

Using Matlab and the System Identification Toolbox to Estimate Time Series Models

Jim McLellan
February, 2004

To estimate time series disturbance models in Matlab using the System Identification Toolbox, you can use the following steps.

1. load data into Matlab from whatever file it is presented. If the file is a text file, use the command:

```
load -ascii filename
```

If the file is .mat file, it is already in Matlab internal format, and you can just use the load command: load filename

To check to see what variables are in the workspace, you can use the “who” or “whos” command (the latter lists the variables and their dimensions). Any variables to be imported into the System Identification toolbox should be column vectors. If the number of rows in the variable (vector) is 1, then you should transpose it. For example, if y was 1x500, then re-assign y as follows:

```
y = y';
```

The prime denotes transpose, while the semi-colon suppresses the listing of the contents of y (which you will want when y is long).

2. To plot a variable, use

```
plot(y)
```

If you have a time index variable as well, you can use “plot(t,y)”.

3. You can mean-centre before starting up the System Id toolbox, or you can mean-centre afterwards (remember that we do all of our analyses with mean-centred data). To mean-centre the series in “y”,

```
ymc = y-mean(y);
```

which will subtract the mean of the column vector “y” from each element in “y”. It’s always a good idea to confirm that you have mean-centred by plotting “ymc”.

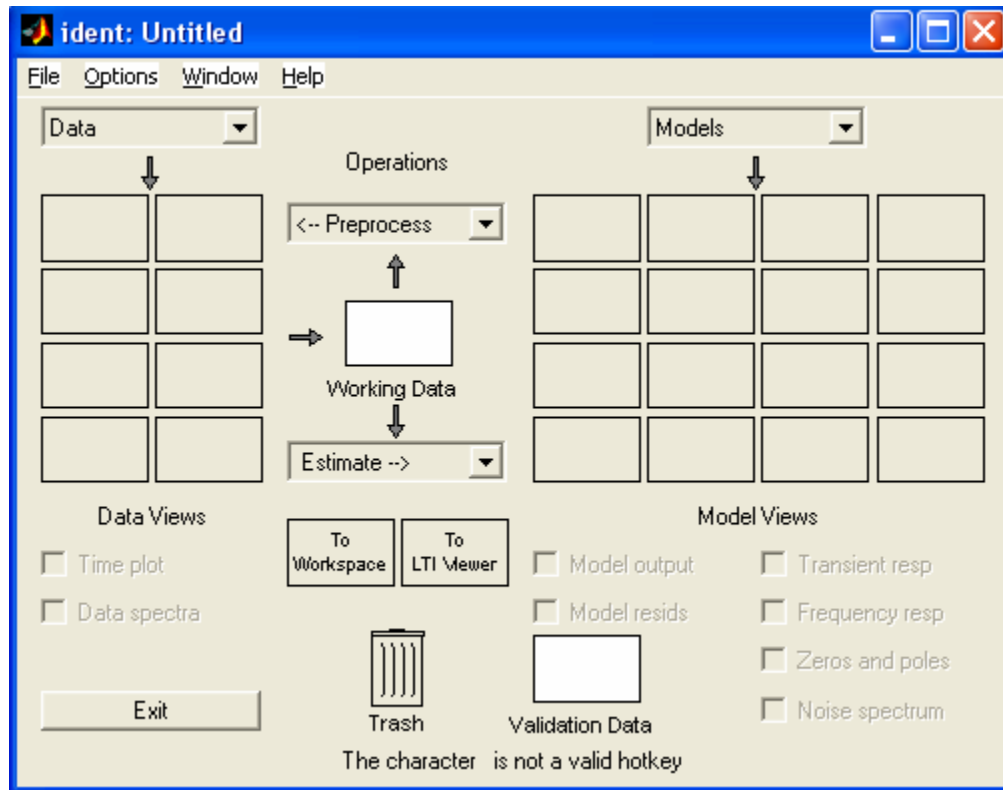
4. To generate autocorrelation and partial autocorrelation plots, you can use the “corrplots” routine available from the course website. Please note that there are a number of routines that you have to download and make accessible to Matlab in order to generate the plots. Unfortunately, the System Id toolbox doesn’t generate such plots. To use the command,

```
corrplots(y,maxlag)
```

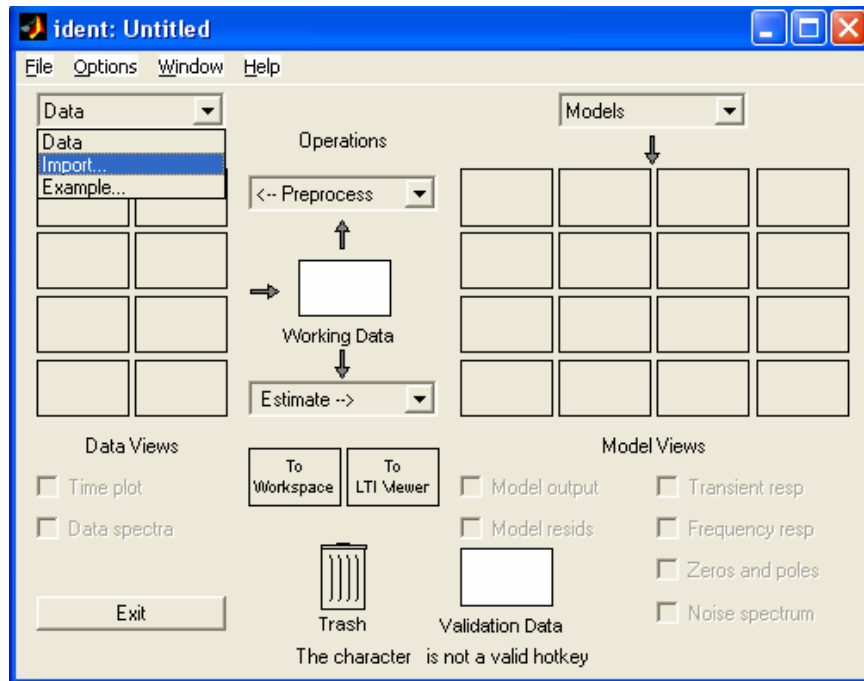
where maxlag is the maximum lag to use in the plot. For example, you could specify

maxlag to be 10. The maximum lag shouldn't be more than about 20% of the data record length (e.g., if you had 100 points, you shouldn't go out more than lag 20 – usually I don't go beyond lag 10).

