Midterm Examination

36-350, Data Mining

14 October 2009

No notes or calculators are allowed. All calculations can be done by hand, possibly (but not necessarily) using the facts on this page. **SHOW YOUR WORK**: partial credit will be based on work; correct answers without work will receive minimal or no credit. If you suspect you have made a mistake but cannot find it, say so, and say why you think there is an error.

Problem	Points
1	15
2	35
3	25
4	25

Possibly Helpful Facts

here, $\mathbf{x}$ is an $m \times 1$ matrix and $\mathbf{A}$ and $\mathbf{B}$ are $m \times m$

$$\begin{aligned}
\frac{d\mathbf{x}^T\mathbf{x}}{d\mathbf{x}} &= 2\mathbf{x} \\
\frac{d\mathbf{x}^T\mathbf{A}\mathbf{x}}{d\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} \\
\mathbf{A}\mathbf{x} = v\mathbf{x} &\Leftrightarrow \mathbf{x} \text{ is an eigenvector of } \mathbf{A} \text{ with eigenvalue } v \\
(\mathbf{A}\mathbf{B})^T &= \mathbf{B}^T\mathbf{A}^T
\end{aligned}$$

1. (15 points in all) Briefly define the following terms (2 pt each). Formulas are OK, but explain what the symbols in them mean.

   (a) Ward's method (of clustering)

   (b) Entropy

   (c) Inverse document frequency

   (d) Cross-validation

   (e) Nearest neighbor classifier

   (f) Dendrogram

   (g) Confusion matrix

2. *Finding reviewers* (35 pts total) Scientific papers submitted to a journal or conference are "peer-reviewed", meaning that they are evaluated by other scientists familiar with work in the area. Journal editors and conference organizers spend a lot of time selecting reviewers, and authors worry about getting good referees.

   Suppose that a journal has a database of the full text of all papers previously published in the journal, along with their authors.

   (a) (3 pts) Explain what the bag-of-words representation for an individual paper would be.

   (b) (2 pts) Explain how to combine the representations of all papers by a given author to get a bag-of-words for that author.

   (c) (15 pts) Describe an algorithm for finding the three authors whose work is most relevant to a given paper, and are not authors of the paper. (You do not have to write code, but be clear about what needs to be done.)

   (d) (5 pts) How could you use principal components analysis of bags of words to simplify and improve this system?

   (e) (5 pts) Describe how to use the bags-of-words to hierarchically cluster authors.

   (f) (5 pts) Describe another algorithm for finding peer reviewers of a paper, using the hierarchical clustering of authors.

3. (25 points in all) `state.x77` is a data set about the United States in 1977, using figures taken from the Census's *Statistical Abstract*. (You will see this again in the homework.) The variables are:

Population	in thousands
Income	dollars per capita
Illiteracy	Percent of the adult population unable to read and write
Life Exp	Average years of life expectancy at birth
Murder	Number of murders and non-negligent manslaughters per 100,000 people
HS Grad	Percent of adults who were high-school graduates
Frost	Mean number of days per year with low temperatures below freezing
Area	In square miles

The summary statistics for these variables will be helpful.

```
> summary(state.x77)
   Population        Income       Illiteracy       Life Exp
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
 1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
 Median : 2838   Median :4519   Median :0.950   Median :70.67
 Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88
 3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60
     Murder          HS Grad          Frost             Area
 Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
 1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
 Median : 6.850   Median :53.25   Median :114.50   Median : 54277
 Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
 3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81162
 Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

We will do two different principal component analyses of this data.

