Chapter 6
Regularizing a Correlation Matrix

This chapter will show how to regularize the correlation and inverse correlation matrices. Many techniques from multivariate analysis, such as classification, are based on a covariance or correlation matrix. The inverse covariance matrix is also known as a *precision matrix*. A regularized estimator reduces the degrees of freedom d of the estimator. Often regularization is done by reducing the number of parameters in the model. For MLR, lasso and ridge regression were regularized if $\lambda > 0$. A covariance matrix of a $p \times 1$ vector $\boldsymbol{x}$ is symmetric with $p + (p-1) + \cdots + 2 + 1 = p(p+1)/2$ parameters. A correlation matrix has $p(p-1)/2$ parameters. We want $n \geq 10p$ for the sample covariance and correlation matrices $\boldsymbol{S}$ and $\boldsymbol{R}$. If $n < 5p$, then these matrices are being overfit: the degrees of freedom is too large for the sample size n, and the matrices may be ill conditioned. Too much regularization results in underfitting. We roughly want d to be such that the matrix is well conditioned for a given n, and the statistical or machine learning technique that used the matrix, such as classification, performs satisfactorily.

6.1 Correlation and Inverse Correlation Matrices

The sample covariance and correlation matrices $\boldsymbol{S}$ and $\boldsymbol{R}$ are given in Definitions 1.13 and 1.14.

Rule of Thumb 6.1. Multivariate procedures based on $\boldsymbol{S}$ or $\boldsymbol{R}$ start to give good results for $n \geq 10p$, especially if the distribution is close to multivariate normal. In particular, we want $n \geq 10p$ for the sample covariance and correlation matrices. For procedures with large sample theory on a large class of distributions, for any value of n, there are always distributions where the results will be poor, but will eventually be good for larger sample sizes. Norman and Streiner (1986, pp. 122, 130, 157) gave this rule of thumb and note that some authors recommend $n \geq 30p$. This rule of thumb is much like

the rule of thumb that says the central limit theorem normal approximation for $\overline{Y}$ starts to be good for many distributions for $n \geq 30$. See the paragraph below Theorem 1.2.

The population and sample correlation are measures of the strength of a **linear relationship** between two random variables, satisfying $-1 \leq \rho_{ij} \leq 1$ and $-1 \leq r_{ij} \leq 1$. Let the $p \times p$ sample standard deviation matrix

$$\boldsymbol{D} = \text{diag}(\sqrt{\text{S}_{11}}, ..., \sqrt{\text{S}_{\text{pp}}}). \tag{6.1}$$

Then

$$\boldsymbol{S} = \boldsymbol{DRD}, \tag{6.2}$$

and

$$\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}. \tag{6.3}$$

The inverse covariance matrix or inverse correlation matrix can be used to find the partial correlation $r_{ij,\boldsymbol{x}(ij)}$ between x_i and x_j where $\boldsymbol{x}(ij)$ is the vector of predictors with x_i and x_j deleted where $i \neq j$. This partial correlation is the correlation of x_i and x_j after eliminating the linear effects of $\boldsymbol{x}(ij)$ from both variables: regress x_i and x_j on $\boldsymbol{x}(ij)$ and get the two sets of residuals, then find the correlation of the two sets of residuals. If $p \geq 3$ and $\boldsymbol{S}^{-1} = (S^{ij})$, then

$$r_{ij,\boldsymbol{x}(ij)} = \frac{-S^{ij}}{(S^{ii}S^{jj})^{1/2}} = \frac{-r^{ij}}{(r^{ii}r^{jj})^{1/2}}.$$

Srivastava and Khatri (1979, p. 53) proved this result. The second equality holds since

$$\boldsymbol{R}^{-1} = \boldsymbol{D}\boldsymbol{S}^{-1}\boldsymbol{D} = (r^{ij}) = (S^{ij}\sqrt{S_{ii}}\sqrt{S_{jj}}). \tag{6.4}$$

The ith diagonal element r^{ii}, called a variance inflation factor, is found by regressing x_i on the remaining predictors $x_1, ..., x_{i-1}, x_{i+1}, ..., x_p$. Then

$$r^{ii} = VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the squared multiple correlation from the regression. See Belsley et al. (1980, p. 93).

