ORF525 - Problem Set 1

Bachir EL KHADIR

February 19, 2016

```r
train.data <- read.csv('data/train.data.csv')
test.data <- read.csv('data/test.data.csv')
```

1. The conversion is necessary because otherwise we would have an unwanted order relation. For three similar houses A, B, C in zipcodes 98001, 98002, 98003, a linear model would be forced to affect a price for the house A that lies between the price for house A and C, which is a bug and not a future of the data itself.

   ```r
   train.data$zipcode <- as.factor(train.data$zipcode)
   test.data$zipcode <- as.factor(test.data$zipcode)
   ```

2. ```r
   # Linear regression
   formula <- price ~ bedrooms + bathrooms + sqft_living + sqft_lot
   linear.model <- lm(formula, data=train.data)
   r2.train <- mean(residuals(linear.model)**2)
   r2.test <- mean(
   (predict(linear.model, test.data) - test.data$price)**2
   )
   ```

   a) ```r
      r2.train <- mean(residuals(linear.model)**2)
      r2.test <- mean(
      (predict(linear.model, test.data) - test.data$price)**2
)
      ```

      R^2 error:

$train$	$test$
6.6139436×10^{10}	6.6430543×10^{10}

   b) ```r
      # Linear regression with zip code
      formula <- update(formula, ~ . + zipcode)
      linear.model.with.zipcode <- lm(formula, data=train.data)
      r2.train.with.zipcode <- mean(
      residuals(linear.model.with.zipcode)**2
)
      r2.test.with.zipcode <- mean(
      (
          predict(linear.model.with.zipcode, test.data)
          - test.data$price
)**2)
      ```
      ```

R^2 error:

$train$	$test$
3.5200371×10^{10}	3.5178693×10^{10}

The addiitional factor (zipcode) explain reduce the R^2-error by almost half. As we will see before, it is one of the most important factor for the prediction of houses price.

3.
```r
library(glmnet)
feature.names <- colnames(train.data)[5:length(train.data)]
formula <- as.formula(paste('~', paste(feature.names, collapse='+')))
X.train <- model.matrix(formula, data=train.data)[,-1]
y.train <- train.data$price
X.test <- model.matrix(formula, data=test.data)[,-1]
y.test <- test.data$price
```

a)
```r
# Plot lasso / ridge path
plot.path <- function(alpha, ...) {
    fit <- glmnet(X.train, y.train, alpha=alpha)
    L <- length(fit$lambda)
    y.train <- fit$beta[, L]
    labs <- names(y.train)
    par(mar=c(4.5,4.5,1,4))
    plot(fit, xvar="norm", label=T)
    vnat=coef(fit)
    vnat=vnat[-1,ncol(vnat)]
    axis(4, at=vnat,line=-.5,label=labs,las=1,
        tick=FALSE, cex.axis=0.5)
}
```

