CPSC 540: Machine Learning

Mixture Models

Mark Schmidt

University of British Columbia

Winter 2018

Last Time: Multivariate Gaussian

- The multivariate normal/Gaussian distribution models PDF of vector x^i as

$$p(x^i|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)\right)$$

  where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ and $\Sigma \succ 0$.
- Last time with showed there is a closed-form MLE for μ:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x^i.$$

- We'll now show the analogous result for MLE of the variance:

$$\Sigma = \frac{1}{n}\sum_{i=1}^{N}\underbrace{(x^i-\mu)(x^i-\mu)^T}_{d\times d}.$$

- So MLE is closed-form and given by sample mean and sample variance.

Maximum Likelihood Estimation in Multivariate Gaussians

- To get MLE for Σ we re-parameterize in terms of precision matrix $\Theta = \Sigma^{-1}$,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^T\Sigma^{-1}(x^i-\mu)+\frac{n}{2}\log|\Sigma| \\
=&\frac{1}{2}\sum_{i=1}^{n}(x^i-\mu)^T\Theta(x^i-\mu)+\frac{n}{2}\log|\Theta^{-1}| && \text{(ok because } \Sigma \text{ is invertible)} \\
=&\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left((x^i-\mu)^T\Theta(x^i-\mu)\right)+\frac{n}{2}\log|\Theta|^{-1} && (y^TAy=\mathrm{Tr}(y^TAy)) \\
=&\frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}((x^i-\mu)(x^i-\mu)^T\Theta)-\frac{n}{2}\log|\Theta| && (\mathrm{Tr}(ABC)=\mathrm{Tr}(CAB))
\end{aligned}$$

- Where the trace $\mathrm{Tr}(A)$ is the sum of the diagonal elements of A.
  - That $\mathrm{Tr}(ABC)$ =$\mathrm{Tr}(CAB)$ when dimensions match is the cyclic property of trace.

Maximum Likelihood Estimation in Multivariate Gaussians

- So in terms of precision matrix Θ we have

$$= \frac{1}{2}\sum_{i=1}^{n} \mathrm{Tr}((x^i-\mu)(x^i-\mu)^T\Theta) - \frac{n}{2}\log|\Theta|$$

- We can exchange the sum and trace (trace is a linear operator) to get,

$$= \frac{1}{2}\mathrm{Tr}\left(\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^T\Theta\right) - \frac{n}{2}\log|\Theta| \qquad \sum_i \mathrm{Tr}(A_iB) = \mathrm{Tr}\left(\sum_i A_iB\right)$$

$$= \frac{n}{2}\mathrm{Tr}\left(\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^T}_{\text{sample covariance } `S'}\right)\Theta\right) - \frac{n}{2}\log|\Theta|. \qquad \left(\sum_i A_iB\right) = \left(\sum_i A_i\right)B$$

Maximum Likelihood Estimation in Multivariate Gaussians

- So the NLL in terms of the precision matrix Θ and sample covariance S is

$$f(\Theta) = \frac{n}{2}\text{Tr}(S\Theta) - \frac{n}{2}\log|\Theta|, \text{ with } S = \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^T$$

- Weird-looking but has nice properties:
  - $\text{Tr}(S\Theta)$ is linear function of Θ, with $\nabla_\Theta \text{Tr}(S\Theta) = S$.
    (it's the matrix version of an inner-product $s^T\theta$)
  - Negative log-determinant is strictly-convex and has $\nabla_\Theta \log|\Theta| = \Theta^{-1}$.
    (generalizes $\nabla \log|x| = 1/x$ for for $x > 0$).
- Using these two properties the gradient matrix has a simple form:

$$\nabla f(\Theta) = \frac{n}{2}S - \frac{n}{2}\Theta^{-1}.$$

Maximum Likelihood Estimation in Multivariate Gaussians

- Gradient matrix of NLL with respect to Θ is

$$\nabla f(\Theta) = \frac{n}{2}S - \frac{n}{2}\Theta^{-1}.$$

- The MLE for a given μ is obtained by setting gradient matrix to zero, giving

$$\Theta = S^{-1} \quad \text{or} \quad \Sigma = S = \frac{1}{n}\sum_{i=1}^{n}(x^i - \mu)(x^i - \mu)^T.$$

- The constraint $\Sigma \succ 0$ means we need positive-definite sample covariance, $S \succ 0$.
  - If S is not invertible, NLL is unbounded below and no MLE exists.
  - This is like requiring "not all values are the same" in univariate Gaussian.

- For most distributions, the MLEs are not the sample mean and covariance.

