Assignment 3.

Given February 9, due February 23.

1. Write a program that computes the Choleski factorization of a symmetric positive definite pentadiagonal matrix. The program should return an error message if it is unable to do this, for example, if the matrix is not positive definite or a parameter is out of range.

2. Write a program that takes a lower triangular matrix, L, and computes LL^t. Use this to check the L computed by 1.

3. Write a program to do forward and back substitution to solve $LL^tu = f$. It should use an externally supplied L. Write a tester that multiplies out LL^tu, given L and u.

4. Write a procedure that takes as input x, x_0, Δx, and computes the value at x of the symmertic cubic b-spline basis function centered at x_0 with uniform knot spacing Δx. The answer is $b(x, x_0, \Delta x)$. Use finite differences (with a step size much smaller than Δx) to check that b is correct. That is, that b, together with its first and second derivatives is continuous, and that the third derivative is constant between knots.

5. We are trying to reconstruct a function $y(x)$ from noisy data. The sample points are uniformly spaced: $x_k = k$, $k = 1, \ldots, n$. the approximating function will be a cubic spline with uniformly spaced knots at $r_j = s \cdot j$. Here s is a positive integer, the knot spacing. The approximating curve will have the form

$$\hat{y}(x) = \sum_{j=0}^{n/s} u_j b(x, r_j, s) \quad .$$

Statisticians use the hat to indicate a hat to indicate a statistical estimator, so $\hat{y}$ is a statistical estimate of the unknown function $y(x)$. Write a program that takes as input the data[1] $Y = (Y_0, \ldots, Y_n)$, and the parameters n and s, and computes the u_j to minimize

$$R^2 = \sum_{k=0}^{n} \left(\hat{y}(x_k) - Y_k\right)^2 \quad .$$

As one check, note that $\hat{y}$ should interpolate the data if $s = 1$. Experiment with other powers of 2 for s and with the three data sets (which have $n = 1024$). Try to get an s that gives an accurate picture of function from the noisy data set. Using least squares splines in this way is one method for smoothing noisy data.

[1] Statisticians often (but not always) use capital letters to indicate sample data that include random noise.

6. It is hard to know in advance what s to use. Taking $s = 1$ fits the data best but has little predictive power for noisy data. We will choose s using a simple form of *cross validation*. The idea of cross validation is to fit the model with part of the data and see how well it predicts the other part. In our situation, we can fit the model by minimizing the "in sample" sum of squares,

$$R_{\text{IS}}^2 = \sum_{k=0}^{n/2} \left(\hat{y}(x_{2k}) - Y_{2k}\right)^2 \quad ,$$

and then checking the predictive power by looking the "out of sample" sum of squares,

$$R_{\text{OS}}^2 = \sum_{k=0}^{n/2-1} \left(\hat{y}(x_{2k+1}) - Y_{2k+1}\right)^2 \quad .$$

The in sample sum of squares will always be smaller for smaller s, but the out of sample sum of squares behaves differently. Use this to select a good s value for each of the three data sets.