MM Optimization Algorithms

Chathuranga Weeraddana

March 2022

Lecture 7: Some Applications (Part 2)

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

Image Denoising

- an $m \times m$ distorted image Y is given
- prior information:
  - original image $X \in \mathbb{R}^{m \times m}$ is usually smooth
  - neighboring pixels values are not very different
  - boundaries of distinct color changes exist
- least-squares:
  - no accounts for neighboring pixels conditions
  - exhibits ringing phenomenon

- total variation denoising
  - accounts for neighboring pixels conditions
  - mitigates the ringing phenomenon

Total Variation Denoising

- problem formulation:

$$\underset{X}{\text{minimize}} \quad \frac{1}{2}\|X-Y\|^2+\lambda\sum_i\sum_j\sqrt{(X_{i,j}-X_{i,j+1})^2+(X_{i,j}-X_{i+1,j})^2}$$

- Newton's method doesn't apply directly $\rightarrow$ reformulate
- a convex reformulation:
  - second-order cone program (SOCP) [1]
- int.-point method applies to the reformulated problem

[1] See §. 4.4.2, *Convex Optimization* by S. Boyd and L. Vandenberghe, 2004.

Apply MM Principle

- we have the following majorization function of the objective: [2]

$$\frac{1}{2}\|X-Y\|^2 + \frac{\lambda}{2}\sum_{i=1}^{m} w_{nij}\left[(X_{i,j}-X_{i,j+1})^2+(X_{i,j}-X_{i+1,j})^2\right]+c_n$$

where c_n is an irrelevant constant and

$$w_{nij} = \frac{1}{\sqrt{\left(X_{i,j}^{(n)}-X_{i,j+1}^{(n)}\right)^2+\left(X_{i,j}^{(n)}-X_{i+1,j}^{(n)}\right)^2+\epsilon}}$$

- the majorization function is quadratic
  - favorable for large scale problems
  - e.g., Landweber's method is applied (see Lecture 3, pp. 9-11)

[2] See Homework 1 → Problem 1 → Part 3.

Factor Analysis

- $y_1, \ldots, y_m \in \mathbb{R}^p$ random samples
- suppose $m \ll p$
  - standard Gaussian model cannot be fitted
  - cannot be modeled even with a single Gaussian
  - ML of the covariance matrix become singular [3]
- factor analysis
  - is a model that capture some of the correlations of data
  - doesn't run into the problem of singular covariance

[3] There are other fixes, e.g., constrain the covariance matrix to be diagonal. Usually those impositions are related to invalid assumptions.

Observation Model

- m independent observations are of the form

$$y_k = \mu + F z_k + u_k \tag{1}$$

- $F \in \mathbb{R}^{p \times q}$: factor loading matrix, typically $q \ll p$
- $z_k \in \mathbb{R}^q$ latent variables
- $u_k \in \mathbb{R}^p$ measurement errors
- z_k and u_k are independent and Gaussian with

$$\mathbb{E}\{z_k\} = 0 \qquad \texttt{Var}\{z_k\} = I$$
$$\mathbb{E}\{u_k\} = 0 \qquad \texttt{Var}\{u_k\} = D$$

where D is a diagonal matrix

- (y_k, z_k) is Gaussian, i.e., $(y_k, z_k) \sim \mathcal{N}\big((\mu, 0), \Omega\big)$, where

$$\Omega = \begin{bmatrix} FF^\mathsf{T} + D & F \\ F^\mathsf{T} & I \end{bmatrix} = \begin{bmatrix} D^{1/2} & F \\ 0 & I \end{bmatrix} \begin{bmatrix} D^{1/2} & 0 \\ F^\mathsf{T} & I \end{bmatrix}$$

- parameters to be estimated $\theta = (\mu, F, D)$

- w.l.g., we assume $\mu = 0$, i.e., $\theta = (F, D)$?

- log-likelihood function of observed data y_k is given by [4]

$$l(\theta) = -\tfrac{1}{2} \ln |FF^\mathsf{T} + D| - \tfrac{1}{2} y_k^\mathsf{T} (FF^\mathsf{T} + D)^{-1} y_k$$

- l is not convex in $F, D \rightarrow$ alternating optimization applies

- now the idea is to find a minorization function to l

[4] Up to an irrelevant constant.

- a meaningful mechanism to maximize l and to compute θ?
  - EM principle
  - MM principle, based on the bounds on
    - $\ln|FF^{\mathsf{T}} + D|$
    - $y_k^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}y_k$

Bounding $\ln |FF^{\mathrm{T}} + D|$

- Schur complement of $(FF^{\mathsf{T}} + D)$ in the matrix Ω is given by

$$I - F^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}F$$

- for clarity let us define G as

$$\begin{aligned} G &= (I - F^{\mathsf{T}}(FF^{\mathsf{T}} + D)^{-1}F)^{-1} \\ &= I + F^{\mathsf{T}}D^{-1}F \end{aligned}$$

