Matt Dickenson `mcd31`
STA561/CS571 — Fall 2013 **Midterm Exam** Due: 16 October, 2013

Notes: I did not work with anyone else on this exam or refer to resources other than the supplied article, course notes, textbook, and course Piazza page.

Problem 1

A We can write out the expected log likelihood as:

$$\begin{aligned}
\mathbb{E}[\ell_c] &= \mathbb{E}[\log \prod_{i=1}^{n} (2\pi)^{p/2} |\Psi|^{-1/2} \exp\{-\frac{1}{2}[x_i - \mu - \Lambda z_i]' \Psi^{-1} [x_i - \mu - \Lambda z_i]\}] \\
&= c - \frac{n}{2} \log |\Psi| - \\
&\qquad \sum_{i=1}^{n} \mathbb{E}[\frac{1}{2}(x_i \Psi^{-1} x_i' - 2 x_i \Psi^{-1} \mu' - 2 x_i \Psi^{-1} \Lambda z_i + \mu \Psi^{-1} \mu' + 2 \mu \Psi^{-1} \Lambda z_i + z_i' \Lambda' \Psi^{-1} \Lambda z_i)] \\
&= c - \frac{n}{2} \log |\Psi| - \\
&\qquad \sum_{i=1}^{n} \mathbb{E}[\frac{1}{2} x_i \Psi^{-1} x_i' - x_i \Psi^{-1} \mu' - x_i \Psi^{-1} \Lambda z_i + \frac{1}{2} \mu \Psi^{-1} \mu' + \mu \Psi^{-1} \Lambda z_i + \frac{1}{2} z_i' \Lambda' \Psi^{-1} \Lambda z_i] \\
&= c - \frac{n}{2} \log |\Psi| - \sum_{i=1}^{n} \{\frac{1}{2} x_i \Psi^{-1} x_i' - x_i \Psi^{-1} \mu' - x_i \Psi^{-1} \Lambda \mathbb{E}[z_i | x_i] + \\
&\qquad \frac{1}{2} \mu \Psi^{-1} \mu' + \mu \Psi^{-1} \Lambda \mathbb{E}[z_i | x_i] + \frac{1}{2} tr(\Lambda' \Psi^{-1} \Lambda E[z_i z_i' | x_i])\}
\end{aligned}$$

From this we can determine that the expected sufficient statistics are $\mathbb{E}[z_i|x_i]$ and $E[z_i z_i'|x_i]$.

B We can derive the maximum likelhihood estimated of μ, Ψ, and Λ by differentiating the expected log likelihood:

$$\begin{aligned}
\frac{\partial \mathbb{E}[\ell_c]}{\partial \mu} = 0 &= -\sum_{i=1}^{n} \{-x_i \Psi^{-1} + \frac{1}{2} \mu^{(new)} \Psi^{-1} + \Psi^{-1}\} \\
&= \sum_{i=1}^{n} x_i \Psi^{-1} - \frac{n}{2} \sum_{i=1}^{n} 2 \mu^{(new)} \Psi^{-1} - \sum_{i=1}^{n} \Psi^{-1} \mathbb{E}[z_i | x_i] \\
\mu^{(new)} \sum_{i=1}^{n} \Psi^{-1} &= \frac{1}{n} \sum_{i=1}^{n} x_i \Psi^{-1} - \frac{1}{n} \sum_{i=1}^{n} \Psi^{-1} \mathbb{E}[z_i | x_i] \\
\mu^{(new)} &= \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[z_i | x_i]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbb{E}[\ell_c]}{\partial \Psi^{-1}} = 0 &= -\frac{n}{2}\Psi^{(new)} - \sum_{i=1}^{n}\{\frac{1}{2}x_i x_i' - x_i\mu' - x_i\Lambda\mathbb{E}[z_i|x_i] + \frac{1}{2}\mu\mu' + \mu\Lambda\mathbb{E}[z_i|x_i] + \frac{1}{2}\Lambda'\Lambda E[z_i z_i'|x_i]\} \\
\frac{n}{2}\Psi^{(new)} &= \sum_{i=1}^{n}\{\frac{1}{2}x_i x_i' - x_i\mu' - x_i\Lambda\mathbb{E}[z_i|x_i] + \frac{1}{2}\mu\mu' + \mu\Lambda\mathbb{E}[z_i|x_i] + \frac{1}{2}\Lambda'\Lambda E[z_i z_i'|x_i]\} \\
\Psi^{(new)} &= \frac{1}{n}\sum_{i=1}^{n}\{x_i(x_i' - 2\mu' - \Lambda\mathbb{E}[z_i|x_i])\} + \frac{1}{n}\sum_{i=1}^{n}\{\mu(\mu' + 2\Lambda\mathbb{E}[z_i|x_i])\} + \frac{1}{n}\sum_{i=1}^{n}\Lambda'\Lambda E[z_i z_i'|x_i]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbb{E}[\ell_c]}{\partial \Lambda} = 0 &= -\sum_{i=1}^{n}\{-x_i\Psi^{-1}\mathbb{E}[z_i|x_i] + \mu\Psi^{-1}\mathbb{E}[z_i|x_i] + \Lambda'\Psi^{-1}E[z_i z_i'|x_i]\} \\
n\Lambda^{(new)}\Psi^{-1}E[z_i z_i'|x_i] &= \sum_{i=1}^{n}\{-x_i\Psi^{-1}\mathbb{E}[z_i|x_i] + \mu\Psi^{-1}\mathbb{E}[z_i|x_i]\} \\
\Lambda^{(new)} &= E[z_i z_i'|x_i]^{-1}\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}[z_i|x_i](\mu - x_i)\}
\end{aligned}$$

