

TO READ THIS DOCUMENT, YOU MUST AGREE TO THE FOLLOWING TERMS AND CONDITIONS:

Jurik Research & Consulting ("JRC") grants you a non-exclusive license to use the accompanying software documentation ("Documentation") in the manner described as follows:

1. **LICENSE GRANT.** As licensor, JRC grants to you, and you accept, a non-exclusive license to use the enclosed Documentation, only as authorized in this agreement.

2. **COPYRIGHT.** This Document is copyrighted and protected by both United States copyright law and international treaty provisions. All rights are reserved. You may not permit other individuals to use the Documentation except under the terms in this agreement. No part of the Documentation may be reproduced or transmitted in any form or by any means, for any purpose other than the purchaser's personal use without the written permission of JRC. You may not remove any proprietary notices or labels on the Documentation. You may print this document only for your personal use.

3. **LIMITED WARRANTY.** Information in the Documentation is subject to change without notice and does not represent a commitment on the part of JRC. The user's sole remedy, in the event a typographical or other error is found in the Documentation within the warranty period, is that JRC will replace the documentation. The above express warranty is the only warranty made by JRC. It is in lieu of any other warranties, whether expressed or implied, including, but not limited to, any implied warranty of merchantability of fitness for a particular purpose. This disclaimer of warranty constitutes an essential part of the agreement. Some jurisdictions do not allow exclusions of an implied warranty, so this disclaimer may not apply to you.

4. **LIMITATION OF LIABILITY.** The user agrees to assume the entire risk of using the software described by the Documentation. In no event shall JRC be liable for any indirect, incidental, consequential, special or exemplary damages or other damages, regardless of type, including, without limitation, damages for loss of profit or goodwill resulting from the use of this documentation. In no event shall JRC be liable for any damages even if JRC shall have been informed of the possibility of such damages, or for any claim by any other party. Some jurisdictions do not allow the exclusion or limitation of incidental or consequential damages, so this exclusion and limitation may not apply to you. JRC's total liability to you or any other party for any loss or damages resulting from any claims, demands or actions arising out of or related to this agreement shall not exceed the license fee paid to JRC for use of the software described by the Documentation.

5. **TITLE.** You acquire no right, title or interest in or to the Documentation. Title, ownership rights, and intellectual property rights shall remain in Jurik Research and/or its respective suppliers.

6. **GOVERNING LAW.** The license agreement shall be construed and governed in accordance with the laws of California. In any legal action regarding the subject matter hereof, you agree to have venue be in the State of California.

7. **COST OF LITIGATION.** In any legal action regarding the subject matter hereof, the prevailing party shall be entitled to recover, in addition to any other relief granted, reasonable attorney fees and expenses of litigation.

8. **TERMINATION.** The license will terminate automatically if you fail to comply with the limitations described herein. On termination, you must destroy all your copies of the Software and Documentation.

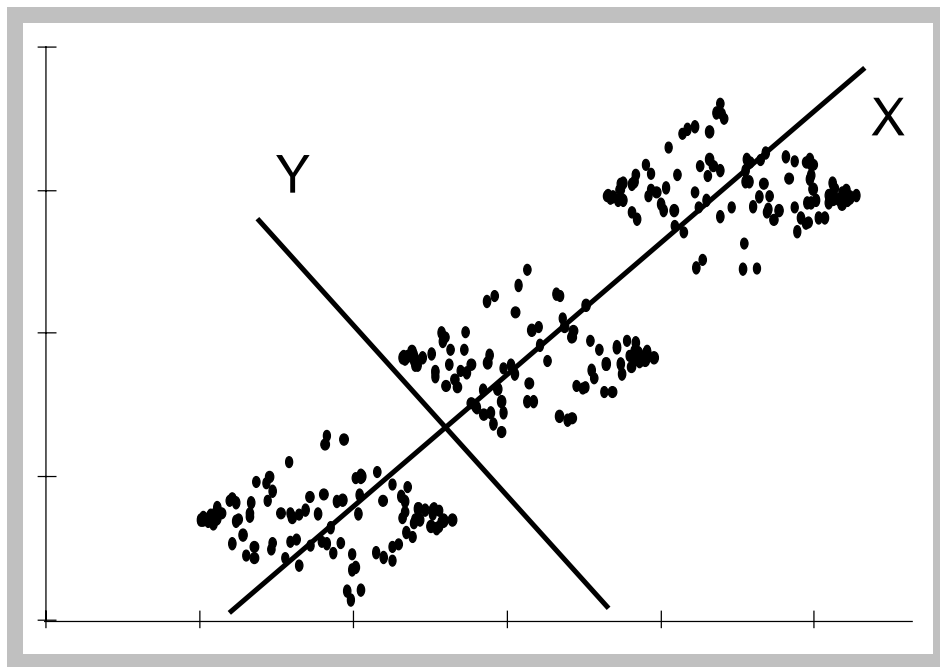
9. **NO WAIVER.** The failure of either party to enforce any rights granted hereunder shall not be deemed a waiver by that party as to subsequent enforcement of rights in the event of future breaches.

10. **MISCELLANEOUS.** If any provision of this Agreement is held to be unenforceable, such provision shall be reformed only to the extent necessary to make it enforceable.

JURIK RESEARCH

DDR 2.2

Decorrelator and Dimension Reducer
Add-In Tool for Microsoft Excel for Windows®



USER'S GUIDE

Requirements

Our tools run inside

- Microsoft 5.0c, under Windows 3.1, 3.11, 95, or NT
- Microsoft Excel 7 and 97 under Win 95 and NT

Installation

1. Using either the Window's Program Manager or Explorer, go to the floppy disk and run JRS_XL.EXE
It will request a password. Press OK. The installer will give you a computer identification number.
Write it down.
2. Get your installation password from Jurik Research Software. Call 323-258-4860 (USA) , fax 323-258-0598
or
E-mail to nfsmith@anet.net. Either way, give your full name, mailing address and computer
identification number. You will then be given a password.
3. Rerun JRS_XL.EXE, this time entering the password. The installer will verify your password. When
approved, it will install documentation and demonstration files into a user specified directory and the
tool(s) into your EXCEL \ XLSTART subdirectory. Read messages in all windows -- they are important.
Scroll down if necessary.
4. Start Excel. The tool(s) will be ready to run from the DATA command menu.