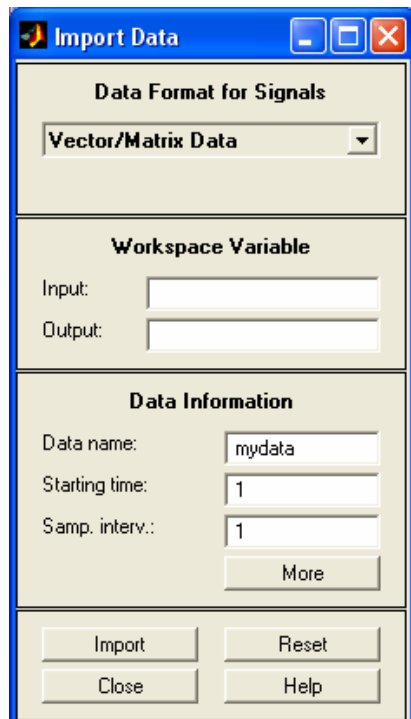
5. To start up the System Id toolbox, use the command “ident”. A GUI (graphical user interface) for the toolbox will appear:



6. The first step is to import data into the identification tool. Do this by selecting “data”, and then “import”,

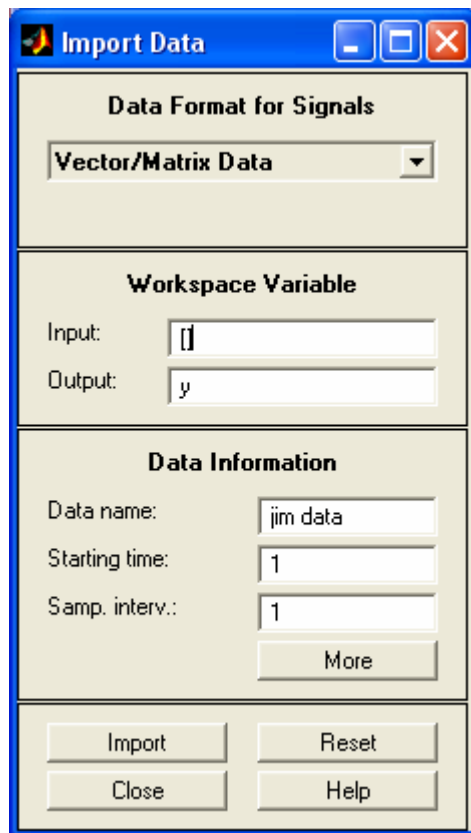


and then the following separate window will appear:



Specify the variable containing your time series as an **output** variable (which it is – it is NOT an input). You can also give it a name. If you know the sampling interval, you can also specify it, or leave the entry at the default of 1 time unit. Note that I have entered

“[]” in the input entry. Adding this tells the tool that there is no input variable (if you don’t do this, you will be asked to confirm that there is no input variable – you can click yes). Once you have filled in the entries, click on “import”.



The image shows a MATLAB 'Import Data' dialog box. It has a blue title bar with the text 'Import Data' and standard window controls. The dialog is divided into four main sections. The first section, 'Data Format for Signals', contains a dropdown menu set to 'Vector/Matrix Data'. The second section, 'Workspace Variable', has an 'Input:' field containing '[]' and an 'Output:' field containing 'y'. The third section, 'Data Information', includes fields for 'Data name:' (containing 'jim data'), 'Starting time:' (containing '1'), and 'Samp. interv.:' (containing '1'), along with a 'More' button. The bottom section contains four buttons: 'Import', 'Reset', 'Close', and 'Help'.

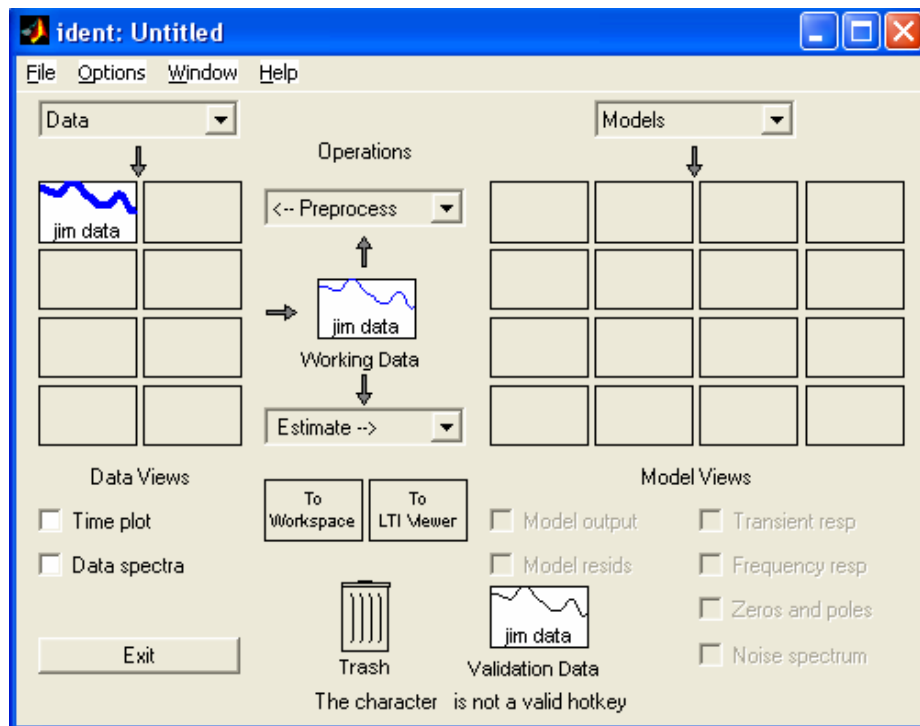
Data Format for Signals	
Format:	Vector/Matrix Data