MAP Estimation in Multivariate Gaussian

- We typically don't regularize μ, but you could add an L2-regularizer $\frac{\lambda}{2}\|\mu\|^2$.
- A classic regularizer for Σ is to add a diagonal matrix to S and use

$$\Sigma = S + \lambda I,$$

which satisfies $\Sigma \succ 0$ by construction (eigenvalues at least λ).

- This corresponds to a regularizer that penalizes diagonal of the precision,

$$\begin{aligned} f(\Theta) &= \mathrm{Tr}(S\Theta) - \log|\Theta| + \lambda \mathrm{Tr}(\Theta) \\ &= \mathrm{Tr}(S\Theta + \lambda\Theta) - \log|\Theta| \\ &= \mathrm{Tr}((S + \lambda I)\Theta) - \log|\Theta|. \end{aligned}$$

- L1-regularization of diagonals of inverse covariance.
  - But doesn't set to exactly zero as it must be positive-definite.

Graphical LASSO

- Recent substantial interest in a generalization called the graphical LASSO,

$$f(\Theta) = \text{Tr}(S\Theta) - \log|\Theta| + \lambda\|\Theta\|_1.$$

  where we are using the element-wise L1-norm.

- Gives sparse off-diagonals in Θ.
  - Can solve very large instances with proximal-Newton and other tricks ("QUIC").

- It's common to draw the non-zeroes in Θ as a graph.
  - Has an interpretation in terms on conditional independence (we'll cover this later).
  - Examples: https://normaldeviate.wordpress.com/2012/09/17/high-dimensional-undirected-graphical-models

Closedness of Multivariate Gaussian

- Multivariate Gaussian has nice properties of univariate Gaussian:
  - Closed-form MLE for μ and Σ given by sample mean/variance.
  - Central limit theorem: mean estimates of random variables converge to Gaussians.
  - Maximizes entropy subject to fitting mean and covariance of data.

- A crucial computation property: Gaussians are closed under many operations.
  1. Affine transformation: if $p(x)$ is Gaussian, then $p(Ax + b)$ is a Gaussian[1].
  2. Marginalization: if $p(x, z)$ is Gaussian, then $p(x)$ is Gaussian.
  3. Conditioning: if $p(x, z)$ is Gaussian, then $p(x|z)$ is Gaussian.
  4. Product: if $p(x)$ and $p(z)$ are Gaussian, then $p(x)p(z)$ is proportional to a Gaussian.

- Most continuous distributions don't have these nice properties.

[1] Could be degenerate with $|\Sigma| = 0$ dependending on A.

Affine Property: Special Case of Shift

- Assume that random variable x follows a Gaussian distribution,

$$x \sim \mathcal{N}(\mu, \Sigma).$$

- And consider an shift of the random variable,

$$z = x + b.$$

- Then random variable z follows a Gaussian distribution

$$z \sim \mathcal{N}(\mu + b, \Sigma),$$

where we've shifted the mean.

Affine Property: General Case

- Assume that random variable x follows a Gaussian distribution,

$$x \sim \mathcal{N}(\mu, \Sigma).$$

- And consider an affine transformation of the random variable,

$$z = Ax + b.$$

- Then random variable z follows a Gaussian distribution

$$z \sim \mathcal{N}(A\mu + b, A\Sigma A^T),$$

although note we might have $|A\Sigma A^T| = 0$.

Marginalization of Gaussians

- Consider partitioning multivariate Gaussian variables into two sets,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix} \right),$$

  so our dataset would be something like

$$X = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & z_1 & z_2 \\ | & | & | & | \end{bmatrix}.$$

- If I want the marginal distribution $p(x)$, I can use the affine property,

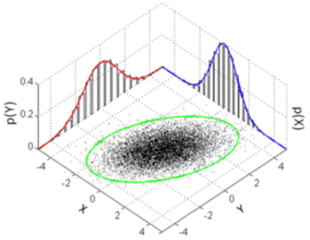
$$x = \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{A} \begin{bmatrix} x \\ z \end{bmatrix} + \underbrace{0}_{b},$$

  to get that

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}).$$

Marginalization of Gaussians

- In a picture, ignoring a subset of the variables gives a Gaussian:



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

- This seems less intuitive if you use usual marginalization rule:

$$p(x) = \int_{z_1} \int_{z_2} \cdots \int_{z_d} \frac{1}{(2\pi)^{\frac{d}{2}} \left| \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix} \right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \left(\begin{bmatrix} x \\ z \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix} \right) \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}^{-1} \left(\begin{bmatrix} x \\ z \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix} \right) \right) dz_d dz_{d-1} \ldots dz_1.$$

Conditioning in Gaussians

- Consider partitioning multivariate Gaussian variables into two sets,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right).$$

- The conditional probabilities are also Gaussian,

$$x \mid z \sim \mathcal{N}(\mu_{x|z}, \Sigma_{x|z}),$$

  where

$$\mu_{x|z} = \mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(z - \mu_z), \quad \Sigma_{x|z} = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}.$$

- "For any fixed z, the distribution of x is a Gaussian".

