CS522 - Homework 2 FAQ and Hints

Tibor Jánosi

March 4, 2005

This document summarizes the answers to questions that arose in connection to homework 2. The explanations and clarifications below have already been given during lectures and during the supplementary meetings that we had over the last few weeks. Some questions arose in questions that some of you asked privately. A systematic written presentation will be useful for many of you.

1. Can I clean my data using program XX?

   You must come to know your dataset well, and to examine it thoroughly for possible errors. Since the there is such a large quantity of data that you must handle, you will not be able to do this by manually examining it. It is OK to use any widely available tool (Matlab, Excel, g/awk, Perl), as long as you tell us what you did, and how you did it. Using Matlab you can programmatically check for any conditions that you want to impose. You must be aware, however, that certain tests will be difficult to perform in an environment that does not support programming.

2. My data reading/cleaning code is very slow in Matlab - can I do something about it?

   The best approach is probably to reduce the number of transactions that your are working on as early as possible. For example, you could read in one day's data at a time. Even better, you can split one day's data into smaller files that correspond to one Treasury only. Then you can process the data per Treasury. You will ultimately need to pull together the resulting data for all Treasuries, but this final set of values will be very small. For example, if you are computing the forward rate curves, you will only have no more than around 200 price points per day.

   Vectorization will also probably help. Try to rely as much as you can on Matlab's implicit loops, rather than writing explicit 'for' loops.

   Finally, you can try to use a computer that has a faster processor and/or has more memory.

3. I found XX bonds that were callable. Is this right? How do I report the results? Should I talk about something else as well?

   What we are looking for are the tests that you performed on your data. If your data set is good, your tests will only detect a small number of 'bad' transactions. If your data is bad, you will see many such failures. The point is that you will not know until you actually look.

   You should tell us what you did (i.e. what tests you have performed), and what you have found. You should tell us the distribution of various eliminated bonds based on leftover bond maturity.

Depending on how much time you can put into this, you can perform pretty sophisticated tests. We do not expect you to test everything that could be testable, but you should undertake reasonable tests. For example, you should not accept records with prices that are too high, or negative.

4. Should we worry about accumulated interest?

   Yes. Prices, as quoted by CRSP and GovPX are clean. You must add accumulated interest.

5. Should we add accumulated interest to bill prices?

   The accumulated interest tries to capture the accumulation of the next coupon payment due as time passes. A bill has no coupons, so it does not have accumulated interest.

   You can also think of a bill that has a 0 coupon. In this case you can 'compute' accumulated interest, but the resulting value will always be trivially 0. This point is not gratuitous; it might be to your advantage to treat bills like bonds that have a 0 coupon. Such an approach will simplify your programs, since you do not have to make an exception for bills.

6. It there a problem with bill prices? In CRSP I see values just under 100, in GovPX I see values around 4-5 for these.

   Well, CRSP gives you actual dollar (clean) prices. GovPX, as its scarce documentation mentions, provides bill prices on a 'discount yield' basis. This pricing convention for bills has been described in the writeup on Treasury markets at page 10. The 'discount yield' or the 'discount basis' is not the same as a continuously compounded yield, the notion of yield that we used in our computations.

7. Which GovPX prices should we use?

   For the purpose of generating forward rate curves you should use the last intra-day traded price for each instrument for which you have data.

   It is important to note that GovPX records are not necessarily in order. It is typical for a few entries to be made at the end of the day; these entries will be physically at the end of the records for the respective day. You can restore the true order if you follow the various time stamps and counter fields that GovPX provides. The per-instrument 'record transaction count' (RTC) field is a very reliable indicator of the true underlying order of transactions.

   This being said, 'last' traded price means 'last in the true ordering' or 'last with respect to the wall clock' not 'last in the file'.

   For this purpose, it does not matter that the last trade was a 'take' or a 'hit.'

8. What is the meaning of the term 'inconsistent with previous trading history'?

Under normal circumstances, for example, the total traded volume for any instrument can only stay constant (if no trading occurs, but other information is distributed, e.g. bid or ask prices), or go up (when an actual trade occurs). A record showing a total trading volume that is decreasing would be incorrect, and should be dropped. For this test we must look at the GovPX records in their true (RTC-based) order.

One could try to be very sophisticated about reconstituting the trading history. Consider the following pattern of total traded volume: 1, 3, 9, 4, 5, 10. A simple solution would involve dropping 4 and 5, as these are both less than the last 'consistent' value we saw before (9). A more sophisticated solution would be to minimize the number of transactions dropped; in this case dropping 9 (a single record) would solve the problem. The simple solution is more than acceptable.

