Kernel Smoothing Toolbox *
for MATLAB

Contents

*Jan Koláček, Dept. of Mathematics and Statistics, Masaryk University, Kotlářská 2, Brno, Czech Republic, kolacek@math.muni.cz, available on *http : //math.muni.cz/ ∼ kolacek/docs/kerns.zip*

1 Kernel regression

1.1 Start menu

The *Start menu* (Figure 1) for kernel regression is called up by command

$>> ksregress$

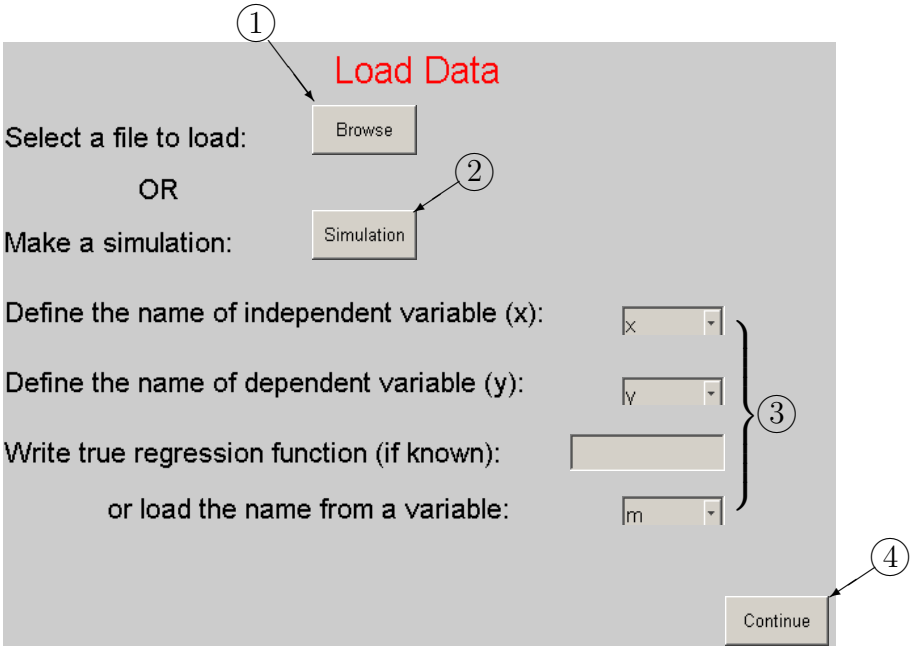


Figure 1: *Start menu*

You can skip this menu by typing input data as an argument
$>> ksregress(x, y)$;
where vectors x and y should have the same length n and they mark x and y axes of measurements. If we know also the right regression function f (for example for simulated data), we can set it as the next argument. For more see $>>$ *help ksregress*.
After execution of this command directly the window on Figure 3 is called up.
In the *Start menu*, you have several possibilities how to define input data. You can load it from a file (button ①) or simulate data (button ②). In fields ③ you can list your variables in current workspace to define input data. If your workspace is empty, these fields are non-active. If you know the true regression function of the model, you can write it to the text field or load it from a variable. If you need to simulate a regression model, press button ② . Then the menu for simulation (Figure 2) is called up.
In the *Simulation menu*, as the first set the regression function. You can write it to the text field ⑤ or load it from a variable. In fields ⑥ specify the interval, number of design points and the variance. You can simulate a regression model by pressing button ⑦ . Then you can save the data to variables or as a file by using buttons ⑧ . If you

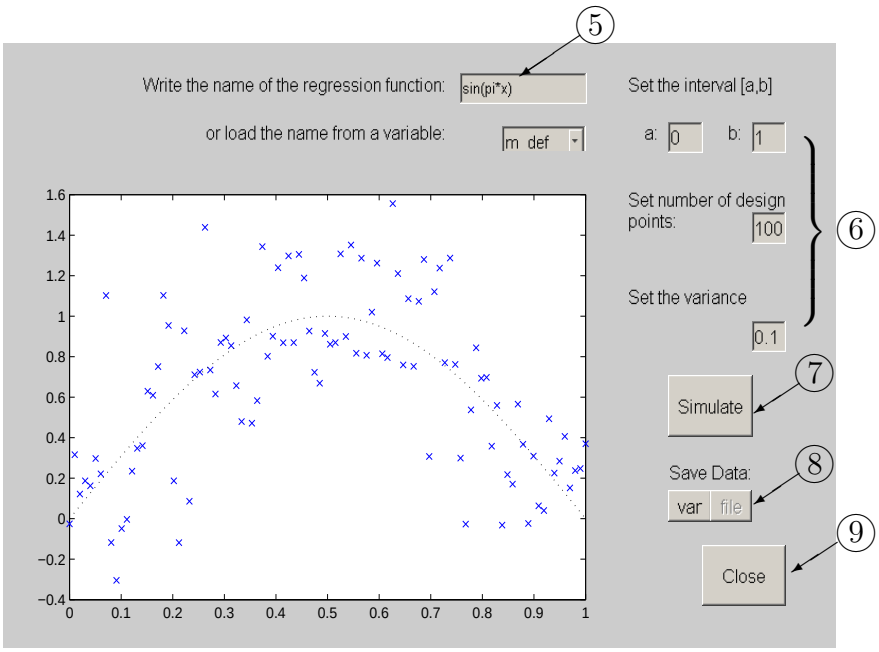


Figure 2: *Simulation menu*

have done the simulation, press button ⑨ . The *Simulation menu* will be closed and you will be returned to the *Start menu*. In this menu, you can redefine the input data. If you want to continue, press button ④ . The menu will be closed and the *Basic menu* (see the next subsection) will be called up.

1.2 Basic menu and setting of parameters

This menu (Figure 3) was called up from the *Start menu* or directly from the command line (see >> *help ksregress*). The values of independent variable x are automatically transformed to interval $[0, 1]$. Symbols $\times$ mark measurements after this transformation. If you want to show the original data, use button ⑪ . The button ⑫ ends the application. Use button ⑩ to continue. Other buttons are non-active.

