validation Error (LSCVE) as obtained by applying `Ker_LSCV_OUT.R` to the Unidimensional and Multidimensional data.

The data stored in the `MATALAB` dataset may be divided in four classes: Input X, Output Y, Environmental Variables Z and Bandwidths. Let us define an observation as a triple (x_{ij}, y_{ij}, z_{ij}) $i \in [1, n]$ $j \in [1, r]$. The function `Ker_LSCV_OUT.R` evaluates the LSCVE of a bandwidth for a given observation i with respect to the sample j. The bandwidth is composed by an univariate bandwidth for the output Y and one univariate bandwidth for every Environmental Variable Z. A set of (multidimensional) bandwidths is generated taking all the possible ordered selections with repetition of unidimensional bandwidths. Unidimensional bandwidths are obtained by multiplying the standard deviation of the corresponding variable (the mean of the standard deviation across all variables for Y) by the following vector of coefficients:

Table 1: Coefficients

0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.5	3

Since an unidimensional bandwidth is adopted for Y, and given that the Unidimensional DGP entails a scalar Z while the Multidimensional one has a two-dimensional Z, the multivariate bandwidth considered are bi-dimensional in the former case and three-dimensional in the latter. Given that the vector of coefficients has nine elements, $9^2 = 81$ and $9^3 = 729$ bandwidths are considered, respectively, for the two DGPs.

Second Step

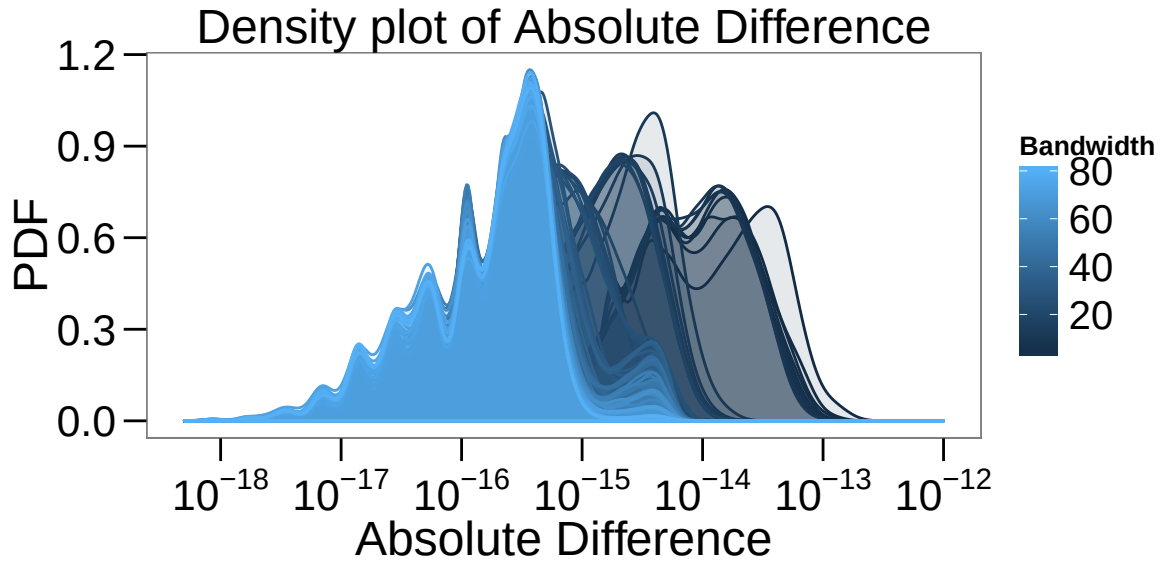
The second script reads in MATLAB the `.mat` files written in the previous step and computes the LSCVEs using `KeR_LSCV_OUT.m`. The output consists in two csv files, `CV_ban_Uni_M_r.csv` and `CV_ban_Mul_M_r.csv`, for every dataset.

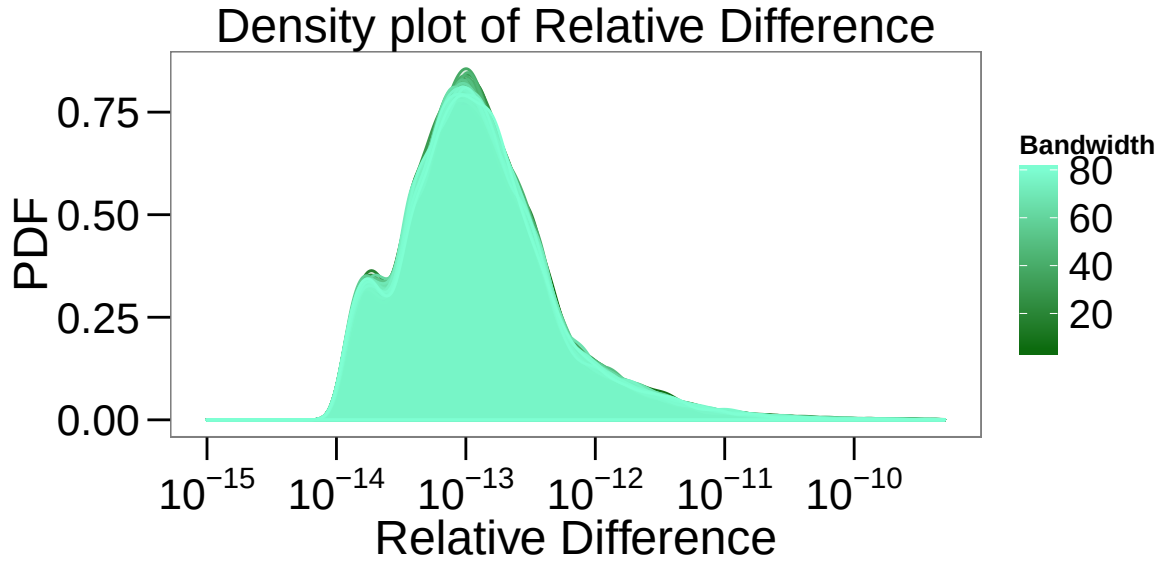
Third Step

In the third step (performed by this script) the LSCVEs obtained by the two functions are compared. The information provided by the present report concerns the possible discrepancies in the values of LSCVEs computed by the two functions in terms of: `NA` elements, `0` elements, Absolute Difference and Relative Difference (the latter one taking as correct the output of `Ker_LSCV_OUT.m`). Furthermore, an assessment of Mismatches in the bandwidths rankings produced by the two functions is provided. The key measures reported are the number of cases where the two functions do not agree on bandwidth ranking, the number of cases in which functions fail to agree on the lowest-LSCVE bandwidth and the difference in the sum of absolute LSCVEs between the sets of lowest-LSCVE bandwidth identified by the two functions.