9. Point (4)[1] asks us to ignore bid/ask prices, but don't we need them for (5)?

   Yes, you do. You should ignore bid-ask prices in GovPX for the purpose of generating forward rate curves. When plotting the graphs in (5), you will need to plot the evolution of the bid-ask prices. This is not a complication. You can drop the bid-ask prices after you have captured the information that you will need for (5).

10. Point (6) and (7) requires us to work with CRSP data, which contains bid and ask prices. Which ones should we use?

    Use either bid or ask prices; mention your choice in your writeup.

11. I have cleaned the data, I created the plots. How do I go about generating the forward rate curve?

    We have to solve an inverse problem. Theory tells us how to compute prices of bonds given a forward rate curve. In reality, however, we observe the bond prices, and we must infer the forward rate curve.

    Talking abstractly, the forward rate curve can be seen as a parameterized function of time. For concreteness, assume that our forward rate curve has the following functional form:

$$f(t) = a + e^{-bt} \tag{1}$$

    Here a and b are parameters, and t represents time from the current date to a date in the future (e.g. the time when a certain cash-flow will occur). If we know the values of the two parameters, then f is fully determined, i.e. given a time t we can fully compute the value of $f(t)$.

    Both in the Svensson and the smoothest forward rate curve model our curves are parameterized.

    Here is, in principle, the structure of the program that you need to write:

[1] These 'points' refer to the homework, not to the items in this writeup.

(a) Read in the data. You will probably find it useful to organize your data as an array of structures, so that each element of the array corresponds to a Treasury, and the fields of the structure describe the respective instrument. Thus, you will likely have fields for maturity date, time and amount of cash flows, and other relevant characteristics.

(b) Choose some initial parameter values. Generally speaking, in non-linear optimization you do not know what initial values of the parameters will work. Briefly, this problem is due to the existence of local minima in the function that you are trying to minimize. To find the true solution you have to start 'close' to the global minimum. What 'close' means is, of course, problem-dependent. In general, unless you have spent a lot of time characterizing your function (and you have been successful) you will not know too well where to start.

Given this, you have a number of alternatives that will likely be successful:

i. You can try random combinations of the parameters, being minimally careful so that these parameters do not involve extreme forward rates. It is unlikely, for example, that you will find the true solution if you choose initial parameters so that the implied forward rate is around 10^{100}.

ii. You can try with a set of simple, or even trivial, forward rate curves, and infer the value of the parameters that correspond to them. You might assume, as a first approximation that the forward rate is constant across time, or that it slopes upward, or that is slopes downward. You can even assume that the curve is broken into two-three pieces with distinct characteristics (this latter idea can be exploited on curves defined on subintervals, like the smoothest forward rate curve).

Let us assume that you decided to try parameters corresponding to constant forward rate curves. In the Svensson model, this would correspond to β_0 being non-zero, while all the other parameters are 0. Now you can try a series of values from a lower limit to a reasonable upper limit. For example, you might vary β_0 from 0.02 to .10 (10%), with a step of .01 (2.5%). You can also infer a reasonable value for β_0 if you compute the continuously compounded yield of the longest-maturity bond and you vary β_0 in an interval centered around this value.

In the case of simple sloping curves, you might try to vary the intercept and the slope of the curve, and infer the value of the respective parameters.

iii. If you have a notion, or an educated guess, about the likely values of the parameters, you can use *meshgrid* to generate a set of points in your parameter space. You can then explore these points systematically; however, this will take a long time. For example, if you assume that each parameter will be sampled at three points, five parameters will results in 3^5 parameter combinations. You might want to randomly sample points in this space, or to explore the 'corners' of your parameter hypercube, or otherwise reduce the number of parameter combinations that you will examine.

(c) Set up your model function. Your model function (named "hat" in the example that you have on the course slides) is the most important component of your homework. Abstractly, the solution for both the Svensson and the smoothest forward rate curve (SFRC) is the same, and it involves the following steps:

i. Retrieve the parameters from the parameter array.

ii. Build your curve. For Svensson this is trivial; if you have the parameter values you have a fully defined curve, for the SFRC you have to solve a system of linear equations first.

iii. Integrate your curve so that you can evaluate expressions like $\int_0^t f(\tau)d\tau$. The functions you are dealing with are easy to integrate analytically; indeed, you do not need to know more than to integrate polynomials, and simple exponential functions.

iv. Use the integral to compute discounted cash flows $ce^{-\int_0^t f(\tau)d\tau}$.

v. Sum the discounted cash flows for each bond to get the computed value of each bond. These are the results that $lsqcurvefit$ needs. You must make sure that the ordering of the market prices, computed prices, and bonds in the bond collection is the same, i.e. that $computed\ price_i$ and $market\ price_i$ correspond to $bond_i$ in the bond collection.

