MM Optimization Algorithms

Chathuranga Weeraddana

March 2022

Lecture 6: Some Applications (Part 1)

Applications

- many applications have already been discussed
  - check for previous lectures
- last two lectures: we discuss a few more applications
  - K-mean clustering with missing information
  - Gaussian estimation with missing data
  - regression
  - total variation denoising of images
  - factor analysis
  - matrix completion

K-Mean Clustering [1]

- m subjects
- each subject i is associated with a vector $y \in \mathbb{R}^d$
- subjects must be assigned to one of K clusters
- $\mu_k \in \mathbb{R}^d$, the center of cluster k
- subjects are assigned to clusters based on proximity
- the set of subjects assigned to cluster k is $\mathcal{C}(\mu_k)$

[1] For application examples, see pp. 70-71, 85, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares* by S. Boyd, 2018.

- Loyd algorithm (1957): key idea
    - choose cluster centers μ_ks arbitrarily
    - centers to clusters: compute $\mathcal{C}(\mu_k)$ [2]
    - clusters to centers: $\mu_k =$ centroid of points in $\mathcal{C}(\mu_k)$
    - iterate above two steps
- we can formulate the problem of K-mean clustering as

$$\text{minimize} \quad f(\mu) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}(\mu_k)} \|y_i - \mu_k\|^2$$

- Loyd alg. $\rightarrow$ not necessarily yield the optimal $\mu_k, \mathcal{C}(\mu_k)$

[2] Ties are broken such that $\cap_k \mathcal{C}(\mu_k) = \emptyset$.

WHAT IF y_iS ARE INCOMPLETE?

- e.g., suppose $y_i \in \mathbb{R}^3$
  - $y_1 = (1, 0.5, \_\_)$, $y_2 = (\_\_, \_\_, 3)$, $y_3 = (0.2, 0.4, 2)$, $\ldots$
  - 1st and 2nd indexes of y_1 is observed
  - 3rd index of y_2 is observed
- let $\mathcal{O}_i$ denote the set of indexes observed in subject i
  - $\mathcal{O}_1 = \{1, 2\}$, $\mathcal{O}_2 = \{3\}$, and $\mathcal{O}_3 = \emptyset$ in the example above
- incomplete data destroy the simple two steps of Loyd alg.

Apply MM Principle

- with incomplete y_is, the objective function is given by

$$f(\mu) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}(\mu_k)} \left[\sum_{j \in \mathcal{O}_i} (y_{ij} - \mu_{kj})^2 \right]$$

- we can simply majorize f as

$$\begin{aligned} f(\mu) &\leq \sum_{k=1}^{K} \sum_{i \in \mathcal{C}(\mu_k)} \left[\sum_{j \in \mathcal{O}_i} (y_{ij} - \mu_{kj})^2 + \sum_{j \notin \mathcal{O}_i} (\mu_{kj}^{(n)} - \mu_{kj})^2 \right] \\ &= g(\mu|\mu^{(n)}) \end{aligned}$$

since $\sum_{j \notin \mathcal{O}_i} (\mu_{kj}^{(n)} - \mu_{kj})^2 \geq 0$

- symmetry of the data is restored
- thus the Loyd algorithm can be applied as it is

Gaussian Estimation with Missing Data

- $y_1, \ldots, y_m \in \mathbb{R}^p$ random sample from a Gaussian distribution
- mean of the Gaussian is $\bar{y}$
- covariance matrix is Ω
- ML estimates of $\bar{y}$ and Ω is given by [3]

$$\bar{y}_{\mathtt{ml}} = \frac{1}{m} \sum_{i=1}^{m} y_i \qquad \Omega_{\mathtt{ml}} = \frac{1}{m} \sum_{i=1}^{m} (y_i - y_{\mathtt{ml}})(y_i - y_{\mathtt{ml}})^{\mathsf{T}}$$

[3] It is assumed that $m \geq p$.

What if y_is are Incomplete?

- e.g., suppose $y_i \in \mathbb{R}^3$
  - $y_1 = (1, 0.5, \_\_)$, $y_2 = (\_\_, \_\_, 3)$, $y_3 = (0.2, 0.4, 2)$, $\ldots$
  - 1st and 2nd indexes of y_1 is observed
  - 3rd index of y_2 is observed
- some components are missing in each y_i
- incomplete data destroy the above simple formulas of $\bar{y}_{\mathtt{ml}}, \Omega_{\mathtt{ml}}$

How to Restore the Symmetry?

- we rely on Schur Compliment Majorization [4]
- more specifically
  - we are given a vector x
  - we have a parameter matrix D given by

$$D = \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}$$

- now it is useful to bound [5]

$$1) \quad x^{\mathsf{T}} A^{-1} x \qquad 2) \quad \ln |A|$$

[4] See Example 4.9.7 of the textbook.

[5] Details of the bounds are given in pp. 20-23.