Some R code illustrating the result for r^{ij} is shown below. The function `lsfit` is used to regress x_1 on x_3 and then regress x_2 on x_3. Note that $\boldsymbol{x}(i = 1, j = 2) = x_3$ once x_1 and x_2 have been deleted since $p = 3$.

```
x <- buxx[,1:3]; z<-solve(cor(x))
z #inverse correlation matrix

               len      nasal    bigonal
len     1.02042523 0.13535798 0.06134196
```

```
nasal   0.13535798 1.02358206 0.08336109
bigonal 0.06134196 0.08336109 1.00931453

out1 <- lsfit(x[,3],x[,1])$resid
out2 <- lsfit(x[,3],x[,2])$resid
cor(out1,out2)
[1] -0.1324439

-z[1,2]/sqrt(z[1,1]*z[2,2])
[1] -0.1324439

zz <- solve(var(x)) #inverse covariance matrix
-zz[1,2]/sqrt(zz[1,1]*zz[2,2])
[1] -0.1324439
```

The *slpack* function `gcor` returns a (generalized) correlation matrix R given a symmetric positive definite matrix C with positive diagonal elements. The matrix D is such that C = D R D. See the following *R* code.

```
> C <- var(buxx)
> R<- cor(buxx)
> R
                len       nasal     bigonal    cephalic
len       1.00000000 -0.12815187 -0.05019157 -0.08359332
nasal    -0.12815187  1.00000000 -0.07480324 -0.08261217
bigonal  -0.05019157 -0.07480324  1.00000000  0.07204296
cephalic -0.08359332 -0.08261217  0.07204296  1.00000000
> out<-gcor(C)
> out$R
            [,1]        [,2]        [,3]        [,4]
[1,]  1.00000000 -0.12815187 -0.05019157 -0.08359332
[2,] -0.12815187  1.00000000 -0.07480324 -0.08261217
[3,] -0.05019157 -0.07480324  1.00000000  0.07204296
[4,] -0.08359332 -0.08261217  0.07204296  1.00000000
> C
                len       nasal     bigonal    cephalic
len      118299.9257 -191.084603 -104.718925 -124.477916
nasal      -191.0846   18.793905   -1.967121   -1.550533
bigonal    -104.7189   -1.967121   36.796311    1.892005
cephalic   -124.4779   -1.550533    1.892005   18.743774
> out$D%*%R%*%out$D
            [,1]        [,2]        [,3]        [,4]
[1,] 118299.9257 -191.084603 -104.718925 -124.477916
[2,]   -191.0846   18.793905   -1.967121   -1.550533
[3,]   -104.7189   -1.967121   36.796311    1.892005
[4,]   -124.4779   -1.550533    1.892005   18.743774
```

6.2 Regularizing a Correlation Matrix

Ridge regression regularizes $\boldsymbol{W}^T\boldsymbol{W} = n\boldsymbol{R}$, which is closely related to regularizing a covariance or correlation matrix. For $\delta \geq 0$, a simple way to regularize a $p \times p$ correlation matrix $\boldsymbol{R} = (r_{ij})$ is to use

$$\boldsymbol{R}_\delta = \frac{1}{1+\delta}(\boldsymbol{R} + \delta \boldsymbol{I}_p) = (t_{ij}) \tag{6.5}$$

where $t_{ii} = 1$ and

$$t_{ij} = \frac{r_{ij}}{1+\delta}$$

for $i \neq j$. Note that each correlation r_{ij} is divided by the same factor $1 + \delta$. If λ_i is the ith eigenvalue of $\boldsymbol{R}$, then $(\lambda_i + \delta)/(1 + \delta)$ is the ith eigenvalue of $\boldsymbol{R}_\delta$. The eigenvectors of $\boldsymbol{R}$ and $\boldsymbol{R}_\delta$ are the same since if $\boldsymbol{R}\ \boldsymbol{x} = \lambda_i\ \boldsymbol{x}$, then

$$\boldsymbol{R}_\delta\ \boldsymbol{x} = \frac{1}{1+\delta}(\boldsymbol{R} + \delta \boldsymbol{I}_p)\ \boldsymbol{x} = \frac{1}{1+\delta}(\lambda_i + \delta)\ \boldsymbol{x}.$$

Note that $\boldsymbol{R}_\delta = \kappa \boldsymbol{R} + (1 - \kappa)\boldsymbol{I}_p$ where $\kappa = 1/(1 + \delta) \in (0, 1]$. See Warton (2008).