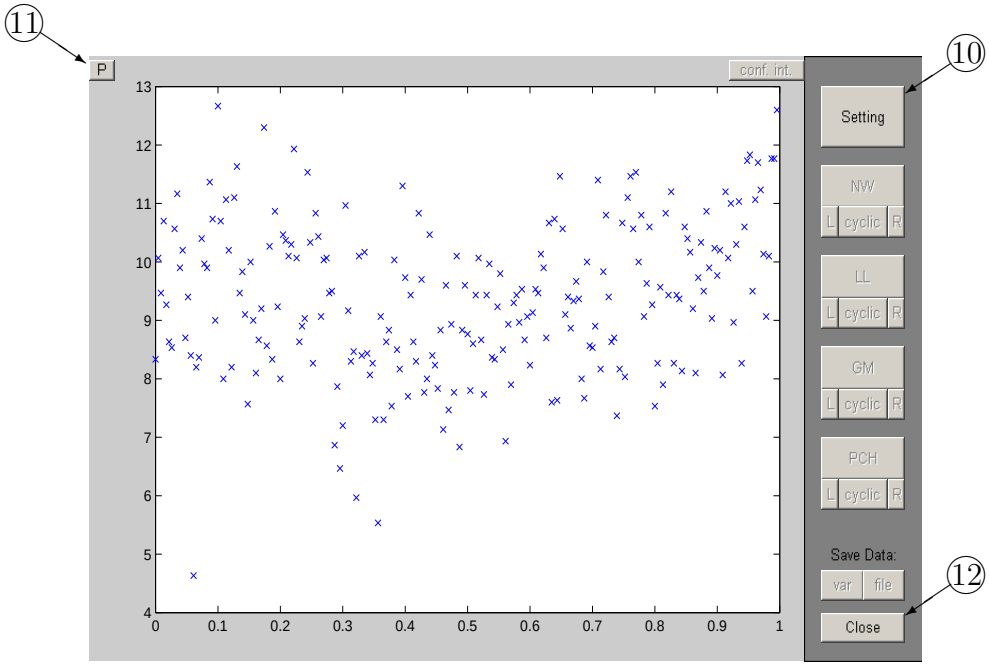


Figure 3: *Basic menu*

The button ⑩ calls up the menu for setting of parameters, which will be used for kernel regression:

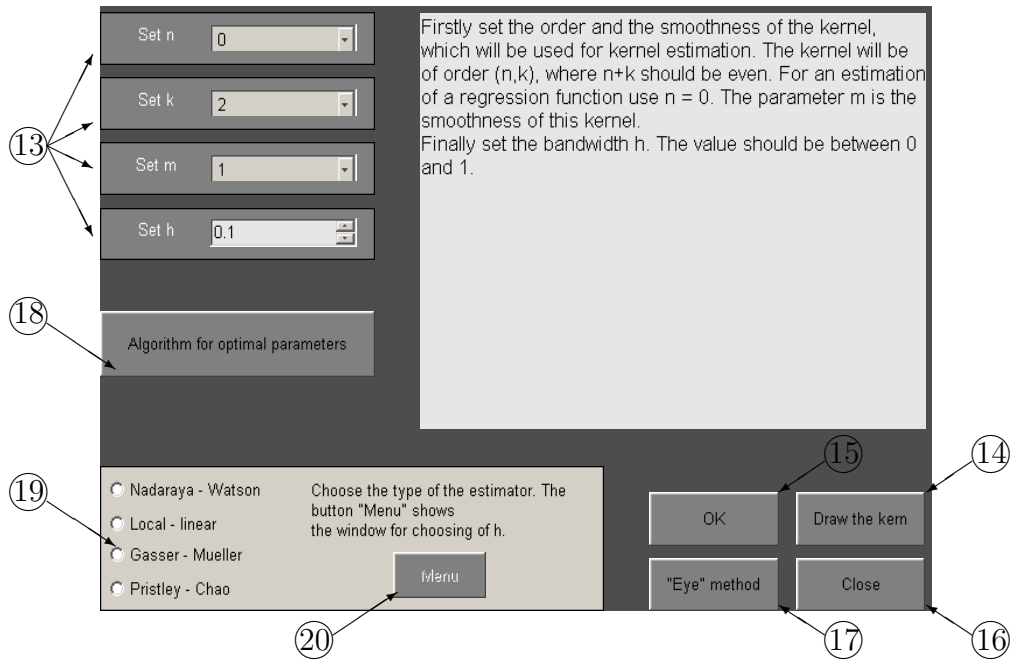
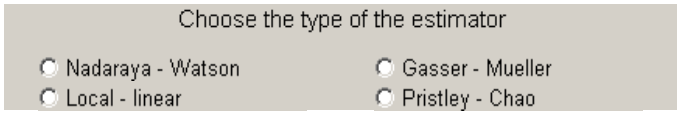


Figure 4: *Setting of parameters*

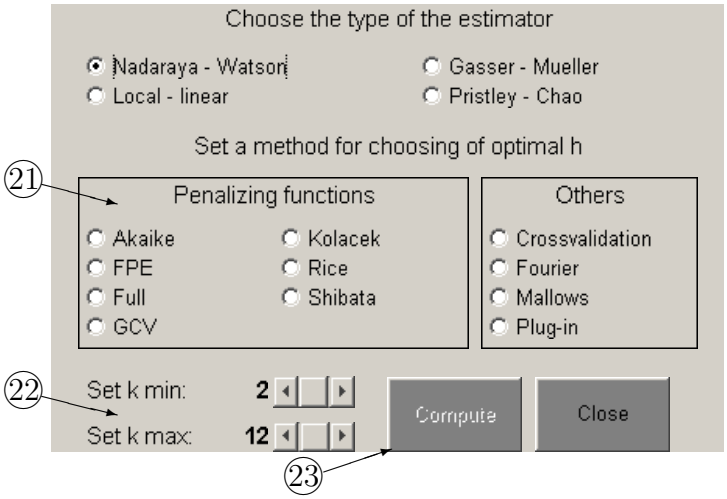
In the array ⑬ , there we can set kernel regression parameters. As the first, set the order of the kernel (n, k), where $n + k$ should be even, for regression estimation $n = 0$ is used. The parameter m is the smoothness of this kernel. If you want to draw the kernel, use button ⑭ . Finally set the bandwidth h. The value should be between 0 and 1. To confirm the setting use ⑮ , to close the window, use ⑯ . The others buttons apply to a choosing of optimal bandwidth h.
The button ⑰ calls up the *"Eye-control" menu* (see page 7), where we can change the value of h and observe the effect upon the final estimate. The button ⑱ starts the algorithm for estimation of optimal kernel order and optimal bandwidth h (see page 6). This algorithm automatically sets the values of optimal parameters in array ⑬ . By selecting one type of kernel estimators in ⑲ , you make active ⑳ . This button calls up the menu for using and comparing of various methods for choosing the optimal smoothing parameter h (see page 8).

1.3 Estimation of optimal parameters

The button ⑱ calls up an algorithm for estimation of optimal order of kernel and optimal bandwidth. Firstly, it is necessary to set the type of kernel estimator:



As the next is a menu for estimation of optimal bandwidth called up:



By choosing one of methods in array ㉑ we make active the button ㉓ , which starts the computation of optimal parameters k and h (we suppose $n = 0, m = 1$). In array ㉒ , there we can set limits for parameter k, default values are $k_{min} = 2$, $k_{max} = 12$. Results of the computation are automatically set in array ⑬ .