Workspace Variable	
Input:	[]
Output:	y

Data Information	
Data name:	jim data
Starting time:	1
Samp. interv.:	1
More	

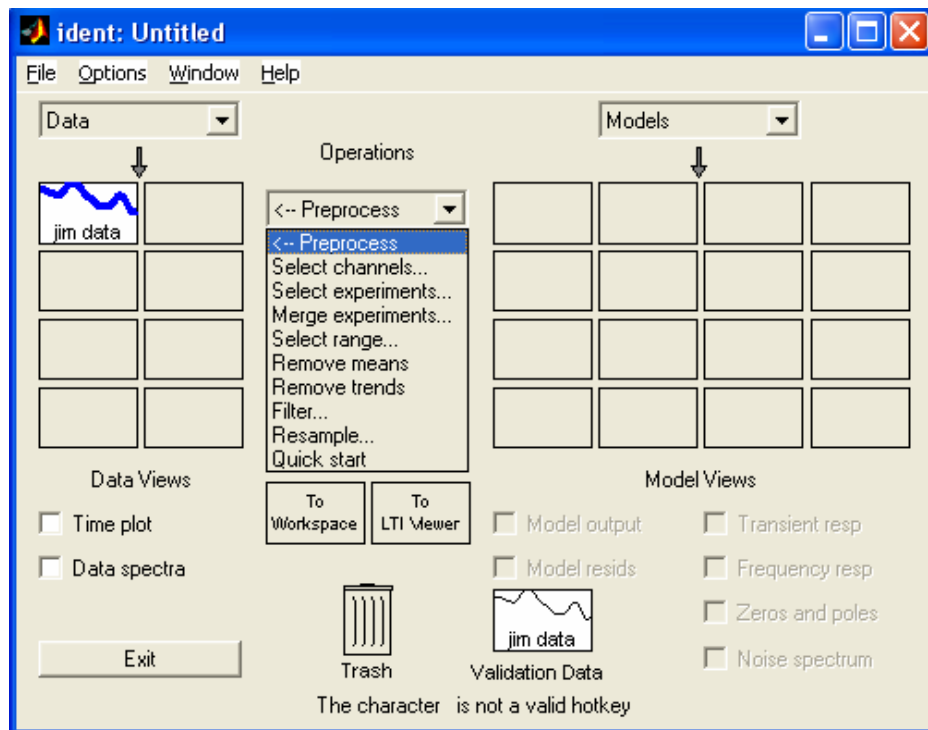
Import	Reset
Close	Help

Your workspace will now look as follows:

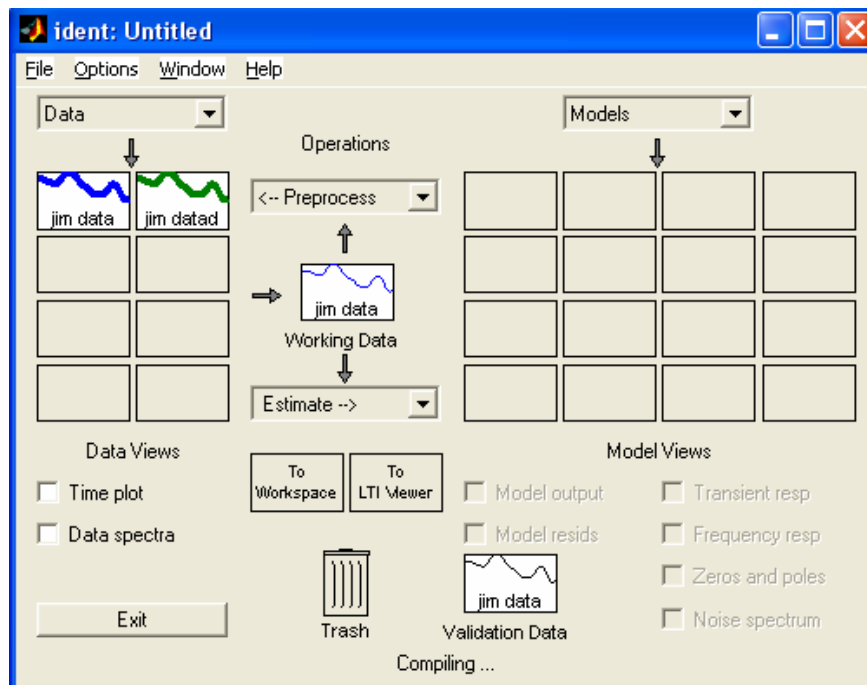


7. Important things to know – notice the “working data” and “validation data” boxes on the GUI. You can specify the dataset to use for estimating your model, or validating your model, by dragging dataset icons from the “data views” panel on the left to the “working data” and “validation data” boxes. Working data is the dataset used to estimate the model. Validation data is data that you use to judge how well the model works – it may be the same as the estimation data, or if you have another dataset collected under the same circumstances, you can use it as a benchmark to test your model by predicted the new data values and comparing them. If there is only one dataset, both “working data” and “validation data” are set to the one dataset.

If you **haven't** mean-centred your data before loading into the toolbox, you can do it using one of the Preprocess commands. Click on the Preprocess menu item to get the following list:

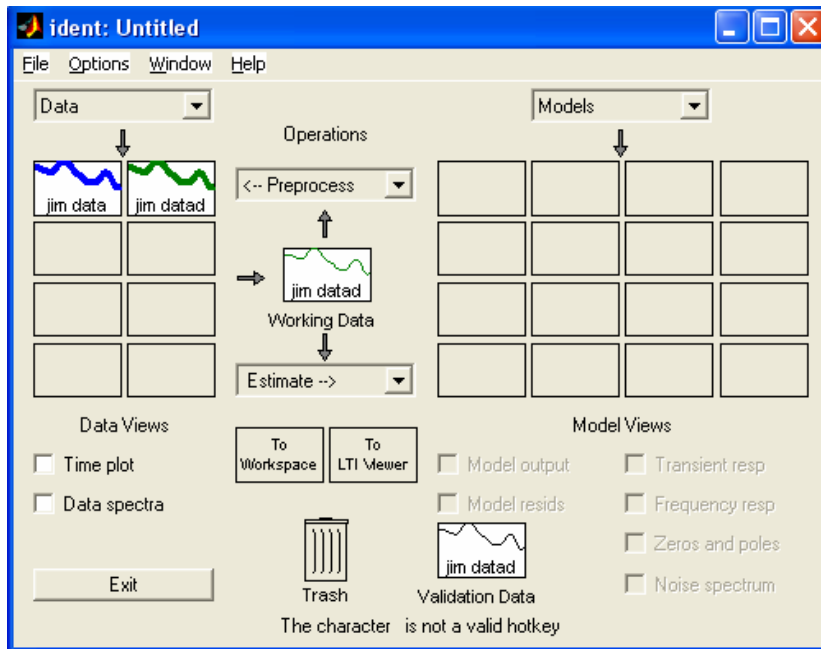


Select “remove means”, and you will now get a new dataset to work with (note that you only need to do this if you did NOT mean-centre your data prior to starting the identification toolbox):



Now, you need to drag the mean-centred dataset to the Working Data and Validation Data boxes. Common mistake – not dragging the mean-centred data to the