The model function can be seen as an abstract function defined on the space of n parameters and Treasuries, with positive real values $f : R^n \times BONDS \rightarrow R_+$. This function answers a simple "question" given a certain set of parameters and a set of Treasuries, what is the computed price of these Treasuries? The model function must implement the theory we have learned about, i.e. getting from parameters to the forward rate curve, from the curve to discounted cash flows, and from here to the computed price of Treasuries.

Let us now look at an example based on the simple function given above in formula [1]:

```
function comp_prices = model(pars, bonds)
  % Retrieve parameters.
  a = pars(1);
  b = pars(2);
  % The curve is fully determined now, since we know a and b.
  % f(t) = a + exp(-b*t). The integral from 0 to t for this
  % function is F(t) = at - exp(-b*t)/b.
  for i = 1 : number_of_bonds
    cf = bonds(i).cash_flows;
    cft = bonds(i).cash_flow_times;
    comp_prices(i) = sum(cf .* exp(-a*cft-exp(-b*cft)/b));
  end
```

For this simple function, the implementation is trivial, and very similar to what you need to do for the Svensson curve. Retrieving the function and integrating it analytically can be done without any explicit coding in Matlab. The SFRC is a bit more complicated, but it is addressed in the next question.

(d) Call function $lsqcurvefit$. If you have dealt with the previous steps, this step is very simple. You must pay attention to provide all the arguments needed, and to capture all the results needed. You should study the meaning of the various options that you can set. Especially in the early stages, when you are not sure whether your model function works, you probably want to relax the parameters governing the precision of your solution (e.g. set $tolFun$ to a higher value); such settings can lead to a great increase in speed and they do not always degrade significantly the quality of the results.

You can think of $lsqcurvefit$ as a smart helper who has the job of finding which parameters are those conducive to the least sum of squared errors (or residuals, i.e. the difference between model and market prices for the same bond). You tell this helper where to start, the helper then "probes" your model function by calling it with various parameter combinations until the solution can not be improved upon.

(e) Finally, you have to process the results to deliver the results the problem asks for.

12. How do choose knot points for the smoothest forward rate curve (SFRC)? How many knot points do we need?

    You want a small number of knot points, probably not more that 5.[2] You should chose knot points to reflect what you know about your data and about the factors that influence the shape of the forward rate curve. Knot points at 1, 2, 5, and (almost) 30 years are likely to work reasonably well. If you have too many knot points (say, 30, at a distance of one year from each other), you will end up matching all prices very closely, but your curves will contain the noise and distortions embedded in your data. Such curves will not reveal too much about the underlying "true" forward rate.

13. I do not understand how to set up the system of equations for the SFRC. What do I need to do?

    We obtained the conditions for SFRC by solving the minimization problem $\int_0^T (f''(\tau))^2 d\tau$ with the price condition $\int_0^{t_i} f(\tau)d\tau = -\log p_i$, $i = 1, n$. The price conditions express the fact that we want the curve to price a default free, one-dollar payment at time

[2] Point t = 0 is an implicit knot point in all curves. The last (rightmost) knot point should be chosen to correspond with the leftover maturity of the longest outstanding bond, which will be close to 30 years for the dataset that you have. Knot points other than the leftmost one (0) and the right-most one will be called "intermediate knot points." Unless we state specifically, "knot point" refers to any intermediate knot point or to the right-most knot point.

t_i in the future to be worth p_i dollars. Evidently, the p_i values should decrease; the farther in time t_i is in the future, the smaller p_i should be, assuming positive interest rates. Thus we must have $0 < p_i < 1$ for all i, and $p_(i-1) < p_i$ if $t_(i-1) < t_i$ (note that $t_0 = 0$).

The solution to our minimization problem consists of a function that is a 4th-degree polynomial on each interval $t_{i-1} < t_i$. More, at each intermediate knot point the function is continuous, together with its first three derivatives. We have also imposed the condition that the second and third derivative at the left and right end of the curve is equal to 0. Adding the price restrictions $\int_0^{t_i} f(\tau)d\tau = -\log p_i$, $i = 1, n$ discussed above, we can write a system of $5n$ equations with $5n$ unknowns to determine the coefficients that determine our forward rate curve (n intervals $\times$ 5 coefficients per interval $= 5n$ coefficients).