1.4 "Eye-control" method

The button ⑰ calls up a window, where we can change the value of parameter h and observe the effect of these changes upon the final estimate:

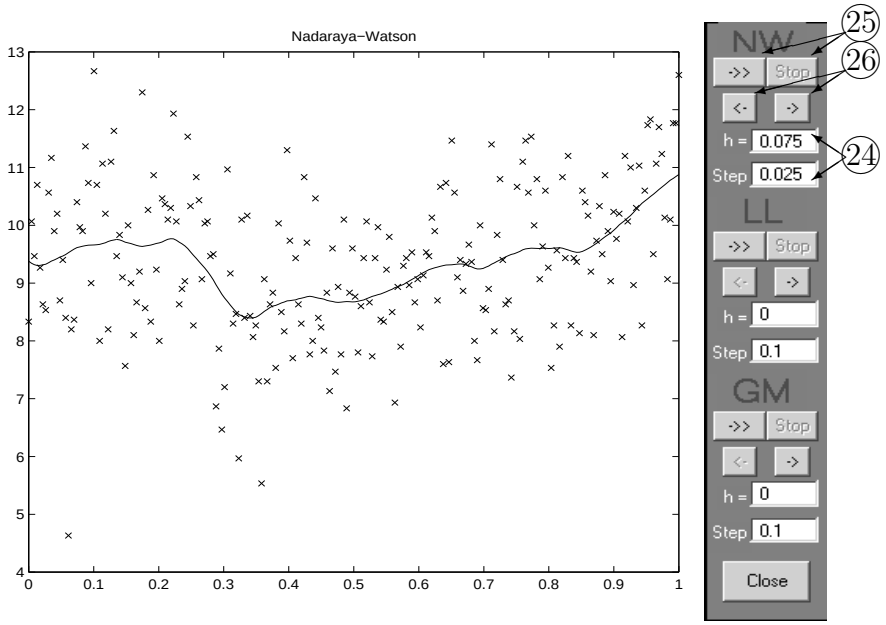


Figure 5: *"Eye-control" menu*

In arrays ㉔ , set the starting value of parameter h and the step (it can be positive or negative number) for the size of changes of h. The left button ㉕ starts a sequence of pictures representing the quality of estimation dependence on h. The right button stops the sequence. You can change the value of h only one step more or less by buttons ㉖ .

1.5 Comparing of methods for bandwidth selection

The button ⑳ calls up a window for using and comparing of various methods for optimal bandwidth selection:

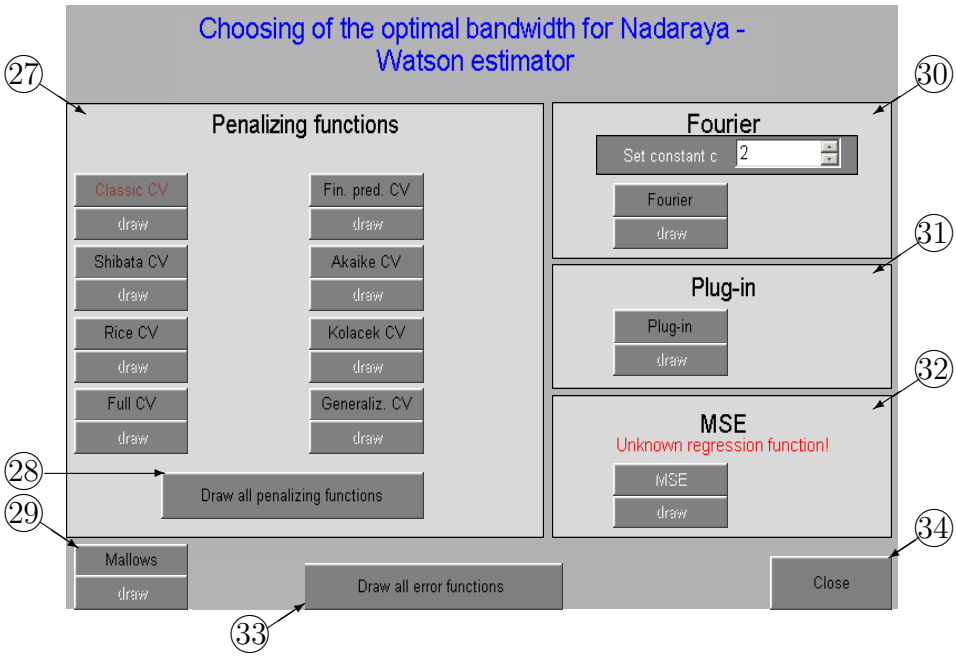


Figure 6: *Methods for optimal bandwidth estimation*

In this window, there are all studied methods for choosing the optimal bandwidth. In array ㉗ , there are presented all penalizing functions and also the cross-validation method (Classic CV). By click on the button with the method's name, the optimal bandwidth is computed by this method. The button "draw" calls up a graph of the minimized error function. To draw all penalizing functions applied up to this time use ㉘ . The button ㉙ represents Mallows method, the button ㉚ marks the method of Fourier transformation and ㉛ marks the plug-in method. If we know the right regression function (for example for simulated data), we can compute the theoretical value of optimal bandwidth as a minimum of the Mean Square Error MSE in array ㉜ . To do this computation, it is necessary to be the *Symbolic Toolbox* installed on your computer. If this toolbox is not installed or if we don't know the regression function, the array is not active. For the graph of all error functions and their minimal values use ㉝ . The button ㉞ closes the application.

1.6 The final estimation of the regression function

If you set all values in window for setting parameters (see page 5) and if you want to go to the final estimation of the regression function, confirm your setting by the button ⑮ . It calls up the *Basic menu*, where all buttons are active already:

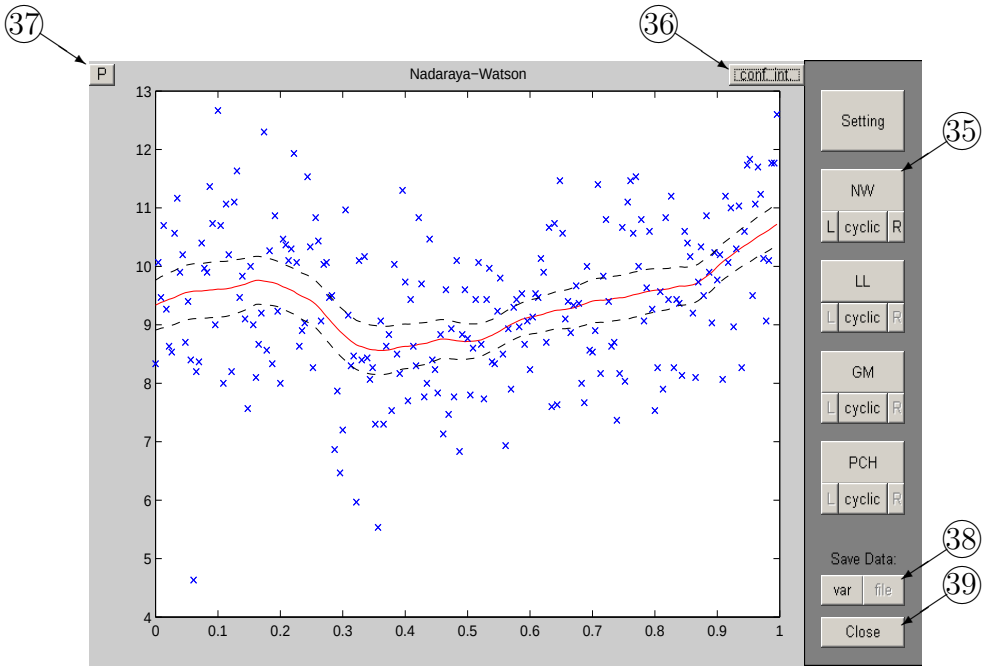


Figure 7: *Final kernel estimate*

By click on the button with the estimator's name (for example ㉟ for Nadaraya–Watson estimator) is drawn the relevant regression estimate (solid line in the figure). The button "cyclic" shows the regression estimate with using the assumption of cyclic model. By using buttons "L" and "R" we get the estimate on the boundary of the interval obtained by using special boundary kernels (L=left, R=right). The button ㊱ draws confidence intervals (dashed). To do this computation, it is necessary to be the *Stats Toolbox* installed on your computer. If this toolbox is not installed, the button is not active. The button ㊲ shows original data and the final estimate (see page 10). You can also save the data to variables or as a file by using buttons ㊳ . The button ㊴ ends the all application.