Following Datta (1995, pp. 250-254), the condition number of a symmetric positive definite $p \times p$ matrix $\boldsymbol{A}$ is $cond(\boldsymbol{A}) = \lambda_1(\boldsymbol{A})/\lambda_p(\boldsymbol{A})$ where $\lambda_1(\boldsymbol{A}) \geq \lambda_2(\boldsymbol{A}) \geq \cdots \geq \lambda_p(\boldsymbol{A}) > 0$ are the eigenvalues of $\boldsymbol{A}$. Note that $cond(\boldsymbol{A}) \geq 1$. A well conditioned matrix has condition number $cond(\boldsymbol{A}) \leq c$ for some number c such as 50 or 500. Hence $\boldsymbol{R}_\delta$ is nonsingular for $\delta > 0$ and well conditioned if

$$cond(\boldsymbol{R}_\delta) = \frac{\lambda_1 + \delta}{\lambda_p + \delta} \leq c,$$

or

$$\delta = \max\left(0, \frac{\lambda_1 - c\lambda_p}{c-1}\right) \tag{6.6}$$

if $1 < c \leq 500$. Taking $c = 50$ suggests using

$$\delta = \max\left(0, \frac{\lambda_1 - 50\lambda_p}{49}\right).$$

This type of regularization is simple, but inverting a $p \times p$ matrix is expensive for large p. It would good to be able to do variable selection with r variables where $n \geq 10r$, and then use the correlation matrix of these variables. Since the t_{ij} are between -1 and 1, $|t_{ij}| < 0.02$ are likely unimportant, and we want a well conditioned matrix, the grid of δ values can be small: e.g. $\delta \in \{0, 0.01, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 2, 3, ..., 20, 40, 50\}$.

The matrix can be further regularized by setting $t_{ij} = 0$ if $|t_{ij}| \leq \tau$ where $\tau \in [0, 1)$ should be less than 0.5. Denote the resulting matrix by $\boldsymbol{R}(\delta, \tau)$. We suggest using $\tau = 0.05$. Note that $\boldsymbol{R}_\delta = \boldsymbol{R}(\delta, 0)$. Using τ is known as

thresholding. We recommend computing $\boldsymbol{I}_p, \boldsymbol{R}(\delta, 0)$ and $\boldsymbol{R}(\delta, 0.05)$ for $c =$ 50, 100, 200, 300, 400, and 500. Compute $\boldsymbol{R}$ if it is nonsingular. Note that a regularized covariance matrix can be found using

$$\boldsymbol{S}(\delta, \tau) = \boldsymbol{D} \ \ \boldsymbol{R}(\delta, \tau) \ \ \boldsymbol{D} \tag{6.7}$$

where $\boldsymbol{D}$ is given by Equation (6.1).

A common type of regularization of a covariance matrix $\boldsymbol{S}$ is to use $\boldsymbol{S}_D = diag(\boldsymbol{S})$ where the ijth element of $\boldsymbol{S}_D = 0$ and $\boldsymbol{S}_D(i, i) = \boldsymbol{S}(i, i)$. The corresponding correlation matrix is the identity matrix, and Mahalanobis distances using the identity matrix correspond to Euclidean distances. These estimators tend to use too much regularization, and underfit. Note that as $\delta \to \infty$, $\boldsymbol{R}_\delta \to \boldsymbol{I}_p$, and $\boldsymbol{I}_p$ corresponds to $c = 1$. Note that $\boldsymbol{S}_D$ corresponds to using $\boldsymbol{R}(\delta = \infty, 0) = \boldsymbol{I}_p$ in Equation (6.6).