```r
plot.path(alpha = 1)
```

```r
plot.path(alpha = 0)
```

Criteria for choosing: The most important variables will be the first to take the value non zero (they best explain the price when we resitrict the norm to be too small), and don't go back to zero when we let the norm to be bigger (no other combination of the variablese explain the price better). The ridge/lasso path agree that the most important variables are:

– Is the house in zipcode 96039?
– Does the house have a waterfromt?
– What is the latitude of the location

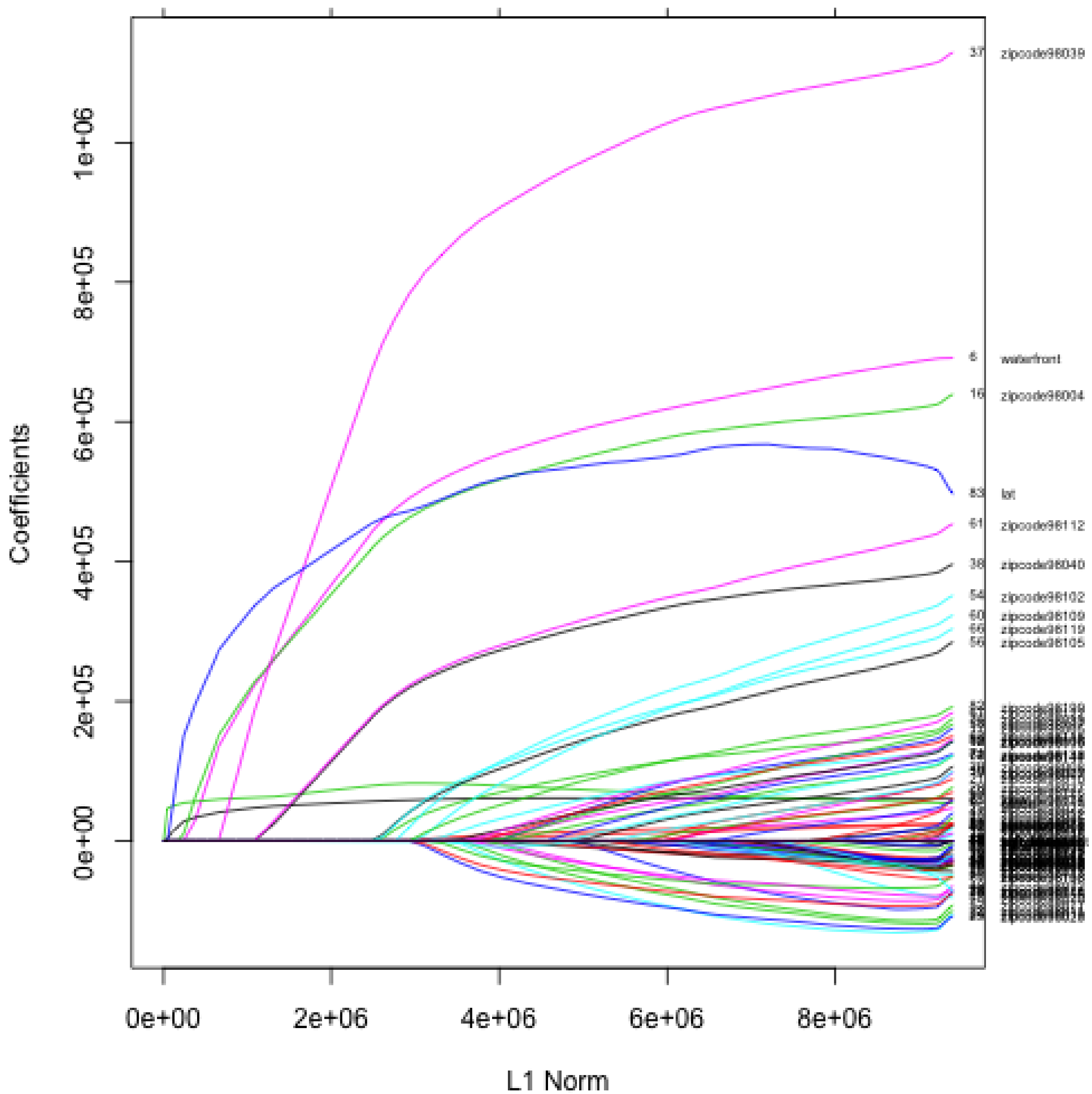


Figure 1: lasso

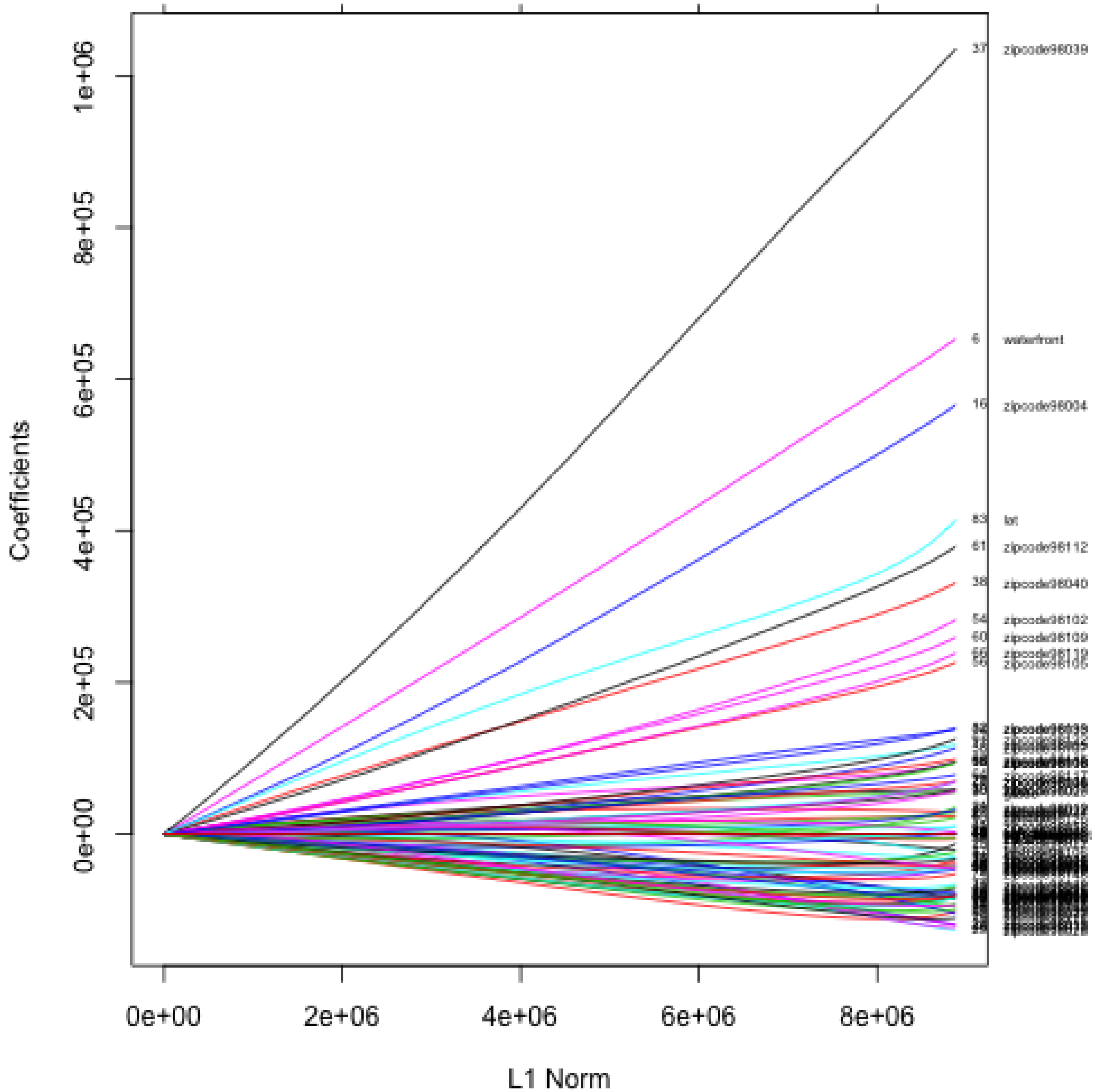


Figure 2: ridge

- Is the house in zipcode 96112?

b)
```
do.cv <- function(alpha) {
    model <- glmnet(X.train, y.train, alpha=alpha)
    cv <- cv.glmnet(X.train,y.train, nfolds=5,
                    alpha=alpha)
    l1.norm.min = sum(abs(coef(cv, s='lambda.min')[-1]))

    par(mfrow=c(2,1))
    plot(model, xvar="norm")
    abline(v=l1.norm.min, lty=2)
    text(x=l1.norm.min*0.95, y=2e5,
         label="best coefficient")
    plot(cv)
    cv
}
```

```
lasso.model.cv <- do.cv(alpha=1)
```

```
ridge.model.cv <- do.cv(alpha=0)
```

c)
```
evaluate.on.test <- function(alpha){
    model <- glmnet(X.train, y.train, alpha=alpha)
    cv <- cv.glmnet(X.train,y.train, nfolds=5,
                    alpha=alpha)
    pred <- predict(cv, X.test, s="lambda.min")
    mse <- mean((pred - y.test)^2)
}

r2.cv.lasso <- evaluate.on.test(alpha=1)
r2.cv.ridge <- evaluate.on.test(alpha=0)
```

R^2 error:

lasso	ridge
2.6825407×10^{10}	2.6905625×10^{10}

4. We regroup the zipcode by the city they are located in, We use the zipcode package to know in which cities are the hosues located:

```
library(ggplot2)
library(zipcode)
library(randomForest)
```

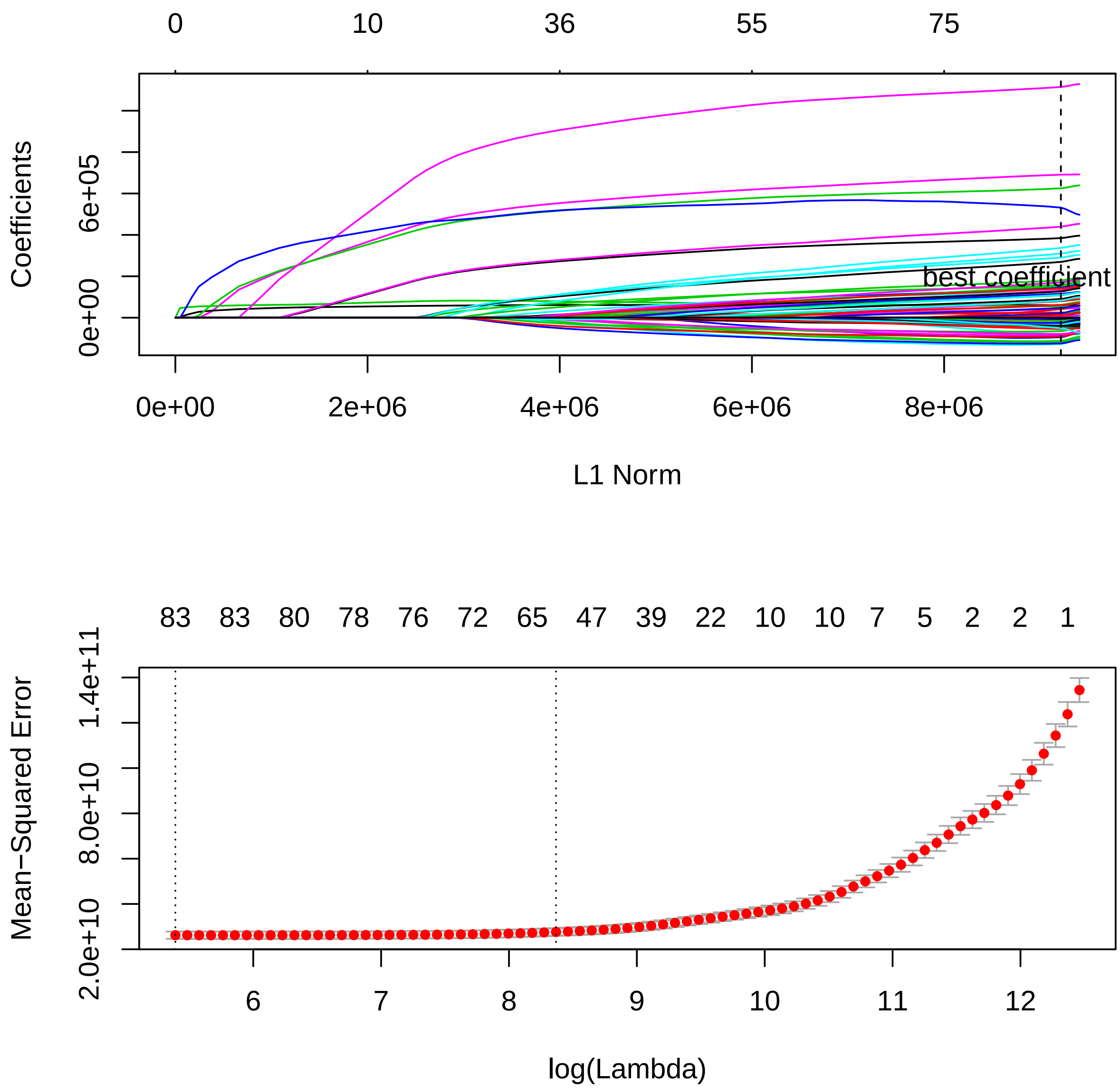


Figure 3: lasso cross validation

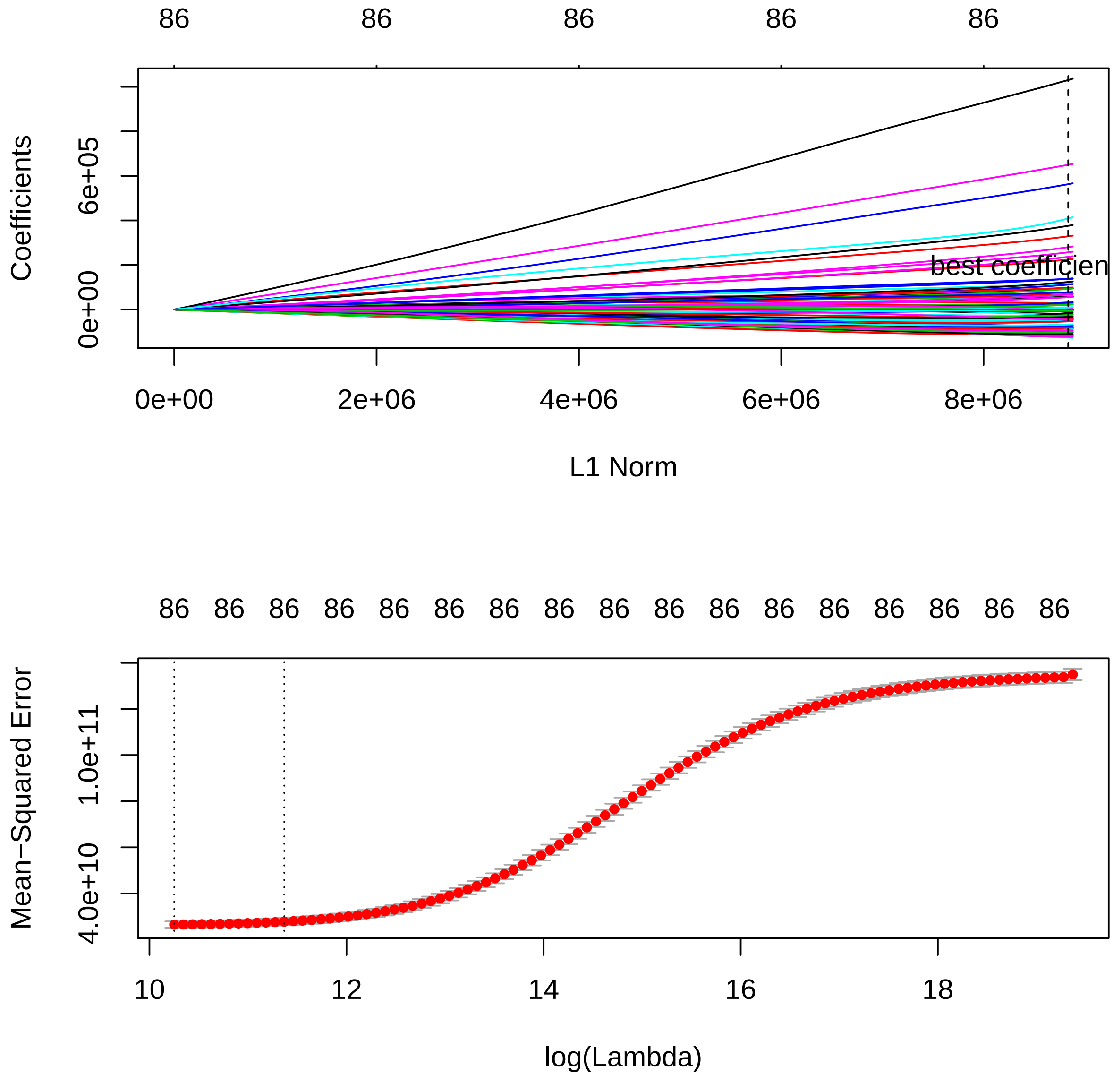


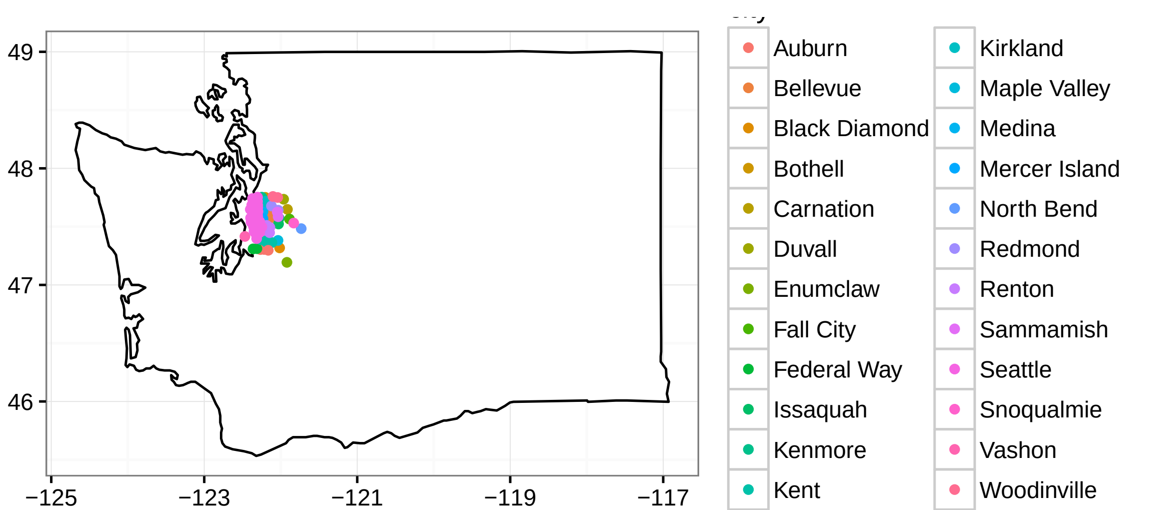
Figure 4: ridge cross validation