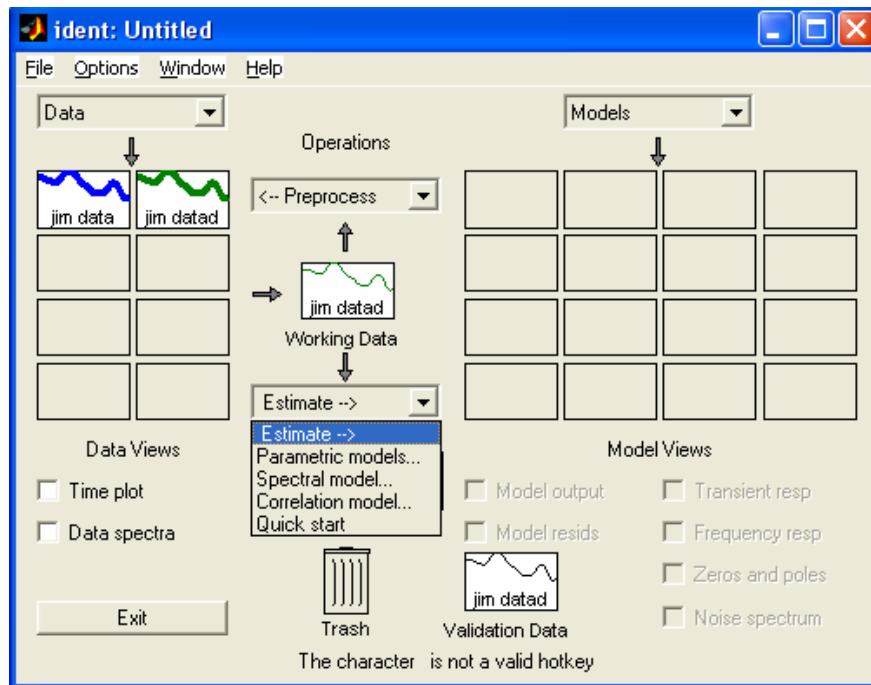
Validation Data box – if you don't do this when you are using mean-centred data to estimate the model, there will be a constant offset, and your residual autocorrelation diagnostics will be awful! (believe me, I have done this on more than one occasion). Your window should look something like this:



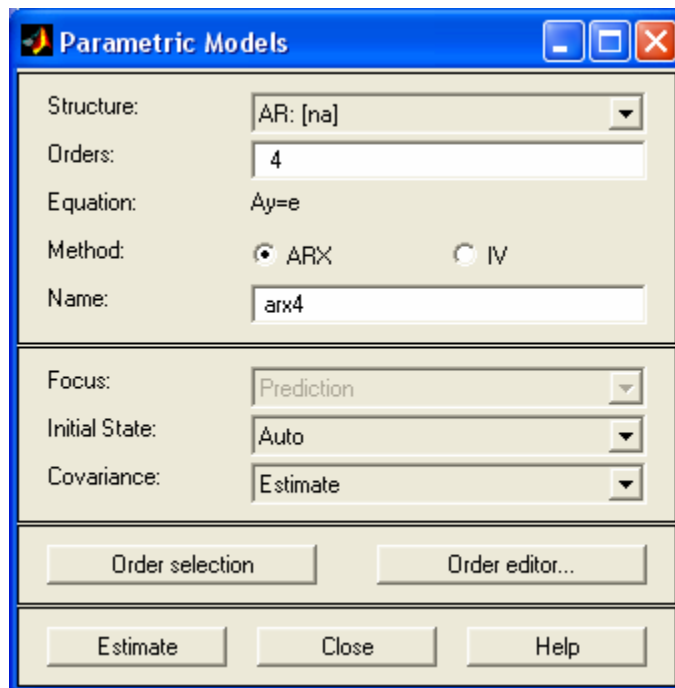
You can see that the mean-centred dataset (jim datad) is now being used for working and validation data.

If you want to see a time plot, you can check off the “time plot” box on the lower left hand corner.

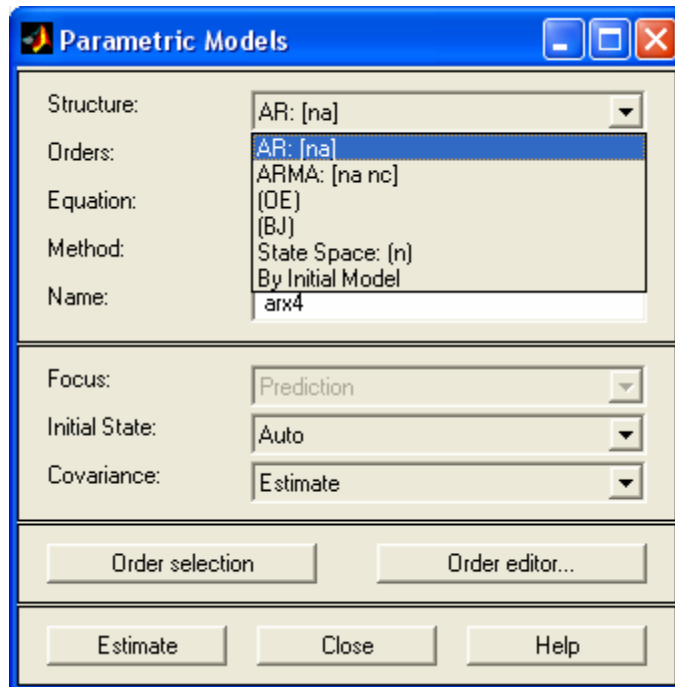
8. Now it's time to estimate models. First, go to the “estimate” menu window, and click on the arrow icon:



Select “parametric models”, and the following window will appear:



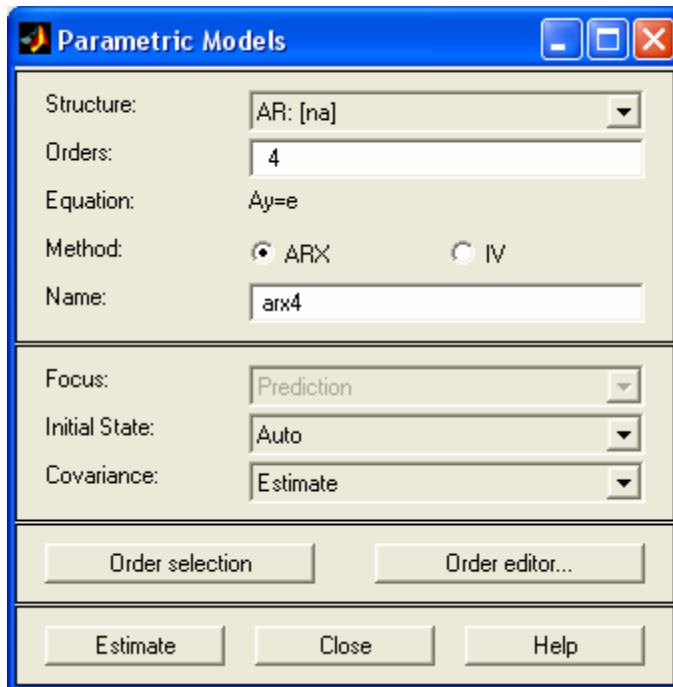
Click on the arrow associated with the “structure” window to choose the type of model:



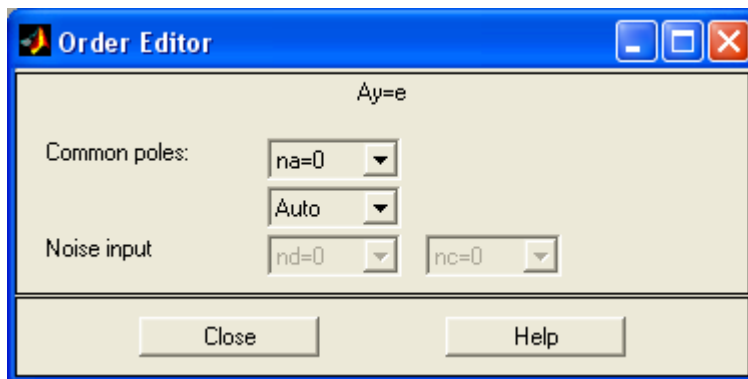
AR is autoregressive, ARMA is autoregressive-moving average; the remaining models are ones that we will use later when we have inputs as well.

The AR model is $A(q) y = e(t)$, where $A(q)$ is a polynomial in the backwards shift operator of order “na”, and $e(t)$ is a white noise input. Similarly, the ARMA model is $A(q) y = C(q) e(t)$ where $A(q)$ and $C(q)$ are polynomials of order “na” and “nc” respectively.

Clicking “estimate” (bottom left hand corner) for the following specification will estimate a 4th order AR model:

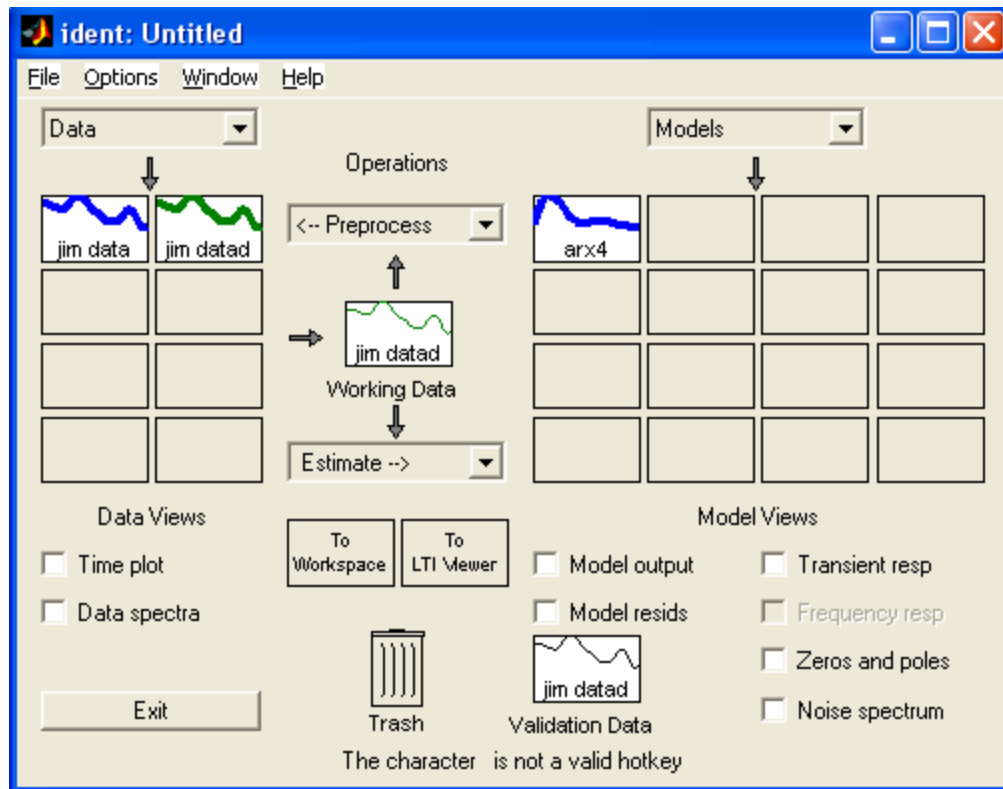


With the AR model, you can also specify a range of models orders to estimate. Click on “order selection” to do this (or just enter 1:5, if you want to estimate AR models AR(1) through AR(5), for example). You can also use the “order editor”, which will show the appropriate layout for a given model type.



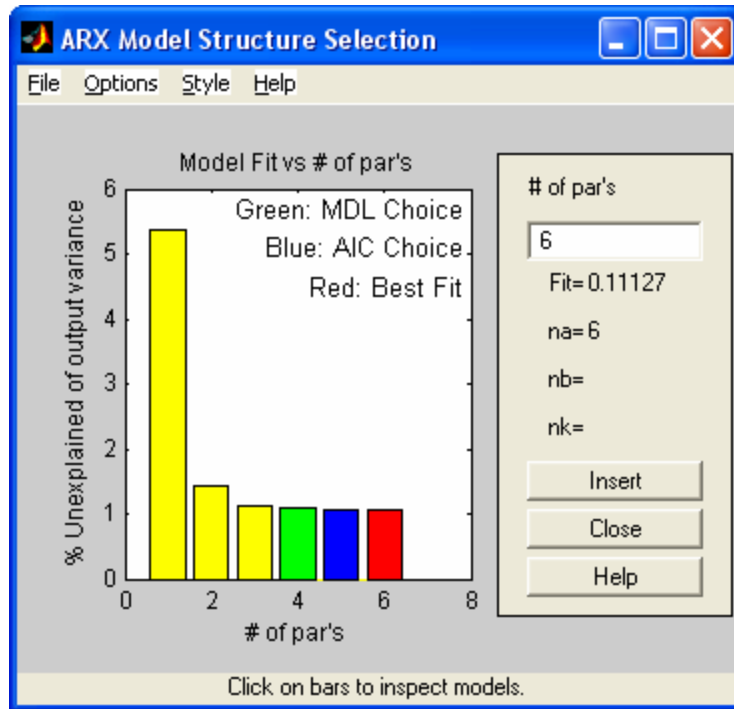
Click “close” once you have specified the model order.