The *slpack* function `corrlar` produces the regularized correlation matrices Rd $= \boldsymbol{R}(\delta, 0)$ and Rt $= \boldsymbol{R}(\delta, \tau)$ given a correlation matrix (e.g. from the function `gcor`), condition number c and threshold tau with $\tau = 0.05$ the default. The value delta $= \delta$ depends on c through Equation (6.6). See the following *R* code.

```
R<- cor(buxx)
corrlar(R,tau=0.05) #well conditioned so no regularization
corrlar(R,tau=0.07)
$Rr #no regularization
                 len       nasal     bigonal    cephalic
len       1.00000000 -0.12815187 -0.05019157 -0.08359332
nasal    -0.12815187  1.00000000 -0.07480324 -0.08261217
bigonal  -0.05019157 -0.07480324  1.00000000  0.07204296
cephalic -0.08359332 -0.08261217  0.07204296  1.00000000
$Rt #two entries changed to 0
                 len       nasal     bigonal    cephalic
len       1.00000000 -0.12815187  0.00000000 -0.08359332
nasal    -0.12815187  1.00000000 -0.07480324 -0.08261217
bigonal   0.00000000 -0.07480324  1.00000000  0.07204296
cephalic -0.08359332 -0.08261217  0.07204296  1.00000000
corrlar(R,c=1.2)
$Rr
                 len       nasal     bigonal    cephalic
len       1.00000000 -0.06378780 -0.02498294 -0.04160871
nasal    -0.06378780  1.00000000 -0.03723343 -0.04112034
bigonal  -0.02498294 -0.03723343  1.00000000  0.03585950
cephalic -0.04160871 -0.04112034  0.03585950  1.00000000
$Rt #too much regularization
```

```
                len      nasal bigonal cephalic
len       1.0000000 -0.0637878       0        0
nasal    -0.0637878  1.0000000       0        0
bigonal   0.0000000  0.0000000       1        0
cephalic  0.0000000  0.0000000       0        1
```

It is also common to analyze analogs of the inverse correlation matrix $\boldsymbol{R}^{-1} = (r^{ij})$ since the r^{ij} are closely related to partial correlations. See the discussion above and below Equation (6.4).

Here is a simple algorithm. If the condition number $cond(\boldsymbol{R}) \leq 500$, let $\boldsymbol{R}_d = \boldsymbol{R}$. Otherwise, let $\boldsymbol{R}_d = \boldsymbol{R}(\delta = 0.01, 0)$. Let $\boldsymbol{A} = \boldsymbol{R}_d^{-1}$ be the analog of $\boldsymbol{R}^{-1}$ to be regularized. Let $\boldsymbol{D}_A = \text{diag}(\sqrt{\text{A}_{11}}, ..., \sqrt{\text{A}_{\text{pp}}})$. Hence $\boldsymbol{A}$ acts like a covariance matrix. Then a generalized correlation matrix $\boldsymbol{R}_I = \boldsymbol{D}_A^{-1}\boldsymbol{A}\boldsymbol{D}_A^{-1}$ is made and regularized with $\boldsymbol{R}_{I,d} = \boldsymbol{R}_I(\delta, 0)$ and $\boldsymbol{R}_{I,t} = \boldsymbol{R}_I(\delta, \tau)$. Then the regularized analogs of the inverse correlation matrix are $\boldsymbol{R}_{INV,d} = \boldsymbol{D}_A\boldsymbol{R}_{I,d}\boldsymbol{D}_A$ and $\boldsymbol{R}_{INV,t} = \boldsymbol{D}_A\boldsymbol{R}_{I,t}\boldsymbol{D}_A$. The *slpack* function `rinvrlar` gets the above two matrices.

```
R<- cor(buxx) #no regularization
rinvrlar(R) #same as solve(R) = R^(-1)
$Rinvd
           [,1]       [,2]        [,3]        [,4]
[1,] 1.02906945 0.14379621  0.05564264  0.09389398
[2,] 0.14379621 1.03181920  0.07779758  0.09165646
[3,] 0.05564264 0.07779758  1.01307222 -0.06190635
[4,] 0.09389398 0.09165646 -0.06190635  1.01988077
$Rinvt
           [,1]       [,2]        [,3]        [,4]
[1,] 1.02906945 0.14379621  0.05564264  0.09389398
[2,] 0.14379621 1.03181920  0.07779758  0.09165646
[3,] 0.05564264 0.07779758  1.01307222 -0.06190635
[4,] 0.09389398 0.09165646 -0.06190635  1.01988077
```

If p is large, then matrix inversion should be avoided if possible: the step $\boldsymbol{A} = \boldsymbol{R}_d^{-1}$ has the expensive $O(p^3)$ complexity. See Friedman et al. (2008) and Hsieh et al. (2011).