```
library(maps)

state.map <- map_data('state', region = 'Washington')
data(zipcode)

areas.of.interest <- data.frame(
    zipcode=levels(train.data$zipcode)
)
areas.of.interest <- merge(areas.of.interest, zipcode,
                           by.x='zipcode', by.y ='zip')
areas.of.interest$region <- substr(areas.of.interest$zipcode,
                                   1, 1)
num.cities <- length(unique(areas.of.interest$city)) #24

# plot map
g = ggplot(data=areas.of.interest)
g = g + geom_polygon( data=state.map, aes(x=long, y=lat,group=group),
                     colour="black", fill="white" )
g = g + geom_point(aes(x=longitude, y=latitude, colour=city))
g = g + theme_bw()
g + labs(x=NULL, y=NULL)
```



The following adds a “city” column to the dataset.

```
# Utility function to find the city of the zipcode
zip.to.city <- function(df) {
    city <- merge(df, zipcode, by.x='zipcode', by.y='zip')$city
    as.factor(city)
}
```

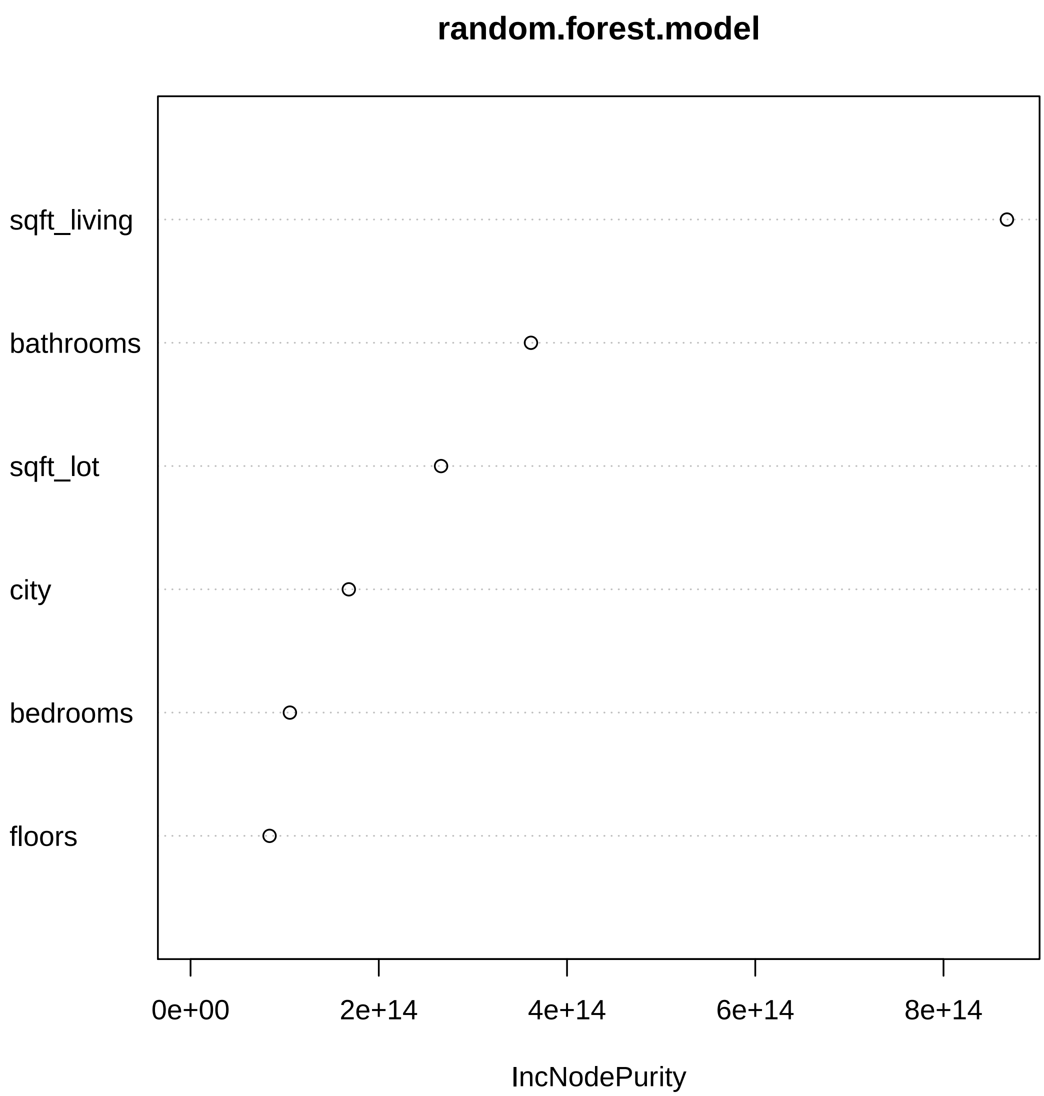
```r
train.data$city <- zip.to.city(train.data)
test.data$city <- zip.to.city(test.data)
```

The following code trains a random forest