Accuracy Assessment, Unidimensional DGP

The Unidimensional dataset consists of 100 repetitions of a sample of 100 observations and 81 bandwidths for a total of $81 \times 100 \times 100 = 810000$ observations. The two routines compute a zero LSCVE in the exact same points for 60084 occurrences (the 7.42 % of cases). Furthermore, the number of NAs is 0. The LSCVEs computed by the two functions show a difference lower than 15th digit for 95472 observations (the 11.79 % of cases. The Maximum Absolute Difference has a magnitude of $2.0961011 \times 10^{-13}$, while the Maximum Relative Difference is 9.3363264×10^{-8}. Density plots of the Absolute and Relative difference are displayed below.

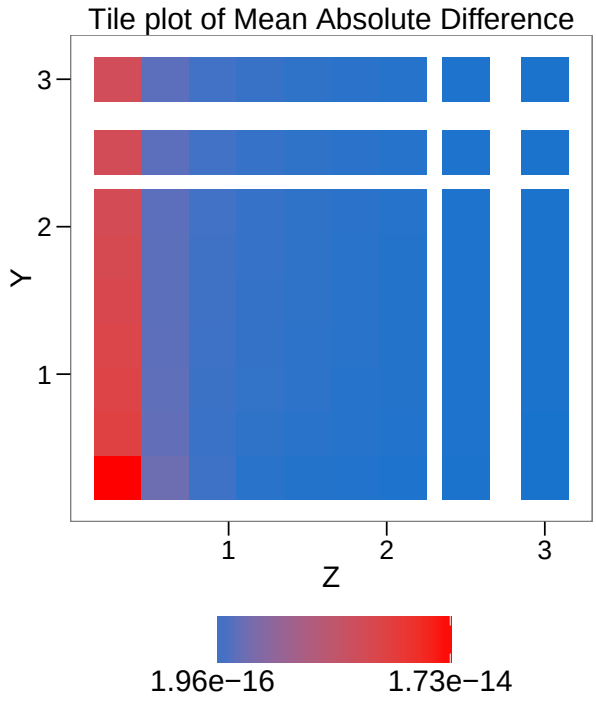


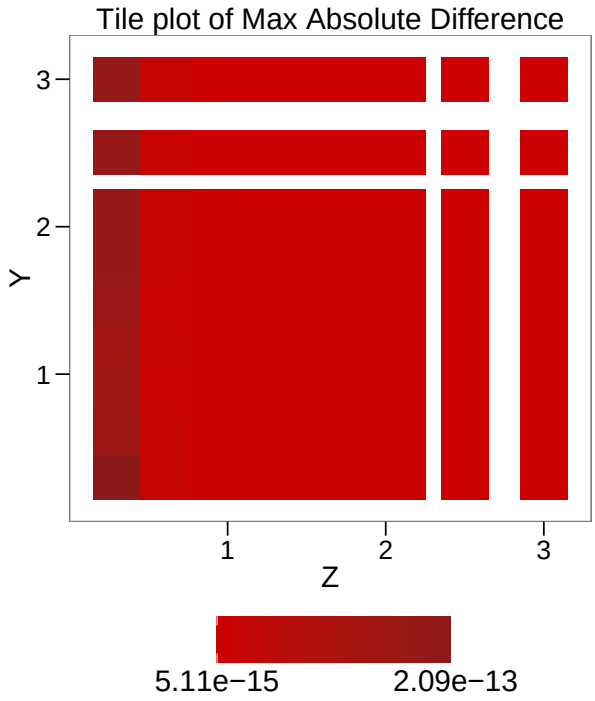


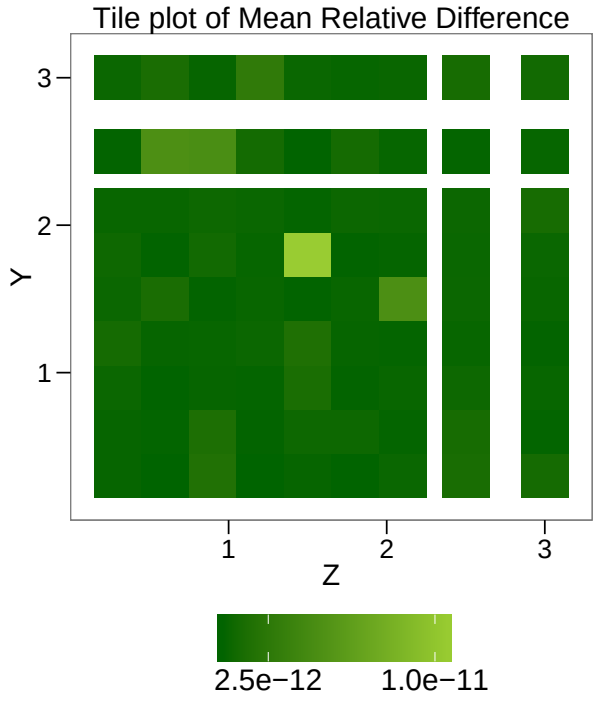
As it can be seen from the graphs, the Absolute Difference varies in a range of 10^{-18} - 10^{-12}, while the