```
> states.pca.1 = prcomp(state.x77,scale.=FALSE)
> states.pca.2 = prcomp(state.x77,scale.=TRUE)
```

The figures following show some displays for these two PCAs, which you will need to use to answer the questions.

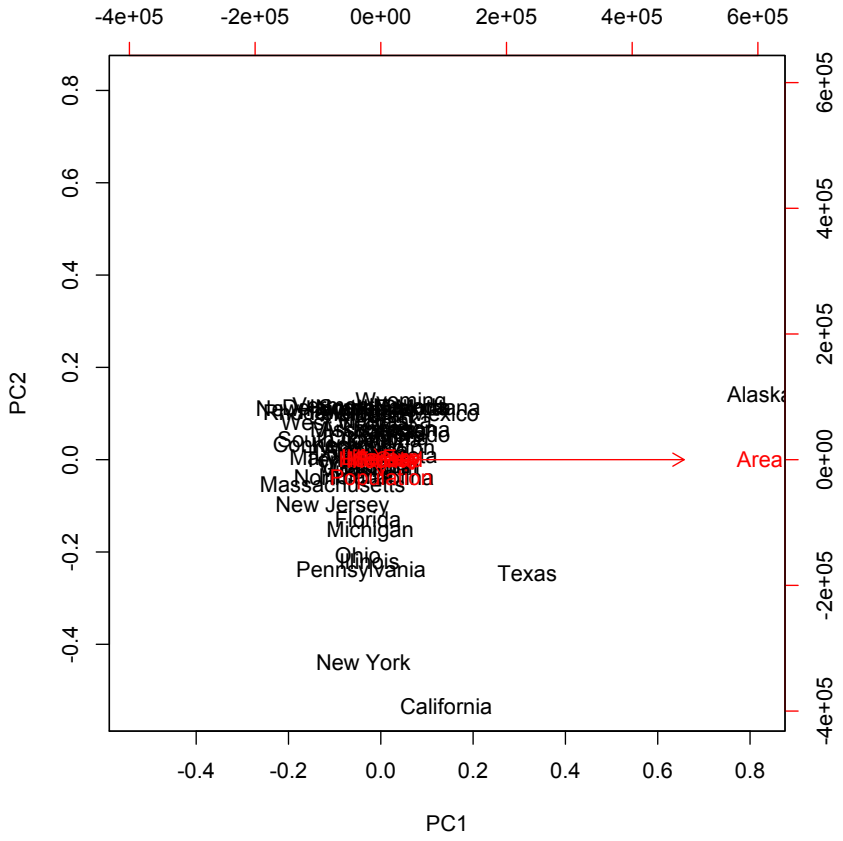

Figure 1: Biplot for `states.pca.1`.

	PC1	PC2
Population	1.18×10^{-03}	-1.00
Income	2.62×10^{-3}	-2.80×10^{-2}
Illiteracy	5.52×10^{-7}	-1.42×10^{-5}
Life Exp	-1.69×10^{-6}	1.93×10^{-5}
Murder	9.88×10^{-6}	-2.79×10^{-4}
HS Grad	3.16×10^{-5}	1.88×10^{-4}
Frost	3.61×10^{-5}	3.87×10^{-3}
Area	1.00	1.26×10^{-3}

Table 1: Projections of the features on to the first two principal components of `states.pca.1`.

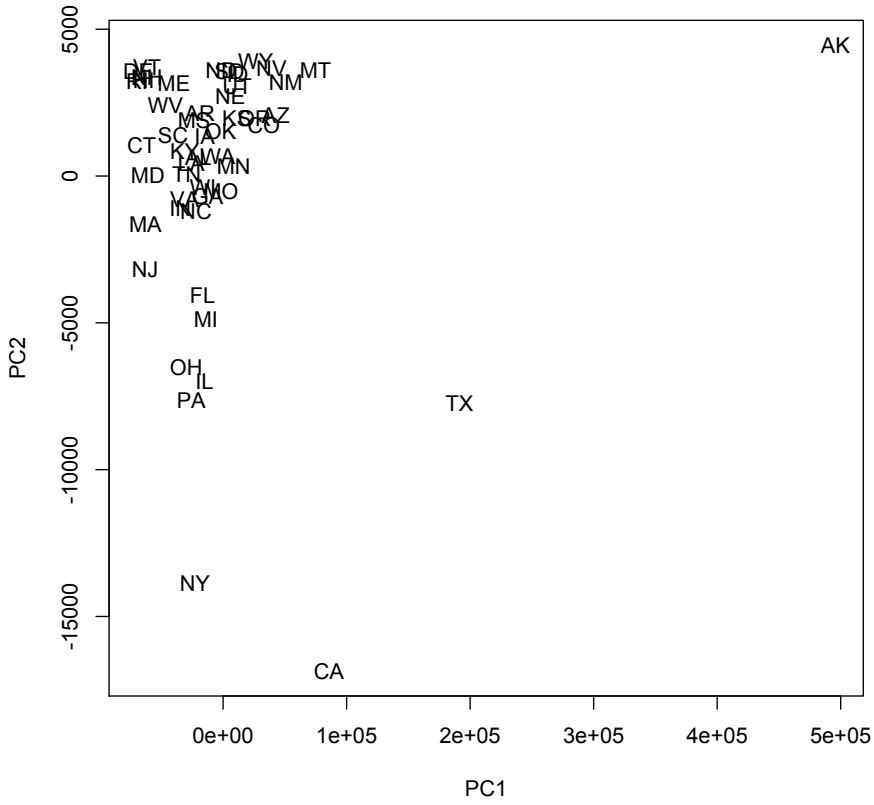