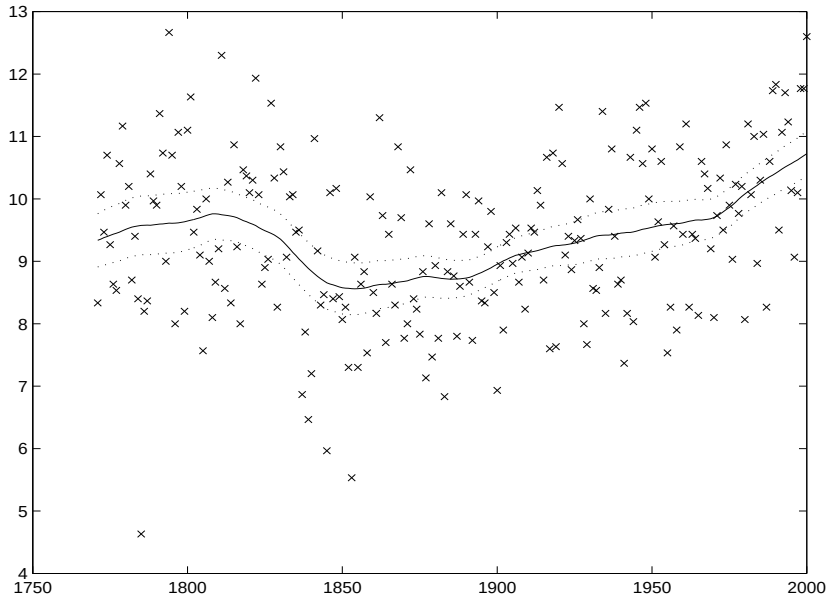


Figure 8: *Original data and the final kernel regression estimate*

2 Kernel quality indexes

2.1 Start menu

The *Start menu* (Figure 9) for kernel estimation of quality indexes is called up by command

$>> ksquality$

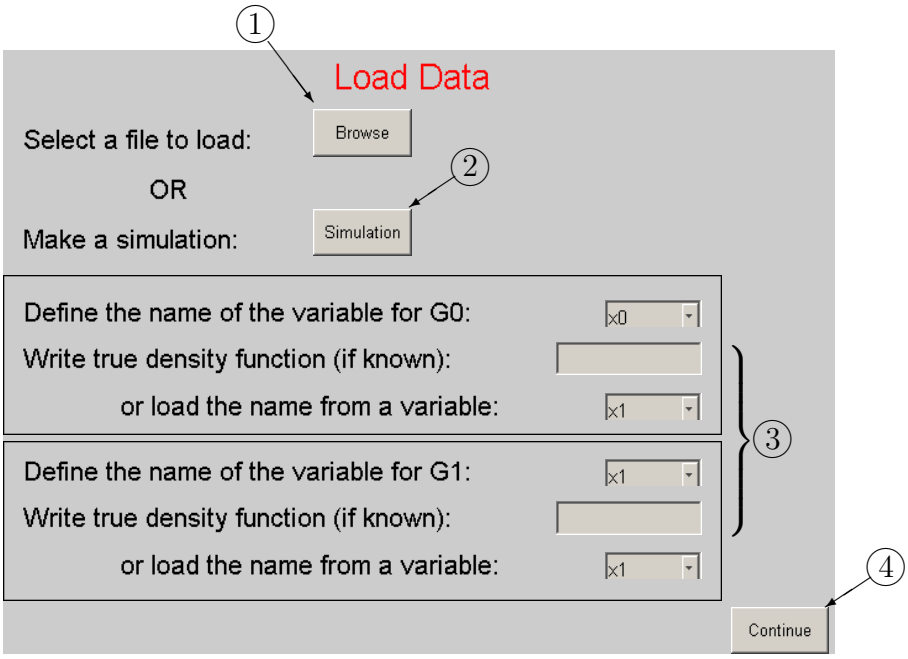


Figure 9: *Start menu*

You can skip this menu by typing input data as an argument
$>> ksquality(x0, x1)$;
where vectors $x0$ and $x1$ are score results for two groups G_0 and G_1. If we know also their densities $f0$ and $f1$ (for example for simulated data), we can set them as next arguments. For more see $>>$ *help ksquality*. After execution of this command directly the window on Figure 13 is called up.
In the *Start menu*, you have several possibilities how to define input data. You can load it from a file (button ①) or simulate data (button ②). In fields ③ you can list your variables in current workspace to define input data. If your workspace is empty, these fields are non-active. If you know the true densities for both groups, you can write them to text fields or load them from selected variables. If you need to simulate values for a model, press button ② . Then the menu for simulation (Figure 10) is called up.

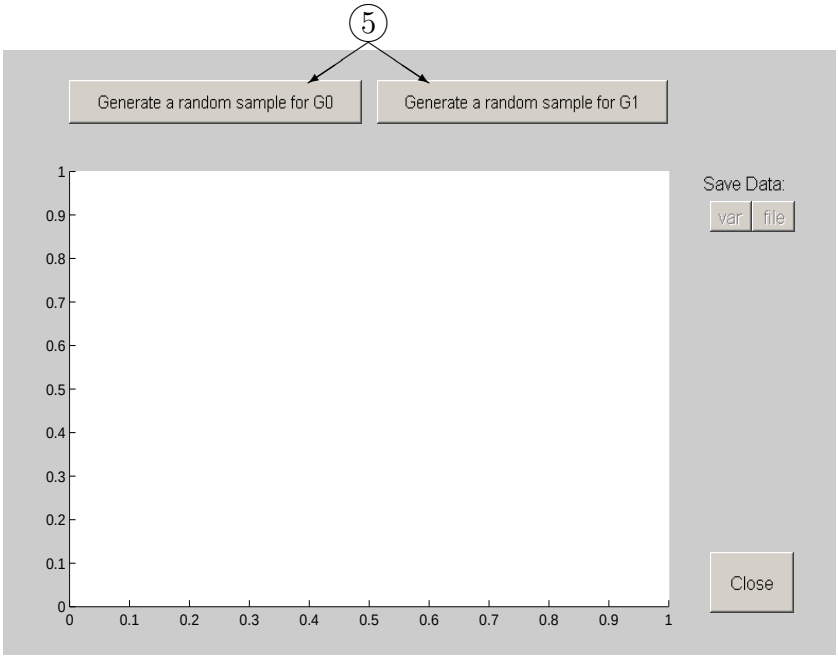


Figure 10: *Simulation menu – start*

In the *Simulation menu*, as the first it is necessary to generate random samples for both groups by buttons (5) . Either of these buttons calls up the *Data generation menu* (Figure 11).