Example 6.1. Let

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}.$$

Then

$$\boldsymbol{R}_{\delta=1} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix} = \boldsymbol{R}(\delta = 1, \tau = 0.1), \text{ and } \boldsymbol{R}(\delta = 1, \tau = 0.2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that for $\boldsymbol{R}_{\delta=1}$, the nondiagonal (nonunit) elements of $\boldsymbol{R}$ are divided by $1 + \delta = 2$.

6.3 Complements

Note that we can regularize robust covariance and correlation matrices such as the `covmb2` estimator $\boldsymbol{C}$ given by Definition 1.16.

There is a lot of recent work on high dimensional covariance matrix or inverse covariance matrix estimation. See Pourahmadi (2011) for a review. Regularizing $\boldsymbol{S}^{-1} = (S^{ij})$ needs the inverse covariance matrix to exist, or a method to compute the S^{ij} directly. It is also possible to regularize a positive definite analog of $\boldsymbol{S}^{-1}$. The inverse covariance matrix is also known as a precision matrix or concentration matrix. Friedman et al. (2008) provides an interesting method: graphical lasso (Glasso) takes a positive semidefinite (possibly singular) covariance matrix estimator as an input, and returns a positive definite one. Then the resulting estimator of the inverse covariance matrix has many of its elements exactly equal to zero. Also see Hastie et al. (2015, ch. 9). Again the robust `covmb2` estimator could be the input. See Croux and Öllerer (2016), which has some useful R code.

Also see Cai et al. (2011), Hsieh et al. (2011), Huang et al. (2006), Ledoit and Wolf (2004), Liu et al. (2003), Naul and Taylor (2017), Rothman et al. (2008), Schäfer and Strimmer (2007), Yu et al. (2017), and Yuan and Lin (2007). There are R packages for graphical lasso: `glasso` and `huge`. The second package appears to be better. See Croux and Öllerer (2016).

Some topics from multivariate analysis are discussed next. These topics often need a covariance or correlation matrix, possibly regularized. Texts on high dimensional multivariate analysis include Fujikoshi, et al. (2010), Izenman (2008), Koch (2014), Pourahmadi (2013), Rish and Grabarnik (2015), and Yao et al. (2015). Also see Hastie et al. (2015, ch. 7, ch. 8).

For high dimensional clustering, see Jin and Wang (2016).

Discrimination analysis when $p > n$ is interesting. See Cai and Liu (2011), Hand (2006), Mai et al. (2012), and Mai and Zou (2013). See Friedman (1989) for regularized discriminant analysis. Witten and Tibshirani (2011) give a LASSO type FDA method useful for $p > n$. See the R package *penalizedLDA*. Also see Xia (2017).

For high dimensional GLM variable selection, see Guo et al. (2017).

For a high dimensional 1 and 2 sample Hotelling's T^2 type tests, see Hyodo and Nishiyama (2017), Gregory et al. (2015), and Feng and Sun (2015).

Methods like ridge regression and lasso can also be extended to multivariate linear regression. See, for example, Obozinski et al. (2011).

For high dimensional outlier detection see section 1.3 of this text, Aggarwal (2017), Agostinelli et al. (2015), Boudt et al. (2017), Öllerer and Croux (2015), and Ro et al. (2015)

For high dimensional principal component analysis, see Croux et al. (2013), Johnstone and Lu (2009), and Zou et al. (1993). Feng and He (2014) give a method for the singular value decomposition that may be useful for principal component analysis.

6.4 Problems

6.1. Suppose

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.4 & 0.8 \\ 0.4 & 1 & 0.5 \\ 0.8 & 0.5 & 1 \end{bmatrix}.$$

a) Find $\boldsymbol{R}_{\delta=1}$.
b) Find $\boldsymbol{R}(\delta = 1, \tau = 0.3)$.

6.2. Suppose

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.6 & -0.4 \\ 0.6 & 1 & 0.9 \\ -0.4 & 0.9 & 1 \end{bmatrix}.$$

a) Find $\boldsymbol{R}_{\delta=1}$.
b) Find $\boldsymbol{R}(\delta = 1, \tau = 0.3)$.

R Problems

For some of the following problems, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/slrhw.txt) into *R*.