- why the aforementioned bounds are useful?
  - the log-likelihood function is based on similar terms
  - we can find a surrogate using the bounds
  - apply MM principle
- for convenience suppose
  - we have one realization y_1 from the distribution
  - the first block x_1 of components of y_1 is observed
  - the second block z_1 of y_1 is missing

- y is Gaussian, i.e., $y_1 \sim \mathcal{N}\big((\bar{x}, \bar{z}), \Omega\big)$, where

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^{\mathsf{T}} & \Omega_{22} \end{bmatrix}$$

- parameters to be estimated $\theta = (\bar{x}, \bar{z}, \Omega)$

- log-likelihood function of observed data x_1 is given by [6]

$$l(\bar{\theta}) = -\tfrac{1}{2} \ln |\Omega_{11}| - \tfrac{1}{2}(x_1 - \bar{x})^{\mathsf{T}} \Omega_{11}^{-1} (x_1 - \bar{x})$$

  where $\bar{\theta} = (\bar{x}, \Omega_{11})$

- note $\to$ l doesn't contain a part of the parameters, i.e.,
  - Ω_{12}, Ω_{22}, and $\bar{z}$

[6] Up to an irrelevant constant.

- a meaningful mechanism to maximize l and to compute θ?
  - EM principle
  - MM principle
    - based on the bounds pointed in page 10
    - see (1), and (2) in pages 20-23

- in particular from (1), and (2) in pages 20-23, we deduce

$$l(\bar{\theta}) \geq -\tfrac{1}{2}\ln|\Omega| - \tfrac{1}{2}\ln|G^{(n)}| + \tfrac{1}{2}\mathtt{Tr}(I) - \tfrac{1}{2}\mathtt{Tr}\left[\Omega^{-1}F^{(n)}\right]$$
$$- \tfrac{1}{2}\begin{bmatrix} x_1 - \bar{x} \\ z_1^{(n)} - \bar{z} \end{bmatrix}^{\mathsf{T}} \Omega^{-1} \begin{bmatrix} x_1 - \bar{x} \\ z_1^{(n)} - \bar{z} \end{bmatrix}$$
$$= -\tfrac{1}{2}\ln|\Omega| - \tfrac{1}{2}\mathtt{Tr}\left[\Omega^{-1}\left(F^{(n)} + \begin{bmatrix} x_1 - \bar{x} \\ z_1^{(n)} - \bar{z} \end{bmatrix}\begin{bmatrix} x_1 - \bar{x} \\ z_1^{(n)} - \bar{z} \end{bmatrix}^{\mathsf{T}}\right)\right]$$
$$- \frac{1}{2}\ln|G^{(n)}| + \frac{1}{2}\mathtt{Tr}(I)$$

- here

$$z_1^{(n)} = \Omega_{12}^{(n)\mathsf{T}}\left(\Omega_{11}^{(n)}\right)^{-1}\left(x_1 - \bar{x}^{(n)}\right) + \bar{z}^{(n)}$$
$$G^{(n)} = \left[\Omega_{22}^{(n)} - \Omega_{12}^{(n)\mathsf{T}}\left(\Omega_{11}^{(n)}\right)^{-1}\Omega_{12}^{(n)}\right]^{-1}$$
$$F^{(n)} = \begin{bmatrix} 0 & 0 \\ 0 & (G^{(n)})^{-1} \end{bmatrix}$$

- now consider all data
  - m realizations $y_1, \ldots, y_m$ (same observed, missing indexes)
  - let l denote the log-likelihood function
  - a minorization function of l is of the form

$$-\frac{m}{2} \ln |\Omega| - \frac{1}{2} \sum_{i=1}^{m} \mathtt{Tr} \left[\Omega^{-1} \left(F^{(n)} + (y_i^{(n)} - \bar{y})(y_i^{(n)} - \bar{y})^{\mathsf{T}} \right) \right]$$

- ML estimates of $\bar{y}$ and Ω is given by

$$\bar{y}^{(n+1)} = \frac{1}{m} \sum_{i=1}^{m} y_i^{(n)}$$

$$\Omega^{(n+1)} = \frac{1}{m} \sum_{i=1}^{m} \left[F^{(n)} + (y_i - y^{(n+1)})(y_i - y^{(n+1)})^{\mathsf{T}} \right]$$

- if the observed, missing indexes are different for y_is
  - $F^{(n)} \leftarrow F_i^{(n)}$
  - permutation matrices are to be introduced accordingly
  - e.g., to $F_i^{(n)} + (y_i^{(n)} - \bar{y})(y_i^{(n)} - \bar{y})^\mathsf{T}$

Regression

- least squares estimation
  - sum of squared deviation is considered
  - well known: suffers from the distorting influence of outliers
- least absolute deviation regression
  - sum of absolute deviation is considered
  - mitigates the impact of outliers

Least Absolute Deviation Regression

- problem formulation:

$$\underset{\beta}{\text{minimize}} \quad \sum_{i=1}^{m} |y_i - a_i^\mathsf{T}\beta|$$

- Newton's method doesn't apply directly $\rightarrow$ reformulate
- a convex reformulation: [7]

$$\begin{array}{ll} \underset{t_1,\ldots,t_m,\beta}{\text{minimize}} & \sum_{i=1}^{m} \; t_i \\ \text{subject to} & y_i - a_i^\mathsf{T}\beta \le t_i \quad i = 1,\ldots,m \\ & y_i - a_i^\mathsf{T}\beta \ge -t_i \quad i = 1,\ldots,m \end{array}$$

- int.-point method applies (i.e., a sequence of Newton's steps)

[7] We use the epigraph problem form of the original problem, see p. 134, *Convex Optimization* by S. Boyd and L. Vandenberghe, 2004.