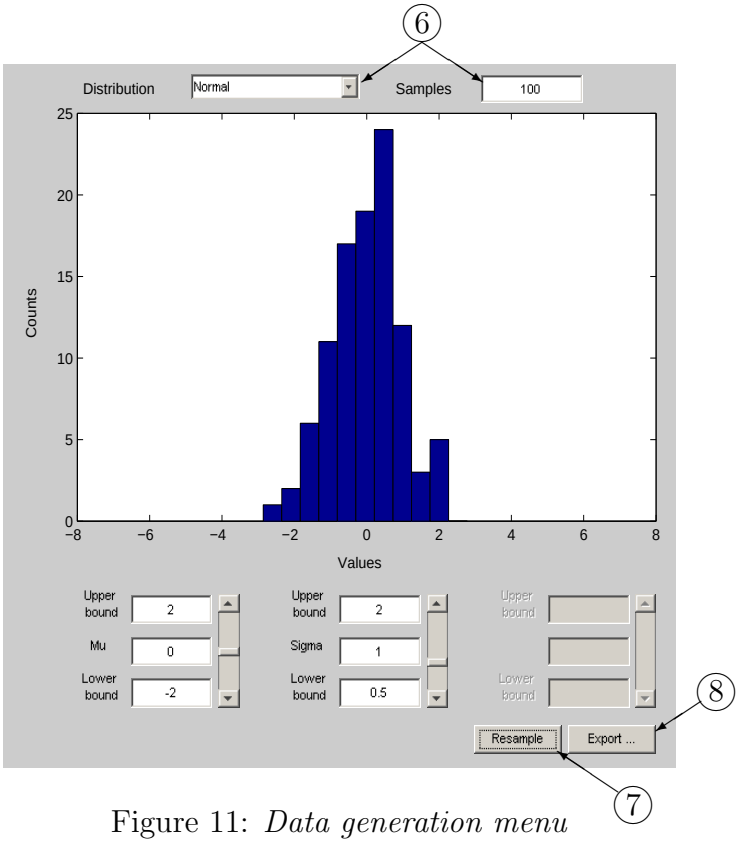


Figure 11: *Data generation menu*

In this menu, you can set the type of distribution and the size of sample (fields ⑥). In the figure, the histogram for the generated sample is illustrated. In the bottom part of menu, you can change parameters of a distribution and make a new sample by ⑦ . If you have done the sample generation, you need to export your data. The button ⑧ calls up a menu, where you specify the name of variable and confirm by pressing "OK". Then the *Data generation menu* will be closed and you will be returned to the *Simulation menu*. After data generation for both groups it seems like Figure 12. In the figure, the histograms of generated samples for both groups are illustrated. The cyan color represents data for G_0 and the red color is for G_1. In this stage you can save the data to variables or as a file by using buttons ⑨ . If you have done the simulation, press button ⑩ . The *Simulation menu* will be closed and you will be returned to the *Start menu* (Figure 9). In this menu, you can redefine the input data. If you want to continue, press button ④ . The menu will be closed and the *Basic menu* (see the next subsection) will be called up.

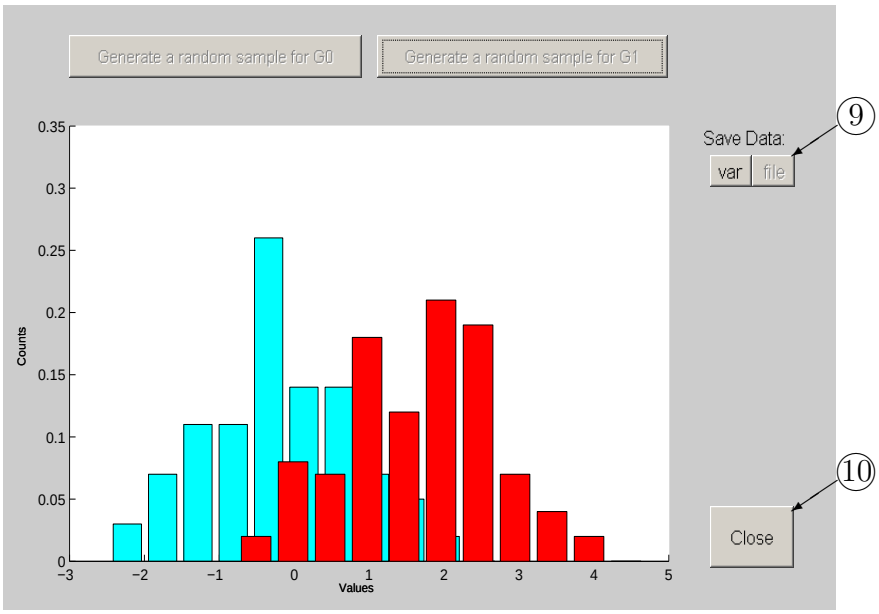


Figure 12: *Simulation menu*

2.2 Basic menu

This menu (Figure 13) was called up from the *Start menu* or directly from the command line (see >> *help ksquality*). At the start of this application, you can see some color symbols in the figure. Blue crosses represent score values for the first group G_0, red circles illustrate score values for the second group G_1.

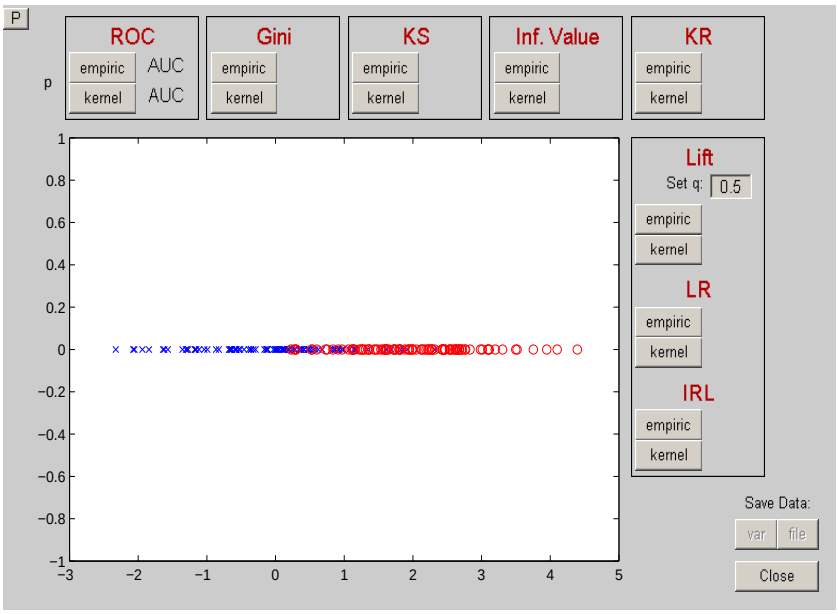


Figure 13: *Basic menu*

In the menu, there are also six fields for computation of quality indexes. For example in the first field you can obtain ROC curve for actual model. You can use the empirical estimate of ROC (press button "empiric") or a kernel estimate (press button "kernel"). At the right hand side of the used button the value of AUC (area under curve) is written. The actual curve is illustrated in the figure, see Figure 14.