Let us assume that our polynomials are defined as given below (i is the index of the interval; we identify the interval with its right-end point t_i):

$f(t) = a_{i4}t^4 + a_{i3}t^3 + a_{i2}t^2 + a_{i1}t + a_{i0}$, $t_{i-1} \le t \le t_i$, $0 < i \le n$

Note that the unknowns are the coefficients a_{ij}, $0 < i \le n$, $0 \le j \le 4$. In our approach we know where the knot points are and what the values of the parameters p_i should be. "All" we need to do is determine the coefficients a_{ij} to know the curve.

We can compute the first three derivatives of the polynomial on interval i as follows:

$$\begin{aligned} f'(t) &= 4a_{i4}t^3 + 3a_{i3}t^2 + a_{i2}t + a_{i1} \\ f''(t) &= 12a_{i4}t^2 + 6a_{i3}t + a_{i2} \\ f'''(t) &= 24a_{i4}t + 6a_{i3} \end{aligned}$$

Let us denote the polynomial on interval i ($[t_{i-1}, t_i]$) by $P_i(t)$.

Now, the continuity conditions at the intermediate knot points t_i, $0 < i < n$, imply that the equalities hold:

$$\begin{aligned} P_i(t_i) &= P_{i+1}(t_i) \\ P_i'(t_i) &= P_{i+1}'(t_i) \\ P_i''(t_i) &= P_{i+1}''(t_i) \\ P_i'''(t_i) &= P_{i+1}'''(t_i) \end{aligned}$$

We can immediately write the following equivalent relations:

$$\begin{aligned} P_i(t_i) - P_{i+1}(t_i) &= 0 \\ P_i'(t_i) - P_{i+1}'(t_i) &= 0 \\ P_i''(t_i) - P_{i+1}''(t_i) &= 0 \\ P_i'''(t_i) - P_{i+1}'''(t_i) &= 0 \end{aligned}$$

For example, the condition on the continuity of the second derivative at t_i can be written as the following equation that must be satisfied by the coefficients:

$$12a_{i4}{t_i}^2 + 6a_{i3}t_i + a_{i2} - 12a_{(i+1)4}{t_i}^2 + 6a_{(i+1)3}t_i + a_{(i+1)2} = 0$$

Analogous relations can be written for the continuity of the function, and its first and third derivative at the intermediate knot points.

At each knot point, except for 0, we also have a price condition imposed by the condition $\int_0^{t_i} f(\tau)d\tau = -\log p_i$. This relationship is problematic, as it potentially implies an integration over several intervals. To avoid this, we rewrite the pricing formulas so that they involve a single interval:

$$\left.\begin{array}{l}\int_0^{t_i} f(t)dt = -\log p_i \\ \int_0^{t_{i-1}} f(t)dt = -\log p_{i-1}\end{array}\right\} \Rightarrow \int_{t_{i-1}}^{t_i} f(t)dt = \log\frac{p_{i-1}}{p_i}$$

Note that $p_0 = 1$ (the price of a sure dollar paid "now" is equal to one dollar).

Let us know integrate P_i from t_{i-1}, the left end of the interval on which P_i is defined, to t_i:

$\int_{t_{i-1}}^{t_i} f(\tau)d\tau = \log\frac{p_i}{p_{i+1}}$

$\frac{1}{5}a_{i4}(t_i^5 - t_{i-1}^5) + \frac{1}{4}a_{i3}(t_i^4 - t_{i-1}^4) + \frac{1}{3}a_{i2}(t_i^3 - t_{i-1}^3) + \frac{1}{2}a_{i1}(t_i^2 - t_{i-1}^2) + a_{i0}(t_i - t_{i-1}) = \log\frac{p_i}{p_{i+1}}$

Remember that we know the value of t_i's (since you chose the knot points), and that the unknowns in this equation are the values of the coefficients a_{ij}.

Finally, $P_1''(0) = 0$, $P_1'''(0) = 0$, and $P_1''(t_n) = 0$, $P_1'''(t_n) = 0$, immediately provide the equations corresponding to the conditions at the end of the curve.

