The goal of this homework is to work through the exact computations and Monte Carlo simulation of the error in a kernel estimator. There are some preliminary steps that are useful for this assignment but these will also be the skeleton of other spatial methods that we use. To run the code in the rest of this handout start up R and

```r
library(fields)
data(COmonthlyMet)
```

**Preliminaries**

- Spatial locations are organized as a matrix where the columns are the different spatial coordinates and the rows are the different spatial locations. `CO.loc` are the locations for the Colorado stations a 376X2 matrix. There are 376 stations total and the 2 coordinates are longitude and latitude. Try

```r
plot( CO.loc)
US( add=TRUE)
```

- For this homework we consider the problem of predicting the average, minimum springtime temperature across Colorado. `CO.tmin.MAM` has the observations reported for the last 103 years. We will use `CO.tmin.MAM.climate` the mean of these annual values. *Climate* here refers to the estimate of a long time average – note we are ignoring any trends or climate change in this exercise.

- The fields function `rdist` finds the (Euclidean) distance between two sets of locations `rdist.earth` works the same but assumes lon/lat co-ordinates and finds the great circle distance in miles. Great circle distance is “as the crow flies”. These functions take two arguments e.g. `rdist(x1, x2)` where x1 and x2 are two sets of locations that need not be the same. This in fact is an important feature that allows us to create cross covariance matrices.

```r
BigD<- rdist.earth( CO.loc, CO.loc)
```

BigD is a 376X376 matrix with each element being the distance between a pair of locations. e.g. BigD[15, 118] is the distance between CO.loc[15,] and CO.loc[118, ]. (In fact Boulder and Longmont.)
To simplify working with these data strip out all the missing values and create a new set of location and values. There are 213 good values (e.g. \( \text{sum(good)} \) below)

```r
good<- !is.na(CO.tmin.MAM.climate)
xHW1<- CO.loc[good,]
yHW1<- CO.tmin.MAM.climate[good]
```

Constructing the weights for a kernel estimator. Predict at the center of the region.

```r
x0<- cbind( -105.6, 40)  # about the center
theta<- 30  # distance between point to predict and all data locations
Dw<- rdist.earth(xHW1,x0)
theta.grid=c(10, 400),
smoothness=.5, m=1, Distance="rdist.earth")
```

This leads to the covariance 825.1 \* \( \exp(-d/187.8) \) and the variance for the measurement error variance is given by 1.01^2. We are assuming the errors are uncorrelated with each other and also with \( g(x) \). The advantage of this functional form is that one can posit covariance between any two locations independent of whether there are observations at these
locations. In a way it is a type of extrapolation. E.g for two locations the formula is

\[ \text{COV}(g(x_1), g(x_2)) = 825.1 \times e^{(-D(x_1,x_2)/187.8)} \]

\( D(x_1,x_2) \) is the great circle distance between the two points. In R code:

\[
\text{with locations } x1 \text{ and } x2 \text{ the covariance is}\]

\[825.1* \exp(- \text{rdist.earth}(x1,x2) /187.8)\].

This code also works if \( x1 \) and \( x2 \) are both matrices of several locations and the result is a cross covariance matrix.

The covariance matrix for the observations (\( \Sigma_V \) below) is given by

\[ \Sigma_V = (K + \sigma^2 I) \]

\( K_{i,j} = \text{COV}(g(x_i), g(x_j)) \). \( I \) is a the identity matrix and this part is added as the contribution from the measure error. In R code:

\[
825.1*\exp(-\text{rdist.earth}(xHW1,xHW1)/187.8) + \text{diag}(1.01^2, 213)\}.

### Problems

1. Find the kernel estimate for the tmin springtime means (\( yHW1 \) above) for several choices of bandwidths at the location \( x_0 \). Just report in a plot or table of \( \hat{g} \) as function of the bandwidth.

2. Find the variance of the estimate at each of these bandwidths. Do this by adapting the formula from class:

\[
U \sim (\mu_U, \Sigma_U) \text{ and } V \sim (\mu_V, \Sigma_V)
\]

\[ \text{COV}(U,V) = \Sigma_{U,V} \]

\[ Z = A_U U + A_V V \]

\[ Z \sim (A_U \mu_U + A_V \mu_V, A_U \Sigma_{U,U} A_U^T + A_U \Sigma_{U,V} A_V^T + A_V \Sigma_{V,U} A_U^T + A_V \Sigma_{V,V} A_V^T) \]

\( U \) is the true value at \( x_0 \) (just a number in this case, i.e. a vector of length 1). \( V \) is the observation vector. \( A_U \) is 1. \( A_V = -w \). \( \Sigma_U \) is 825.1. You also can assume that \( \mu_U = \mu_V = 0 \)

a) Justify the formula \( \Sigma_V = K + \sigma^2 I \) based on the assumptions of the measurement error.

b) Find the variance of \( Z \) for several bandwidths and report the values as a plot of variance against bandwidth.

3. How does your answer in 2 depend on the actual observations (\( yHW1 \))?