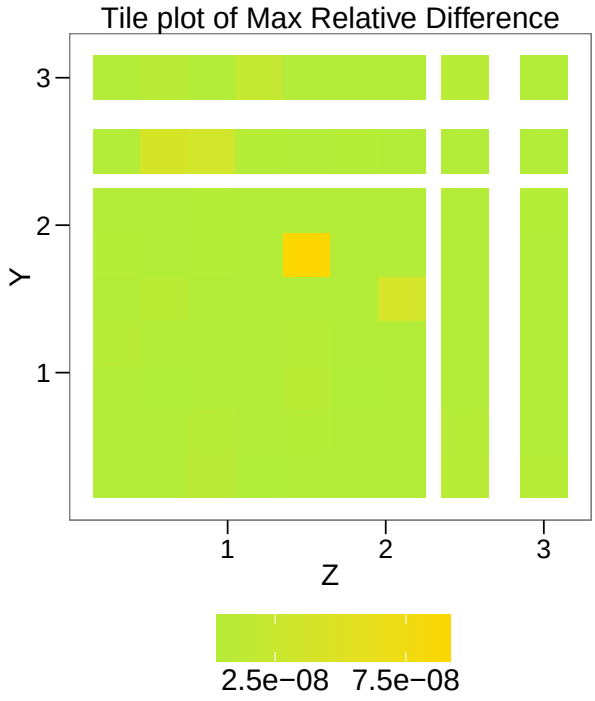
Relative Difference varies in a range of 10^{-15} - 10^{-10}. Notably, while the Absolute Difference shows some variations with respect to the bandwidth, the normalization with respect to the LSCVE almost removes this variability leading to distributions that are essentially overlapped. Hence, the differences between the LSCVEs computed by the two scripts are extremely small both in absolute and in relative terms and their relative impact (on LSCVE computed with `MATLAB` function) tends to be stable across the different bandwidths.

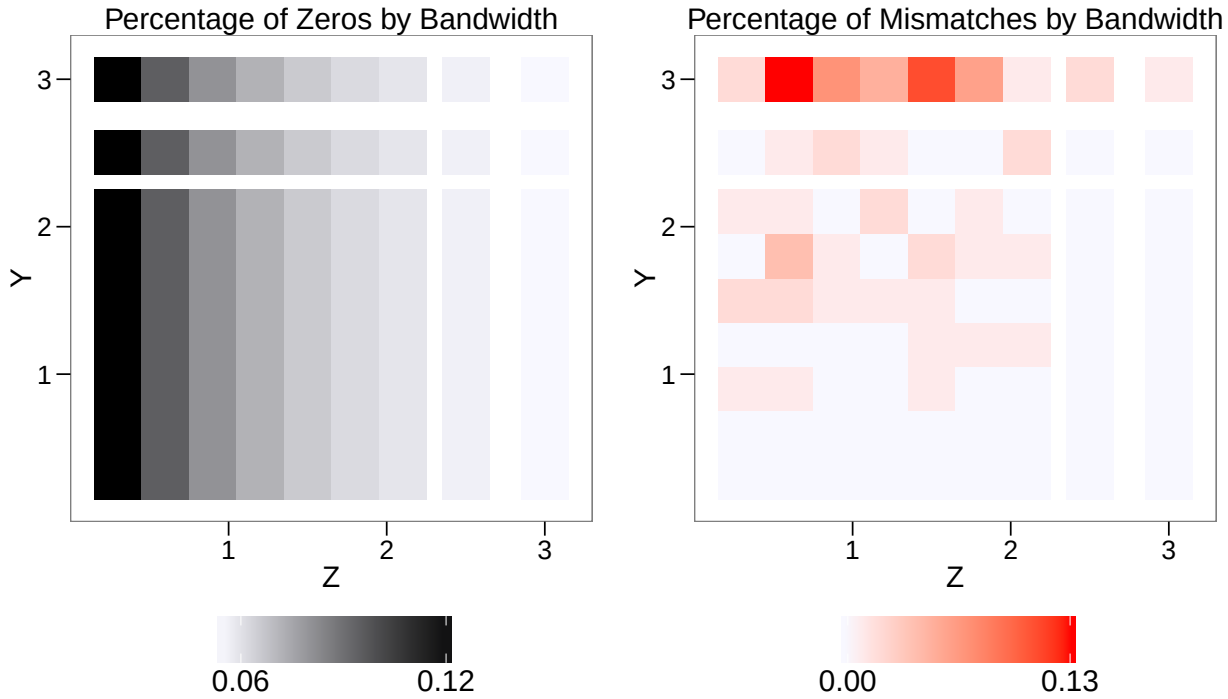
In order to further investigate how the differences between the LSCVEs computed by the two routines vary with bandwidths, some tile plots are reported. In the first and second plot, Medium and Maximum Absolute Difference are considered; as the variable goes from lower to higher values, the colour of tiles change from blue to red. The third and fourth graph show Medium and Maximum Relative Difference; a growth in the two variables is represented in the plot with a transition of the tile colour from green to yellow. In the fifth tileplot an increase in the relative percentage of Zero-differences is represented by darker tiles, while in the sixth a higher relative Percentage of Mismatch in the lowest-LSCVE bandwidth leads to a variation of the colour of the tile from white to red. The fifth plot provides information on the number of observations where the two functions LSCVEs have a non-measurable difference and, hence, on the number of Relative Difference observations for every bandwidth. The sixth graph is crucial for the purposes of the present report. In fact, in an applicative setting the function `Ker_LSCV_OUT` is used to identify the bandwidth entailing the lowest Least Squares Cross Validation Error for an observation. Hence, when evaluating a set of bandwidths on a sample, its outcome is intended to be the set of bandwidths that minimizes the LSCVE according to a certain criterion (in the present report, two criteria are considered: the minimum mean LSCVE across observations and minimum-LSCVE for every observation). Hence, any meaningful measure of the correctness of the `R` function has to assess the impact of the cases (if any) where it fails to select the same minimum-LSCVE bandwidth identified by the `MATLAB` function. From this point of view, Absolute and Relative Difference may be interpreted as proxies of the validity of the `R` function, since directly related to bandwidth rankings. However, the information provided by these measures is not enough to infer nor changes in the rank of bandwidths nor changes of minimum-LSCVE bandwidths.


