Data Science Methods: Assignment 1

André Couder (2070121) Daniel Redel (2102630)
Frederico Sandmann (2071361) Xander Smid (2093581)

March 5, 2023

Question 1

Consider the covariance matrix $\mathbf{R}$ defined by:

$$\mathbf{R} = \frac{X'X}{n} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Question 1.a)

Calculate the eigenvalues of R. Show your calculations

First, we will calculate the eigenvalues λ. Recall that the eigenvalue λ of $\mathbf{R}$ is a scalar such that the following equation has a *nontrivial solution*:

$$\mathbf{R}a = \lambda a$$

To find the solution we can re-arrange as follow:

$$\begin{aligned} \mathbf{R}a &= \lambda a \\ \mathbf{R}a - \lambda a &= 0 \\ (\mathbf{R} - \lambda \mathbf{I}_p)a &= 0 \\ \det(\mathbf{R} - \lambda \mathbf{I}_p) &= 0 \end{aligned}$$

Now we can solve our problem:

$$\det\left(\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} - \lambda\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right) = 0$$
$$\det\left(\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix} - \begin{bmatrix}\lambda & 0\\ 0 & \lambda\end{bmatrix}\right) = 0$$
$$\det\left(\begin{bmatrix}2-\lambda & 1\\ 1 & 2-\lambda\end{bmatrix}\right) = 0$$

When calculating that determinant we get:

$$\begin{aligned}(2-\lambda)(2-\lambda) - 1 \times 1 &= 0\\ 4 + \lambda^2 - 2\lambda - 2\lambda - 1 &= 0\\ \lambda^2 - 4\lambda + 3 &= 0\\ (\lambda - 1)(\lambda - 3) &= 0\\ \lambda^* &\in (1; 3)\end{aligned}$$

We find that there are two possible eigenvalues that solves the equation: 1 and 3.

Question 1.b)

Based on the eigenvalues, find the eigenvectors a of R that satisfy $a'a = 1$. Show your calculations.

Using $\lambda = 3$, we solve:

$$\begin{bmatrix}2 & 1\\ 1 & 2\end{bmatrix}\begin{bmatrix}a_{11}\\ a_{12}\end{bmatrix} = 3\begin{bmatrix}a_{11}\\ a_{12}\end{bmatrix}, \text{ s.t. } a_1'a_1 = 1$$

Which is just an equation system:

$$\begin{aligned}2a_{11} + a_{12} &= 3a_{11}\\ a_{11} + 2a_{12} &= 3a_{12}\\ a_{11}^2 + a_{12}^2 &= 1\end{aligned}$$

We solve:

$$\begin{aligned}a_{12} &= a_{11}\\ a_{11} &= a_{12}\end{aligned}$$

We replace this into our PCA restriction:

$$\begin{aligned} a_{11}^2 + a_{11}^2 &= 1 \\ 2a_{11}^2 &= 1 \\ a_{11}^2 &= \frac{1}{2} \\ a_{11} = a_{12} &= \pm\frac{1}{\sqrt{2}} \end{aligned}$$

Thus, the loading vector of *first principal component* for this data set is defined by:

$$a_1 = \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

We repeat the exercise, but now using $\lambda = 1$:

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = 1 \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix}, \text{ s.t. } a_1'a_1 = 1, a_2'a_2 = 1$$

This lead to the following system of equations:

$$\begin{aligned} 2a_{21} + a_{22} &= a_{21} \\ a_{21} + 2a_{22} &= a_{22} \\ a_{21}^2 + a_{22}^2 &= 1 \end{aligned}$$

We solve:

$$\begin{aligned} a_{21} &= -a_{22} \\ a_{22} &= -a_{21} \end{aligned}$$

And we replace this into our PCA restriction:

$$\begin{aligned} a_{21}^2 + a_{22}^2 &= 1 \\ a_{21}^2 + (-a_{21})^2 &= 1 \\ 2a_{21}^2 &= 1 \\ a_{21}^2 &= \frac{1}{2} \\ a_{21} = \frac{1}{\sqrt{2}}, a_{22} &= -\frac{1}{\sqrt{2}} \end{aligned}$$

Thus, the loading vector of *second principal component* for this data set is defined by:

$$a_2 = \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

Question 1.c)

What are the first PC loadings on each variable? Explain

The loadings of the first principal components is the eigenvector that correspond to the largest eigenvalue, which in this case is $\lambda = 3$. Recall that the eigenvalue λ_j denotes the **amount of variability captured along that dimension**, $\text{Var}(Z_j)$.

On the other hand, the a_{11}, a_{12} elements of a_1 are called the ***loadings*** and a_1 is the ***first loading vector***. In our case, these elements are defined by:

$$\begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

- a_{11} is the loading of *covariate* X_1 on the *first principal component* $Z_1 = a_1\mathbf{X}$
- a_{12} is the loading of *covariate* X_2 on the *first principal component* $Z_1 = a_1\mathbf{X}$.

Question 2

True model:

$$x_{it} = \alpha_i f_t + \epsilon_{it} \quad \forall\, i = 1 \dots P \;;\; \forall\, t = 1 \dots N$$

$$F = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} \quad ; \quad \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_P \end{bmatrix}$$

Objective Function:

$$\min_{\alpha, F} O(\alpha, F) \;=\; \sum_{i=1}^{P} \sum_{t=1}^{N} (x_{it} - \alpha_i f_t)^2$$

Question a)

Add to objective function (2) the constraint that $F'F = 1$, and solve for F given α.