9. If you have specified a single model order to be estimated, and you click “estimate”, the following change will appear on the GUI:



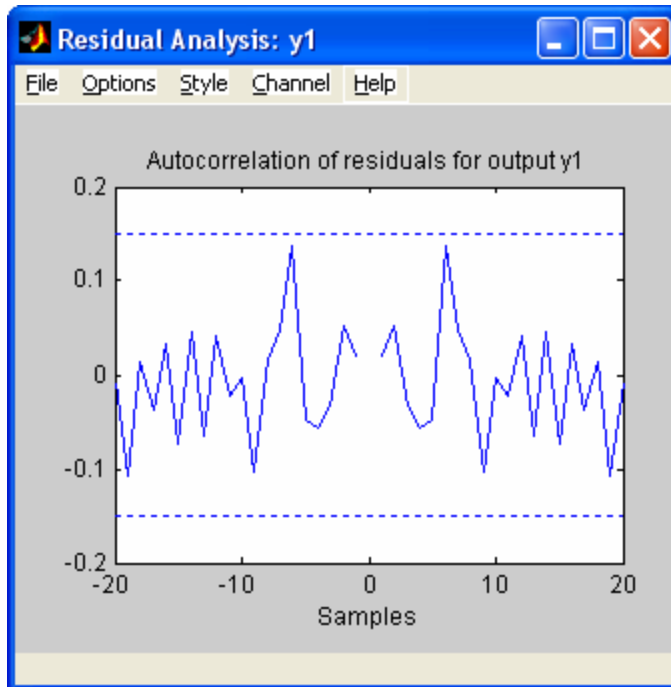
Note that there is now a model (arx4) in the “model view” palette on the right hand side of the GUI window. As you estimate models, they will be added here. If the curve on the icon is thick, it is “active”, i.e., it will appear in any of the diagnostic windows that you bring up (more on this later). If you don’t want it to appear on these windows, simply single click on the icon, and the curve on the icon will become thin, telling you that it is inactive as far as diagnostic windows are concerned.

If you selected a range of orders to be estimated for the AR model structure, you will get the following window:



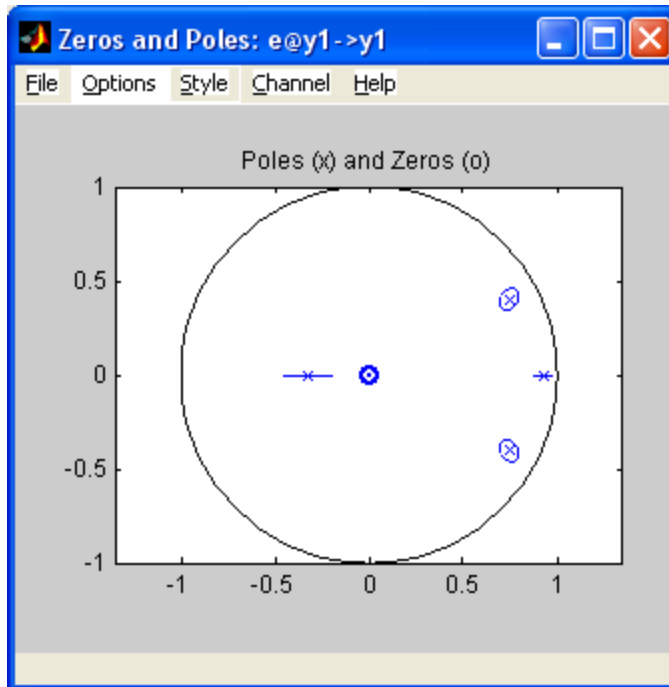
The vertical axis on the axis tells you the % unexplained variance (i.e., residual variance as a percentage of total variation in the data). One or more bars may be highlighted as MDL or AIC choice, or as “Best Fit”. MDL and AIC are selection criteria a bit reminiscent of adjusted R-squared in conventional regression, in that they account for residual variance but include a penalty on the number of parameters used. # of pars is the number of parameters, and corresponds to the order of the AR model. To add these to your set of working models, click on a bar, and then click on “insert”. You will see the model appear in the Model View palette.

10. It's diagnosis time. To select diagnostic windows, check off the boxes, and the windows automatically appear. Once checked off, a window will remain (but may be minimized – simply click on the icon along the bottom of the screen to retrieve it again). As you activate/deactivate models, the windows will change correspondingly. The most useful one is the “model resids” window, which generates an autocorrelation plot (with hypothesis limits to decide whether the autocorrelation is significant). The goal is NOT to have any significant autocorrelation remaining in your residuals:

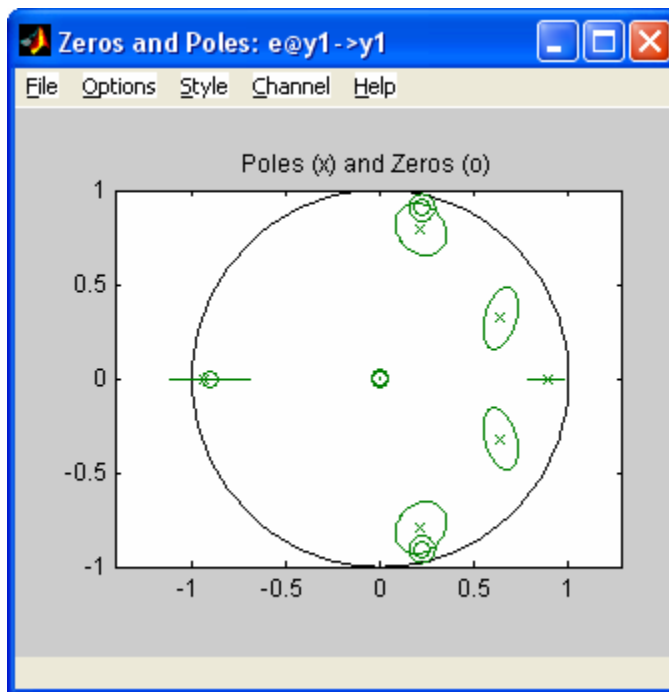


This plot is promising – the hypothesis limits are the dashed lines, and we can see that none of the autocorrelations is significant. The horizontal axis (labelled “samples”) is the lag for the autocorrelation. If you want to change the significance level, you can use the Options window menu item. Clicking File|Copy Figure will copy the figure to the clipboard, and you can then paste it into a Word document (or some other program, if you wish).