Tile plot of Mean Absolute Difference
Y
Z
1
2
3
1.96e−16
1.73e−14


Tile plot of Max Absolute Difference
Y
Z
1
2
3
5.11e−15
2.09e−13


Tile plot of Mean Relative Difference
Y
Z
1
2
3
2.5e−12
1.0e−11


Tile plot of Max Relative Difference
Y
Z
1
2
3
2.5e−08
7.5e−08



Since the univariate bandwidths are computed by multiplying the standard deviation of the variable times a coefficient, a bijection relates the set of multivariate bandwidths with the set of coefficients vectors. Hence, every tile of the tileplots represents a multivariate bandwidth and is identified by the coefficient used to generate its univariate components. In the first plot, it can be noted that the Mean Absolute Difference increases the lower the coefficients of both Y and Z. A similar pattern, although less marked, is displayed in the second plot, the one of Maximum Absolute Difference. The third and fourth plot show an almost homogeneous distribution of Mean and Maximum Relative Differences with respect to coefficients and their values range in the intervals $10^{-12} - 10^{-11}$ and 10^{-8} respectively. The relative percentage of Zero valued LSCVEs tends to decrease the higher it is the coefficient of Z while seems to be not affected by the coefficient of Y; hence, the number of the observations for the Relative difference increases with the Z coefficient. Nevertheless, the latter effect is almost negligible since the maximum value of the percentage of Zero is just 0.12%. Finally, the tileplot of Mismatches shows an increasing gradient in Y and a slightly decreasing one in Z. Notably, a remarkable part of the Mismatches affects the bandwidths with maximal coefficient in Y. However, the magnitude of these mismatches is low and the maximum one is 0.13%.

Assessment of mismatches in bandwidth ranking

In this section we provide an assessment of the mismatches between the rankings of bandwidths induced by the two functions. To assess the similarity of the rankings, Pearson's correlation coefficient, Spearmans' correlation coefficient, Kendall's rank correlation coefficient and Dot Product (numerator of Cosine Similarity) are provided. Pearson's correlation coefficient is a measure of linear dependance of two variables, while Spearman's correlation coefficient measures the extent to which both variables can be described by monotonic function. Kendall coefficient is increasing in the relative number of concordant pairs among the vectors and decreasing in the discordant ones. Finally, cosine similarity is a geometrical distance of the two vectors in the n-dimensional Euclidean space. It measures the angle between the two vectors when both have the origin as initial point and can be shown to be equivalent to Pearson's correlation coefficient for centered data. In the following, the above-mentioned coefficients will be referred to as CRS (Coefficients of Rank-Similarity). In case of mismatches between the rankings induced by the two functions, further checks are performed.

The rankings follow two criteria: the minimum mean LSCVE error across observations and the minimum

LSCVE for every observation. The minimum mean LSCVE criterion ranks the bandwidths of a given sample based on the mean error across observations. First, mean LSCVE are computed for every bandwidth in every sample and, second, for every sample the bandwidth are ranked in ascending order from the one with minimum mean-LSCVE. Third, CRS of the rankings induced by the two functions are evaluated for every sample to measure the magnitude of the mismatches. Finally, the mean of the CRS across all the samples is produced

Table 2: Mean CRS across all observation in every sample

Pearson	Spearman	Kendall	Dot Prod.
1	1	1	1

The table above shows that the two functions generate identical rankings when the minimum mean LSCVE criterion is adopted. Conversely, when rankings of the bandwidths are computed observation-wise, some mismatches arise. Following the latter criterion, bandwidths are ranked in ascending order starting from the one entailing the lowest LSCVE on the single observation. To provide a measure of the entity of the mismatches, the mean of the CRS across every observation is reported.