Notes

In the installed directory, you will find the following files ...

- | | |
|-----------------|--|
| 1. LEGALESE.TXT | Legal notices and warranties. |
| 2. ORDRFORM.HLP | A printable order form for all products we sell. |
| 3. CATALOG.HLP | An online catalog of all products we sell. |

In each installed "xxx_DEMO" subdirectory, you will find the following files ...

1. All the necessary demonstration XLS files.
2. A new VBA module, showing how to control a tool using Excel's Visual Basic.

Passwords

If you upgrade to a new computer, you will need a new password to install these tools. If you want to run them on additional computers, you will need additional passwords. Call Jurik Research Software (323-258-4860) for details.

What the DECORRELATOR / DIMENSION REDUCER (DDR) is all about

Brief Description

If you are building a model whereby each data fact-record contains numerous input (independent) variables and you can arrange the fact-records as rows in a spreadsheet, then the Decorrelator / Dimension Reducer, **DDR**, is for you. On the same or another spreadsheet, DDR creates a new data set, arranged in the same number of rows and columns as the original data set, but with two important differences:

- All the columns are decorrelated.
- All the columns are ranked according to the strength of their information content.

This helps models in two ways:

- Models learn faster with decorrelated variables than correlated ones.
- Models learn faster when uninformative input variables are deleted.

Why Decorrelate and Reduce?

Knowing the future sure has its advantages, especially where making a profit is concerned. This is understood most clearly in financial exchanges where competing traders speculate on the rise and fall of market prices. Successful traders have several traits in common, and probably most important is knowing where to look for good information. They know that even the best forecast models are useless if the chosen indicators are not relevant.

Collecting Relevant Indicators - When an aspiring forecaster has no idea which indicators to use, he usually constructs models by feeding them lots of data, data that might have any relation to the desired forecast. For example, a model intended to forecast gold prices might be fed historical precious metal prices as well as estimates of its future supply and demand. Since gold is used a lot in jewelry, and its demand is a function of the public's perceived ability to buy jewelry, then additional indicators for the model may include estimates of the consumer confidence index and related indices. This collection of indicators could swell very quickly.

A Well Kept Secret - Suppose your model's input consists of 100 different indicators. Most beginners think that regression models receiving a large number of indicators will perform better than models receiving a small number. Surprisingly, smaller regression models frequently outperform larger ones! The statistical world refers to this counter-intuitive behavior as the "phenomenon of multi-collinearity". It says that models prefer uncorrelated indicators, and that feeding a large number of *mutually correlated* indicators to a model typically **DEGRADES** its performance.

The Secret Explained - To understand the "phenomenon of multi-collinearity", let's suppose we have some data records arranged as rows, with each record containing a few input variables. For each record, we also have the target output value. As an example, the input variables could be a person's height, weight, and shoe size; the target value could be the person's life expectancy. We would like a regression model that provides us with coefficients for calculating target values (life expectancy) from the corresponding input variables (height, weight and shoe size).

Recalling high school algebra, a model with fewer records than variables is "**underconstrained**", resulting in not one but an infinite number of sets of coefficients. Although each set would correctly calculate the target values, they may all produce different answers on new input data! This would be unacceptable.

In contrast, the more likely case in the real world is to have more records than variables. Models based on such data are *typically* "**overconstrained**" whereby no set of coefficients can deliver a perfect answer for every record. In such a case, we must simply accept a set of coefficients that offer performance with low overall error, usually a set

delivering least mean square error. Standard regression, like the one embedded into Microsoft Excel, is designed to deliver least mean square error between a model's output and true target values.

The second sentence in the preceding paragraph is very important. The key word is "typically" because databases with more records than variables are sometimes **underconstrained**, or very nearly so. This situation occurs when at least one input variable can be closely approximated by other input variables. For example, suppose the illustrated table is a collection of five records (rows), each with three input variables and one target value. Now suppose that for any record in this database, a regression model can determine the target by multiplying the three input variables with coefficients (1, 3, -2) respectively.

A	B	C	Target
1.2	3.4	4.6	2.2
0.9	2.2	3.1	1.3
1.8	1.5	3.3	-0.3
2.5	2.7	5.2	0.2
2.1	1.9	4	-0.2

Let's see if this is true. In the first row, input data A=1.2, B=3.4 and C=4.6.

$$\begin{aligned}\text{Model's Output} &= 1 \times A + 3 \times B + -2 \times C \\ &= 1 \times 1.2 + 3 \times 3.4 + -2 \times 4.6 \\ &= 1.2 + 10.2 + -9.2 \\ &= 2.2\end{aligned}$$

The model's output matches the target value of 2.2. The coefficients (1, 3, -2) work just as flawlessly for the other four records. One might believe, then, that these are the best coefficients for the model. They are not! Amazing at it may seem, there are an infinite number of equally good coefficient sets to this modeling problem. Some other equally good sets are (-1, 1, 0) and (0, 2, -1) and (3, 5, -4) and (7000, 7002, -7001)!

This is bad news. To see why, let's compare the performance of a model using coefficients (7000, 7002, -7001) and a model using (-1, 1, 0). The input data to both models will be a slightly modified version of record #1 in the example database: A=1.2, B=3.4, C=5. Multiplying the input data by coefficients (-1, 1, 0) produces the output value 2.2. However, multiplying the input data by coefficients (7000, 7002, -7001) produces the output -2798.2!! Although both coefficient sets produced identical results with the original database, they also produced drastically different results with new data. This is simply unacceptable.

The reason why this amazing experiment was possible is because column C of the database does not represent a truly "independent" variable. In fact, each value in column C could be calculated by adding the corresponding values in columns A and B. Simply put, $C = A + B$. As a rule of thumb, interdependence among input variables seriously degrades the ability of regression models (including neural nets) to perform reliably with new data.