- For a careful discussion of Gaussians, see the playlist here:
  - https://www.youtube.com/watch?v=TC0ZAX3DA88&t=2s&list=PL17567A1A3F5DB5E4&index=34

Product of Gaussian Densities

- Let $f_1(x)$ and $f_2(x)$ be Gaussian PDFs defined on variables x.
  - Let (μ_1, Σ_1) be parameters of f_1 and (μ_2, Σ_2) for f_2.

- The product of the PDFs $f_1(x)f_2(x)$ is proportional to a Gaussian density,

$$\text{covariance of} \quad \Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}.$$

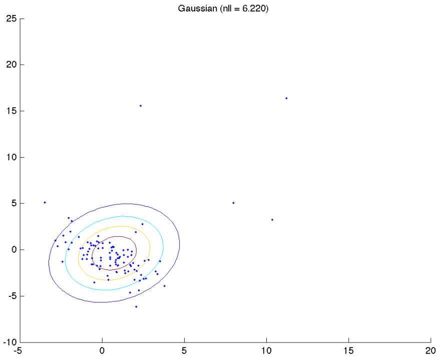
$$\text{mean of } \mu = \Sigma\Sigma_1^{-1}\mu_1 + \Sigma\Sigma_2^{-1}\mu_2,$$

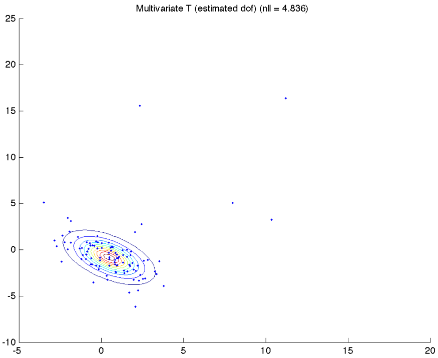
  although this density may not be normalized (may not integrate to 1 over all x).

- But if we can write $p(x) \propto f_1(x)f_2(x)$ then this density must be normalized, so x is Gaussian with the above mean/covariance.
  - Special case: if $\Sigma_1 = I$ and $\Sigma_2 = I$ then $\mu = \frac{\mu_1+\mu_2}{2}$ and $\Sigma = \frac{1}{2}I$.

Problems with Multivariate Gaussian

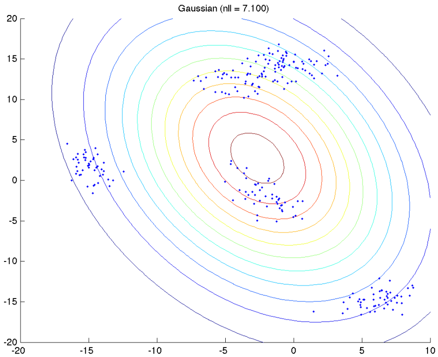
- Why not the multivariate Gaussian distribution?
  - Still not robust, may want to consider multivariate Laplace or multivariate T.
    - These require numerical optimization to compute MLE/MAP.





Problems with Multivariate Gaussian

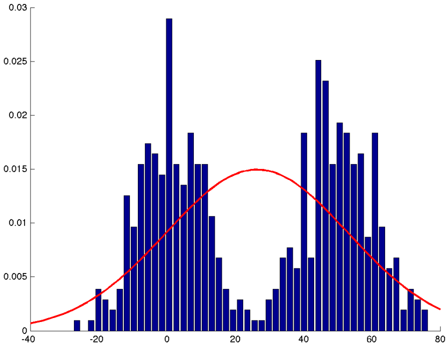
- Why not the multivariate Gaussian distribution?
  - Still not robust, may want to consider multivariate Laplace of multivariate T.
  - Still unimodal, which often leads to very poor fit.



Outline

1 Gaussian for Multi-Modal Data

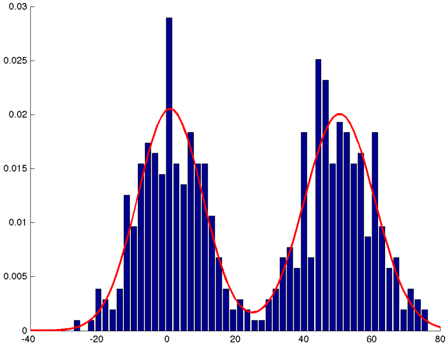
- Major drawback of Gaussian is that it's uni-modal.
  - It gives a terrible fit to data like this:



- If Gaussians are all we know, how can we fit this data?