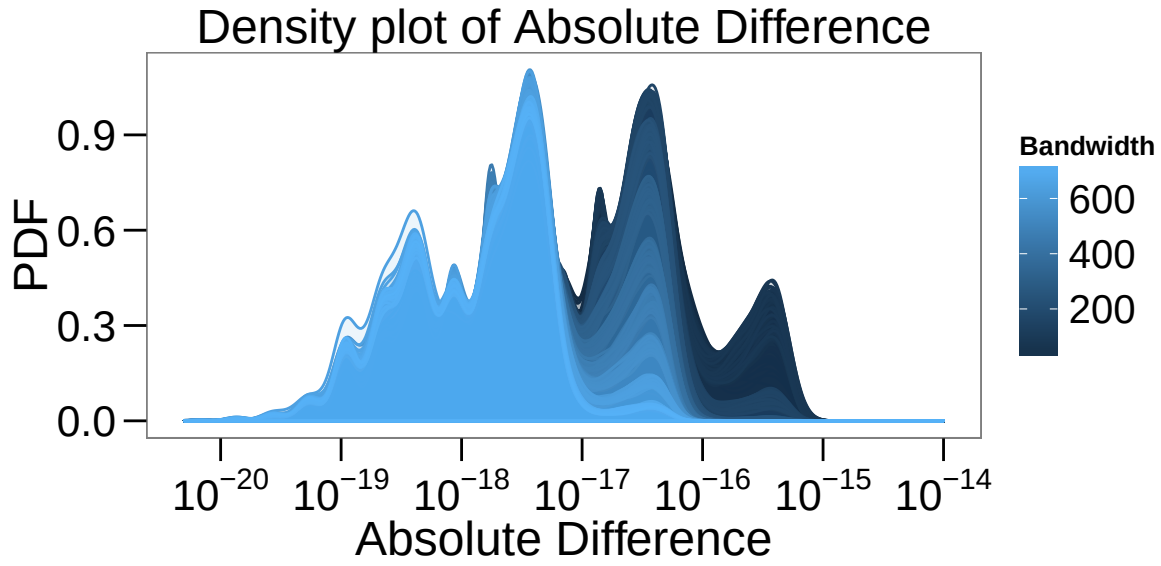
Table 3: Mean CRS of every observation

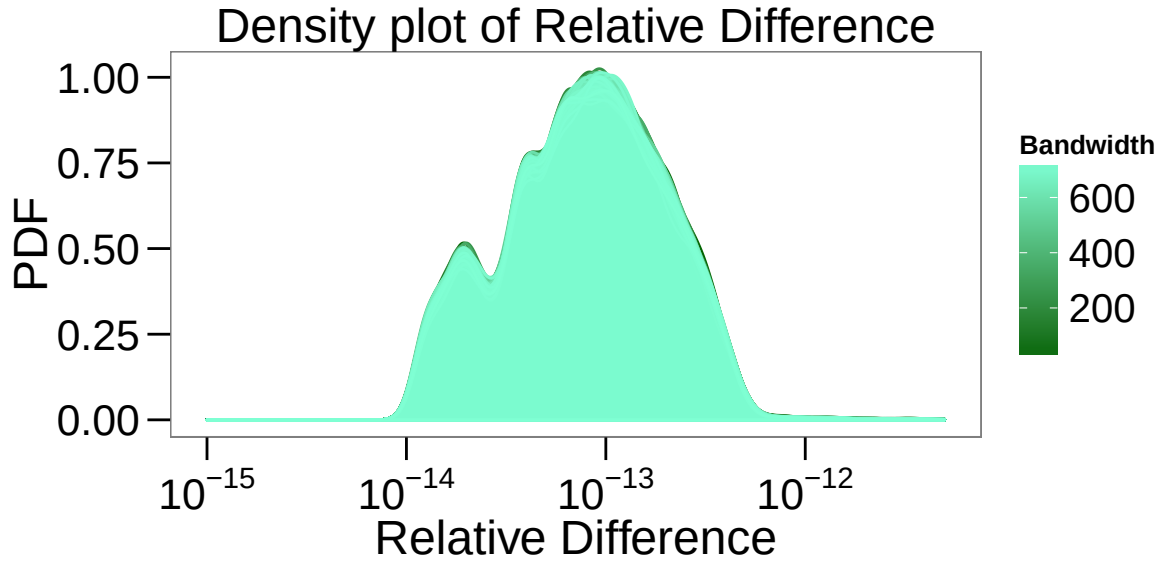
Pearson	Spearman	Kendall	Dot Prod.
0.999842	0.999842	0.998339	0.999961

The relative percentage of mismatch is 8.17% while the mismatches affecting the best bandwidth (the one with lowest LSCVE) are the 0.008% with an impact of the latter on the former of 0.099%. To measure the impact of mismatches, the sum of absolute differences between the set of minimum-LSCVE bandwidths determined by the two functions is computed:$3.1740684 \times 10^{-13}$. In percentage terms, the impact of the latter variable on the total error of the minimum-LSCVE bandwidths set (computed using the `MATLAB` function) is $9.1739011 \times 10^{-14}$%. Thus, the difference in the LSCVE incurred when using `Ker_LSCV_OUT.R` instead of `Ker_LSCV_OUT.m` is negligible when considering data from the unidimensional DGP.

Accuracy Assessment, Multidimensional DGP

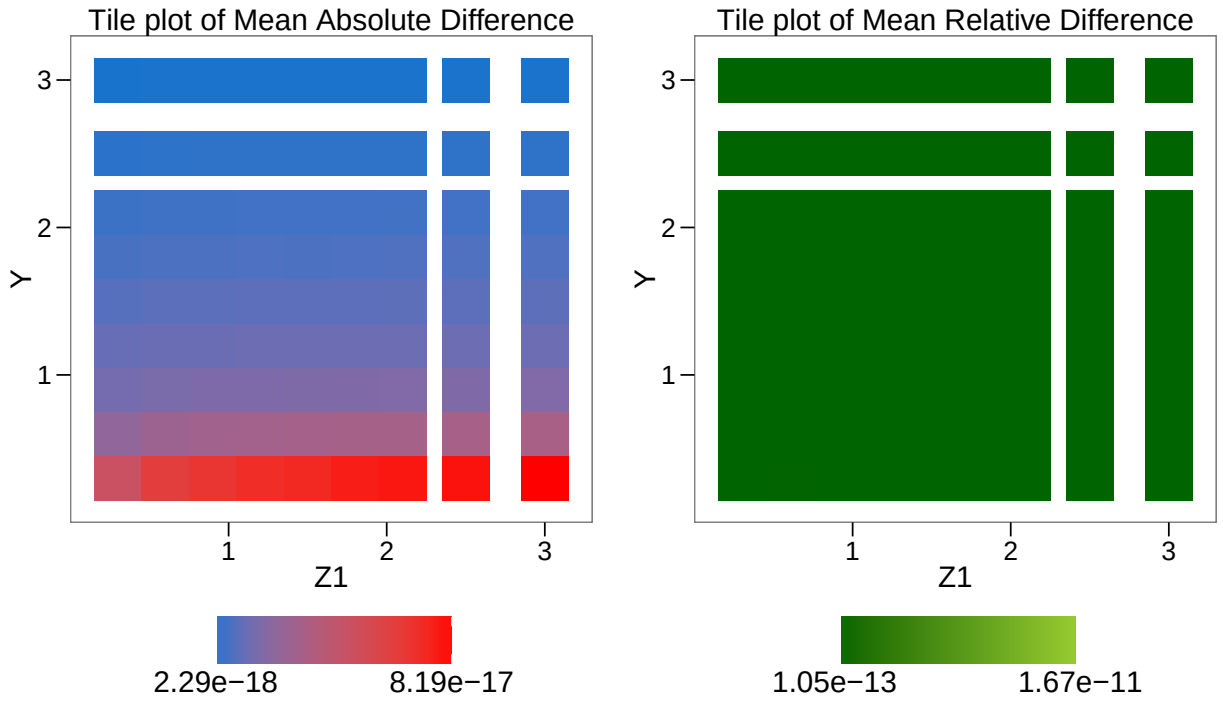
The Multidimensional dataset consists of 100 repetitions of a sample of 100 observations and 729 bandwidths for a total of $729 \times 100 \times 100 = 7290000$ observations. The two routines compute a zero LSCVE value in the exact same points (as shown by diffZeroMul=0), for 729036 occurrences (the 10 % of cases). Furthermore, the number of NAs is 0. The LSCVEs computed by the two functions show a difference lower than 15th digit for 1119427 observations (the 15.36 % of cases). The Maximum Absolute Difference has a magnitude of $1.1657342 \times 10^{-15}$, while the Maximum Relative Difference is 1.4979294×10^{-7}. Density plots of the Absolute and Relative Differences between LSCVEs computed with the `MATLAB` and `R` functions are displayed below.