```r
formula <- price ~ sqft_living + sqft_lot +
    bedrooms + bathrooms + floors + city
random.forest.model <- randomForest(formula, data=train.data)
```

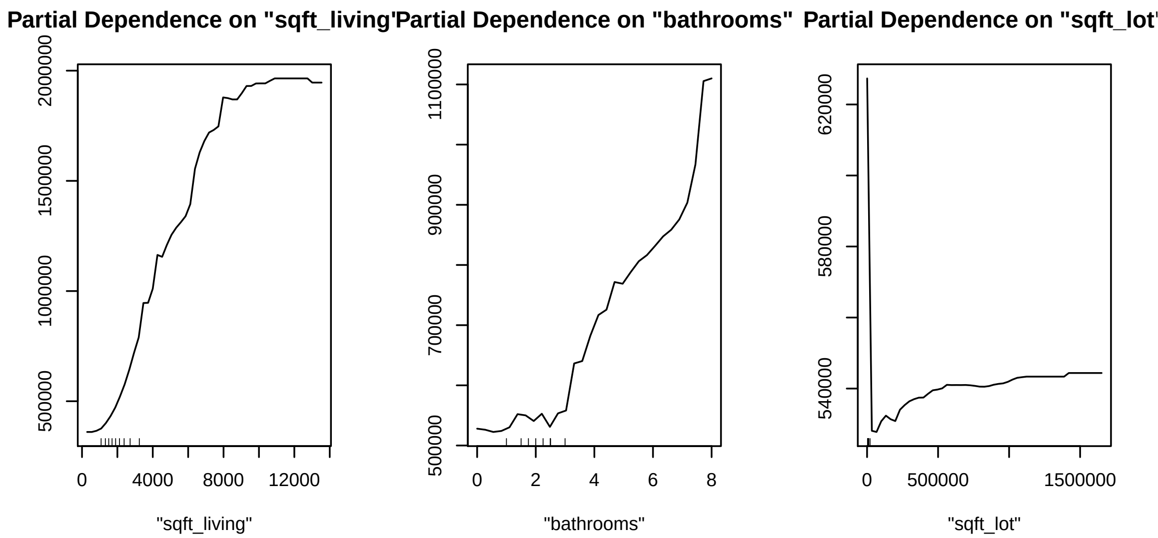
Here we plot variable importance:

```r
varImpPlot(random.forest.model)
```



and partial importance for each variable:

```r
par(mfrow=c(2, 3), xpd=NA)
partialPlot(fit, train.data, x.var="sqft_living")
partialPlot(fit, train.data, x.var="bathrooms")
partialPlot(fit, train.data, x.var="sqft_lot")
```



5. We use a lasso estimator computed in question c) using cross validation.

```r
# Convert donald trump to a dummmy variable understandable by glmnet
donald.trump <-  list(bedrooms=8, bathrooms=25,
                      sqft_living=50000,
```

```
                              sqft_lot=225000, floors=4,
                              condition=10, grade=10,
                              waterfront=1, view=4,
                              sqft_above=37500, sqft_basement=12500,
                              yr_built=1994,yr_renovated=2010,
                              lat=47.627606, long=-122.242054,
                              sqft_living15=5000,sqft_lot15=40000,
                              zipcode98039=1)
trump.dummy <- 0*X.train[1,]
for(entry in names(donald.trump))
    trump.dummy[entry] <- donald.trump[entry]
trump.dummy <- t(unlist(trump.dummy))