The Zeros and Poles window is useful to check for pole-zero cancellation. Sometimes when you have numerator and denominator orders specified higher than necessary, you may get cancellation of a zero and a pole (one or more), effectively producing a lower-order relationship. Parsimony (having as few parameters as possible) is the name of the game, so you want to be on the lookout for this. Poles are shown as “x”, and zeros as “o” (the bold “o” at the centre just marks the origin, as far as I can tell!) It is useful to put confidence intervals on the pole and zero locations (recognizing that these are generated approximately). If there is overlap between confidence intervals, then the poles/zeros may be the same. Click on the Options menu item, and then select “Show confidence limits”. The default confidence level is 99% - 95% is more typically the norm, and you can change this by using the “Set confidence level” item from the same sub-menu. The plot will now look as follows:



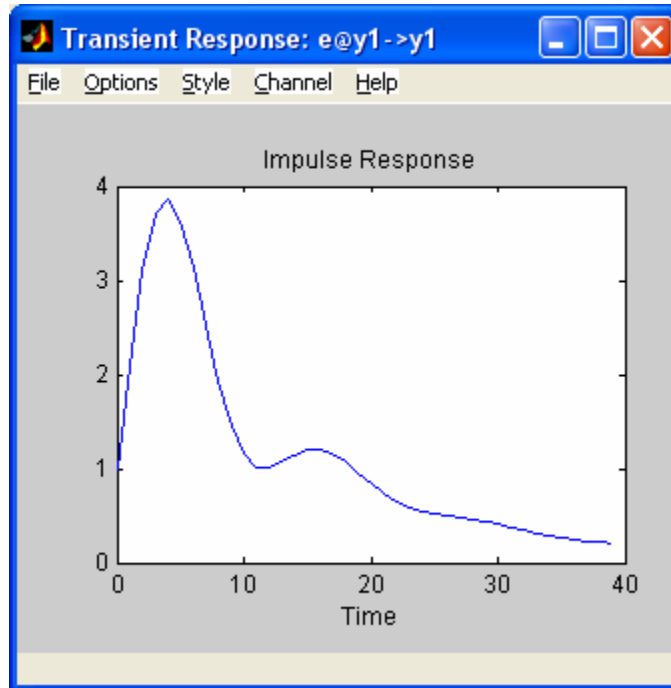
Note that poles on the real (horizontal) axis have confidence intervals that are lines, while those with complex parts off the horizontal axis have intervals that are ellipses. This is because there is a real and imaginary part, and each has some statistical uncertainty. For the example that I have shown here, there are no zeros, so we can't go looking for any overlap. However, consider the following:



Here, I (rather foolishly) estimated an ARMA model with 6th order AR term, and 3rd order MA term. You can see that two of the zeros and two of the poles have overlapping

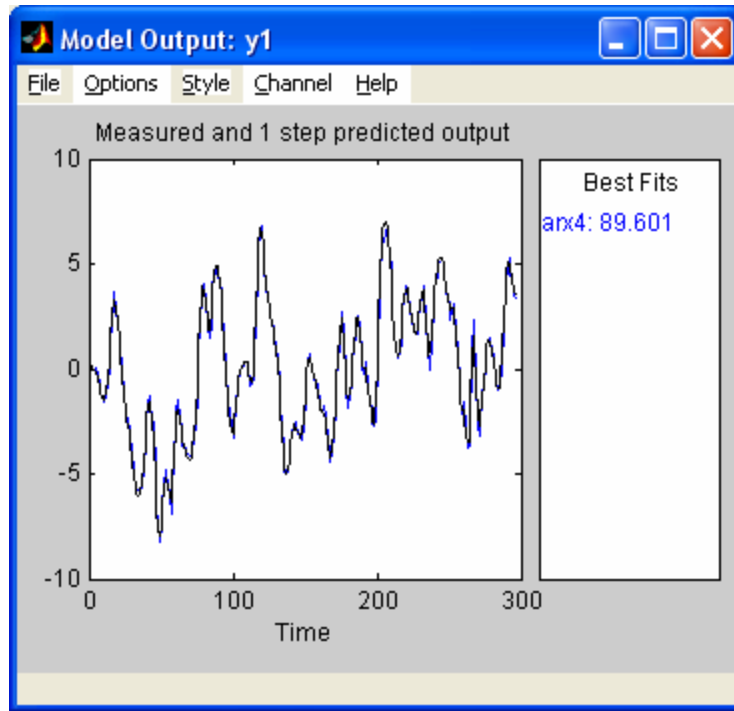
confidence regions. With a result like this, you should consider dropping the order of the numerator and denominator down by 2 each, i.e., make the MA 1st order, and the AR 4th order.

Checking off the “transient response” box produces a step response plot for the estimated model. It isn’t that useful – it shows what the associated responses look like for your active models (recall that we are looking at time series models, so the notion of a step response isn’t very natural). You can show the impulse response by going to the Options menu item and selecting impulse response (this will be of more interest, since it represents the infinite MA form of the model).



Checking off “Model Output” produces a plot that shows the validation data (usually the working data) and the predicted values on the same plot, each as a time series. Along the right hand margin is the list of active models, with a performance function that is essentially the R-squared for each model. In this instance, larger is better, and the active models are listed in descending order of quality of fit (by this measure).

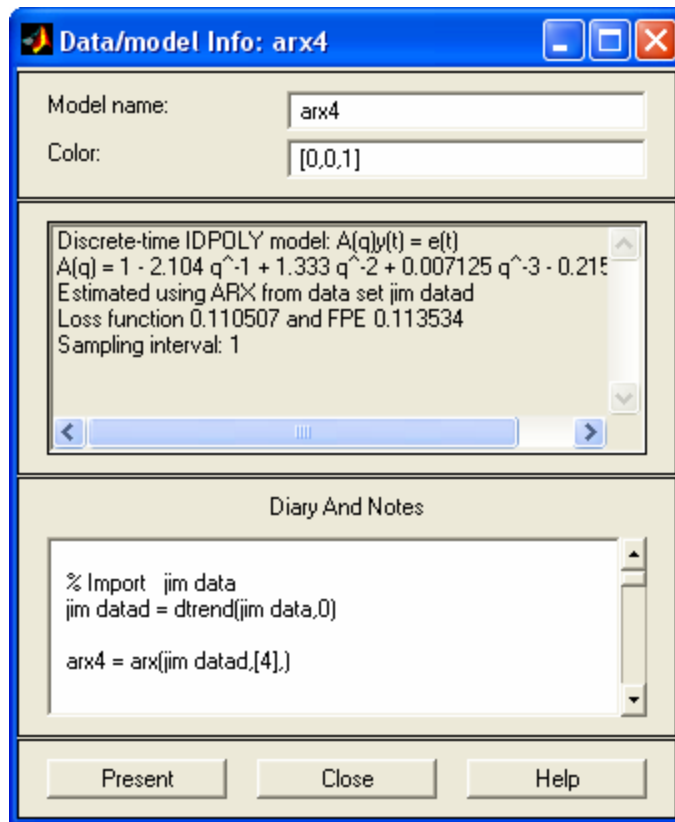
By default, the Model Output window shows the 5-step ahead predictions, while the estimation is usually done with 1-step ahead predictions. Consequently, I typically go to the Options menu item and change the predictions to 1-step ahead, and use this as a basis. In this way, the residuals should be pure random noise, and the predicted time series should follow the validation series closely (but not exactly – there is a random error component just like in conventional regression).