- Rewrite $O(.)$ for simplicity:

$$\begin{aligned} O(\alpha, F) &= \sum_{i=1}^{P}\sum_{t=1}^{N}(x_{it} - \alpha_i f_t)^2 \\ &= \sum_{i=1}^{P}\sum_{t=1}^{N} x_{it}^2 \; - 2\sum_{i=1}^{P}\sum_{t=1}^{N} x_{it}\alpha_i f_t \; + \sum_{i=1}^{P}\alpha_i^2 \sum_{t=1}^{N} f_t^2 \\ &= \sum_{t=1}^{N} x_t' x_t \; - 2\alpha' XF \; + \; \|\alpha\|^2 \|F\|^2 \\ &= \sum_{t=1}^{N} \|x_t\|^2 \; - 2\alpha' XF \; + \; \|\alpha\|^2 \|F\|^2 \end{aligned}$$

- Use Lagrangian with constraint $\|F\|^2 = 1$ and solve for F given α

$$\begin{aligned} \mathcal{L}(F, \lambda) &= \sum_{t=1}^{N} \|x_t\|^2 \; - 2\alpha' XF \; + \; \|\alpha\|^2 \|F\|^2 \; - \; \lambda(\|F\|^2 - 1) \\ \frac{\partial}{\partial F}\mathcal{L}(F, \lambda) &= -2X'\alpha \; + \; 2(\|\alpha\|^2 - \lambda)F = 0 \\ &\Leftrightarrow F = \frac{1}{(\|\alpha\|^2 - \lambda)} X'\alpha \\ \frac{\partial}{\partial \lambda}\mathcal{L}(F, \lambda) &= \|F\|^2 - 1 = 0 \\ &\Leftrightarrow \|F\|^2 = 1 \\ &\Leftrightarrow \left(\frac{1}{\|\alpha\|^2 - \lambda}\right)^2 \underbrace{\alpha' XX'\alpha}_{V} = 1 \\ &\Leftrightarrow \lambda^2 \; - \; 2\lambda\|\alpha\|^2 \; + \; (\|\alpha\|^4 - V) = 0 \end{aligned}$$

After using the quadratic formula, we get:

$$\begin{aligned} \lambda &= \|\alpha\|^2 \pm \sqrt{V} \\ \implies F &= \pm\frac{1}{\|X'\alpha\|} X'\alpha \end{aligned}$$

Is the solution linear in α?

- The solution for F is not linear in α , indeed, as F is a vector divided by its norm, assuming X' and α entries can take any values in $\mathbb{R}$, F would take all the values on a N-Sphere of radius 1 around the origin in $\mathbb{R}^N$. Put more simply, $\|X'\alpha\|$ is linear in α as well as $X'\alpha$ thus F is constant with respect to any scaling of α . Let be some scalar μ , then

$$\frac{1}{\|\mu X'\alpha\|}X'\mu\alpha = \frac{1}{\mu\|X'\alpha\|}\mu X'\alpha = F$$

Is it the solution of a particular OLS regression problem or not? Explain.

- Note that F can be seen as a set of t coefficients of the t different Least Squares regressions $\frac{\|\alpha\|^2}{\|X'\alpha\|}x_t = f_t\alpha$ with t endogenous variables (the t columns of X) and always the same unique regressor α . I.e. for each endogenous variable x_t :

$$\begin{aligned} f_t &= \frac{1}{\|X'\alpha\|}\alpha' x_t \\ &= \frac{\alpha'\alpha}{\sqrt{\alpha' XX'\alpha}}\frac{\alpha' x_t}{\alpha'\alpha} \end{aligned}$$

- And where F, the vector of estimated coefficients of the **system of OLS equations** is constrained to be of length 1, or equivalently, where the set of endogenous variables x_t is scaled by the ratio between variance of regressor $\|\alpha\|^2$ and length of the covariance vector $\|X'\alpha\|$.
- The same analyse can be made for the N rows of X where α_i would be the coefficients and F the covariate.

Question b)

Now solve (2) for F, a given α that satisfies $\alpha'\alpha = 1$.

- same as before but no restrictions and $\|\alpha\| = 1$, i.e.

$$O(F) = \sum_{t=1}^{N} \|x_t\|^2 \; - 2\alpha' XF \; + \; \|F\|^2$$

$$\frac{\partial}{\partial F} O(F) = -2X'\alpha \; + \; 2F = 0$$

$$\Leftrightarrow F = X'\alpha$$

Explain how the solution $\hat{F}$ can be interpreted as a regression coefficient, and which regression that is.

- In this case, F can be directly interpreted as a vector of OLS regression coefficients where each entry f_t of F is the coefficient of the regression $x_t = \alpha f_t + e_t$, where $f_t = \alpha' x_t = \frac{\alpha' x_t}{\alpha' \alpha}$ as $\|\alpha\|^2 = 1$

Question c)

Continue with b) and given $\hat{F}$, solve (2) for α, assuming $a'a = 1$.

- Continuing with $\hat{F} = X'\alpha$ and assuming $\|\alpha\|^2 = 1$, we now solve for α
- the objective function becomes[1]:

$$\begin{aligned} O(\alpha) &= \sum_{t=1}^{N} \|x_t\|^2 \; - 2\alpha' X\hat{F} \; + \; \|\alpha\|^2 \|\hat{F}\|^2 \\ &= \sum_{t=1}^{N} \|x_t\|^2 \; - 2\alpha' XX'\alpha \; + \; \alpha' XX'\alpha \\ &= \sum_{t=1}^{N} \|x_t\|^2 \; - \alpha' XX'\alpha \end{aligned}$$

- Thus minimizing wrt α amounts to maximizing $\|\hat{F}\|^2 = \alpha' XX'\alpha$ given the constraint that $\|\alpha\|^2 = 1$:

[1] Note that the "baseline" for this data is calculated between Jan 3 - Feb 6, 2020 and its calculated by finding the median mobility for every individual day of the week during the period. Both figures show the percent change in mobility in comparison to the baseline day of the week.