2 Gaussians for Multi-Modal Data

- We can fit this data by using two Gaussians



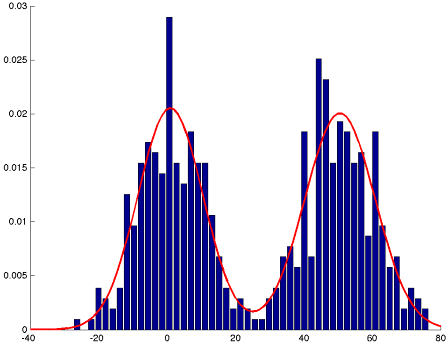
- Half the samples are from Gaussian 1, half are from Gaussian 2.

Mixture of Gaussians

- Our probability density in this example is given by

$$p(x^i \mid \mu_1, \mu_2, \Sigma_1, \Sigma_2) = \frac{1}{2} \underbrace{p(x^i \mid \mu_1, \Sigma_1)}_{\text{PDF of Gaussian 1}} + \frac{1}{2} \underbrace{p(x^i \mid \mu_2, \Sigma_2)}_{\text{PDF of Gaussian 2}},$$

  - We need the $(1/2)$ factors so it still integrates to 1.

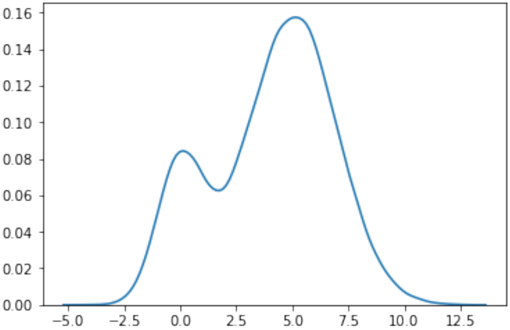


Mixture of Gaussians

- If data comes from one Gaussian more often than the other, we could use

$$p(x^i \mid \mu_1, \mu_2, \Sigma_1, \Sigma_2, \pi_1, \pi_2) = \pi_1 \underbrace{p(x^i \mid \mu_1, \Sigma_1)}_{\text{PDF of Gaussian 1}} + \pi_2 \underbrace{p(x^i \mid \mu_2, \Sigma_2)}_{\text{PDF of Gaussian 2}},$$

where π_1 and π_2 and are non-negative and sum to 1.

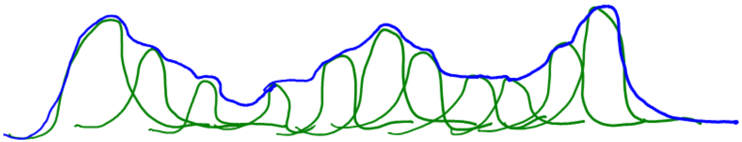


Mixture of Gaussians

- In general we might have mixture k Gaussians with different weights.

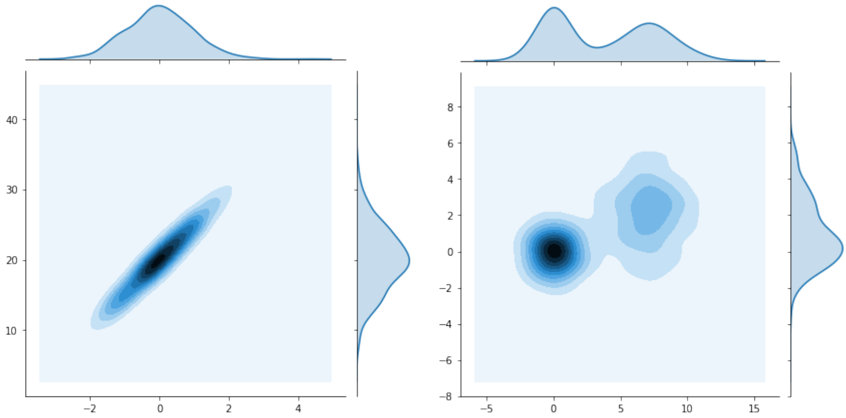
$$p(x \mid \mu, \Sigma, \pi) = \sum_{c=1}^{k} \pi_c \underbrace{p(x \mid \mu_c, \Sigma_c)}_{\text{PDF of Gaussian } c},$$

  - Where the π_c are non-negative and sum to 1.
  - We can use it to model complicated densities with Gaussians (like RBFs).
    - "Universal approximator": can model any continuous density on compact set.