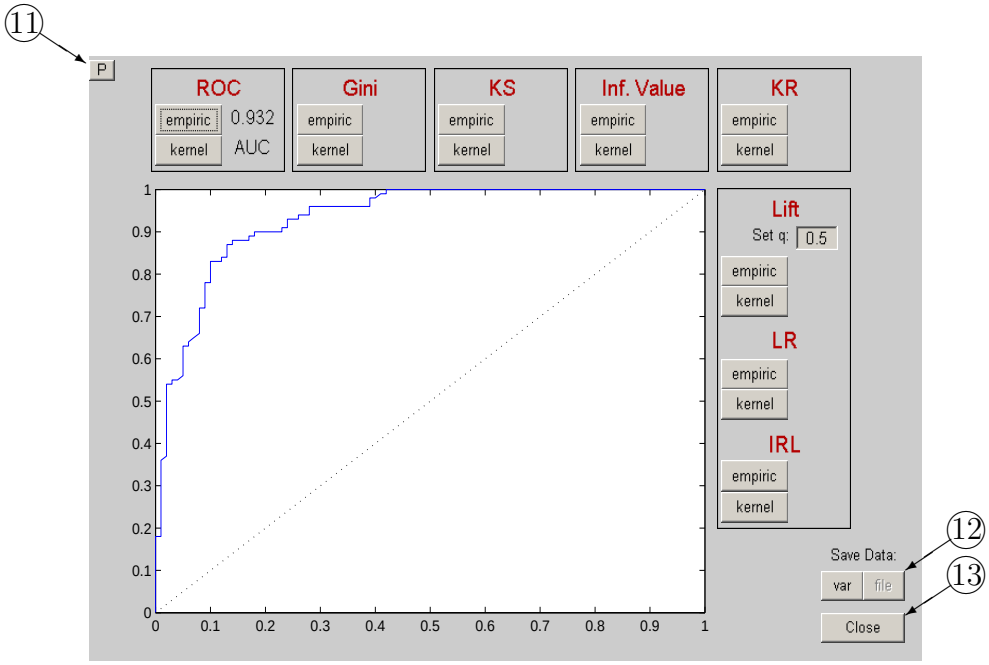


Figure 14: *Basic menu*

If you want to show only the actual curve, use button (11) (see Figure 15). You can also save the data to variables or as a file by using buttons (12) . The button (13) ends the application.

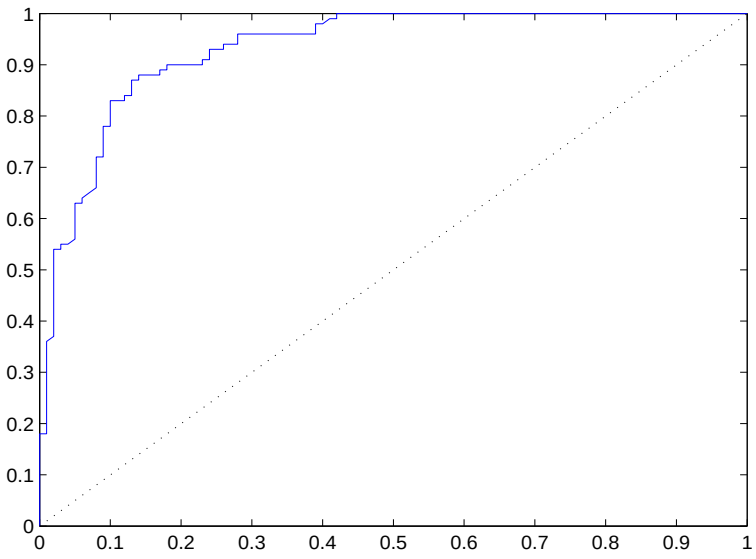


Figure 15: *The actual curve*

3 Two-dimensional density estimation

3.1 Start menu

The *Start menu* (Figure 16) for kernel estimation of two-dimensional density is called up by command

$>> ksbivardens$

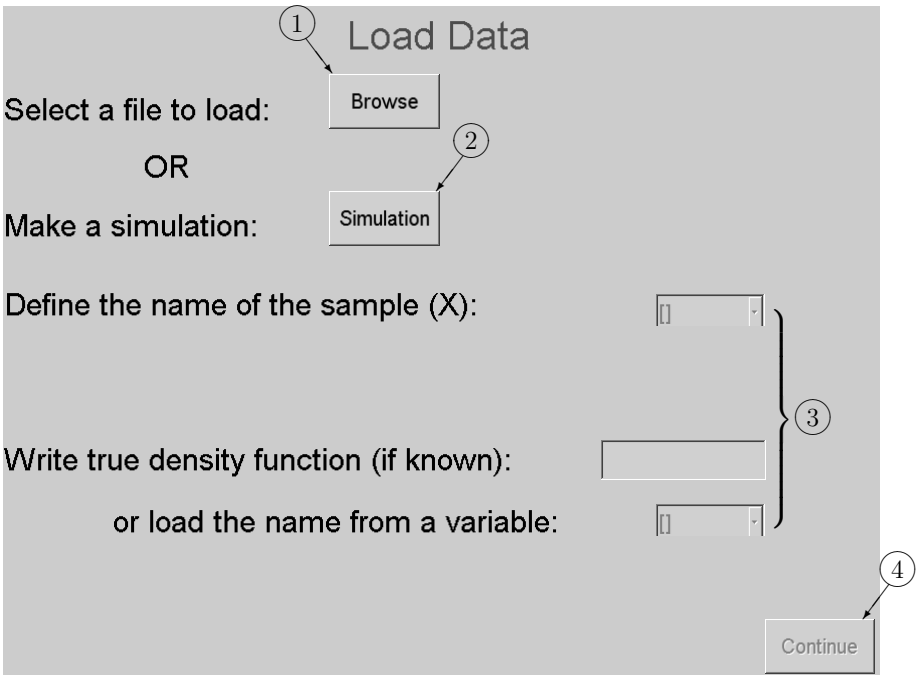


Figure 16: *Start menu*

You can skip this menu by typing input data as an argument
$>> ksregress(X)$;
where the matrix X should have the size $2 \times n$, where n is the sample size. If we know also the original density f (for example for simulated data), we can set it as the next argument. For more see $>>$ *help ksbivardens*. After execution of this command directly the window on Figure 3 is called up.
In the *Start menu*, you have several possibilities how to define input data. You can load it from a file (button ①) or simulate data (button ②). In fields ③ you can list your variables in current workspace to define input data. If your workspace is empty, these fields are non-active. If you know the true density of the sample, you can write it to the text field or load it from a variable. If you need to simulate a sample, press button ② . Then the menu for simulation (Figure 17) is called up. This application generates a random sample from a two-dimensional normal mixture density.