$$
\begin{aligned}
\mathcal{L}(\alpha, \lambda) &= \alpha' XX'\alpha \; - \; \lambda(\|\alpha\|^2 - 1) \\
\frac{\partial}{\partial \alpha}\mathcal{L}(\alpha, \lambda) &= 2XX'\alpha \; - \; 2\lambda\alpha = 0 \\
&\Leftrightarrow \frac{X\hat{F}}{\lambda} = \alpha \quad \text{and} \quad (XX' - \lambda I_P)\alpha = 0 \\
\frac{\partial}{\partial \lambda}\mathcal{L}(\alpha, \lambda) &= \|\alpha\|^2 - 1 = 0 \\
&\Leftrightarrow \frac{1}{\lambda^2}\hat{F}'X'X\hat{F} = 1 \\
&\Leftrightarrow \lambda = \pm\|X\hat{F}\| \\
&\Leftrightarrow \alpha = \pm\frac{1}{\|X\hat{F}\|}X\hat{F}
\end{aligned}
$$

- Thus α is an eigenvector of XX' and λ its corresponding eigenvalue. Therefore to maximize $O(\alpha)$, λ must be the biggest eigenvalue of XX'. Furthermore:

$$
\begin{aligned}
XX'\alpha &= \lambda\alpha \\
\alpha' XX'\alpha &= \lambda\alpha'\alpha = \lambda = \|F\|^2
\end{aligned}
$$

- and we can rewrite:

$$
\begin{aligned}
\frac{\partial}{\partial \alpha}\mathcal{L}(\alpha, \lambda) &= 2XX'\alpha \; - \; 2\lambda\alpha = 0 \\
&\Leftrightarrow X\hat{F} = \lambda\alpha = \|\hat{F}\|^2\alpha \\
&\Leftrightarrow \alpha = \frac{X\hat{F}}{\|\hat{F}\|^2}
\end{aligned}
$$

- the eigenvalue corresponding to α is $\|F\|^2$ i.e. the variance of F as in PCA.
- And

$$
\begin{aligned}
\|F\|^2 &= \alpha' XX'\alpha = \|XF\| \\
&= \sqrt{F'X'XF} = \sqrt{\alpha' XX'XX'\alpha} \\
&= \sqrt{\alpha' V\Sigma^2 V'\alpha} \\
&= \sqrt{(1, 0, \dots, 0)\Sigma^2(1, 0, \dots, 0)'} \\
&= \lambda_1
\end{aligned}
$$

Question d)

Suppose that this $\hat{F}, \hat{\alpha}$ where the PCA solutions. Do you see any implications for efficient computation of the first few principal components?

The set of solutions $(\hat{F}, \hat{\alpha})$ requires, to be computed by Eigen decomposition or **Singular Value Decomposition**, the finding of the roots of the polynomial of degree up to $N+1$, $p(\lambda) = \det(XX' - \lambda I_P)$, (depending on the rank of the error matrix Σ). Therefore if N is large, one might instead decide to use the **NIPALS algorithm**. This process is computationally more efficient and can be used because

- $\hat{F} = X'\alpha$ is a linear combination of X' and, furthermore,
- $\hat{\alpha} = \frac{1}{\|\hat{F}\|^2} X\hat{F}$ is a linear combination of X as well

This is how to compute the NIPALS algorithm:

1. Standardize X' , i.e. divide each column of X' by its norm.
2. Because $\hat{F}$ is a linear combination of X' initialize it by a simple combination:

$$\hat{F}^{(I)} = (X')_{\cdot\ 1}$$

   Where $(X')_{\cdot\ 1}$ is the first column of X' .
3. To get an estimate of the loadings, minimize the sum of squared errors of the model $(X')_{\cdot\ j} = \alpha_j^{(I)} \hat{F}^{(I)} + e_{\cdot\ j}$ i.e.:

$$\alpha_j^{(I)} = \frac{\hat{F}^{(I)} (X')_{\cdot\ j}}{\hat{F}^{(I)'} \hat{F}^{(I)}}$$

4. Standardize the loadings, $\alpha_j^{(II)} = \frac{1}{\|\alpha\|} \alpha_j^{(I)}$
5. Improve the scores estimates by considering the scalars $\hat{F}_i^{(II)}$ as coefficients this time:

$$\hat{F}_i^{(II)} = \frac{(X')_{i\ \cdot}\ \alpha^{(II)}}{\alpha^{(II)'} \alpha^{(II)}} = (X')_{i\ \cdot}\ \alpha^{(II)}$$

6. Now substitute $\hat{F}^{(II)}$ in step 2. and repeat the process until convergence, that is, $\hat{F}^{(n+1)} - \hat{F}^{(n)} < \epsilon$ for some ϵ arbitrarily small.
7. To obtain the next principal component, we need to deflate X and start all over again from step 1. This step removes the variability captured from PC1: $E_1 = X' - \hat{F}\alpha'$

Question 3

Question a)