What is even more disheartening is that similar damaging effects would occur even if column C was simply *correlated* to the sum of A+B! The greater the correlation, the more pronounced the effects. This is why you should consider giving models decorrelated input variables. It minimizes a model's divergent response to new data.

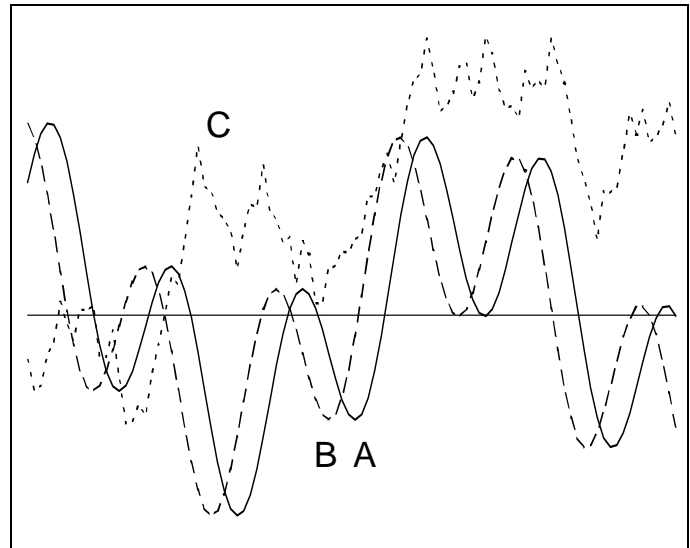
The Bad News - Unfortunately, financial indicators are highly correlated with each other, causing much frustration among those trying to model market behavior. This phenomenon forces all modelers to consider trying various combinations of two or more input variables until the best combination is found. Do you know how many possible combinations you can have with 100 indicators? About 10^{30} , equivalent to 100,000 times the number of atoms in a liter of water! Even with just ten indicators, there are over one thousand combinations to try! Yes, over ONE THOUSAND. Examining the effectiveness of all these combinations could take you a very, very long time!

Cutting Corners - Some modeling tools try to get around this problem by performing correlation analysis between pairs of input variables. This practice is based on the assumption that if two variables are correlated, then you do not need both, and so one of them can be eliminated. This popular practice can easily lead to a dead end. Here's a simple example to illustrate why. Suppose your input consists of three time-series:

- A) the daily high tide level near San Francisco,
- B) the daily high tide level near Los Angeles,
- C) the price of apples in China.

Since signals A and B are very similar, (and therefore highly correlated), one might be tempted to eliminate either one from the set of inputs to the model. But if you do, then the model could never create desired output signal D if its formula is $D = A - B$. In other words, a model can only calculate (A minus B) when both A and B are present! Therefore removing inputs on the basis of correlation can leave you with insufficient data and a non-working model.

Is there a better way to reduce the number of inputs to a model? YES! Professional forecasters have better ways to reduce the number of input variables and large companies can afford to pay their fees.

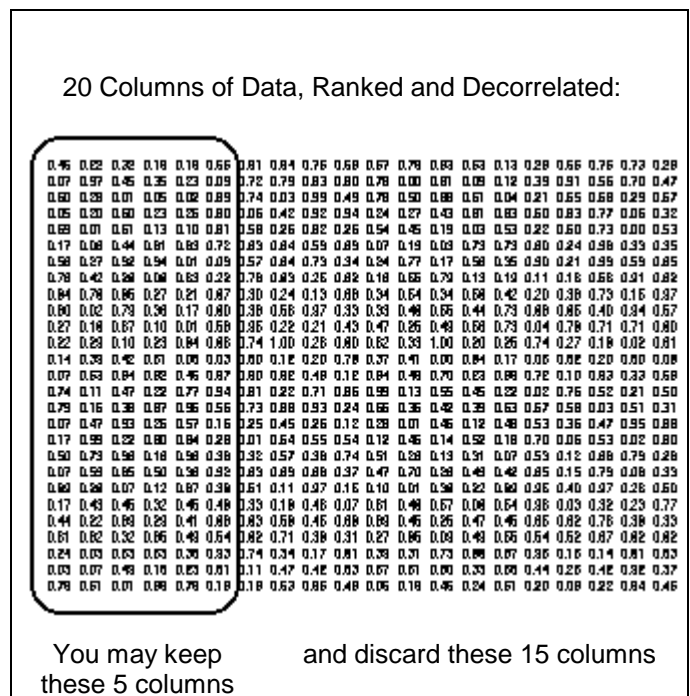


This put individuals at a disadvantage. Until now, that is. With DDR, you now have access to the same powerful technique used by professional forecasters.

Suppose you arrange your model's data so that each indicator fills a separate column and each data case fills a separate row. DDR will take your data, and produce a new data array the same size as the old but with two important differences. (See illustration on the right.)

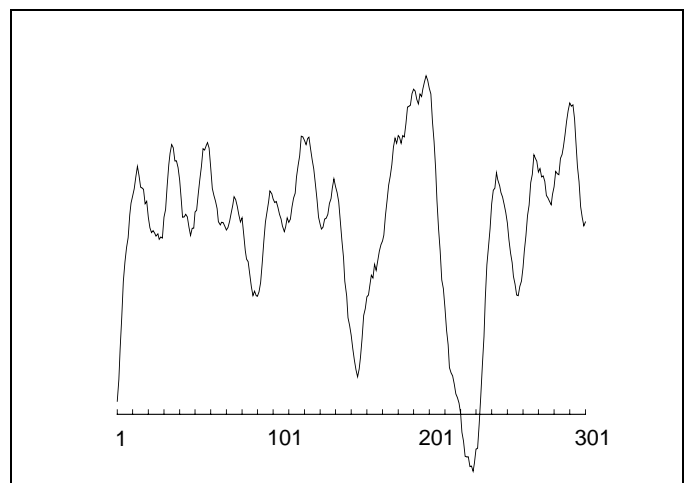
First, ***all the new columns will be completely uncorrelated with each other!*** This will very likely enhance a model's performance.