Apply MM Principle

- we have the following majorization: [8]

$$\sum_{i=1}^{m} |y_i - a_i^\mathsf{T}\beta| \le \frac{1}{2}\sum_{i=1}^{m} w_{ni}(y_i - a_i^\mathsf{T}\beta)^2 + c_n$$

  where $w_{ni} = 1/|y_i - a_i^\mathsf{T}\beta^{(n)}|$ and c_n is an irrelevant constant

- the majorization function is quadratic
  - favorable for large scale problems
  - caveat:
    - w_{ni} can be zero
    - let $w_{ni} = 1/\sqrt{|y_i - a_i^\mathsf{T}\beta^{(n)}|^2 + \epsilon}$

[8] See Homework 1 → Problem 1 → Part 1.

Appendix

Bounding $(x - \bar{x})^\intercal A^{-1}(x - \bar{x})$

- we have

$$
\begin{aligned}
f(x) &= \inf_z \; g(x,z) \\
&= \inf_x \; \begin{bmatrix} x - \bar{x} \\ z - \bar{z} \end{bmatrix}^\mathsf{T} \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix}^{-1} \begin{bmatrix} x - \bar{x} \\ z - \bar{z} \end{bmatrix} \\
&= (x - \bar{x})^\mathsf{T} A^{-1} (x - \bar{x}) \\
&= \begin{bmatrix} x - \bar{x} \\ B^\mathsf{T} A^{-1}(x - \bar{x}) \end{bmatrix}^\mathsf{T} \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix}^{-1} \begin{bmatrix} x - \bar{x} \\ B^\mathsf{T} A^{-1}(x - \bar{x}) \end{bmatrix}
\end{aligned}
$$

- from the last two equations we get for all $A, B, C, \bar{x}$, and $\bar{z}$

$$
\begin{aligned}
(x-\bar{x})^{\mathsf{T}} A^{-1}(x-\bar{x}) &= \begin{bmatrix} x-\bar{x} \\ B^{\mathsf{T}} A^{-1}(x-\bar{x}) \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}^{-1} \begin{bmatrix} x-\bar{x} \\ B^{\mathsf{T}} A^{-1}(x-\bar{x}) \end{bmatrix} \\
&\leq \begin{bmatrix} x-\bar{x} \\ z^{(n)}-\bar{z} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}^{-1} \begin{bmatrix} x-\bar{x} \\ z^{(n)}-\bar{z} \end{bmatrix} \qquad (1)
\end{aligned}
$$

with $z^{(n)} = B^{(n)\mathsf{T}} \left(A^{(n)}\right)^{-1} \left(x-\bar{x}^{(n)}\right) + \bar{z}^{(n)}$

- note that the inequality holds with equality when

$$
A = A^{(n)},\ B = B^{(n)},\ C = C^{(n)},\ \bar{x} = \bar{x}^{(n)},\ \bar{z} = \bar{z}^{(n)}
$$

Bounding $\ln|A|$

- the Schur complement of A in the matrix D is given by

$$C - B^{\mathsf{T}} A^{-1} B$$

- for clarity let us define G as

$$G = (C - B^{\mathsf{T}} A^{-1} B)^{-1}$$

- we have the following determinant identity

$$|D| = |A| \times |C - B^{\mathsf{T}} A^{-1} B| = |A|/|G|$$

- now we can bound $\ln|A|$ as follows:

$$\begin{aligned}\ln|A| &= \ln|D| + \ln|G| \\ &\le \ln|D| + \ln|G^{(n)}| + \mathtt{Tr}\big[(G^{(n)})^{-1}(G - G^{(n)})\big] \\ &= \ln|D| + \ln|G^{(n)}| - \mathtt{Tr}(I) + \mathtt{Tr}\big[(G^{(n)})^{-1}G\big] \\ &= \ln|D| + \ln|G^{(n)}| - \mathtt{Tr}(I) + \mathtt{Tr}\big[D^{-1}F^{(n)}\big] \qquad (2)\end{aligned}$$

- the last equality follows from that $F^{(n)} \triangleq \begin{bmatrix} 0 & 0 \\ 0 & (G^{(n)})^{-1} \end{bmatrix}$ and

$$D^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BGB^{\mathsf{T}}A^{-1} & -A^{-1}BG \\ -GB^{\mathsf{T}}A^{-1} & G \end{bmatrix}$$

- note that the inequality holds with equality when

$$A = A^{(n)},\ B = B^{(n)},\ C = C^{(n)}$$