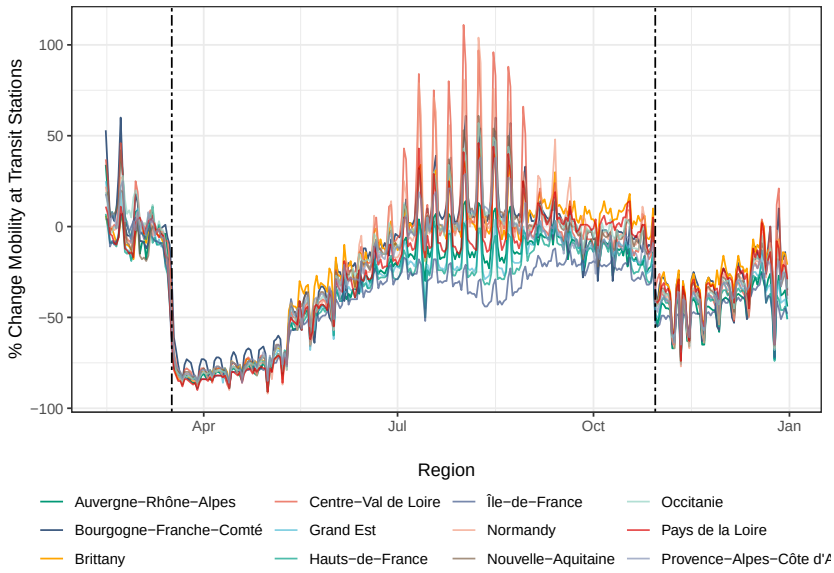

Figure 1: Evolution of Mobility at Transit Stations, by Region

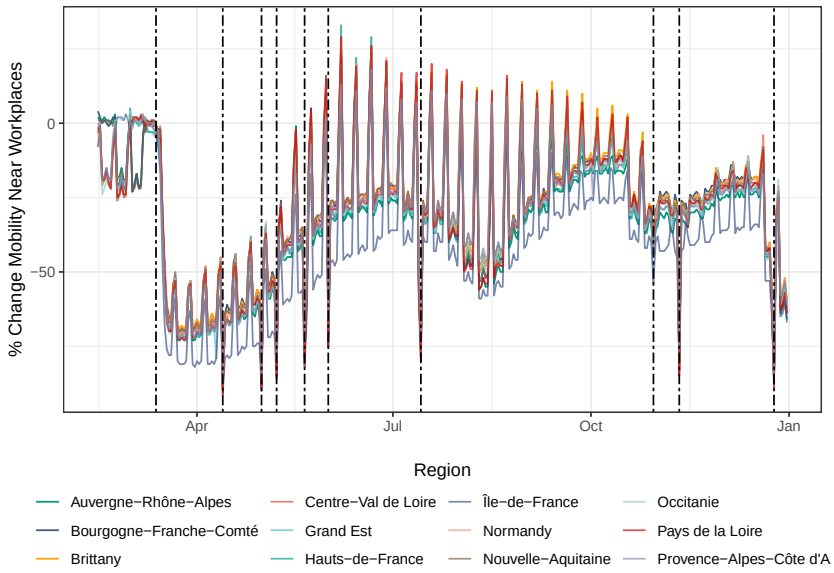

Figure 2: Evolution of Mobility near Workplaces, by Region

Do you see any co-movement between regions? Explain.

We can see from both Figure 1 and Figure 2 that the movement between regions is highly correlated. Indeed, all those regions are from the same country, same economy and same social structure. This implies that their economies and the way people live in those regions are overall structurally very similar. It is then no surprise that shocks like lockdowns, holidays and weekends (that have significant impact of the data of each region) trigger similar reactions for every region[2].

Do you see any seasonality? Explain. Are the two lockdowns visible? Explain.

Regarding seasonality, we observe recurring "peaks" in both figures. These are especially clear when it comes to the mobility near workplaces (Figure 2) throughout the year, although the summer peaks in the transit mobility also reach especially high values. These peaks happened repeatedly during weekends, which confirms the presence of week seasonality. This pattern does not mean that there was more mobility during weekends, but that, unlike weekdays where mobility was probably reduced a lot due to the pandemic (i.e. more frequent remote work before and after lock-downs and straight up restrictions during lock-downs), during weekends this drop was less noticeable when compared to baseline. This might be due to the fact that mobility in our data was already lower during weekends than during weekdays so it did not have so much "space" to fall. Furthermore, during the summer we often observe values above 0 during weekends, meaning that there was more mobility in those weekends than baseline. This correspond to the french summer vacation period when people leave for holidays (the departure mainly happens during weekends as people then leave for entire weeks during this period). This might also be due to people taking extra advantage of weekends (to go out) given that they were mostly stuck at home during 2020 and due to improving weather outside.

Are the two lockdowns visible? Explain. There are eight visible public holidays in the mobility data for workplaces. Can you identify them?

The two lockdowns are clearly visible in both figures Figure 1 and Figure 2, especially the first lockdown, where we notice a sharp drop in mobility. The first lockdown begins on the 17th of March (even though the drop in mobility starts 1 or 2 days before in some regions). The 1st lockdown ended on May 11th, although the recovery in mobility was more gradual, especially in mobility near workplaces (Figure 2), possibly due to remote work. The second lockdown started on October 30th as can be seen quite clearly by the mobility drop near transit stations. The second lock-down was less restrictive (and mobility was already lowered in comparison to baseline) as can be clearly seen in the figures where the fall is much less noticeable. This second lock-down ended on December 15th but it is not as noticeable in the graph.

[2]Note that the "baseline" for this data is calculated between Jan 3 - Feb 6, 2020 and its calculated by finding the median mobility for every individual day of the week during the period. Both figures show the percent change in mobility in comparison to the baseline day of the week.