house.donald.trump <- predict(lasso.model.cv,
                              trump.dummy,
                              s='lambda.min')
```

The price estimate is **12.31M $**.

Q.2

- $\|Y-\theta\|_2^2 + 4\tau^2\|\theta\|_0 = \sum(y_i-\theta_i)^2 + 4\tau^2 1_{\theta_i\neq 0} = \sum_i f(\theta_i)$ Where $f : \theta \to (y-\theta)^2 + 4\tau^2 1_{\theta\neq 0}$, eg

$$f(\theta) = \begin{cases} y^2 & \text{if } \theta = 0 \\ (y-\theta)^2 + 4\tau^2 & \text{if } \theta \neq 0 \end{cases}$$

  The problem is linearly separable, we can minimize on each variable θ_i independetly:

  - If $|y| > 2\tau$, then $y^2 \geq 4\tau^2$ and $(y-\theta)^2 + 4\tau^2 \geq 4\tau^2 = f(y)$.
  - If $|y| \leq 2\tau$, then $f(0) = y^2 \leq 4\tau^2 \leq y^2 + (y-\theta) = f(\theta) \forall \theta \neq 0$.

  So $\arg\min \|Y-\theta\|_2^2 + 4\tau^2 = \hat{\theta}^{\text{hard}}$.

- $$\|Y-\theta\|_2^2 + 4\tau\|\theta\|_1 = \sum(y_i-\theta_i)^2 + 4\tau|\theta_i| = \sum_i g(\theta_i)$$

  The problem is linearly separable, we can minimize on each variable θ_i independetly. $g : \theta \to (y-\theta)^2 + 4\tau|\theta|$, eg

$$g(\theta) = \begin{cases} g_1(\theta) = (y-\theta)^2 + 4\tau\theta & = (\theta - (-2\tau + y))^2 + 2\tau(\tau + y) & \text{if } \theta \geq 0 \\ g_2(\theta) = (y-\theta)^2 - 4\tau\theta & = (\theta - (2\tau + y))^2 + 2\tau(\tau + y) & \text{if } \theta \leq 0 \end{cases}$$

  - * If $|y| \leq 2\tau$, then g_1 is increasing on $[0, \infty)$, so it has a minimum at 0, and the minimum is $y^2 = g(0)$.
    * g_2 is decreasing on $(-\infty, 0]$ so it has a minimum at 0.
    * c/c $\arg\min g$ in this case is 0.
  - * If $y \geq 2\tau$ then g_1 has minimum at $y - 2\tau > 0$ and the minimum is $y^2 - (2\tau - y)^2$.
    * g_2 is decreasing and has a minimum at 0, $g_2(0) = y^2 \geq g(y-2\tau)$ with equality only if $y - 2\tau = 0$
    * c/c $\arg\min g$ in this case is $y - 2\tau$.
  - if $y \leq -2\tau$, by a similar argument, $\arg\min g$ in this case is $y + 2\tau$.

  So $\arg\min \|Y-\theta\|_2^2 + 4\tau\theta = \hat{\theta}^{\text{hard}}$.

Q.3

1. (a) $M(x,y,z)$ is of rank 1, so $\begin{pmatrix} x & y \end{pmatrix}$ and $\begin{pmatrix} y & z \end{pmatrix}$ are colinear, so $\exists \lambda \in \mathbb{R}$ $\begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} y & z \end{pmatrix}$ or $\begin{pmatrix} y & z \end{pmatrix} = \lambda \begin{pmatrix} z & y \end{pmatrix}$
   i. Let's assume that $\begin{pmatrix} x & y \end{pmatrix} = \lambda \begin{pmatrix} y & z \end{pmatrix}$, and we can deduce the other case by symmetry. In this case $x = \lambda^2 z, y = \lambda z$. Which means that $y^2 = \lambda^2 zz = xz$.
   ii. When $\begin{pmatrix} y & z \end{pmatrix} = \lambda \begin{pmatrix} x & y \end{pmatrix}$, by a similar argument, $y^2 = xz$

   The op norm is the biggest eigen value in absolute value, in this case since one of the eigen values is 0 (because the determinant is 0) the op norm is: $|tr(M(x,y,z))| = |x+z|$, this is equal to 1 only if $1 = |x+z|^2 = x^2 + z^2 + 2xz = x^2 + z^2 + 2y^2$ Letting $t = \sqrt{y}$, we have that: $\{(x, \sqrt{2}y, z) | \operatorname{Rank}(M) = 1, \|M\|_{op} = 1\} \subseteq \{(x,t,z) : x^2 + t^2 + z^2 = 1, |x+z| = 1\}$
   For a matrix $M(x,y,z)$ such that $x^2 + z^2 + 2y^2 = 1$ and $|x+z| = 1$, it is easy to see that:
   - $\|M\|_{op} = |\operatorname{tr}(M)| = |x+z| = 1$

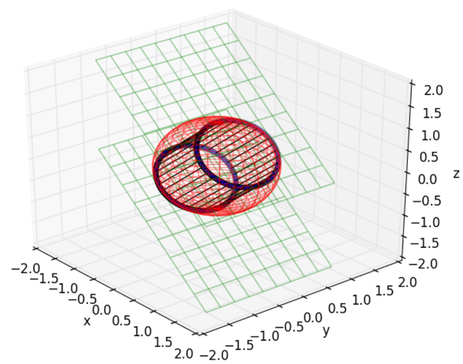


Figure 1: Shapes of the two sets

- $y^2 = \frac{1-x^2+z^2}{2} = \frac{(x+z)^2-x^2-y^2}{2} = xz$, so $\det(M) = 0$, therefore M cannot have rank 2. It cannot have rank 0 either because $x^2+2y^2+z^2 = 1 \implies$ one of the coefficient is not 0.

Therefore we have: $\{(x, \sqrt{2}y, z) | \operatorname{Rank}(M) = 1, \|M\|_{op} = 1\} = \{(x,t,z) : x^2+t^2+z^2 = 1, |x+z| = 1\}$
$x^2+t^2+z^2$ describes the unit sphere, $|x+z| = 1$ describe the union of the two hyperplanes $x+z = \pm 1$. Therefore this set is the union of intersection of sphere with two hyperplanes, e.g a union of two circles.
Or in parametric form:

$$\begin{cases} x &= \frac{1}{2} + \cos\theta \\ t &= \frac{\sqrt{2}}{2}\sin\theta \\ z &= -\frac{1}{2} - \cos\theta \end{cases}$$

$$\begin{cases} x &= -\frac{1}{2} + \cos\theta \\ t &= \frac{\sqrt{2}}{2}\sin\theta \\ z &= -\frac{3}{2} - \cos\theta \end{cases}$$

(b) $M := M(x,y,z)$ is symmetric, let's call σ_1, σ_2 its eigen values, we know that:

$$\begin{aligned} \|M\|^* &= |\sigma_1| + |\sigma_2| \\ tr(M) &= x+z = \sigma_1+\sigma_2 \\ tr(M^TM) &= x^2+z^2+2y^2 = \sigma_1^2+\sigma_2^2 \\ \det(M) &= xz - y^2 = \sigma_1\sigma_2 \end{aligned}$$

Therefore $(\sigma_1-\sigma_2)^2 = \sigma_1^2+\sigma_2^2-2\sigma_1\sigma_2 = tr(M^TM) - 2\det(M) = x^2+z^2+2y^2-2(xz-y^2) = (x-z)^2+4y^2$ Since $\|M\|_*^2 = \max\{|\sigma_1+\sigma_2|^2, |\sigma_1-\sigma_2|^2\}$, $\|M\|_*^2 = \max\{(x+z)^2, (x-z)^2+4y^2\}$, and therefore

$$\|M\|_* \leq 1 \iff \begin{cases} (x+z)^2 \leq 1 \\ (x-z)^2+4y^2 \leq 1 \end{cases}$$

This can be rewritten in the orthornomal basis:

$$\begin{cases} u &= \frac{x-z}{\sqrt{2}} \\ v &= y \\ w &= \frac{x+z}{\sqrt{2}} \end{cases}$$

$$\begin{cases} |w| \leq 2 \\ u^2 + v^2 \leq \frac{1}{4} \end{cases}$$

This describe the part of the cylinder with axe v radius $r = \frac{1}{2}$ ($u^2+v^2 \leq \frac{1}{4}$) contained between the two planes $w = \pm 2$ In the (x, t, x) this describes the cylinder whose axe is $(1, 0, 1)$, and contained between to the planes $x + z = \pm 1$.

