PCP — <LOD Penalty - Boston data

Lizzy Gibson

12/9/2019

Contents

Data

- 20 air pollution species
- 2,478 days (~6.7 years)

```
# Read air pollution data
mixture <- readMat("./Data/mixtures_data.mat")

mixture_data <- as.data.frame(mixture) %>% as_tibble() %>%
  select(Al, As, Ba, bc, Br, Ca, Cl,
         Cr, Cu, Fe, K,  Mn,  Ni,  Pb,  S,  Se,  Si,
         Ti,  V, Zn) %>%
  drop_na()

mixture_data
```

```
## # A tibble: 2,478 x 20
##        Al      As     Ba    bc      Br     Ca     Cl      Cr     Cu     Fe
##     <dbl>   <dbl>  <dbl> <dbl>   <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl>
##  1 0.0194 0       0      0.682 2.00e-3 0.0231 0.0027 5.00e-4 0.0054 0.026
##  2 0.0263 0.00120 0.0206 0.372 1.51e-2 0.0428 0.0239 9.00e-4 0.0056 0.0686
##  3 0.0281 0       0.0099 0.418 0.      0.0332 0.0044 5.00e-4 0.0026 0.0369
##  4 0.0117 0.0005  0.0076 0.361 1.80e-3 0.0167 0.0022 0.      0.0033 0.0211
##  5 0.0102 0.0019  0.0022 0.359 6.00e-4 0.0145 0.0034 8.00e-4 0.003  0.0346
##  6 0.0387 0.0008  0.0256 1.09  0.      0.0261 0.0043 4.00e-4 0.0033 0.103
##  7 0.0217 0.0019  0.0037 0.552 0.      0.0214 0.0018 6.00e-4 0.0037 0.0651
##  8 0.0118 0       0.0026 0.641 8.00e-4 0.0154 0.0022 8.00e-4 0.0035 0.047
##  9 0.0376 0.0015  0.0157 1.01  1.60e-3 0.0326 0.0079 1.50e-3 0.009  0.137
## 10 0.0401 0       0.0295 0.791 4.00e-4 0.0254 0.0019 0.      0.0049 0.0546
## # ... with 2,468 more rows, and 10 more variables: K <dbl>, Mn <dbl>, Ni <dbl>,
## #   Pb <dbl>, S <dbl>, Se <dbl>, Si <dbl>, Ti <dbl>, V <dbl>, Zn <dbl>
```

```
summary(mixture_data)
```

```
##        Al                As                   Ba                  bc
##  Min.   :0.00000   Min.   :0.0000000   Min.   :0.000000   Min.   :0.0539
##  1st Qu.:0.02950   1st Qu.:0.0000000   1st Qu.:0.000700   1st Qu.:0.3802
##  Median :0.04195   Median :0.0003000   Median :0.006600   Median :0.5734
```

```
##  Mean   :0.04803   Mean   :0.0005764   Mean   :0.008541   Mean   :0.6480
##  3rd Qu.:0.05958   3rd Qu.:0.0010000   3rd Qu.:0.013200   3rd Qu.:0.8354
##  Max.   :0.37540   Max.   :0.0073000   Max.   :0.210600   Max.   :2.9620
##        Br                  Ca                 Cl                Cr
##  Min.   :0.0000000   Min.   :0.00040   Min.   :0.0000   Min.   :0.0000000
##  1st Qu.:0.0000000   1st Qu.:0.01953   1st Qu.:0.0013   1st Qu.:0.0001000
##  Median :0.0002000   Median :0.02645   Median :0.0024   Median :0.0004000
##  Mean   :0.0008227   Mean   :0.02887   Mean   :0.0163   Mean   :0.0004283
##  3rd Qu.:0.0013000   3rd Qu.:0.03537   3rd Qu.:0.0051   3rd Qu.:0.0007000
##  Max.   :0.0151000   Max.   :0.11810   Max.   :1.6739   Max.   :0.0040000
##        Cu                  Fe                 K                 Mn
##  Min.   :0.000000   Min.   :0.00250   Min.   :0.00040   Min.   :0.0000000
##  1st Qu.:0.001500   1st Qu.:0.03800   1st Qu.:0.02340   1st Qu.:0.0000000
##  Median :0.003000   Median :0.05370   Median :0.03260   Median :0.0000000
##  Mean   :0.003319   Mean   :0.05925   Mean   :0.04120   Mean   :0.0004234
##  3rd Qu.:0.004700   3rd Qu.:0.07408   3rd Qu.:0.04578   3rd Qu.:0.0003000
##  Max.   :0.099800   Max.   :0.30470   Max.   :3.58110   Max.   :0.0105000
##        Ni                  Pb                 S                 Se
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.0008   Min.   :0.0000000
##  1st Qu.:0.000625   1st Qu.:0.003200   1st Qu.:0.4545   1st Qu.:0.0000000
##  Median :0.001500   Median :0.005000   Median :0.7238   Median :0.0000000
##  Mean   :0.002125   Mean   :0.005378   Mean   :0.9675   Mean   :0.0001374
##  3rd Qu.:0.002775   3rd Qu.:0.007200   3rd Qu.:1.1676   3rd Qu.:0.0000000
##  Max.   :0.022900   Max.   :0.037700   Max.   :6.6640   Max.   :0.0074000
##        Si                 Ti                 V                  Zn
##  Min.   :0.00190   Min.   :0.000000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.03673   1st Qu.:0.001900   1st Qu.:0.000900   1st Qu.:0.00510
##  Median :0.05740   Median :0.002800   Median :0.001900   Median :0.00810
##  Mean   :0.06806   Mean   :0.003292   Mean   :0.002677   Mean   :0.01008
##  3rd Qu.:0.08555   3rd Qu.:0.004300   3rd Qu.:0.003400   3rd Qu.:0.01238
##  Max.   :0.41810   Max.   :0.076600   Max.   :0.051000   Max.   :0.15710
```

Relative error

Overall

Low rank and sparse

Values > and < LOD

Resulting SVD solutions

Loading vectors vs. original data

Loading vectors vs. solution w/ no <LOD

Score vectors vs. original data

Score vectors vs. solution w/ no <LOD