As expected, mobility near workplaces falls significantly during holidays and these drops can be seen clearly in the graph (Figure 2). While there were 11 public holidays in France in 2020 one is outside the scope of our data and 2 holidays fell on weekends. The two that fell on weekends are not noticeable in the graph because mobility on weekends was likely lower in the baseline anyway. Public Holidays in 2020 were:

- **January 1, 2020**: New Year's Day - not in the graph (data starts in February 15)
- **April 13, 2020**: Easter Monday
- **May 1, 2020**: Labor Day
- **May 8, 2020**: Victory in Europe Day
- **May 21, 2020**: Ascension Day
- **June 1, 2020**: Whit Monday (Pentecost Monday)
- **July 14, 2020**: Bastille Day (French National Day)
- **August 15, 2020**: Assumption of Mary - can't see this one (this falls on a weekend)
- **November 1, 2020**: All Saints' Day - can't see this one (this falls on a weekend)
- **November 11, 2020**: Armistice Day
- **December 25, 2020**: Christmas Day

Question b)

> **Plot the first two principal components together with the observations for each case, i.e. plot both loadings and scores. Explain what you see.**

Before plotting the principal components with their corresponding observations, we will take a look at the loadings of the first two principal components of each dataset. These loadings can be seen in Figure 3:

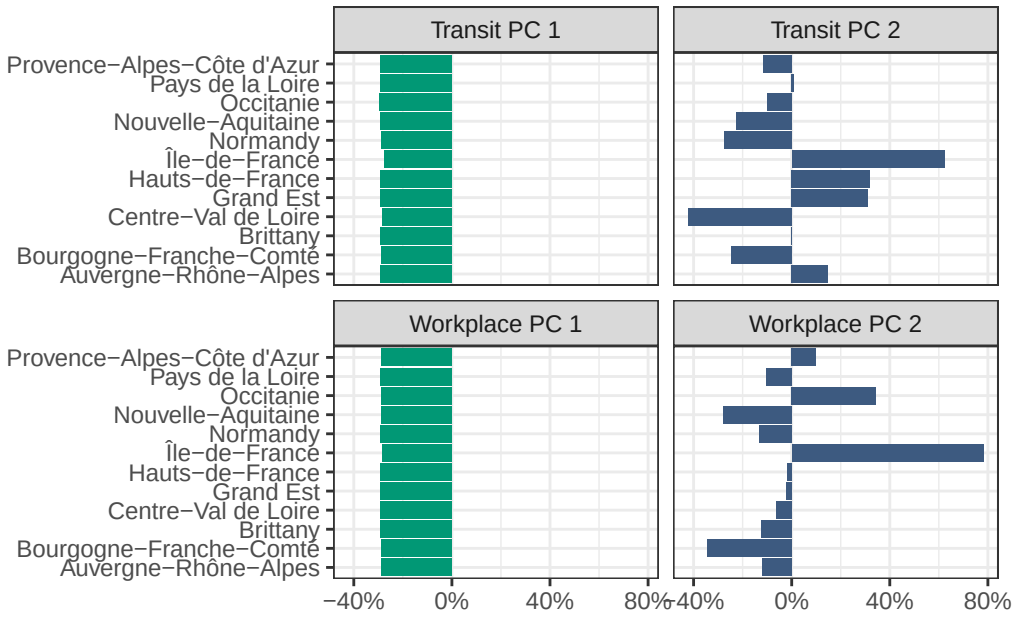

Figure 3: PCA Loadings

In both transit and work data, we observe that all the variables -representing the different regions in France- have similar PC loadings for the 1st Principal component. This happens because, as we saw above, the variables are closely correlated to each other and, as such, PC1 captures this common variance. Additionally, all variables are measuring changes in the same underlying construct, namely, mobility in regions of France which are much alike and equivalently affected by COVID lockdowns, holidays and day of the week, as expressed before. Thus, we can say that the first component roughly corresponds to a measure of common mobility.

On the other hand, the second principal component loadings are much more heterogeneous. As most of the common variance was captured in PC1, the loadings of PC2 no longer have the same sign. Interestingly, for the second PC we observe that Île-de-France has the largest loading, suggesting that Paris may have stronger responses to shocks than other regions. Indeed, regarding work mobility changes, we can see from Figure 2 that Île de France reacted significantly more strongly to the lockdowns than any other region, and consistently displayed the most negative changes compared to baseline, despite also mimicking the background common trend. Île de france is also the only region that didn't experience a return to baseline regarding transit mobility in Summer. Maybe the restrictions in this region that is the most densely populated of France were slightly stronger than those of the other regions. Thus, as such, the second PC mainly captures it's variance.

Now we can plot the first two PCA's in a biplot:

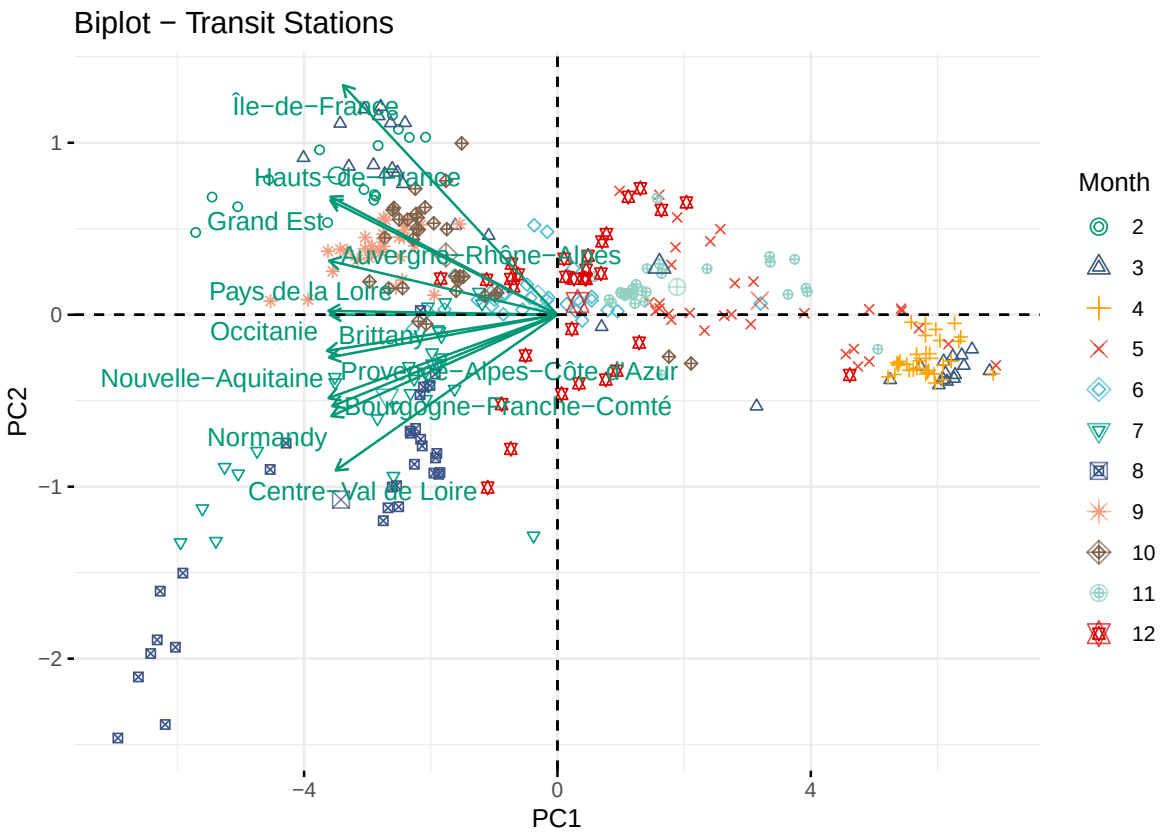

Figure 4: Biplot - Transit Stations

This Biplot represents a 2D space spanned by the first PC (horizontal axis) and the second one (vertical axis) Each point's dimension in this space is a linear combination of all regions for a specific date (the first and second PC scores). The green arrows, on the other hand, indicate the first two principal component loading vectors α_1 and α_2 that define the two linear combinations.

In Figure 4, dates with large positive points on the first PC (x-axis) represents the dates with low common transit mobility across regions in France. We observe that the first months of the lockdowns such as the (half of) March and April have very similar PC1 scores, meaning low mobility. On the other hand, July, August and September (although with a lesser degree) represents greater common mobility across regions, which is in line with the higher mobility variation found Figure 1 in during these months.