2. If $A = uv^T$ has rank ≤ 1, A has at most one non null singular value, then $\|A\|_{op} = |\operatorname{tr}(A)| = |v^T u| = \|A\|_*$, if $\|A\|_{op} \leq 1$, then $\|A\|_* \leq 1$. The nuclear norm is a norm (fact proven in the next exercise), so the unit ball is convex, and therefore: $\operatorname{conv}\{uv^T : \|uv^T\|_{op} \leq 1\} \subseteq \{X : \|X\|_* \leq 1\}$.

   Let $X \in \mathbb{R}^{d_1 \times d_2}$ st $\|X\|_* \leq 1$ and let $U\Lambda V^T$ be its SVD. Then $\Lambda = \sum_{i=1}^d \sigma_i(X) e_i^T e_i$ where $(e_i)_i$ is the canonical basis of R^{d^2}. Therefore $X = \sum_{i=1}^d \sigma_i(X) \underbrace{U e_i^T e_i V^T}_{\text{Rank}=1} + (1 - \underbrace{\sum_{i=1}^{d} \sigma_i(X)}_{\|X\|_{op} \leq 1})0 \in \operatorname{conv}\{uv^T : \|uv^T\|_{op} \leq 1\}$. c/c $\operatorname{conv}\{uv^T : \|uv^T\|_{op} \leq 1\} = \{X : \|X\|_* \leq 1\}$.

3. Two remarks:

   - For A, B sqaure matrices of the same shape, $\|AB\|_F^2 = \operatorname{tr}(ABB^T A^T) = \operatorname{tr}(A^T ABB^T) = \operatorname{tr}(BB^T A^T A) = \|B^T A^T\|_F^2$
   - If O is orthogonal, the $\|AO\|_F^2 = \operatorname{tr}(AOO^T A^T) = \operatorname{tr}(AA^T) = \|A\|_F^2$, $\|OA\|_F = \|A^T O^T\|_F = \|A^T\|_F = \|A\|_F$.

   Let $X = U\Lambda V^T$ be the SVD of X.

   $\|X - XZ\|_F^2 = \|X(I - Z)\|_F^2 = \|U\Lambda V^T(I - Z)\|_F^2 = \|\Lambda V^T(I - Z)\|_F^2 = \|\Lambda V^T V(V^T V - V^T Z V)V^T\|_F^2 = \|\Lambda(I - V^T Z V)\|_F^2$.

   Since V is invertible, let's do the invertible change of variable $Y = V^T Z V$. Moreover, V is orthogonal, so the singular values of Y and Z are the same as well as their nuclear norm. The problem can thus be reduced to:

$$\min_Y \|Y\|_* + \frac{\tau}{2}\|\Lambda(I - Y)\|_F^2$$

   Moreover, we have that:

$$\begin{aligned}
\|\Lambda(I - Y)\|_F^2 &= \sum_{ij} [e_i^T \Lambda(I - Y)e_j]^2 \\
&= \sum_{ij} [\Lambda_i e_i^T (e_j - Y e_j)]^2 \\
&= \sum_{ij} [\Lambda_i(\delta_{ij} - Y_{ij})]^2 \\
&= \sum_{i \neq j} \underbrace{\Lambda_i^2 y_{ij}^2}_{\geq 0} + \sum_i \Lambda_i^2 (1 - y_{ii})^2 \\
&\geq \sum_i \Lambda_i^2 (1 - y_{ii})^2 \\
&= \|\Lambda(I - diag(Y))\|_F^2
\end{aligned}$$

And, for any matrix X:

$$\langle diag(Y)\,,\,X\rangle = \sum_i y_i i x_i i \leq \sum_{u\neq j} y_{ii}x_i i + \sum_{i\neq j} y_{ij}x_i i = \langle diag(Y)\,,\,\tilde{X}\rangle$$

Where $\tilde{X}_{ii} = X_{ii}$ and $\tilde{X}_{ij} = Y_{ij}$ when $i \neq j$ Which proves that $\|diag(Y)\|_{op} \leq \|Y\|_{op}$

As a result $\|Y\|_* + \frac{\tau}{2}\|\Lambda - \Lambda Y\|_F^2 \geq \|diag(Y)\|_* + \frac{\tau}{2}\|\Lambda - \Lambda diag(Y)\|_F^2$ with strict inequality if Y is not diagonal, so we can resitrict the search for diagonla matrices only.

If Y is diagonal, then $\|diag(Y)\|_* + \frac{\tau}{2}\|\Lambda - \Lambda diag(Y)\|_F^2 = \sum_i \underbrace{|y_i| + \Lambda_i^2(1-y_{ii})^2}_{\phi(y_{ii})}$

Minimizing the quadratic function $\Phi : y \to |y| + \frac{\tau}{2}\Lambda_i^2(1-y)^2$ leads to the only solution $y = (1 - \frac{1}{\tau\Lambda_i^2})^+$ Therefore $Y = \max(0, I - \frac{1}{\tau}\Lambda^{-2})$ Doing the inverse change of variable we get that $Z = V^T \max(0, I - \frac{1}{\tau}\Lambda^{-2})V$.

Q.4

4.1 (a) Let $Y = U\Lambda V^T$ be the SVD decomposition of Y, then $\langle Y, UVT\rangle = \mathrm{tr}(V\Lambda U^T U V^T) = \mathrm{tr}(V\Lambda V^T) = \mathrm{tr}(\Lambda) = \|Y\|_*$.

$\langle Y, X\rangle = \mathrm{tr}(Y^T X) = \mathrm{tr}(V\Lambda U^T X) = \mathrm{tr}(\Lambda U^T X V) = \sum \Lambda_{ii}(U^T X V)_{ii} = \sum \Lambda_{ii} \underbrace{u_i^T X v_i}_{\leq \|X\|_{op}} \leq$

$\|X\|_{op}\|\Lambda\|_* \leq \|Y\|_*$.

c/c $\|Y\|_* = \max_{\|X\|_{op}\leq 1}\langle Y, X\rangle$

4.2 For $\alpha \in (0,1)$, $\|\alpha Y + (1-\alpha)Z\|_* = \max_{\|X\|_{op}\leq 1} \alpha\langle Y, X\rangle + (1-\alpha)\langle Z, X\rangle \leq \alpha \max_{\|X\|_{op}\leq 1}\langle Y, X\rangle + (1-\alpha)\max_{\|X\|_{op}\leq 1}\langle Z, X\rangle \leq \alpha\|Y\|_* + (1-\alpha)\|Z\|_*$.

Where we used the fact $\max A + B = \max A + \max B$.

4.3 $\Rightarrow$ Let's suppose $Z \in \partial\|A\|_*$, then $\|B\|_* \geq \|A\|_* + \langle Z, B - A\rangle$ for all B.

- For $B = 2A$, $\|A\|_* \geq \langle Z, A\rangle$.
- For $B = 0$, $0 \geq \|A\|_* - \langle Z, A\rangle \implies \langle Z, A\rangle \geq \|A\|_*$.

So $\langle Z, A\rangle = \|A\|_*$

Let u, v two vectors of norm 1, then:

By the triangular inequality of the nuclear norm: $v^T Z u = \langle Z\,,\,uv^T\rangle \leq \|X_0 + uv^T\|_* - \|X_0\| \leq \|uv^T\|_*$.

But $u^T v$ is of rank 1, so $\|u^T v\|_* = |\,\mathrm{tr}(uv^T)| = |\langle u\,,\,v\rangle| \leq 1$, which proves that $\|Z\|_{op} \leq 1$

$\Leftarrow$ Let Z, such that $\|Z\|_{op} \leq 1$ and $\langle Z, A\rangle = \|A\|_*$, then by 4.1 $\forall X$ $\langle \frac{Z}{\|Z\|_{op}}, X\rangle \leq \|X\|_*$ so that $\langle Z\,,\,X\rangle \leq \|X\|_*\|Z\|_{op} \leq \|X\|_*$ and: $\|X\|_* - \|A\|_* \geq \langle Z, X\rangle - \langle Z, A\rangle$, which means that $Z \in \partial\|A\|_*$

4.4 (a)

$$\begin{aligned}
Z \in \partial\|A\| &\implies \|Z\|_{op} = 1, \|A\|_* = \|\Lambda\|_* = \langle Z \,,\, A\rangle \\
&\implies \|\Lambda\|_* = \mathrm{tr}(Z^T U\Lambda V^T) = \mathrm{tr}((U^T Z V)^T \Lambda) = \sum_i (U^T Z V)_{ii}\Lambda_i \\