Secondly, DDR ranks the new columns according to how well they explain all the input data. DDR typically ***boils down 50 correlated indicators to only 14 with very little loss of information!*** This ranking helps you decide which columns to throw away, giving you additional room to employ more of your favorite indicators!



As depicted in the sample spreadsheet, the user may discover that the first five decorrelated columns contain almost all the information offered by the entire array, allowing the other 15 columns to be discarded.

To demonstrate the power of DDR, we prepared three models to forecast future values of a simulated financial time series. This time series consisted of three kinds of market forces: periodic cycles, aperiodic chaos, and random impulses. It is a composite of sinusoidal curves, the famous Mackey-Glass chaotic time series and Brownian noise. A small portion is illustrated on the left.



For all three models, data consisted of 1500 cases (rows) wherein each case contained 21 independent variables (columns) and a single forecast 10 bars into the future. The variables included past values of the time series as well as moving average values and other relevant indicators. The first model utilized the popular linear regression method. The second model was a neural network trained on the same data. For the third model, we had DDR decorrelate the data. We then trained the neural net on *only the first five columns* of DDR's output.

The results were astounding: the table on the right shows each model's average forecast error. Standard regression on all 21 variables gave an average percentage error more than twice as large as a neural net using the same data. However, after using DDR, the neural net only needed the first 5 columns to produce equivalent performance! DDR puts the secret of professional forecasting in your hands!

Model #1	Error
Simple Regression on 21 columns of original data	15.0%
Model #2	
Neural Net on 21 columns of original data	6.4%
Model #3	
Neural Net on 5 columns produced by DDR	6.4%

NOTE: Because you may not have access to neural net modeling tools, this manual will guide you through a similar experiment using only the linear regression function available in Excel and your DDR add-in module.

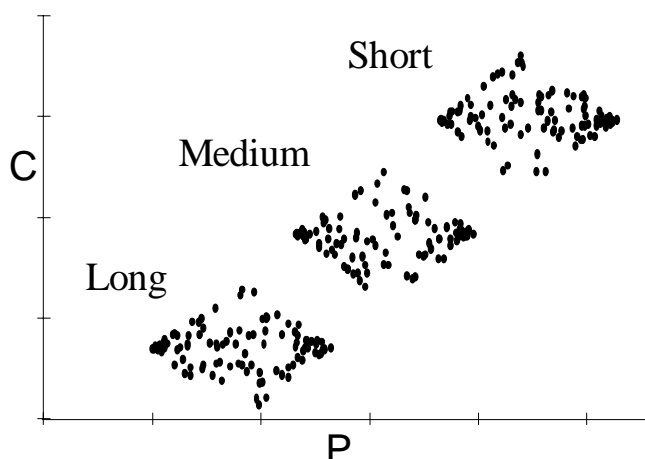
The Theoretical Basis of DDR

(The following section is optional. You may skip this section and proceed to "How to Use DDR".)

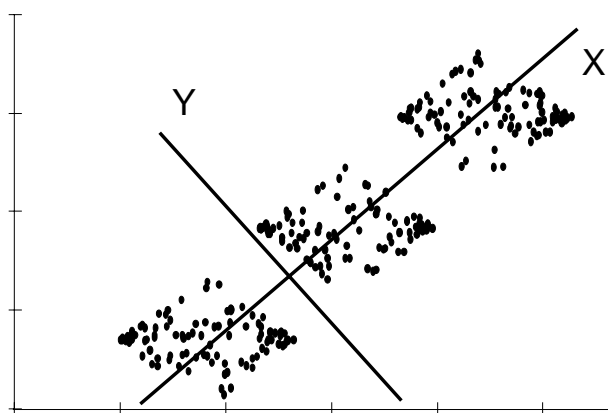
It may seem like magic that in most cases, models based on just a few of the new columns created by DDR perform just as well as models using all the original data! This is not an illusion; it is based on sound principles in mathematics. Consider the following example.

Suppose a long term medical research study measured blood pressure and cholesterol levels on 1,000 people and later recorded their age at death. Points in the figure represent data records, with axis P for pressure and C for cholesterol. The plot reveals three distinct groups, each group having a unique life expectancy.

The plot also shows that you can not use just blood pressure readings or just cholesterol readings to distinguish which group any point would belong. This is because some neighboring groups overlap and share similar blood pressure measurements. Other points overlap and share similar cholesterol measurements. Therefore both are required to determine which group a point belongs.



We might also conclude that if an insurance company built a complex model that needed estimates on life expectancy, the model would also need both measurements. However, models typically perform better with a few key variables than a broad spectrum of many variables. So would it be possible to combine the measurements of blood pressure and cholesterol into one variable that can successfully discriminate among the three life expectancy groups?



The adjacent figure shows one way to do this is. Two new axes are made: X and Y. Axis X travels through the centers of the three groups and axis Y lies perpendicular to X. We can now represent each point in the graph by giving either its P-C coordinates or its X-Y coordinates. The advantage to using these X-Y coordinates is that only the X axis serves to determine which life expectancy group a point belongs. The Y axis value serves no purpose. Therefore, concerning the life insurance model, we can represent information on forecasted life expectancy with only one variable, X, instead of both P and C.

In summary, DDR views each column of the original data as a separate axis, so that 21 columns represent 21 axes. DDR then creates a new set of axes and uses them to evaluate each point's new set of coordinates. DDR chooses the new axes so as to attain all the desirable properties mentioned in the beginning of this manual.

How to Use DDR

AUTOMATICALLY LOADED

Whenever you start Excel, it automatically loads DDR, ready for use. DDR is accessed by the "DDR 2" command in the "Data" menu.

INPUT REGION & INPUT SHEET

In an Excel worksheet select all the cells (rows and columns) containing your input data for a model. All the columns should be contiguous (next to each other with no unwanted, intervening columns). This manual refers to the block of data as the **input region** and the worksheet containing this region is called the **input sheet**.