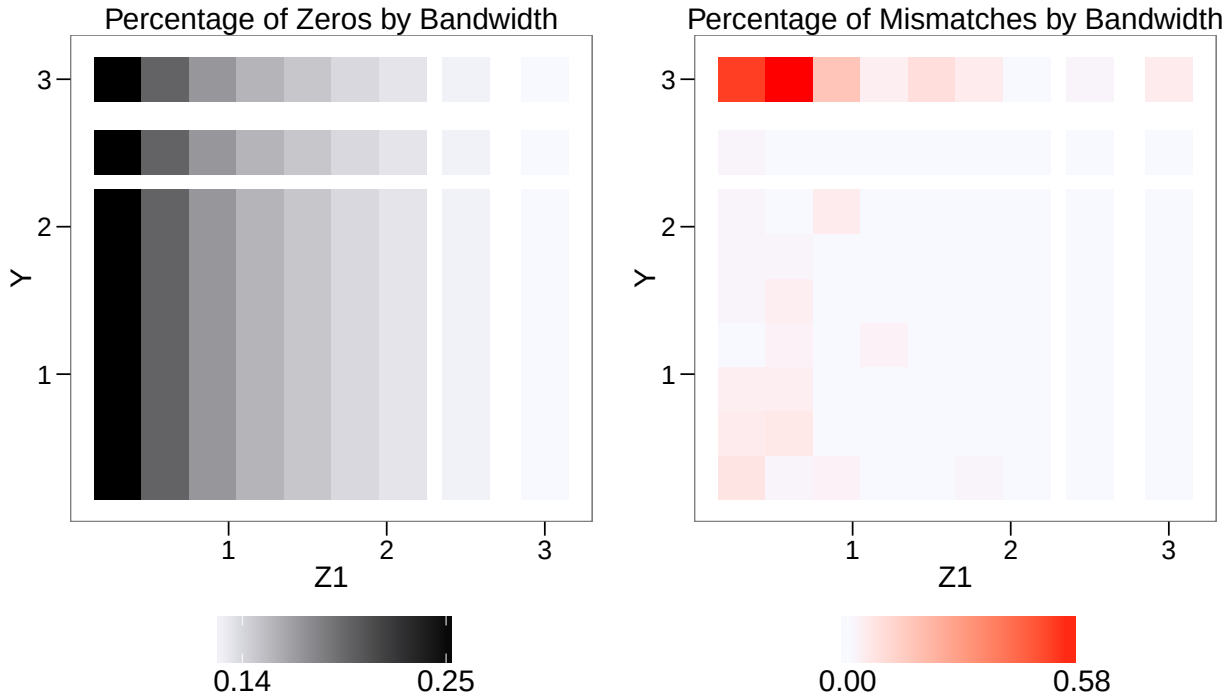




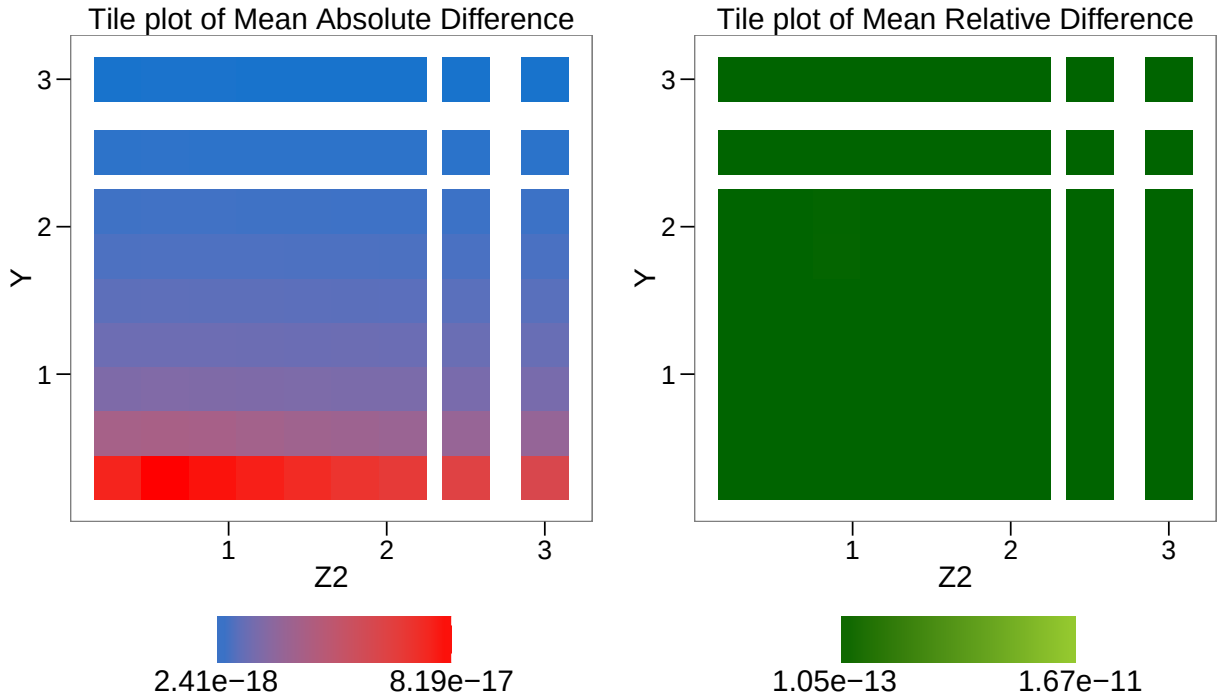
As it can be seen from the graph, the Absolute Difference varies in a range of 10^{-21} - 10^{-14} while Relative