Figure 2: Projections of the states on to the first two principal components of `states.pca.1`.

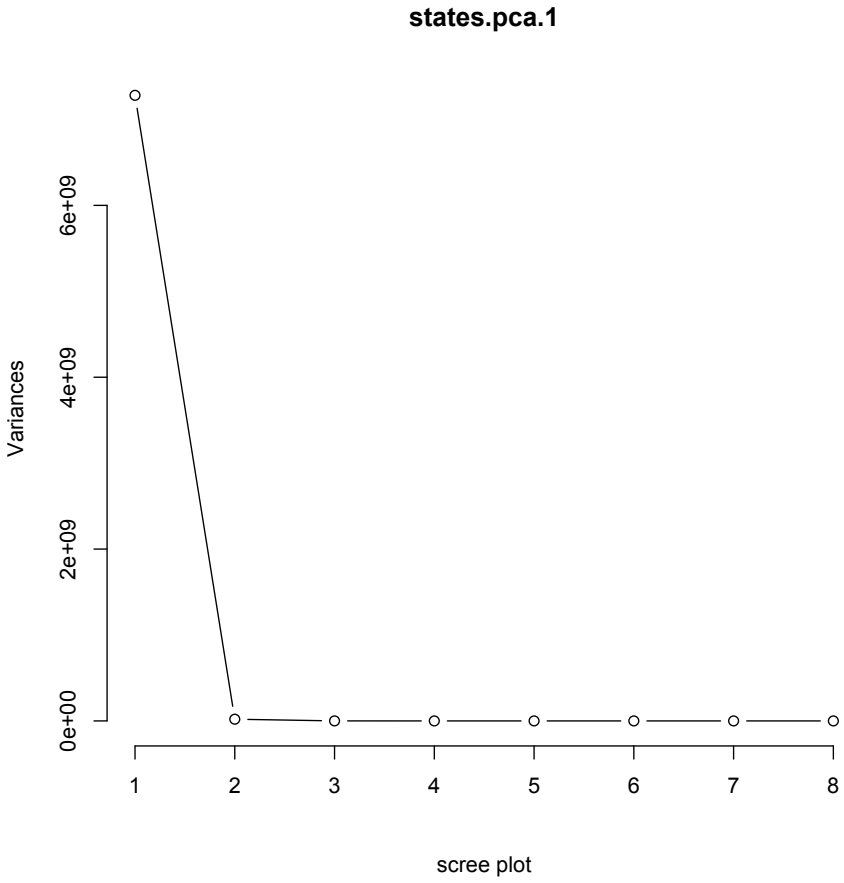

Figure 3: Scree plot for `states.pca.1`.

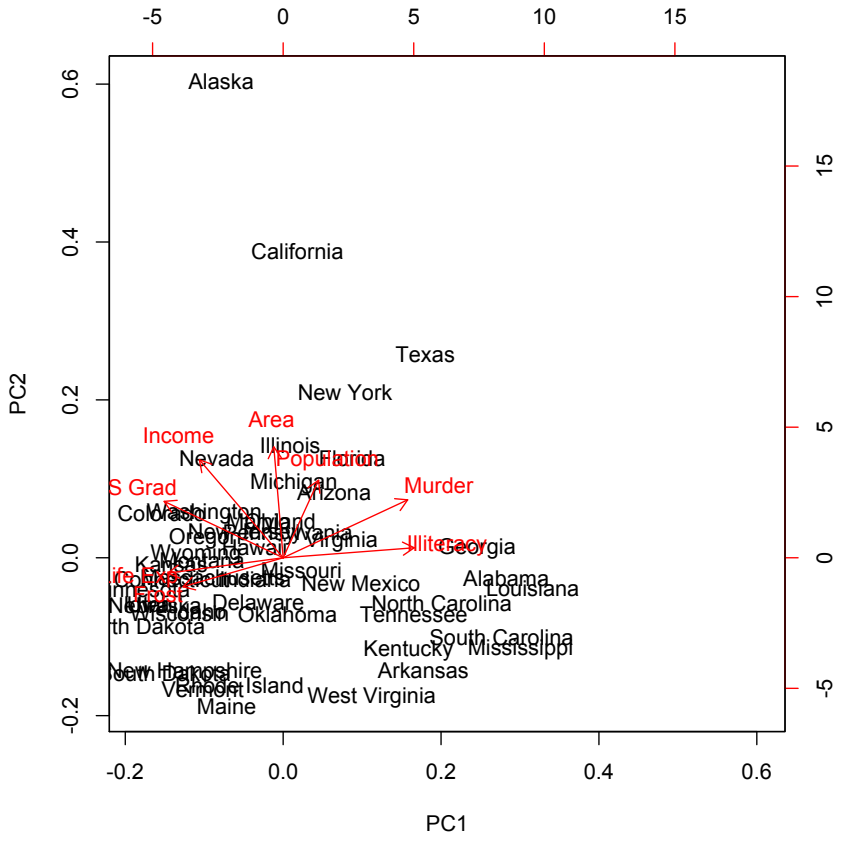

Figure 4: Biplot for `states.pca.2`.

	PC1	PC2
Population	0.1260	0.4110
Income	−0.2990	0.5190
Illiteracy	0.4680	0.0530
Life Exp	−0.4120	−0.0817
Murder	0.4440	0.3070
HS Grad	−0.4250	0.2990
Frost	−0.3570	−0.1540
Area	−0.0334	0.5880

Table 2: Projections of the features on to the first two principal components of `states.pca.2`.

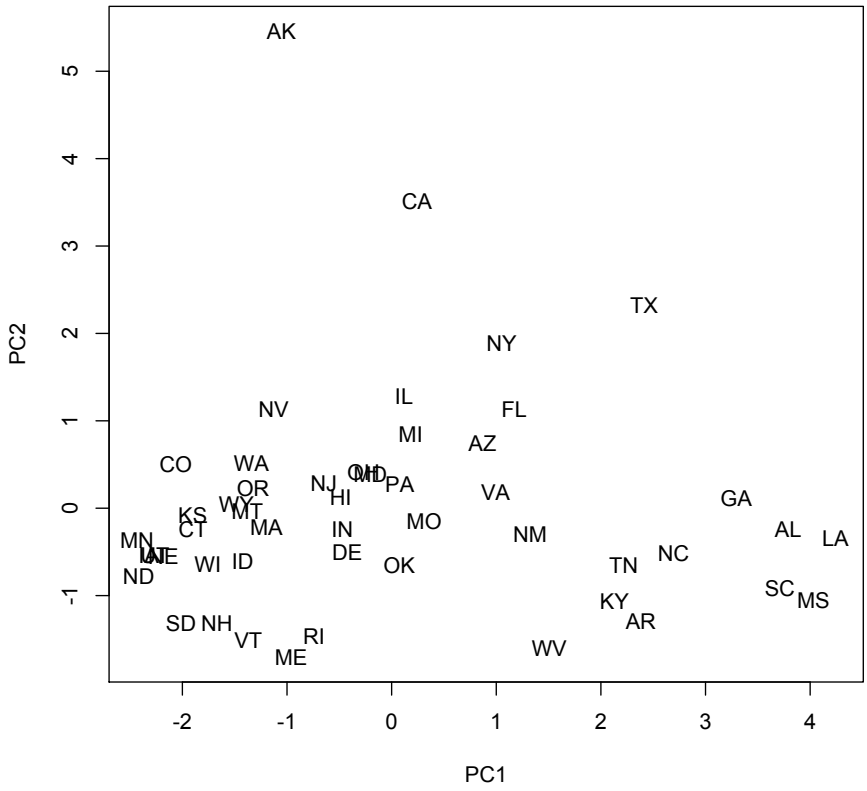

Figure 5: Projections of the states on to the first two principal components of `states.pca.2`.

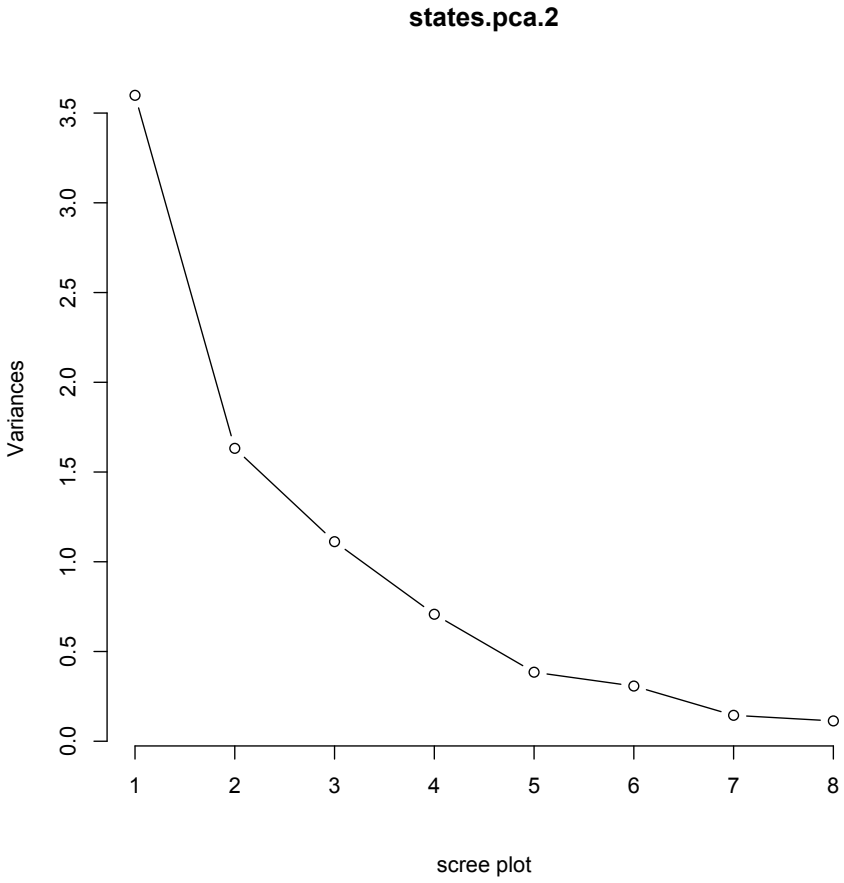