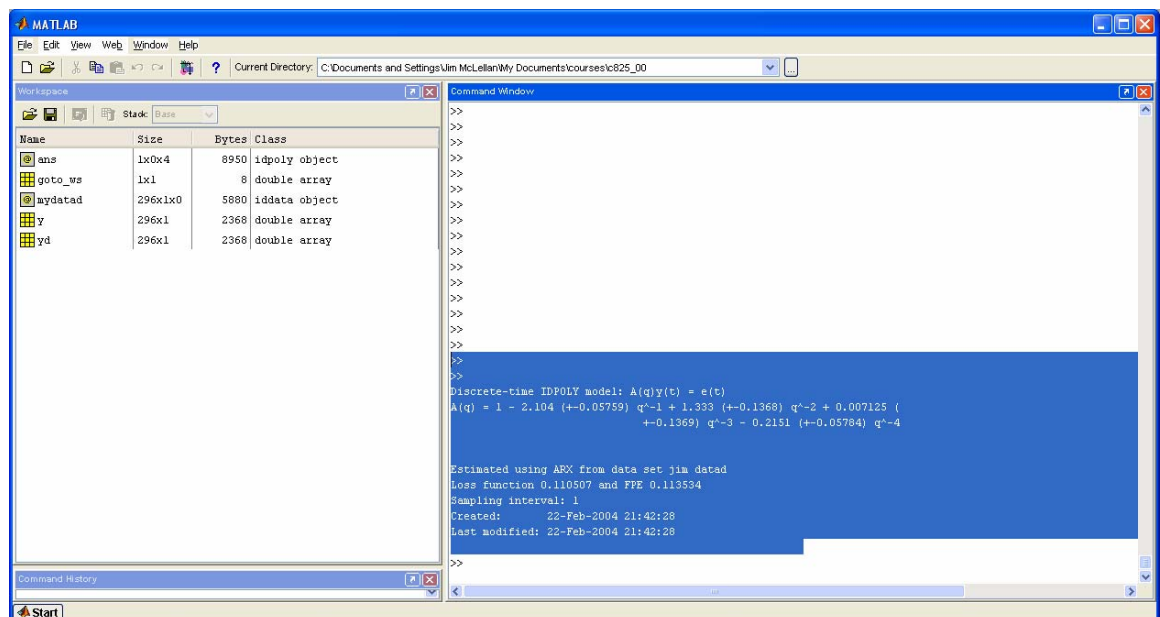
The fit looks encouraging here! You can zoom in on the graph by moving the cursor to the graph area, and clicking with the left mouse button (each click zooms in more). To zoom out, hold ctrl down, and then single click on the window area with the left mouse button.

These are the most useful diagnostic windows, particularly for time series model-building. The noise spectrum window shows the frequency content of the residuals, which can be used to guide the type of disturbance model. More on that later. (For time series models, the spectrum shown is simply that of the time series, y).

11. Quantitative diagnostics – you can generate quantitative diagnostics by moving the cursor over the model that you are interested in, and **right-clicking** the mouse. The following window will appear:



This window lists the actual parameter estimates, and it tells you the loss function and FPE (final prediction error) values. Clicking “Present” generates a summary in the **Matlab window** (your original Matlab session, from which you entered the “ident” command):



I have highlighted where you will see the information. For example, the output for the arx4 model that I have been working with is shown below. The quantities in parentheses besides each estimated parameter is the standard error (i.e., the standard deviation) of each parameter estimate. 95% confidence limits for each parameter are roughly given by 2*standard error for each parameter estimate. Looking below, the 1st, 2nd and 4th lag parameters are statistically significant, while the 3rd lag (q^{-3}) parameter is not, since ± 2 *standard error for this parameter encompasses 0. The approach is the same as you would use in conventional (linear) regression – confidence intervals encompassing 0 indicate terms that should be considered for deleting from the model. In this instance, I would drop the order of the AR model down from 4 to 3, and re-estimate.

A couple of other points – these standard errors are frequently approximate (whenever there are MA terms in the model). MA terms in fact turn the estimation problem into a nonlinear regression problem (to see this, think about doing long division of the MA terms to produce an infinitely-long AR model, which could be considered as a conventional lagged regression problem. The MA parameters would divide any AR parameters present, and would appear nonlinearly in the resulting expression regardless of whether there were AR terms present). Secondly, the parameter estimates are likely to be correlated, so all of the warnings that you heard in your linear regression course about marginal (single) confidence intervals vs. joint confidence regions apply here as well. In general, parameter estimates in time series models will be correlated.

Discrete-time IDPOLY model: $A(q)y(t) = e(t)$

$$A(q) = 1 - 2.104 (+0.05759) q^{-1} + 1.333 (+0.1368) q^{-2} + 0.007125 (+0.1369) q^{-3} - 0.2151 (+0.05784) q^{-4}$$

Estimated using ARX from data set jim datad

Loss function 0.110507 and FPE 0.113534

Sampling interval: 1

Created: 22-Feb-2004 21:42:28

Last modified: 22-Feb-2004 21:42:28

12. Saving your work – you can save your identification session by clicking File|Save session as on the identification toolbox GUI. When you come back to your work later, you pull up the System Identification toolbox, click on File|Open, and you will have your identification session back with all data and active models that you had before. You should really save regularly as you go through an estimation session so that you have a fallback to refer to.
13. You can delete datasets and/or models from the identification session by selecting and holding the left mouse button down on the icon and dragging it to the Trash bin. The bin will then bulge to show that it has some “garbage” in it. The bin gets emptied automatically when you exit the session.
14. And finally... Building a model – selecting a model is an iterative procedure of using pre-screening diagnostics (auto- and partial autocorrelation plots) to suggest a time series model structure, estimating models, assessing diagnostics, and re-estimating until you get a **set** of diagnostics that look good. Don't rely on a single number or diagnostic.

Particularly important diagnostics include the residuals autocorrelation plot (want no residual autocorrelation), statistically significant parameter estimates, model predictions that match the data, and no pole-zero cancellations.

15. Good luck.