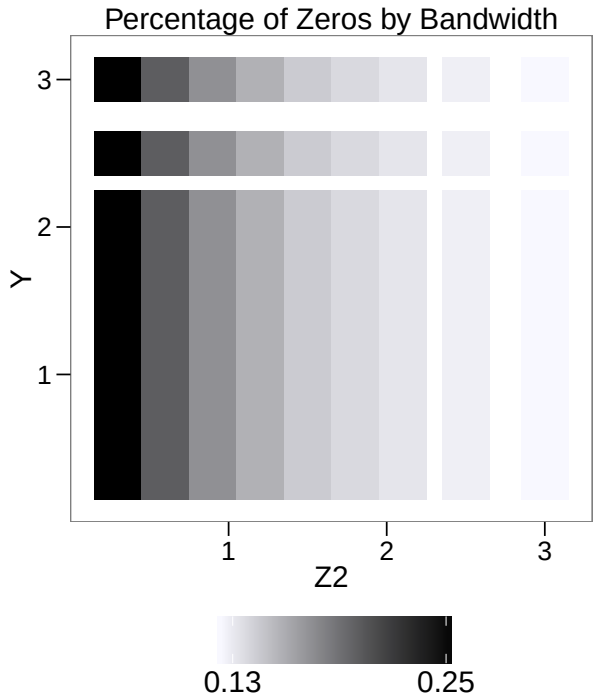
Difference varies in the range 10^{-15} - 10^{-12}. What has been noted in the Unidimensional case holds in the Multidimensional one: even if the Absolute Difference shows some variations with respect to the bandwidth used to compute the LSCVE, the normalization with respect to the LSCVE almost removes this variability. Moreover, also in this case the differences between the LSCVE computed by the two routines are extremely small both in absolute and in relative terms and their relative impact on the LSCVE tends to be stable across the different bandwidths.
Closely following the Unidimensional outline, some tile plots are reported to provide more information regarding the role of the bandwidths on the differences between the LSCVEs computed by the two scripts. Given that the bandwidths of this section are three dimensional, the analysis takes into account all the possible two-dimensional interactions among the variables one at time. To remove the third dimension from the graphical representations, Mean, Minimum and Maximum values have been considered with analogous results. Here are reported the results obtained using Maximum values with respect to the omitted dimension. Furthermore, given that the behaviour of Maximum Absolute and Maximum Relative differences closely resembles the one of their Mean counterpart, only the latter is displayed.

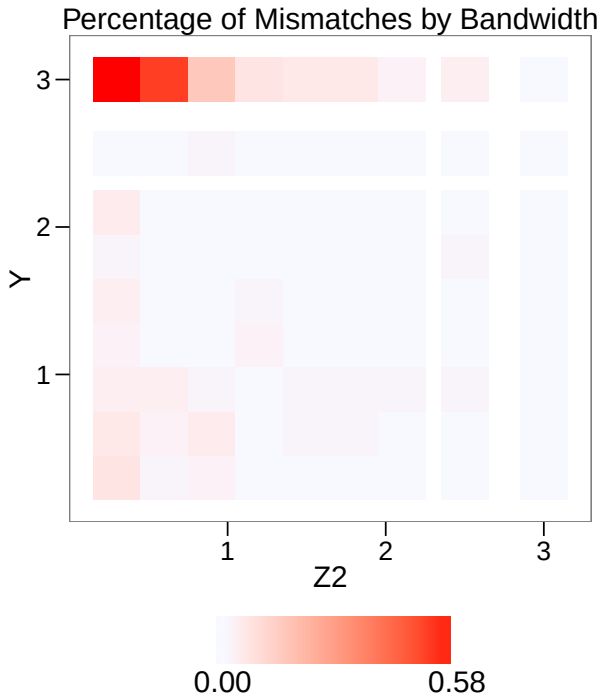




The tileplot of Mean Absolute Difference shows a stronger decreasing gradient in Y while a milder increasing one can be noticed in $Z1$. Mean Relative Difference are homogeneously distributed with respect to bandwidth and are in the order of $10^{-13} - 10^{-11}$ ($10^{-12} - 10^{-7}$ for the Maximum). The third plot shows a marked decreasing gradient in $Z1$ while Y seems to have no effect. However, the relative percentage of Zeros is never higher than 0.16%, so the representativity of Relative Differences measures holds. Finally, the Mismatch tileplot shows a decreasing gradient with respect to $Z1$ while an higher Percentage of Mismatches occurs for extreme values of the bandwidth of Y. Nevertheless, the Percentage of Mismatches does not raise above the 0.11%