C Now we derive the expected sufficient statistics, using fact that data and factors are jointly normal:

$$\begin{aligned}
\mathbb{E}[z|x-\mu] &= \Lambda'(\Psi + \Lambda\Lambda')^{-1}x' \\
\mathbb{E}[z|x] &= \mu + \Lambda'(\Psi + \Lambda\Lambda')^{-1}x' \\
\mathbb{E}[zz'|x] &= Var(z|x) + \mathbb{E}[z|x]E[z|x]' \\
&= I + [x(\mu + \Lambda'(\Psi + \Lambda\Lambda')^{-1})][(\mu + \Lambda'(\Psi + \Lambda\Lambda')^{-1})x']
\end{aligned}$$

D Now the EM pseudocode, written with $\mathcal{R}$ syntax for matrix operations and indexing:

```
initialize mu[1:p] = sample(x)
initialize psi = var(x - mu)
initialize lambda = matrix(epsilon)
repeat
  for each example i=1:N do
    expected_z[i] = mu + t(lambda) %*%
      (psi + lambda %*% t(lambda)) %*% t(x)
    expected_z_squared[i] = I + (x %*% t(lambda) %*%
      (psi + lambda %*% t(lambda))) %*%
      (t(lambda) %*% (psi + lambda %*% t(lambda)) %*% t(x))
```

```
11   end
12   for each factor k = 1:K do
13     for each feature j=1:p do
14       mu[k][j] = mean(x[,j]) + mean(expected_z[,j])
15       psi[k][j, j] = mean(x%*%(t(x)-2*t(mu)-lambda*mean(expected_z))) +
16         mean(mu %*% (t(mu) + 2 lambda*mean(expected_z))) +
17         mean(t(lambda) %*% lambda * mean(expected_z_squared))
18       lambda = factor_loadings(x, z)
19      end
20   end
21 until converged
```

I would set K by cross-validation, holding out subsets of X each time.

E I would assess convergence by the change in each $\mu_{1:p}$ between iterations. If these are not changing, it suggests that the underlying factors are not changing.

Problem 2

Modeling mean values of each feature in the model is better than mean-centering each of the features before performing factor analysis because it takes into account $E[z_i|x_i]$, leveraging the information in the factor analysis to compute feature means. This is useful because the some features will be more strongly correlated with some factors than others.

Problem 3

One problem where a mixture of factor analyzers would be preferable to a simpler factor analysis model is in analyzing voting behavior. For example, some observers of Turkish politics have argued that Turkey is turning its back on the West (specifically the US and EU) in favor of the East, as represented by Russia and Iran. This claim often references voters' preferences for the ruling AKP party, which has reduced the country's efforts to join the EU in favor of a "zero-problems" policy with its immediate neighbors including Iran.

It would be difficult to draw inferences about this debate directly from the individual features that may influence a voter's choice of political party (e.g. gender, education level, income), so factor analysis is preferable to identify certain latent dimensions (e.g. religiosity, favorability toward the West). However, individuals likely differ in the extent to which each of these latent factors motivates them. Thus, a mixture of factor analyzers would be useful.

The parameters of this model are Ψ and Λ as before, μ_j as the mean of the j^{th} factor analyzer (of a total of J), and the vector π of mixing proportions. Ψ and Λ help to

assess the covariance structure but are not of primary interest for the applied problem described above. μ helps to give a baseline for each factor analyzer. The main parameter of interest is π, which indicates the weight given to each factor analyzer in the model.

If much weight is given to a model in which factors appear to be correlated with an East-West divide (i.e. a large value of π_j), that would lend credibility to one side of the Turkish debate. On the other hand, low weight on such factor analyzers would indicate that the East-West dimension is not a primary cleavage in contemporary Turkish politics. In this way, a mixture of factor analyzers could help shed light on an important political question.