OUTPUT REGION & OUTPUT SHEET

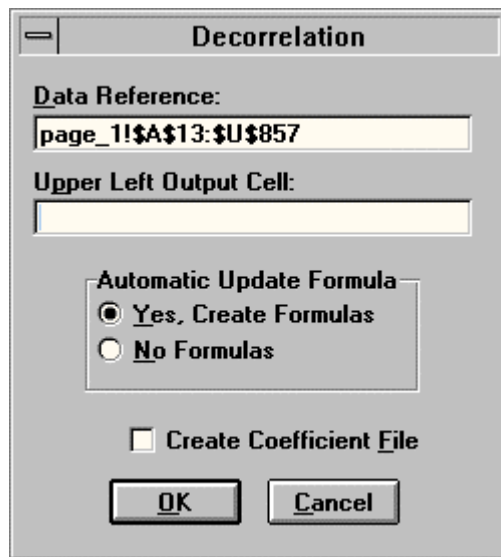
DDR will produce an output data array the same size as the input data array. DDR will place this array into an **output region** on an **output sheet**. You can have both input and output regions be on the same sheet, thereby making the input and output sheets one and the same. If you choose to do this, keep the two regions sufficiently apart so they do not overlap.

BRINGING UP DDR

Bring up the tool dialog box by selecting the "DDR 2" command in the DATA menu.

Excel 5 users -- DDR is not available when the Excel 4 optional menu is displayed. To access DDR, display the Excel 5 standard menu.

As shown in the figure below, the dialog has a several fields and selections. The user may move forward to each next field and selection in the dialog by pressing the TAB key and move backward by pressing the SHIFT-TAB keys.



We now discuss the various fields and selections.

DATA REFERENCE

This field designates the region of cells containing two or more columns of data to be decorrelated. When you select DDR for the first time during a session with Excel, and no multi-cell region has been highlighted just prior to its selection, then the Data Reference field will appear empty. The dialog's default for this field is to use the most recently selected (highlighted) region of cells during your current session with Excel. Thus when you highlight a region of cells before selecting this tool, the first field in the dialog will automatically show the region's location.

The user may change the designated region of cells to any other region as follows:

- 1) activate the dialog's field by clicking the mouse on the "Data Reference" field or press the TAB key until the field becomes highlighted, and
- 2) modify its contents by either typing in the region's address reference or by highlighting a new region on the spreadsheet.

NOTE: For reasons explained later, we strongly recommend you position the top row of your input data region below row 2.

DEMONSTRATION

The best way to see how the tool works is to follow an example. All instructions for the demonstration (demo) are in italic format.

Locate the file DDR_DEMO.XLS in the directory DDR_DEMO, produced during the installation process of DDR. (Hint: It is probably in C:\JRS or C:\JRS32). Open the file using Excel.

The contents of the spreadsheet are explained in Appendix A in the back of this manual. For this demo, pretend you want to forecast the time series (presented in column 1) ten rows into the future (farther down the column). The forecast values are located in the "Target Data" column.

For your demo, bring up DDR by selecting the "DDR 2" command in the DATA menu. Then highlight input region A13:U857 (r13c1:r857c21). A quick way to do this is to click on cell A13 (r13c1), press CTRL-SHIFT-⇒ then CTRL-SHIFT-↵. All cells to be processed should now be highlighted and the Data Reference Field of the dialog box should contain \$A\$13:\$U\$857 or r13c1:r857c21.

UPPER LEFT OUTPUT CELL

This field designates the location of the upper-left cell of DDR's output region. The output region will have the same number of rows and columns as the input region. The output region could be on the same spreadsheet as the input region or on a separate spreadsheet. **Prior to using DDR, the sheet containing the designated output region must be in a workbook that has a filename.** So, if you have just created a new worksheet, save it first onto disk so that it will have a filename. Then you can use DDR.

For convenience, the default value of the "upper-left output cell" field is whatever cell address you designated the last time DDR was executed during your current session with Excel. If this is the first time DDR is being used during this session with Excel, then the field defaults to being blank.

The user may change the designated cell to any other cell as follows:

- 1) activate the dialog's field by clicking the "Upper Left Output Cell" field or by pressing the TAB key until the field becomes highlighted, and
- 2) modify its contents by either typing in the location of the cell containing the time series title, or by selecting (highlighting) any cell you wish to designate.

NOTE: The chosen cell location must not be anywhere within the input region. Doing so will cause an error message stating that this location would have caused DDR to write over the input region. If you position the chosen cell location to the left of the input region (not recommended), make sure the output region and input region do not overlap.

NOTE: The chosen cell location must also not be in rows 1 or 2 of the spreadsheet. DDR reserves these two rows for column titles (to be described later). If you designate the upper-left cell to be in either row 1 or 2, DDR will automatically move it down to row 3.

SUGGESTION - Position the upper left output cell so that its row number is the same as the row number of the upper left cell of the input region. This way the two regions are aligned, row for row. Although this is not necessary for DDR to work properly, doing so will surely make working with new rows or data much easier.

SUGGESTION - Since the output data array does not write into rows 1 or 2, we recommend arranging the input data so that its top row is also not located in rows 1 or 2.

For your demo, designate the upper left cell by clicking on purple colored cell at location \$AB\$13 (r13c28). Note that the input data also starts on row 13.

AUTOMATIC UPDATE FORMULA

You have an option to let DDR insert into the last row of the output array the same formulas used by DDR to calculate the values in each column of DDR's output. These formulas refer to a hidden coefficient matrix auto-matically created by DDR. Having this row of formulas is handy, as you can copy them down to as many additional rows as you want. The advantage to doing so is that when you append additional data to the bottom of the original Data Reference region, the formulas automatically generate new rows in DDR's output region. This way the user can avoid having DDR reprocess all the data every time an additional row is to be calculated. The default value in the dialog box is to have update formulas inserted. If you do not want them, simply select the "No Formulas" option.