Figure 6: Scree plot for `states.pca.2`.

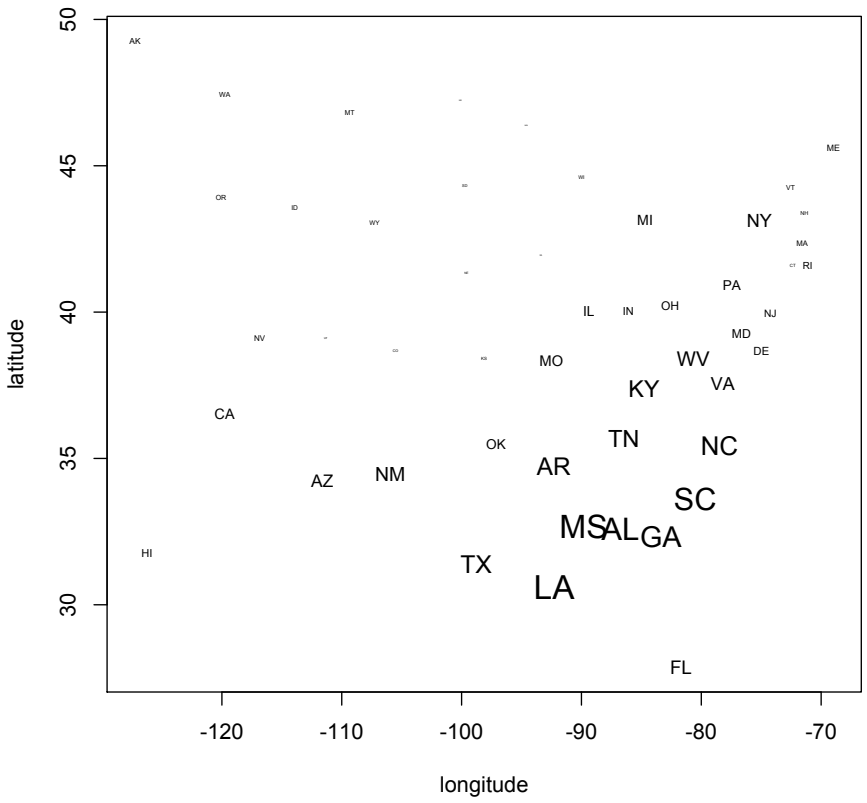

Figure 7: States in their geographic locations, with name size being proportional to the projection on to the first component of `states.pca.2`.

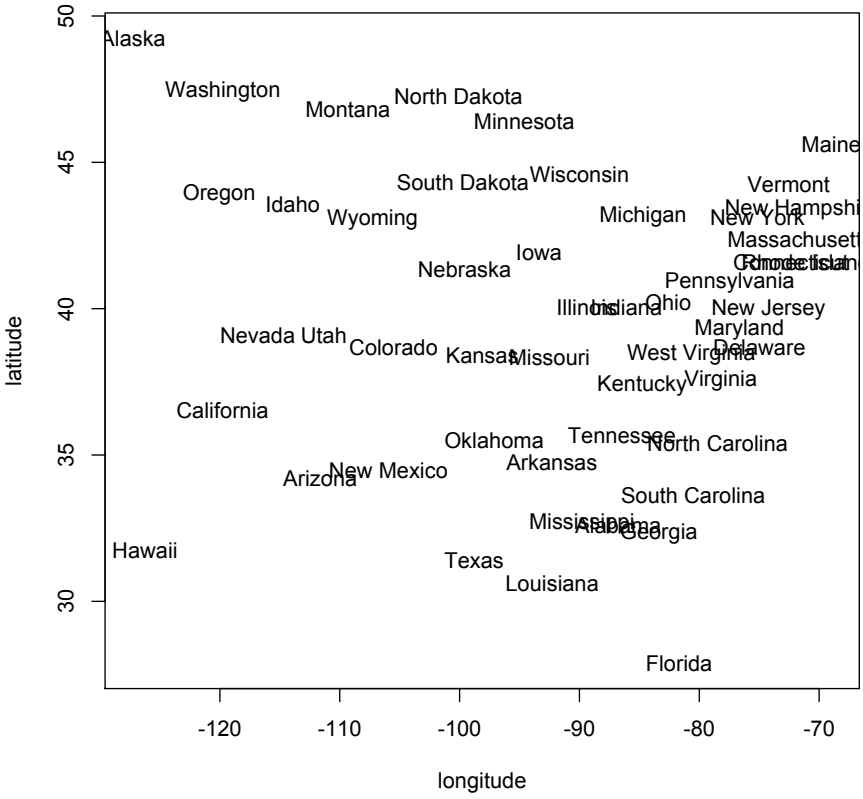

Figure 8: Geographic centers of the states (according to R).

(a) (3 pts) How does the command to create `states.pca.1` differ from that creating `states.pca.2`? What do they do differently?

(b) (6 pts) Describe, in words, the first two principal components of `states.pca.1`

(c) (6 pts) Describe, in words, the first two principal components of `states.pca.2`