&\implies \sum_i \underbrace{\Lambda_i(1 - u_i^T Z v_i)}_{\geq 0} = 0 \\
&\text{(because } |u_i^T Z v_i| \leq \|u_i\|\|v_i\|\|Z\|_{op} \leq 1) \\
&\implies \forall i \; u_i^T Z v_i = 1
\end{aligned}$$

Let's complete the family $(u_i)_{i\leq r}$ into $(u_i)_{i\leq d_1}$ an orthonormal basis of $\mathbb{R}^{d_1}$. Then, for $i \leq r$, $1 \geq \|Zv_i\|^2 = \sum_{j=1}^{d_1}(u_j^T Z v_i)^2 \geq 1 + \sum_{j\neq i} \underbrace{(u_j^T Z v_i)^2}_{=0} \geq 1$

So $u_j^T Z v_i = \delta_{ij}$ and $Zv_i = \sum_{j=1}^{d_1} u_j^T Z v_i u_j = u_i^T Z v_i u_i = u_i$. In matrix form: $ZV = U$, and using a similar arguement $U^T Z = V$.

Let $W := Z - UV^T$, then the last equations can be written as $U = (W + UV^T)V = WV + UV^T V = WV + U \implies WV = 0$ and similarly $U^T W = 0$.

Let $x \in \mathbb{R}^{d_2}$, and let $x = x_1 + x_2$ be a decomposition according to $\mathbb{R}^{d_2} = im(V) + im(V)^\perp$, and let $y \in \mathbb{R}^r$ such that $x_1 = Vy$. (note that $\|x\|^2 = \|x_1\|^2 + \|x_2\|^2$)

Then $\|Wx\|^2 = \|\underbrace{ZV}_{U} y + Zx_2 - U\underbrace{V^T V}_{I_r} y - U\underbrace{V^T x_2}_{=0}\|^2 = \|Uy + Zx_2 - Uy\|^2 = \|Zx_2\|^2 \leq \|x_2\|^2 \leq \|x\|^2$

c/c $Z = U^T V + W, \|W\|_{op} \leq 1$, $WV = U^T W = 0$.

(b) Now take Z of the form $UV^T + W$, and let's prove that $\|Z\|_{op} \leq 1$ and $\langle Z \,,\, A\rangle = \|A\|_*$ so we can conclude using the last question.

$$\begin{aligned}
\langle UV^T + W \,,\, A\rangle &= \mathrm{tr}(VU^T U\Lambda V^T) + \mathrm{tr}(W^T U\Lambda V^T) \\
&= \mathrm{tr}(\Lambda) \\
&= \|A\|_*
\end{aligned}$$

Let $x \in \mathbb{R}^{d_2}$, then :

$$\begin{aligned}
\|UV^T x + Wx\|^2 &= \|UV^T x\|^2 + \|Wx\| && \text{(because } im(U) \perp im(W)) \\
&= \|V^T x\|^2 + \|Wx\|^2 && \text{(Because } U \text{ is an isometrie)}
\end{aligned}$$

Let's write $x = x_1 + x_2$ according to the decomposition $\mathbb{R}^{d_2} = im(V) + im(V)^\perp$, and let $y \in \mathbb{R}^r$ such that $x_1 = Vy$. (note that $\|x_1\| = \|y\|$) $V^T x = V^T x_1 + V^T x_2 = V^T V y = y$, $Wx = WVy + Wx_2 = Wx_2$ so $\|(UV^T + W)x\|^2 = \|y\|^2 + \|Wx_2\|^2 \leq \|x_1\|^2 + \|x_2\|^2 = \|x\|^2$, which proves that $\|UV^T + W\|_{op} \leq 1$. Which proves that $Z \in \partial\|A\|_*$.

4.5

- Let $Z \in \partial\|A\|_*$, then Z can be written as $Z = UV^T + W$ where $U^T W = WV = 0$ and $\|W\|_{op} \leq 1$.
  $UV^T = U\frac{V}{2}^T + \frac{U}{2}V^T \in T$.
  Let's now prove that $W \perp T$, so that we can conclude that $\Pi_{T^c}(Z) = W$.
  Let $X \in \mathbb{R}^{d_2\times r}, Y \in \mathbb{R}^{d_2\times r}$:
  $\langle UX^T + Y^T V \,,\, W\rangle = \mathrm{tr}(UX^T W) + \mathrm{tr}(Y^T V W) = \mathrm{tr}(W^T U X^T) = 0$.
  As a conclusion $\Pi_T(Z) = UV^T, \Pi_{T^c}(Z) = W$, $\|\Pi_{T^c}(Z)\|_{op} = \|W\|_{op} \leq 1$

- Let Z be such that $\Pi_T(Z) = UV^T$ and $\|\Pi_{T^c}(Z)\|_{op} \leq 1$.
  Let's call $W := Z - UV^T$, so that $Z = W + UV^T$
  $\Pi_T(W) = \Pi_T(Z) - \Pi_T(UV^T) = UV^T - UV^T = 0$, so $W = \Pi_{T^c}(W) = \Pi_{T^c}(Z)$ and as a result $\|W\|_{op} \leq 1$, and for any $X \in \mathbb{R}^{d_2 \times r}, Y \in \mathbb{R}^{d_2 \times r}$ we have that: $\mathrm{tr}(W^T UX) + \mathrm{tr}(W^T Y^T V) = 0$.
  For $Y = 0$, we get that $W^T U$ is orthogonal to all matrices in $\mathbb{R}^{d_2 \times r}$, so $W^T U = 0$. Using a similar argument $VW = 0$
  Which proves $Z \in \partial\|W\|_*$.

4.6 Let's prove the hint first:

Let f be a convexe non-negative function, and consider the following problem: $\min_X f(X)$ st $B = \mathcal{A}(X)$.

Slater's condition are satisfied: The problem is feasible, the constraints are linear, f is convexe. X is an optimal solution iff there exists $P \in \mathbb{R}^{d_1 \times d_2}$ such that the KKT condition is verified: $0 \in \partial_X(f(X) - \langle P \, , \, \mathcal{A}(X) - B \rangle)$

But:

$$\begin{aligned}\partial_X(f(X) - \langle P \, , \, \mathcal{A}(X) - B \rangle) &= \partial_X f(X) - \partial_X \langle \mathcal{A}^*(P) \, , \, X \rangle \\ &= \partial_X f(X) - \mathcal{A}^*(P)\end{aligned}$$

So $0 \in \partial_X(f(X) - \langle P \, , \, \mathcal{A}(X) - B \rangle) \iff \exists P \mathcal{A}^*(P) \in \partial_X f(X)$. Which proves the hint.

Back to $f = \|\|_*$.

X_0 is then clearly a solution to the problem because

- there exsits P such taht
- $\mathcal{A}(X_0) = B$
- $\Pi_T(\mathcal{A}^*(P)) = UV^T$, $\|\Pi_{T^c}(\mathcal{A}^*(P))\|_{op} \leq 1$, which means $\mathcal{A}^*(P) \in \partial_X \|X_0\|_*$.

Let's now prove uniqueness (I have not finished this part yet, those are the intermediate results I have proved).

Take $X = U' \Lambda' V'^T$ to be an optimal solution to the minimization problem. Then $B = \mathcal{A}(X) = \mathcal{A}(X_0)$, which means that $\mathcal{A}(X - X_0) = 0$

Also, there exists Q in $\mathbb{R}^{d_1 \times d_2}$ such that $\mathcal{A}^*(Q) \in \partial_X \|X\|_*$.

$\|X\|_* = \|X_0\|_* \implies \langle X \, , \, \mathcal{A}^*(Q) \rangle = \langle X_0 \, , \, \mathcal{A}^*(P) \rangle \implies \langle \mathcal{A}(X) \, , \, Q \rangle = \langle \mathcal{A}(X_0) \, , \, P \rangle$ But $\mathcal{A}(X) = \mathcal{A}(X_0) = B$, so $\langle B \, , \, P \rangle = \langle B \, , \, Q \rangle$.

$\mathcal{A}$ injective on $T \implies null(\mathcal{A}) \cap T = \{0\} \implies Im(\mathcal{A}^*)^\perp \cap T = \{0\} \implies Im(\mathcal{A}^*) + T^c = \mathcal{R}^{d_1 \times d_2}$