COEFFICIENT FILE

For the demo, let the automatic update formula remain enabled.

After DDR has processed your data and (if requested) placed automatic formulas in the bottom output data row, it can also create a special coefficient file containing specifications that allow other software applications to decorrelate data the same way as the live formulas would. For now, the only software application capable of using the coefficient matrix file is **TradeStation**, by Omega Research. There, users can decorrelate data in real time for possibly feeding the results to neural nets or any other mathematical process that TradeStation allows.

OK BUTTON

To start DDR, click on the OK button. *For the demo, press OK now.*

STATUS BAR

The time spent by DDR may take only a few seconds for small input data regions and much longer for very large regions. DDR's speed is faster if your computer has a floating processor unit (FPU). During long waits, the status bar along the bottom of the spreadsheet screen displays what DDR is currently doing.

TITLES ROW

When DDR writes its results to the specified output region, it also gives a title to each column in that region. These titles are placed in the first row of each column. Each title is a concatenation of "D_" and the relative column number within the region: D_1, D_2, D_3, ... and so on.

OUTPUT DATA and RELATIVE INFORMATION CONTENT

The output columns of DDR's decorrelation process will typically not resemble the original reference data columns. Nonetheless, there is no loss of information in the conversion. This is because there will always exist a linear regression that can perfectly convert the output region data back to the original input data.

Each output column of DDR uses **all** the input columns, in varying degrees, for its evaluation. Therefore, regardless of which columns of DDR's output you intend to use for modeling, you must feed DDR during modeling the same columns (input features) as were used during your creation of DDR.

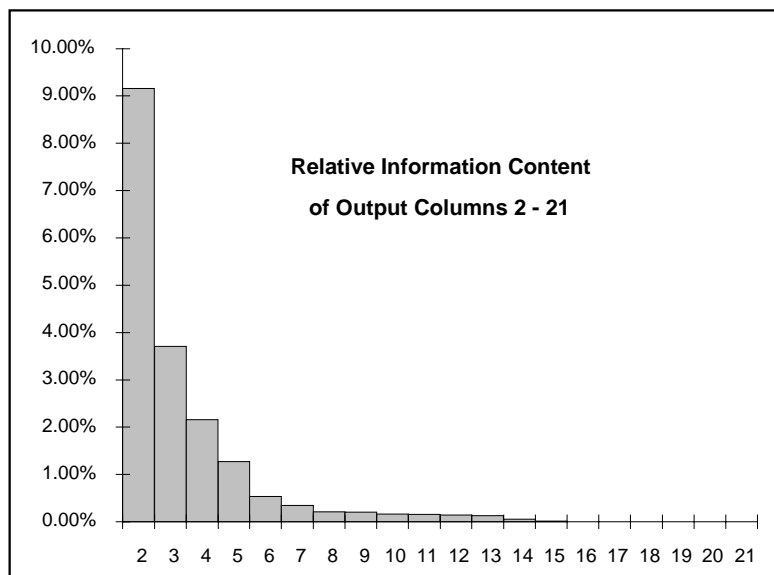
Some of DDR's output columns play a more significant role than others. In other words, some output columns bear more information about the original data than other columns. This is especially true when the input columns are mutually correlated, a property known as multi-collinearity.

Because multi-collinearity may be the case with your data, DDR arranges the output columns according to how much information each column possesses. The leftmost column of the output region bears the most information and the rightmost column bears the least. The relative amount of information contained in each output column is enumerated in the row just under the column titles (row 2). The relative amounts are shown as percentages, and all the percentages in the row add up to 100%. The relative information content helps you select which output columns to include as data to a model. Typically, you would benefit by selecting only columns with a large percentage.

SUGGESTION - To visualize the relative contributions of each output column, create a bar chart of the percentages.

For your demo, a bar chart of the relative information percentages for columns 2 through 21 would look like the chart below. It appears that the majority of the information lies in columns 1 through 5. Columns 6, 7, and 8 together add only another 1.1%. Columns 16 through 21 do not offer anything, their values are zero.

Using the AUTOMATIC FORMULAS



If your DDR dialog box had the “Automatic Update Formula” checked, then the last row of DDR’s output is a row of formulas. These formulas offer a quick way to convert new rows of data in the input region to new rows of data in the output region.

These formulas look for data in the corresponding input row. You may copy down the row of formulas for as many additional rows as you like. Each new formula row will look for data in the corresponding input row.

If no data exists in a corresponding input row, the formula row will fill its output cells with the standard Excel message “#NUM!”. When you fill the input row with data, the output row will automatically update.

For the inquisitive user: the formulas multiply values in the input row with coefficients located on a hidden sheet in the same workbook that holds the output sheet. The sheet is hidden so that you cannot accidentally alter its values.

For your demo, the added formula row is colored yellow and is located at AB857:AV857 (r857c28:r857c48). Highlight and copy this row of cells down to the next row immediately beneath it.. Notice that values in all the output cells of this new output row are “#NUM!”. To see what happens when you append a new row of data to the input region, copy the blue row located at A852:U852 (r852c1:r852c21). Paste the data into the gray row located at A858:U858 (r858c1:r858c21). Automatically the new formula row updates itself.

Note also that the last 6 columns of the output region contain zeroes. This is because DDR determined that at least 6 dimensions of the original data was totally redundant and had no useful information to offer.

SUGGESTION - If you add more rows on a daily basis, then eventually you will have lots of rows being updated by formulas. At this point, we advise reusing DDR over all the data so as to update the coefficient file.

Modeling with DDR's Output

(The following section is optional. You do not need to read it to use DDR. Nonetheless, we advise doing so.)

OPEN THE SHEETS

DDR's output data is typically fed to a model, such as a neural network or a standard regression model. To access DDR's output, you need to open the output sheet.

SUGGESTION - If the output sheet contains formulas that refer to data on the input sheet, we advise opening up the input sheet too. This makes Excel run much faster. The quickest way to do so is to double-click on one of the cells containing the update formula.