(d) (5 pts) Would you rather use `states.pca.1` or `states.pca.2` for further analysis? Pick one and explain your choice. (A choice with no or inadequate reasoning will get little or no credit.)

(e) (5 pts) Figure 7 shows the states in their geographic locations, with the size of the label being proportional to the projection on to the first component (as per `states.pca.2`). What does this suggest about the interpretation of that component?

(f) (5 pts extra credit) Figure 8 shows where R thinks the states are located (using `states.centers`). Does anything look odd about the figure? Would you add these latitude and longitude values as features?

4. (25 points in all) In local linear embedding, we obtain an $n \times n$ matrix $\mathbf{w}$, where w_{ij} is the weight on $\vec{x}_j$ we use to reconstruct $\vec{x}_i$. Each row of $\mathbf{w}$ sums to one. We then try to find coordinates $y_1, y_2, \ldots y_n$ which minimize

$$\Phi(\mathbf{Y}) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} w_{ij} y_j \right)^2$$

where $\mathbf{Y}$ is the $n \times 1$ matrix of y_i values (this is the $q = 1$ case, for simplicity). In the notes, we showed that this is the same as minimizing

$$\Phi(\mathbf{Y}) = \mathbf{Y}^T \mathbf{M} \mathbf{Y}$$

where

$$\mathbf{M} = ((\mathbf{I} - \mathbf{w})^T (\mathbf{I} - \mathbf{w}))$$

(a) (2 pts) Show that $\mathbf{M}$ is a symmetric matrix.

(b) (5 pts) Show that $\mathbf{1}$ is an eigenvector of $\mathbf{M}$, and that its eigenvalue is zero.

(c) (3 pts) Show that $\Phi(\mathbf{Y}) = \Phi(\mathbf{Y} + c\mathbf{1})$, where c is any constant and $\mathbf{1}$ is the $n \times 1$ matrix whose entries are all 1s. (*Hint:* one way is to use the previous two parts.)

(d) (2 pts) Show that $\Phi(\mathbf{Y})$ is minimized by $\mathbf{Y} = 0$.

(e) (3 pts) To avoid the trivial solution of setting all the y_i to zero, we impose the constraint that $n^{-1} \sum_{i=1}^{n} y_i^2 = 1$. We use a Lagrange multiplier to enforce this constraint; write down the Lagrangian (modified objective function) for the constrained minimization problem.

(f) (10 pts) Show that a solution $\mathbf{Y}$ to the constrained minimization problem must be an eigenvector of $\mathbf{M}$.