- last equality $\rightarrow$ classic Woodbury matrix identity

- we can bound $\ln|FF^{\mathsf{T}}+D|$ as follows:

$$\begin{aligned}\ln|FF^{\mathsf{T}}+D| &= \ln|\Omega| + \ln|G| \\ &\leq \ln|\Omega| + \ln|G^{(n)}| + \mathtt{Tr}\big[(G^{(n)})^{-1}(G-G^{(n)})\big] \\ &= \ln|\Omega| + \ln|G^{(n)}| - \mathtt{Tr}(I) + \mathtt{Tr}\big[(G^{(n)})^{-1}G\big] \\ &= \ln|\Omega| + \ln|G^{(n)}| - \mathtt{Tr}(I) + \mathtt{Tr}\big[\Omega^{-1}H^{(n)}\big] \\ &= \ln|D| + \mathtt{Tr}\big[F^{\mathsf{T}}D^{-1}F(G^{(n)})^{-1}\big] + r_n\end{aligned}$$

- where $r_n = \mathtt{Tr}(G^{(n)})^{-1} + \ln|G^{(n)}| - \mathtt{Tr}(I)$

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

$$\Omega^{-1} = \begin{bmatrix} D^{-1} & -D^{-1}F \\ -F^\mathsf{T} D^{-1} & I + F^\mathsf{T} D^{-1} F \end{bmatrix}, \quad \text{and} \quad \ln|\Omega| = \ln|D|$$

- note that the inequality holds with equality when

$$F = F^{(n)},\ D = D^{(n)}$$

Bounding $y_k^T(FF^T + D)^{-1} y_k$

- from the partial minimization result, for all F and D

$$
\begin{aligned}
y_k^\mathsf{T}(FF^\mathsf{T}+D)^{-1}y_k &= \begin{bmatrix} y_k \\ F^\mathsf{T}(FF^\mathsf{T}+D)^{-1}y_k \end{bmatrix}^\mathsf{T} \Omega^{-1} \begin{bmatrix} y_k \\ F^\mathsf{T}(FF^\mathsf{T}+D)^{-1}y_k \end{bmatrix} \\
&\le \begin{bmatrix} y_k \\ z_k^{(n)} \end{bmatrix}^\mathsf{T} \begin{bmatrix} D^{1/2} & 0 \\ F^\mathsf{T} & I \end{bmatrix}^{-1} \begin{bmatrix} D^{1/2} & F \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} y_k \\ z_k^{(n)} \end{bmatrix} \\
&= \left\| \begin{bmatrix} D^{-1/2} & -D^{-1/2}F \\ 0 & I \end{bmatrix} \begin{bmatrix} y_k \\ z_k^{(n)} \end{bmatrix} \right\|^2 \\
&= \|D^{-1/2}y_k - D^{-1/2}Fz_k^{(n)}\|^2 + s_n \\
&= \left(y_k - Fz_k^{(n)}\right)^\mathsf{T} D^{-1} \left(y_k - Fz_k^{(n)}\right) + s_n
\end{aligned}
$$

with $z_k^{(n)} = F^{(n)\mathsf{T}}\left(F^{(n)}F^{(n)\mathsf{T}}+D^{(n)}\right)^{-1}y_k$ and $s_n =$ constant

- note that the inequality holds with equality when

$$F = F^{(n)},\ D = D^{(n)}$$

- now consider all data
  - m realizations $y_1, \ldots, y_m$
  - let l denote the log-likelihood function
  - a minorization function of l is of the form (up to a constant)

$$-\frac{m}{2}\left[\ln|D| + \mathtt{Tr}\left[D^{-1}F(G^{(n)})^{-1}F^\mathsf{T}\right]\right] - \frac{1}{2}\sum_{i=1}^{m}\left(y_k - Fz_k^{(n)}\right)^\mathsf{T} D^{-1}\left(y_k - Fz_k^{(n)}\right)$$

- we need to find D and F that maximize the above function

- maximizing w.r.t. F for fixed $D = D^{(n)}$
  - the minorization function is quadratic with respect to F
  - compute the gradient $\rightarrow$ make it zero to yield

$$F^{(n+1)} = \left[\sum_{k=1}^{m} y_k z_k^{(n)\mathsf{T}}\right] \left[m\left(G^{(n)}\right)^{-1} + \sum_{k=1}^{m} z_k^{(n)} z_k^{(n)\mathsf{T}}\right]^{-1}$$

- here we use the fact that

$$\nabla_X \mathrm{Tr}[BXCX^\mathsf{T}] = BXC + B^\mathsf{T}XC^\mathsf{T}$$

and

$$\nabla_X \mathrm{Tr}[BX^\mathsf{T}] = B$$

- maximizing w.r.t. D for fixed $F = F^{(n+1)}$
  - perform the usual variable transformation $D = E^{-1}$
  - the resulting function is cocave in E
  - compute the gradient $\rightarrow$
    - make it zero to yield a non-diagonal matrix $\hat{D}$
    - pick only the diagonals of $\hat{D}$ to compute D
- here we use the fact that

$$\nabla_X \ln |X| = X^{-1}$$

and

$$\nabla_X \mathrm{Tr}[XA] = A^{\mathsf{T}}$$

- in particular, we get

$$d_{ii}^{(n+1)} = \Big[F^{(n+1)} (G^{(n)})^{-1} F^{(n+1)\mathsf{T}} + \tfrac{1}{m} \textstyle\sum_{k=1}^{m} \big(y_k - F^{(n+1)} z_k^{(n)}\big)\big(y_k - F^{(n+1)} z_k^{(n)}\big)^{\mathsf{T}} \Big]_{ii}$$

and $d_{ij}^{(n+1)} = 0$ for all $i \neq j$, where $d_{ij} = [D]_{ij}$ for all i, j