REGRESSION on DDR's OUTPUT DATA

Always use the first column produced by DDR for your model. It contains an enormous amount of information. As for the other columns, you can visualize the relative information content of each column by creating a chart of their percentage values.

To show how well DDR reduced the dimensionality of the input data, in this section you will use Excel to build a linear regression model to make 10-day forecasts of the time series using just the first 8 columns of DDR's output. We will then see how well its output correlates with the correct forecast.

For comparison, we already created a linear regression model using all 21 columns of the input data. The correlation between this model's forecast and the correct value is shown in cell W5 (r5c23) to be 0.84.

We will now build a linear regression model using just the first 8 columns of DDR's output. First, create a new spreadsheet and then return to your demo worksheet. You can do this quickly by selecting the "New" command under the "File" menu, click on OK, then select DDR_DEMO.XLS under the "Window" menu.

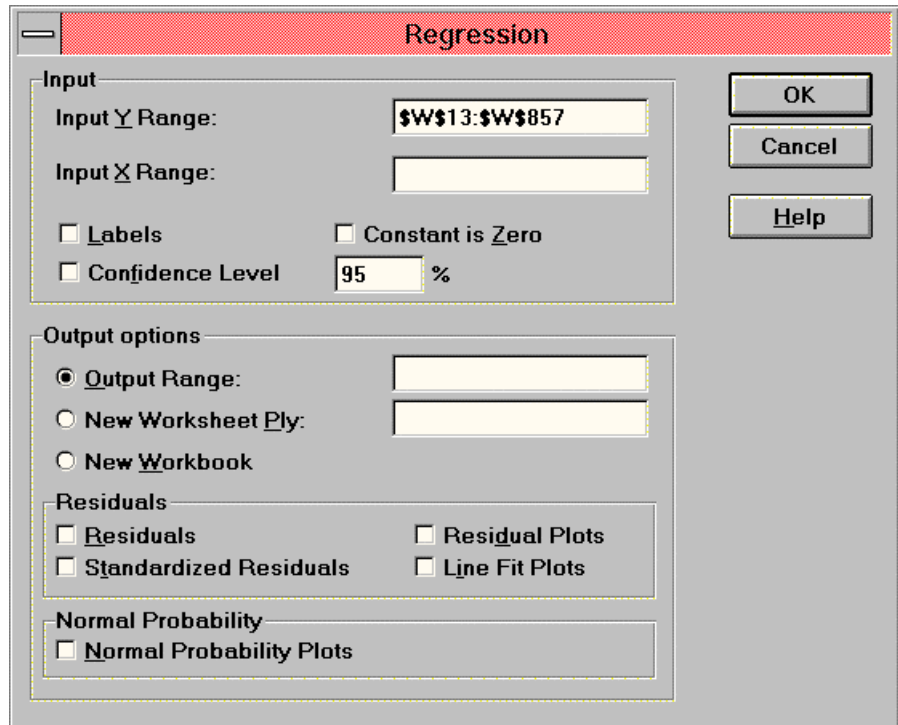
Bring up the regression dialog box. To do this, select the "Data Analysis..." command in the "Tools" menu of Excel. (If your menu does not show this command to be available, consult your Excel user manual on how to use the Add-In Manager to load in these tools.) A menu of tools appears. Scroll down and double-click on the word "Regression". The regression dialog box appears as shown on the next page.

Note the target data location is in the "Input Y Range" field. To fill it in, select all the target (correct) forecast values. An easy way to do this is to click on the cell located at W13 (r13c23) and then press CTRL-SHIFT-↓. Your dialog should now look like that on the next page.

Click the mouse on the next field, called "Input X Range". This field specifies the input data to the regression model. For this we will use the first 8 columns of DDR's output. An easy way to do this is to click on cell AB857 (r857c28), then drag the mouse across 7 more cells so that 8 cells along this row are now highlighted. Press CTRL-SHIFT-→. The "Input X Range" field of the dialog should now read \$AB\$13:\$AI\$857 or r13c28:r857c35.

Press the TAB key several times until the "Output Range" field in the dialog box is highlighted. This field specifies where you want the regression analysis to

post its results. For this we will use the separate spreadsheet you recently created. Select the blank spreadsheet under the “Window” menu. Click on cell B1 (r1c2). The “Output Range” field in the dialog box should now show the location of cell B1.



The image shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' set to '\$W\$13:\$W\$857' and 'Input X Range' is empty. There are checkboxes for 'Labels', 'Constant is Zero', and 'Confidence Level' (set to 95%). The 'Output options' section has 'Output Range' selected with an empty text box, and options for 'New Worksheet Ply' and 'New Workbook'. The 'Residuals' section has checkboxes for 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots'. The 'Normal Probability' section has a checkbox for 'Normal Probability Plots'. On the right side, there are 'OK', 'Cancel', and 'Help' buttons.

Press OK and, in a few seconds, Excel will write its results onto the blank spreadsheet. Return now to the formerly blank spreadsheet and note that the regression results come in three parts. The first part, shown below, gives overall performance measures.

Regression Statistics	
Multiple R	0.8307
R Square	0.6900
Adjusted R Square	0.6871
Standard Error	0.1370
Observations	845

Note that “Multiple R” is the correlation between the target data and what this regression model would have produced. The correlation value is 0.83. This value is very close to the correlation value of 0.84 attained by the model built on using all 21 columns of input data. This proves that the first 8 columns produced by DDR contains almost all useful information that could be extracted from the original 21 columns.

The third section of the regression results table shows, among other things, the coefficients you can use to actually produce the regression model's forecasts. This would enable you to compare its results with the target data on a chart. You may refer to your Excel user manual for instruction on how to do this.

APPENDIX

A Description of the Data in file DDR_DEMO.XLS

DDR_DEMO.XLS is a Microsoft Excel spreadsheet containing 21 columns of input data for DDR.

The first column contains a chaotic time series with some added Brownian noise to better approximate a financial time series. In the next 10 columns, all within the group labeled "MGBN," each successive column has the series shifted down by two additional rows. As a result, any row in the MGBN group contains eleven "snapshots" of the time series looking back in time.