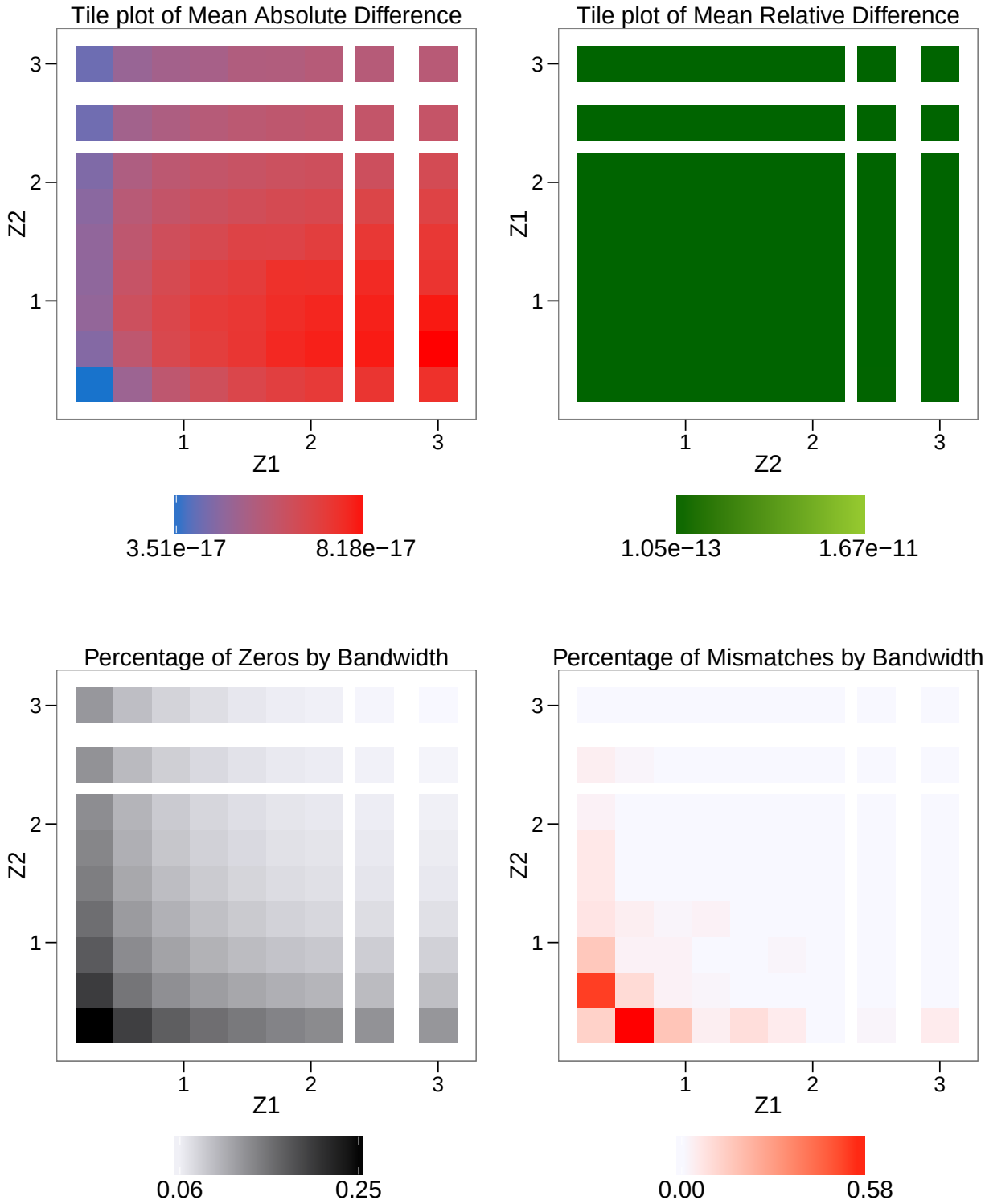






Medium Absolute difference shows a similar pattern to the one above. The decreasing gradient in Y still holds while $Z2$ seems to have no impact. Relative Difference graph almost reproduces the plots of Y $Z1$ with an homogeneous distribution with respect to bandwidths and analogous ranges of variation. The relative percentage of Zeros displays a marked decreasing gradient in $Z2$ while seems to be unaffected by Y. Since the maximum value of percentages of Zero is 0.16%, the representativity of Relative Differences measures

holds in this case also. Finally, the Mismatch tileplot shows a decreasing gradient with respect to $Z2$, while a higher percentage of mismatches occurs for extreme values of the bandwidth of Y.



The tileplot of Mean Absolute Difference relative to $Z1$,$Z2$ displays an increasing gradient in the former dimension and a decreasing one in the latter. Again, Relative Differences are homogeneously distributed with respect to the bandwidths and have analogous ranges of variation with respect to the previous cases. Finally,

the percentage of Zeros and the one of Mismatches decreases along the two dimensions of Z.

Assessment of mismatches in bandwidth ranking

In this paragraph we report an assessment of the mismatches in the rankings of bandwidths analogous to the one performed in the Unidimensional case.

When considering a minimum mean LSCVE criterion to rank the bandwidths, the rankings generated by the two functions are identical.

Table 4: Mean CRS across all observation in every sample

Pearson	Spearman	Kendall	Dot Prod.
1	1	1	1

When rankings of the bandwidths are computed observation-wise, some mismatches arise.

Table 5: Mean CRS of every observation

Pearson	Spearman	Kendall	Dot Prod.
0.999918	0.999918	0.99867	0.99998

The relative percentage of mismatch is 28.8%, more than three times the one of the Univariate case (likely due to the increased number of bandwidths). The mismatches affecting the best bandwidth (the one with lowest LSCVE) are the 0.027% with an impact of 0.094% In this setting, the sum of Absolute Differences between the set of minimum-LSCVE bandwidths is : $4.9958093 \times 10^{-14}$. The latter value has an impact of $3.2693351 \times 10^{-13}$% on the sum of absolute LSCVE of the best bandwidth set determined with the `MATLAB` function. Hence, also in the case of Multidimensional DGP, the difference in the LSCVE incurred when using `Ker_LSCV_OUT.R` instead of `Ker_LSCV_OUT.m` is negligible.

Conclusions

Absolute and Relative Difference Maximums are both computed in the univariate case and are of the order of 10^{-13} and 10^{-8}, respectively. Conversely, the highest percentage of mismatches in ranking occurs in the Multidimensional case: 28.8%; much higher than its univariate counterpart 8.17%. The cause of this loss in performance is likely to be related to the increased number of bandwidths considered. Nevertheless, it should be taken into account that the mismatches affecting the lowest-LSCVE, although higher than for the unidimensional data, are still very low: 0.027% . To sum up, two measures are proposed as indicative of `Ker_LSCV_OUT.R` :

- The maximum Percentage of Mismatches for a single bandwidth (occurring in the Multidimensional case) is 0.58%.
- The relative impact of the increase in absolute sum of LSCVE incurred when using `Ker_LSCV_OUT.R` instead of `Ker_LSCV_OUT.m` is in the range of 10^{-13}% (again, occuring in the Multidimensional case).

Hence, the reported evidence strongly supports the validity of the function `Ker_LSCV_OUT.R`.