High scores on the second PC (y-axis) is harder to interpret, but it could be argued that larger PC2 scores represents the extra mobility variation of île de France, because it has the highest loading.

```
Warning: ggrepel: 4 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

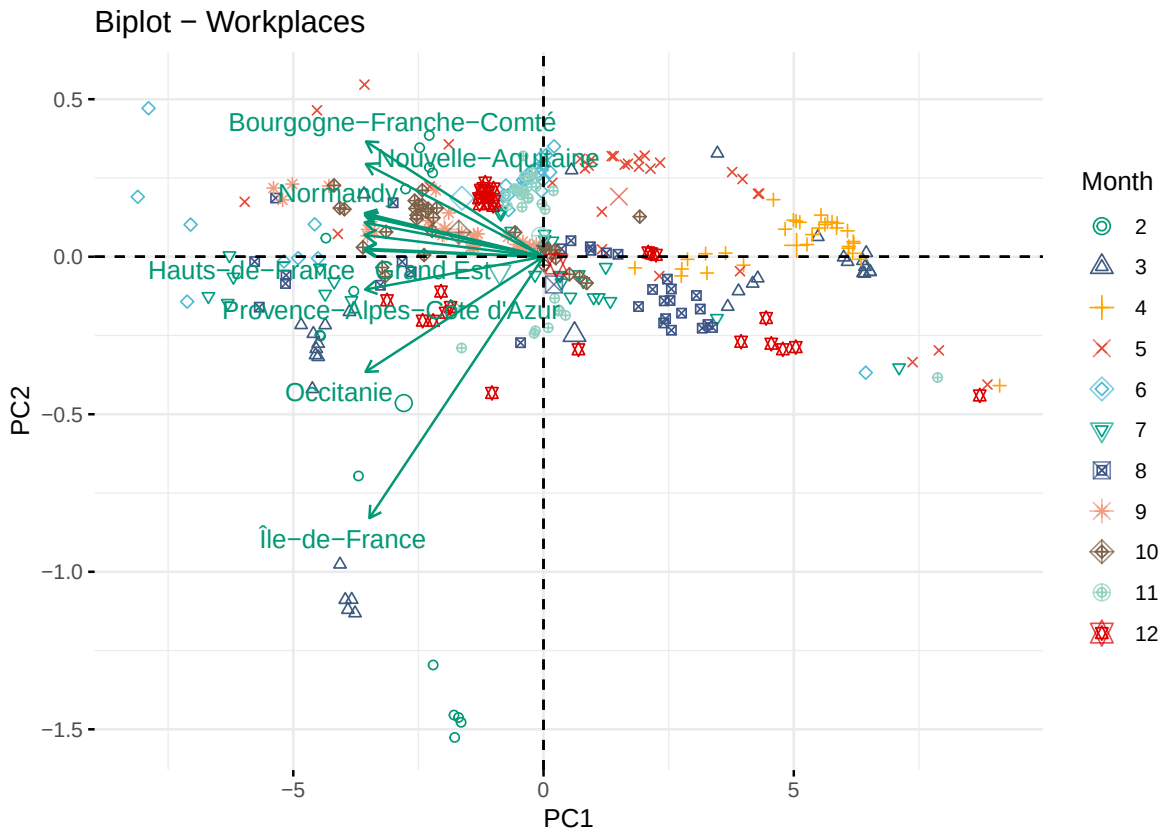

Figure 5: Biplot - Workplaces

Figure 5 we also see larger PC1 scores for the months around the lockdown periods (for example, March and April). Interestingly, we also observe common greater workplace mobility in August. The second principal component, however, is not only harder to interpret, but also accounts for a very small proportion of the variance, as we will see in the next section.

Question c)

Plot the proportion of variance explained as a function of the number of principal components, for each of the two datasets. Explain what you see.

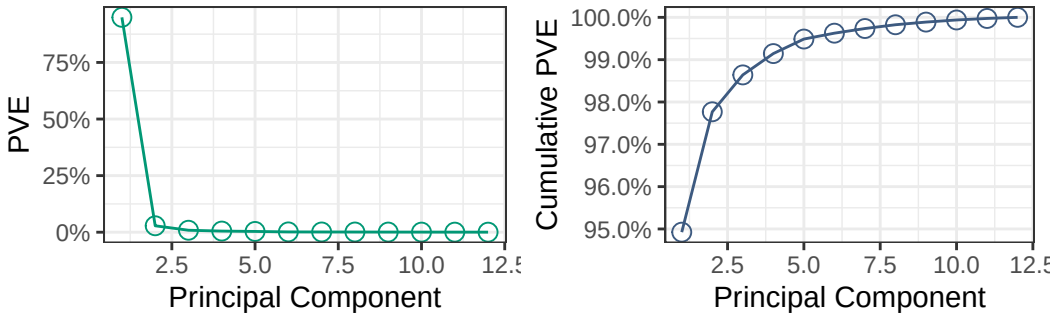

Figure 6: Screeplot - Transit Stations

Figure 6 presents a Scree plot that shows the proportion of the variance explained (left-hand side) and the cumulative proportion of variance explained (right-hand side) by each PC for the transit mobility dataset. As expected from the large difference in the variances from PC1 to PC2, the 1st PC explains most of the variation in the data (95% of the variance). As such only PC1 should be kept, as the others are not very relevant in terms of explaining power. The 2nd PC explains only around 2.8% of the variance.

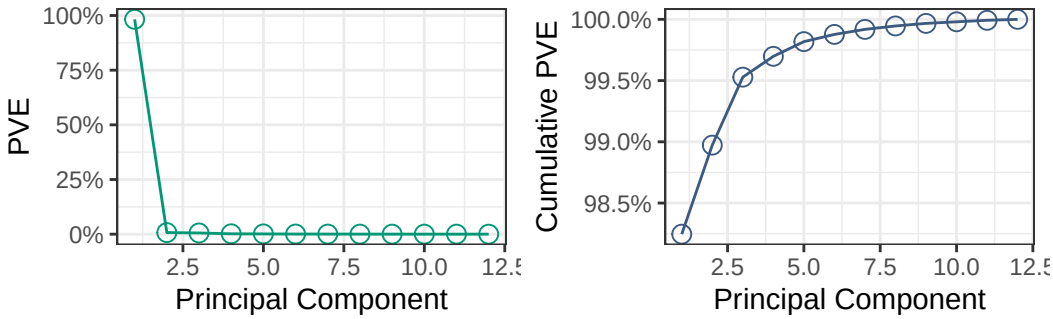

Figure 7: Screeplot - Worplaces

Figure 7 shows the same Scree plot of (Cumulative) PVE's but for the work mobility dataset. Just as above, the 1st PC explains a large part of the variance in the data, 98.2% of the variance to be precise. On the other hand, the second principal component accounts for only around 0.7% of the variance.

Additionally, notice that in both cases only the first eigenvalue is above 1 ($\lambda^{\text{transit}}1 = 3.375; \lambda^{\text{work}}1 = 3.433$) and since data is scaled this is a further indication that we should only use the first principal component, since the others do not even explain the variance of a single variable.

Question d)

Do you see any co-movement? Explain.

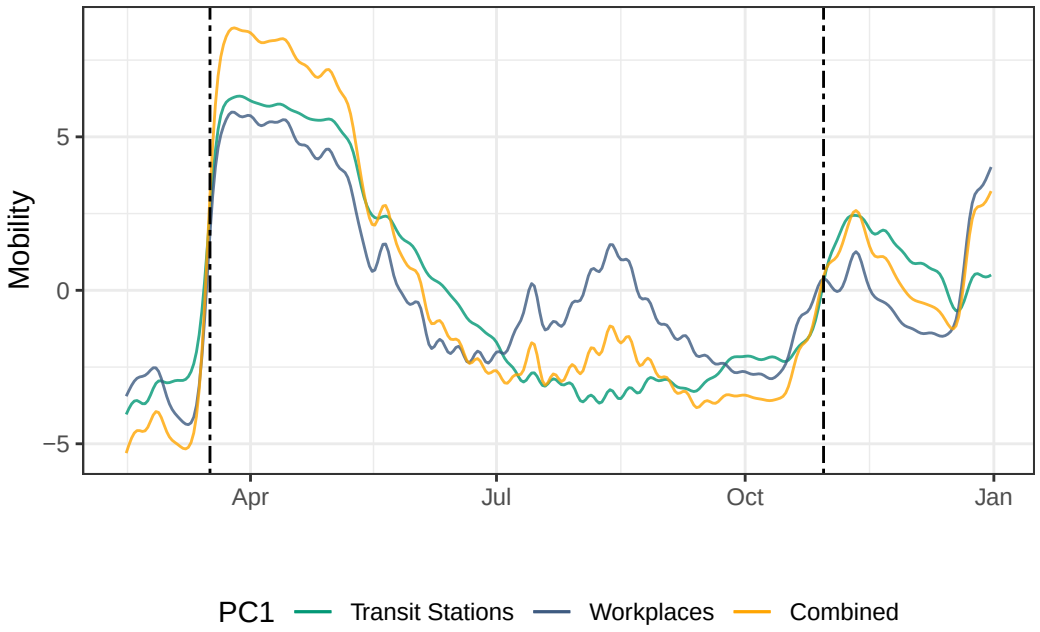

Figure 8: Evolution of Mobility

Co-movement definitely exists as we can see in Figure 8. But transit and work mobility seem to diverge quite a bit especially around summer time. This makes sense since workplace mobility decreases during summer as people take holidays, but this decrease does not affect mobility near transit station and as such this asymmetry causes the graphs to move in different directions. When doing a correlation matrix, we see that there is indeed a large correlation between the lines (a correlation that improves with the smoothing of the lines). This is because both transit station mobility and mobility near work are influenced by many of the same factors (especially lock-downs) and this creates movements in the PC scores in the same direction (loadings are in the same direction too, otherwise PC scores would move in opposite directions).

The mobility of the combined data also co-moves with the other lines for much of the same reasons. The common variability between mobility near workplaces and mobility near transit stations is reflected in the combined data mobility trend.