The next four columns are exponential moving averages (EMA) of the time series in column 1. The EMA filters out some noise as well as simulates simple technical indicators on market price activity. The bar-length of each EMA filter are 5, 10, 15 and 20.

The last six columns contain another technical indicator (MACD) measuring the difference between pairs of columns in the EMA group. For example, the first column in the MACD group took the difference between the 5-bar and 10-bar EMA filters.

The column labeled "Target Forecast" is the same chaos time series of column 1, shifted upward 10 rows. This is done so that the target is the time series forecasted 10 bars into the future.

The box labeled "Correlation r on Regression Model Forecasts" contains two statistical measurements. The first measures the correlation between the "Target Forecast" series and a forecast series produced by a standard regression model on the input data. The second measures the correlation between the "Target Forecast" series and a forecast series we produced by a standard regression on the first eight columns of DDR's output. Note how similar the two measurements are, indicating that **DDR has successfully squeezed almost all the useful information of the original 21 columns into just 8 columns.**

This user manual shows, step-by-step using Excel's regression tool, how you can verify the second measurement. We determined the first correlation measurement using a more sophisticated program because Excel's regression can only process 18 input columns and we are using 21.

Calling DDR

from

Excel's Visual Basic for Applications

The following information is for advanced users who want to maximize the power of DDR by incorporating it within either user-defined subroutines or functions.

DDR may be called from Excel's Visual Basic for Applications (VBA). This powerful capability can be used to...

- search for optimal number of DDR output columns to use for modeling
- automate DDR's operation as part of an automated trading system

The following pages provide instructions on how to embed DDR in an Excel VBA subroutine.

INTRODUCTION

In your DDR installation directory (eg. C:\JRS\DDR_DEMO) the workbook DDR_VBA.XLS contains a working example of how to use Excel's VBA to operate DDR automatically. It contains one spreadsheet and one VBA module sheet.

To show DDR's effectiveness at dimension reduction, the data in column 3 in the workbook's data sheet is designed to be highly correlated to the sum of columns 1 and 2. Consequently, the three columns truly offer only two columns worth of real information. Let's see if DDR can extract two columns worth from these three columns.

The VBA routine will make DDR read data from columns 1-3 on a sheet and output to three other columns (cols 5-7) on the same sheet. You can run this example by executing the menu command TOOLS / MACRO... and selecting the VBA subroutine named "**DDRCall**".

DDRCall assumes the following:

1. There is data in the region A5:C500, in a worksheet named "data" in an open Excel 5 workbook named "DDR_VBA.XLS".
2. The three input data columns have titles in the first row.
3. Both the workbook containing input data for DDR and the workbook set up to receive output from DDR are currently open in Excel. In this example, workbook DDR_VBA.XLS will serve for both input and output.
4. The path to your XLSTART subdirectory is D:\msoffice\excel\xlstart. If this is not true for your system, you **MUST** edit the code accordingly. This will enable the "register" command to find the file JRS_XL.DLL.

CALLING PARAMETERS

DDRCall uses 3 input parameters:

1. **Input Range Reference:** Specify the complete name of the range containing input data.

In the example code below, input data is specified to be in the column range r5c1:r500c3, in a worksheet named "data" in the input workbook DDR_VBA.XLS. The complete range name is "[DDR_VBA.XLS]Data!r5c1:r500c3".

2. **Output Cell Reference:** Specify the cell to be the upper left corner of DDR's output array.

In the example code below, the upper-left cell is located at r5c5 in sheet "data" in workbook DDR_VBA.XLS. The full reference is "[DDR_VBA.XLS]Data!r5c5".

3. **Formula Output:** Specify whether or not you want the bottom row of DDR's output to contain live formulas that can be copied down the spreadsheet. TRUE=Yes, FALSE= No.

In the example code below, the bottom row will have live formulas.

```

Dim DDRFunc As Long

Sub DDRCall()

    Application.ScreenUpdating = False
    DDRFunc = ExecuteExcel4Macro _
        ("register(""D:\MSOFFICE\EXCEL\XLSTART\JRS_XL32.xll"", ""Decor"", ""JRRA"")")

    '*** For Excel v5.0 or a Windows 3.1 environment, use the following line
    '
    'DDRFunc = ExecuteExcel4Macro _
    ' ("register(""C:\EXCEL5\XLSTART\JRS_XL.xll"", ""Decor"", ""IRRA"")")

    ' call DDR with 3 parameters:
    ' Input Range reference = [DDR_VBA.XLS]Data!r5c1:r500c3
    ' Output cell reference = [DDR_VBA.XLS]Data!r5c5
    ' Update formula Option = TRUE (update formulas are desired)

    ExecuteExcel4Macro ("call(" & DDRFunc & _
        ", [DDR_VBA.XLS]Data!r5c1:r500c3" & _
        ", [DDR_VBA.XLS]Data!r5c5" & _
        ", TRUE)")

    ExecuteExcel4Macro ("UNregister(" & DDRFunc & ")")
End Sub

```

Excel VBA code calling DDR

IF YOU FIND A BUG . . . YOU WIN

If you discover a legitimate bug in any of our preprocessing tools, please let us know! We will try to verify it on the spot. If you are the first to report it to us, you will receive the following two coupons redeemable toward your acquisition of any of our preprocessing tools:

- a \$50 discount coupon
- a free upgrade coupon

You may collect as many coupons as you can.

You may apply more than one discount coupon toward the purchase of your next tool.

Our Referral Reward Policy

We understand the competitive nature of trading and your desire to keep some things secret, such as our tools, for example. We are proud that users of our products have attained results that were previously unattainable. So it is our policy to reward those who have put forth the supreme effort needed to recommend our fine products to other traders.

When we receive an order from someone wanting our tools, and he states on the order form that the order was based on your recommendation, we will credit you \$50 toward your next purchase of our products or upgrades. You may eventually accumulate enough credits to get our next upcoming tool free!