To take a specific example, consider the case when we have two (non-zero) knot points, so that $t_0 = 0 < t_1 = 1 < t_2 = 2$.

	a_{14}	a_{13}	a_{12}	a_{11}	a_{10}	a_{24}	a_{23}	a_{22}	a_{21}	a_{20}		RHS
f'(0)			2								=	
f''(0)		6									=	
p(1)	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1						=	$\log\frac{1}{p_1}$
f(1)	1	1	1	1	1	-1	-1	-1	-1	-1	=	
f'(1)	4	3	2	1		-4	-3	-2	-1		=	
f''(1)	12	6	2			-12	-6	-2			=	
f'''(1)	24	6				-12	-6				=	
p(2)						$\frac{31}{5}$	$\frac{15}{4}$	$\frac{7}{3}$	$\frac{3}{2}$	1	=	$\log\frac{p_1}{p_2}$
f'(2)						48	12	2			=	
f''(2)						48	6				=	

In the example above, the leftmost column hints to the type of constraint that the respective equation expresses; for example, the line labeled f_1''' expresses the continuity condition of the third derivative of the forward rate curve at the point

$t = 1$. We have marked explicitly the columns that correspond to the unknown coefficients a_{ij}. All values that have not been explicitly given are 0.[3]

You will note that the left-hand side (LHS) of the system of equations only depends on the position of the knot points, while the right-hand side (RHS) of the system depends only on the parameters p_i. Since you choose the knot points, you can determine the LHS "by hand," or you can write a short function that will determine the appropriate coefficients given the series of knot points. The RHS is variable (i.e. it depends on the parameters that are provided to the model function), but it has a very simple form.

Important note: We presented above the version of forward rates in which polynomials are based at $t = 0$. As we have pointed out, it is possible to write polynomials that are based at the left end of their associated intervals. Such polynomials have the advantage that their associated coefficients are smaller on intervals far from $t = 0$. Perhaps more importantly, the system of equations for determining these coefficients that have more zeros than for the version we have presented. You can use *mkpp* to evaluate these polynomials.

14. How do I determine the initial value of the parameters for the SFRC?

    See the description of the model function above. Let us assume that you want to set the initial guesses to imply a constant forward rate curve at the level of 5%. You would then set parameters to have values $p_i = exp(-0.05 * t_i)$, $1 \leq i \leq n$.

    To take another possibility, you could set the forward rate to be uniformly upward sloping, say, starting at 2% at 0 and 5% at $t = 30$ years. The equation of the forward rate curve corresponding to these assumptions is

    $f(t) = 0.02 + (0.05 - 0.02)t/30 = 0.02 + 0.001t$

    We immediately have the following:

    $F(t) = \int_0^t f(\tau)d\tau = 0.02t + 0.0005t^2$

    If you have chosen knot points at times t_i, $1 \leq i \leq n$, then the values of the respective parameters p_i will be given by the following formulas:

    $p_i = exp(-0.02t_i - 0.0005t_i^2)$, $1 \leq i \leq n$.

15. I have determined the coefficients, how do I use them?

    If you know the coefficients, then you know the polynomial that determines the forward rate curve on each interval. You have seen in class how such polynomials can be used; look up functions *polyval* and *polyint* for details. It is thus easy to evaluate and integrate polynomials once their coefficients become available. Note that if you have a cash flow at time t, and you integrate the forward rate curve from 0 to t, you will need to integrate (in general) over several intervals.

[3] Note, however, that the example might contain typos; if you notice any, please let me know.

We already know the following:

$$\int_{t_{i-1}}^{t} f(\tau)d\tau = \frac{1}{5}a_{i4}(t^5-t_{i-1}^5)+\frac{1}{4}a_{i3}(t^4-t_{i-1}^4)+\frac{1}{3}a_{i2}(t^3-t_{i-1}^3)+\frac{1}{2}a_{i1}(t^2-t_{i-1}^2)+a_{i0}(t-t_{i-1})$$

Let us denote $\int_{t_{i-1}}^{t} f(\tau)d\tau$ by $I_i(t)$. Now, if $t_i \leq t \leq t_{i+1}$, then we have the following relation:

$$\int_{t_0}^{t} f(\tau)d\tau = \sum_{j=1}^{i} I_j(t_j) + I_{i+1}(t)$$

In other words, you need to integrate on the entire interval for all intervals that end at, or before t_{i-1}, and you need to integrate on the partial interval $[t_i, t]$. This can be done trivially using *polyint* and *polyval*; all you need is to set up the appropriate loop.

Good luck with your homework!