The PC scores of the combined data seem to be mostly in between the values of the other 2 lines, but during lockdown it's higher, likely due to the fact that it is the point where all the data (transit and workplace mobility) co-vary in the strongest way. The combined data has a stronger correlation with the work mobility data during the summer because correlation within data during that period is stronger for mobility near workplaces than for mobility near transit stations. Additionally, notice that the second PC for the combined data is more relevant than the second PC for the 2 data sets when done individually. This is to be expected, since the

correlation within the 2 data sets (work and transit mobility) is larger than for the combined data set leaving unexplained variance to be captured by the second PC.

What can you say about the lockdowns? Show them in the plot and explain.

When lockdowns are imposed, co-movement seems to be more noticeable and PC scores increases, indicating that larger common variability is captured by PC1, which is to be expected as during lockdowns mobility decreased "rapidly" in both data sets. This rapid shift is captured by PC1 very well. While less noticeable, in the second lock-down one can still see it there.

What can you say about the public holidays? Show them in the plot and explain.

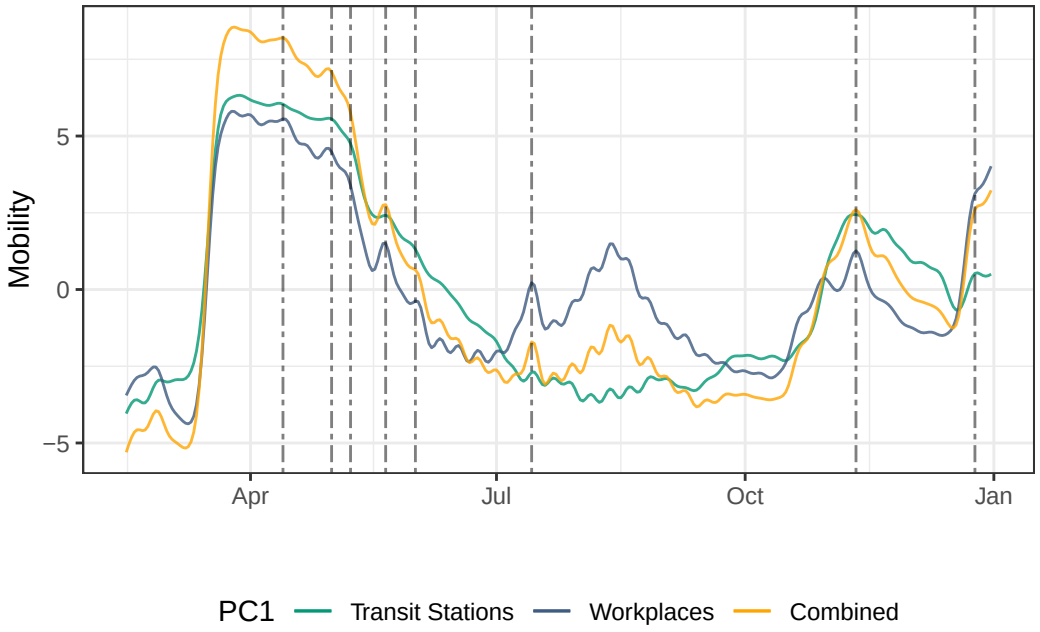

Figure 9: Evolution of Mobility - Public Holidays

Similarly some holidays are very noticeable in Figure 9, especially in the work mobility plot. Since during holidays (that fall on weekdays) mobility decreases significantly near workplaces and loadings are negative, we can see local maximums of the PC1 scores during these days. Due to the (7 day) smoothing, this isn't as easy to see as in the non-smooth figure (Figure 2), but some can still be seen quite clearly, namely:

- **May 21, 2020**: Ascension Day - Thursday

- **July 14, 2020**: Bastille Day (French National Day) - Tuesday
- **November 11, 2020**: Armistice Day - Wednesday

Notice that from the 8 holidays we identified in Figure 2, 5 are on Monday or Friday. These become less noticeable when we smooth out the graph (across 7 days). The remaining 3 holidays are in the middle of the week, and these holidays are more noticeable. Likely this is because of the bandwidth size of the kernel smoothing function applied.If we are on a Wednesday the smoothing function gives larger weights to Thursday and Tuesday than it gives to the weekends. When we are on a Monday or Friday the weight that the smoothing function gives to weekends is larger. Since in the mobility near workplace data weekends and holidays move in opposite directions (as explained in question *a*) this creates and "interference" when the data is smoothed out. As such holidays that are in the middle of the week don't get as much "interference" from weekends as the ones on Monday or Friday. This causes the holidays that fall in the middle of the week to be more noticeable than the ones that fall near weekends.[3]

[3] Additionally notice that if a holiday falls on a Thursday for example people may take an extra day off on a friday to enjoy a 4 day vacation. (This seems to be the case in our data.) This can lead to lower workplace mobility on friday too and further support the fact that after smoothing holidays falling during the middle of the week are more noticeable.