Mixture of Gaussians

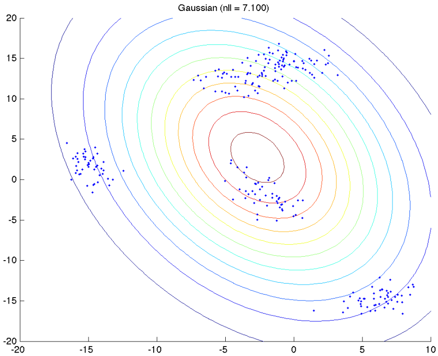
- Gaussian vs. mixture of 2 Gaussian densities in 2D:

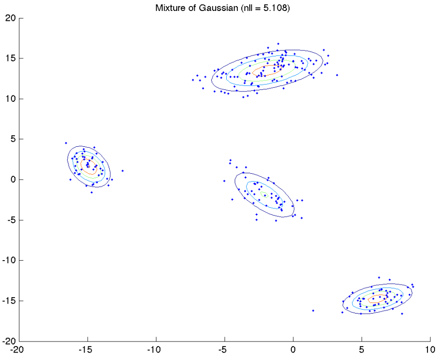


- Marginals will also be mixtures of Gaussians.

Mixture of Gaussians

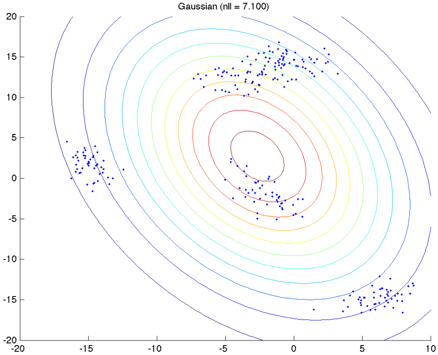
- Gaussian vs. Mixture of 4 Gaussians for 2D multi-modal data:

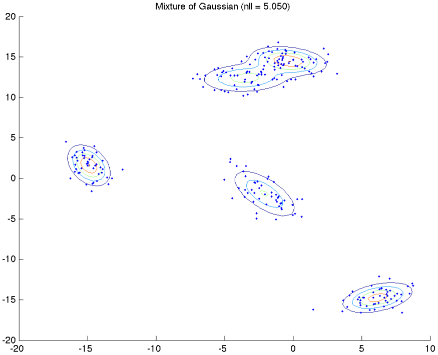




Mixture of Gaussians

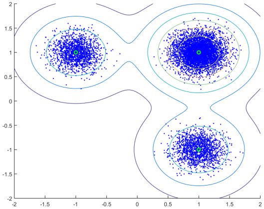
- Gaussian vs. Mixture of 5 Gaussians for 2D multi-modal data:





Mixture of Gaussians

- How a mixture of Gaussian "generates" data:
  1. Sample cluster c based on prior probabilities π_c (categorical distribution).
  2. Sample example x based on mean μ_c and covariance Σ_c.



- We usually fit these models with expectation maximization (EM):
  - EM is a general method for fitting models with hidden variables.
  - For mixture of Gaussians: we treat cluster c as a hidden variable.

Summary

- Multivariate Gaussian generalizes univariate Gaussian for multiple variables.
  - Closed-form MLE given by sample mean and covariance.
  - Closed under affine transformations, marginalization, conditioning, and products.
  - But unimodal and not robust.
- Mixture of Gaussians writes probability as convex comb. of Gaussian densities.
  - Can model arbitrary continuous densities.
- Next time: dealing with missing data.

Positive-Definiteness of Θ and Checking Positive-Definiteness

- If we define centered vectors $\tilde{x}^i = x^i - \mu$ then empirical covariance is

$$S = \frac{1}{n}\sum_{i=1}^{n}(x^i-\mu)(x^i-\mu)^T = \sum_{i=1}^{n}\tilde{x}^i(\tilde{x}^i)^T = \tilde{X}^T\tilde{X} \succeq 0,$$

  so S is positive semi-definite but not positive-definite by construction.
- If data has noise, it will be positive-definite with n large enough.
- For $\Theta \succ 0$, note that for an upper-triangular T we have

$$\log|T| = \log(\text{prod}(\text{eig}(T))) = \log(\text{prod}(\text{diag}(T))) = \text{Tr}(\log(\text{diag}(T))),$$

  where we've used Matlab notation.
- So to compute $\log|\Theta|$ for $\Theta \succ 0$, use Cholesky to turn into upper-triangular.
  - Bonus: Cholesky will fail if $\Theta \succ 0$ is not true, so it checks constraint.