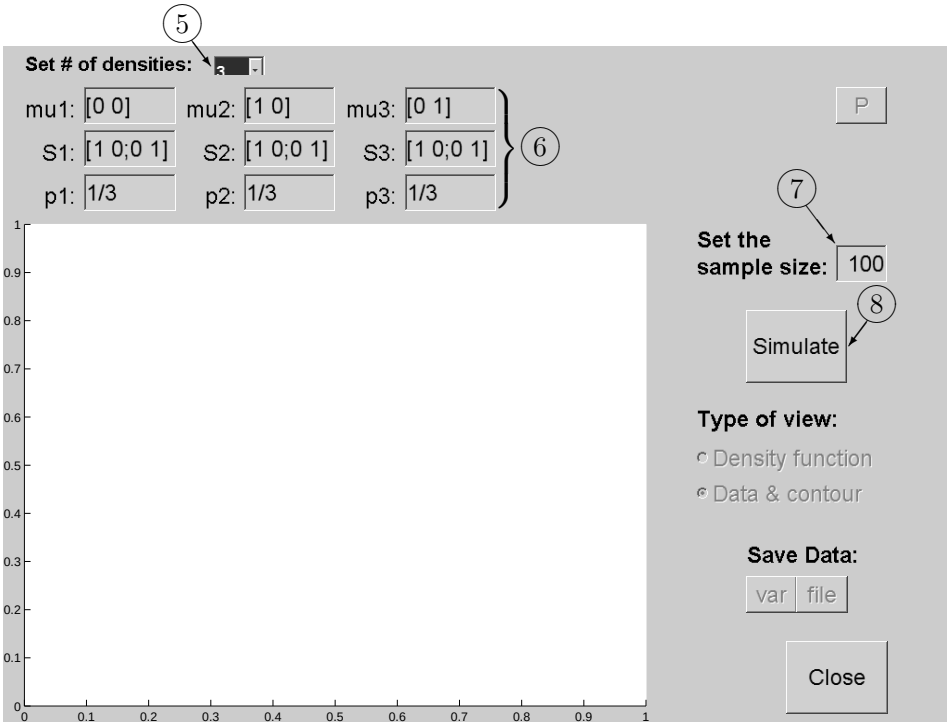


Figure 17: *Simulation menu*

In the *Simulation menu*, as the first set the number of components of the mixture by ⑤ . You can choose from 1 to 5 components of the normal mixture density. In fields ⑥ specify the parameters (mean, variance and weight) of each component. By ⑦ specify the sample size. You can simulate a sample by pressing button ⑧ and then see the result (Figure 18).

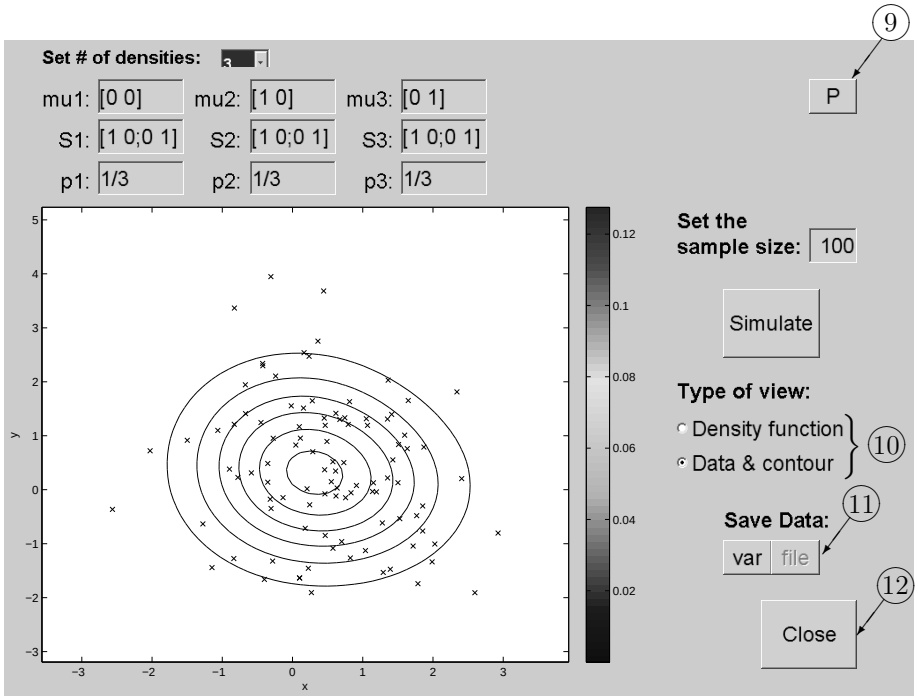


Figure 18: *Simulation menu – results*

By click on ⑨ you can print the current plot to the new figure. In fields ⑩ you switch the type of view between the data plot with contours and the 3-d plot of the density function. You can also save obtained data to variables or as a file by using buttons ⑪ . If you have done the simulation, press button ⑫ . The *Simulation menu* will be closed and you will be returned to the *Start menu*. In this menu, you can redefine the input data. If you want to continue, press button ④ . The menu will be closed and the *Basic menu* (see the next subsection) will be called up.

3.2 Basic menu and setting of parameters

This menu (Figure 19) was called up from the *Start menu* or directly from the command line (see >> *help ksbivardens*). If the original density is known, in ⑭ you can switch the type of view between the data plot with contours and the 3-d plot of the density function. By clicking on ⑮ you can show or hide contours of original density. If the original density is unknown, only the data are plotted. Use button ⑬ to continue.

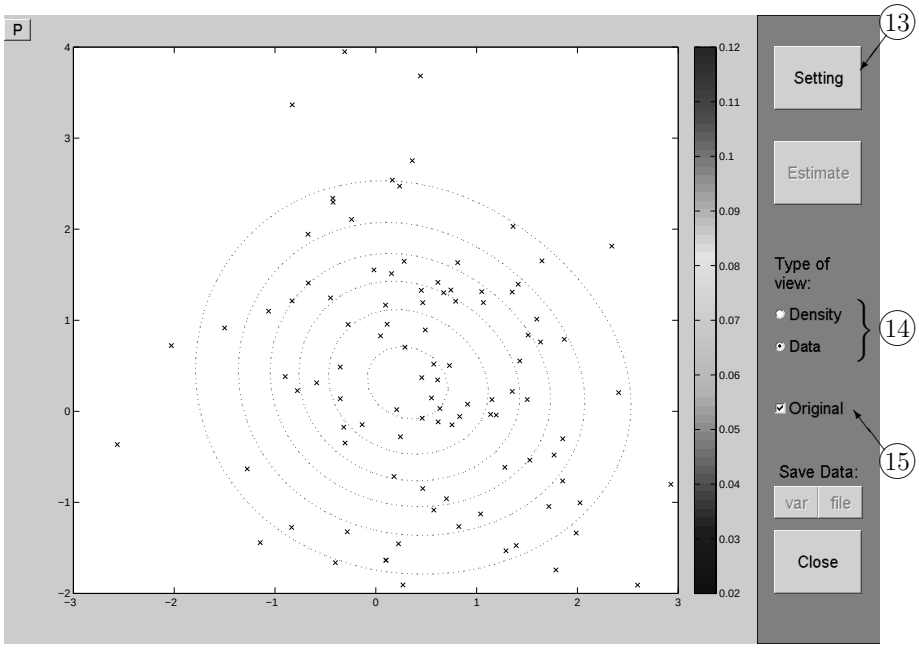


Figure 19: *Basic menu*

The button ⑬ calls up the menu for setting of parameters, which will be used for bivariate kernel density estimation:

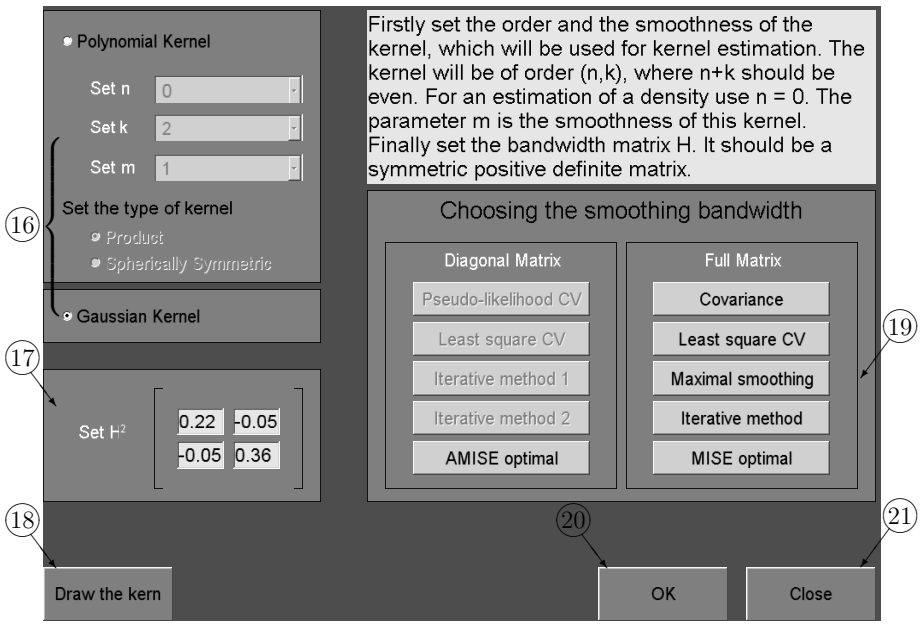


Figure 20: *Setting of parameters*

In the array ⑯ , you can set a type of the kernel. Implicit setting is the Gaussian kernel or it can be changed to polynomial kernel. In this case, set the type of kernel (product or spherically symmetric) and set the order of the kernel (n, k), where $n + k$ should be even, for density estimation $n = 0$ is used. The parameter m is the smoothness of this kernel. If you want to draw the kernel, use button ⑱ . In the array ⑰ , specify the bandwidth matrix H^2. The matrix should be symmetric and positive definite. If these conditions are not satisfied, the application writes an error message. Implicit setting is based on multiple of the covariance matrix estimate (see ...). Terms of the bandwidth matrix can be set manually or you can use one of buttons in the array ⑲ . There are two groups of methods:

1. **Diagonal matrix** – in this case, the diagonal bandwidth matrix is supposed. Because of computational aspects, used methods are developed for product polynomial kernels. For other types of kernel buttons are not active. There are five buttons:

   - *Pseudo-likelihood CV* represents the Pseudo-likelihood cross-validation method described in ...
   - *Least square CV* ...

- *Iterative method 1* ...
- *Iterative method 2* ...
- *AMISE optimal* – this button is active only in the case of known original density, it finds the AMISE optimal diagonal bandwidth matrix

2. **Full matrix** – in this case, the full bandwidth matrix is supposed. Used methods are developed for Gaussian kernel. There are also five buttons:
   - *Covariance* represents the method based on multiply of the covariance matrix estimate, see ...
   - *Least square CV* ...
   - *Maximal smoothing* – maximal smoothing principle described in ...
   - *Iterative method* ...
   - *MISE optimal* – this button is active only in the case of known mixture of normal distributions as an original density. It estimates the MISE optimal bandwidth matrix.

To confirm the setting use ⑳ , to close the window, use ㉑ .

3.3 Final estimation of the density

If you set all values in window for setting parameters (see page 20) and if you want to go to the final estimation of the density, confirm your setting by the button ⑳ . It calls up the *Basic menu*, where all buttons are active already:

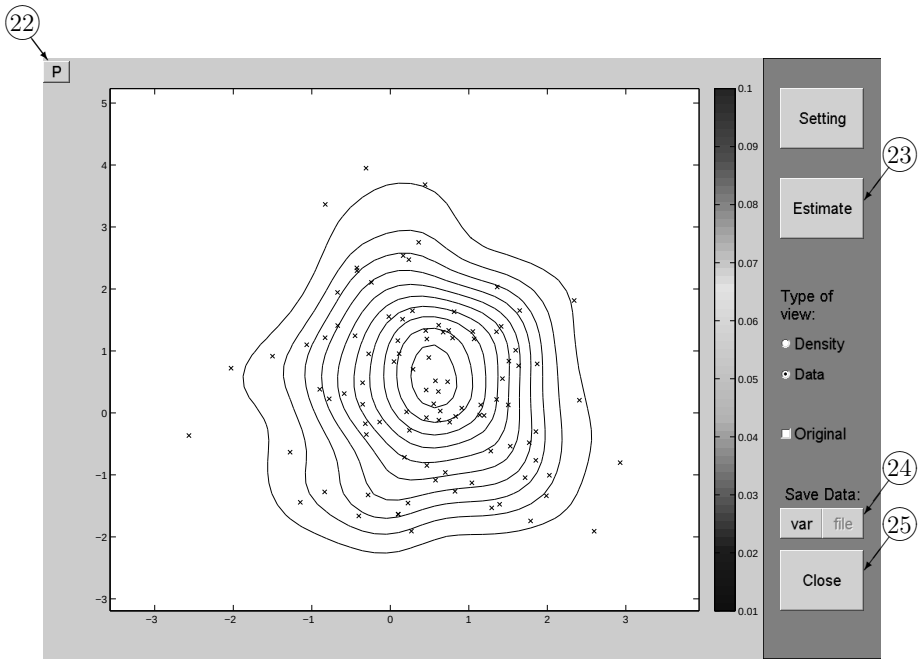


Figure 21: *Final kernel density estimate*

By click on the button ㉓ the relevant kernel density estimate is drawn. You can again add contours for the known original density by ⑮ and switch between types of view in ⑭ . By click on ㉒ you can print the current plot to the new figure. You can also save the data to variables or as a file by using buttons ㉔ . The button ㉕ ends the all application.