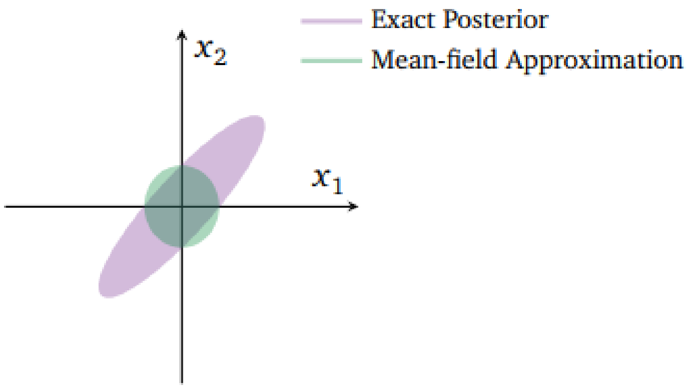
Gaussian Mixture Model + Variational Inference

3/21/2019 4:14 PM
Leo173701

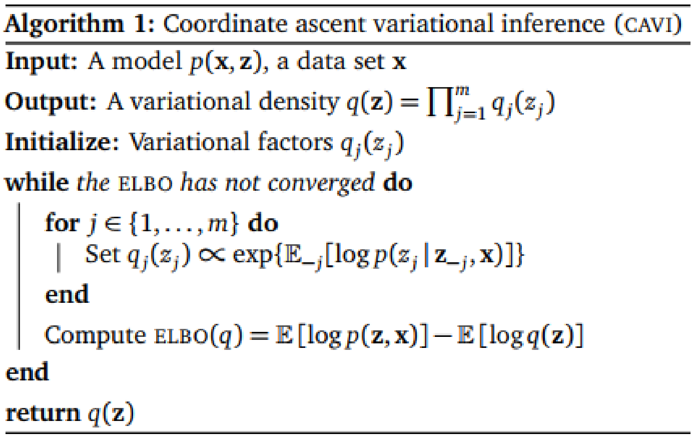
2.3 Visualizing the mean-field approximation.



2.4 Coordinate ascent mean-field variational inference algorithm

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})\right]\right\}. \tag{17}$$

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})\right]\right\}. \tag{18}$$



```
Algorithm 1: Coordinate ascent variational inference (CAVI)
Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$
Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while the ELBO has not converged do
    for $j \in \{1, \dots, m\}$ do
        Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$
    end
    Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$
```

$$\text{ELBO}(q_j) = \mathbb{E}_j\left[\mathbb{E}_{-j}\left[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})\right]\right] - \mathbb{E}_j\left[\log q_j(z_j)\right] + \text{const.} \tag{19}$$

3. A complete example: Bayesian mixture Gaussian

K: quantity of clustering

n: quantity of data

$\mathbf{c} = c_{1:n}$, where c_i is an indicator K-vector , $c_4 = [0,0,0,1]$

M: quantity of dimensions, and it's a scalar

$\boldsymbol{\mu} = \boldsymbol{\mu}_{1:K}$, real-valued mean parameters, $\boldsymbol{\mu}_k = \left(\mathbf{m}_k , \mathbf{s}_k^2\right) = (\mathbf{m}_k, \mathbf{V}_k)$, k=1,2,3,…,K

where $\mathbf{m}_k$ has a shape of $1 \times$ M, and $\mathbf{V}_k$ has a shape of $\mathbf{M} \times \mathbf{M}$

$\boldsymbol{X} = (\boldsymbol{x}_i)$, $\mathbf{X}$ is the dataset with a shape of n × M, $i = 1,2,3, \ldots, n$

$\boldsymbol{x}_i$ is the data

3.1 Prior distributions:

$$p(\mu) = \text{Normal}\left(\mu|0, \sigma^2\right) = \frac{1}{\sqrt{2\pi\ \sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$$

$$p(c) = \frac{1}{K}, \quad \text{in details:} \quad P(c_{ik}) = \frac{1}{K}, \ \ for\ k = 1,2,3, \ldots, K$$

3.2 Observations:

$$p(x_i \,|\, c_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(x_i \,|\, \mu_k)^{c_{ik}}$$

$$p(\boldsymbol{X}|\boldsymbol{c}, \boldsymbol{\mu}) = \prod_{i=1}^{n} \prod_{k=1}^{K} p\left(x_i|\mu_k\right)^{c_{ik}}$$

3.3 Joint distribution:

$$p(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{\mu}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} \prod_{k=1}^{K} p\left(x_i|\ \mu_k\right)^{c_{ik}}$$

In a more clear writing,

$$p(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{\mu}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i|c_i, \boldsymbol{\mu})$$

3.4 From Bayesian formula we know the posterior distribution:

$$p(\boldsymbol{c}, \boldsymbol{\mu}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{\mu})}{p(\boldsymbol{X})}$$

4. variance inference

4.1 $\quad q(\boldsymbol{c}; \varphi_{ik}) q(\boldsymbol{\mu}; \boldsymbol{m}_k, \boldsymbol{S}_k) \Rightarrow p(\boldsymbol{c}, \boldsymbol{\mu}|\boldsymbol{X})$

Now it's time to update the posteriors

4.2 The variational density of the mixture assignments

Mean-field VI for $q(\boldsymbol{c}; \varphi_{ik})$

$$q^*(c_i; \varphi_i) \propto \exp\left\{\log p(c_i) + \mathbb{E}\left[\log p(x_i \,|\, c_i, \mu); \mathbf{m}, \mathbf{s}^2\right]\right\}. \qquad (22)$$

4.2.1

$\exp(\log p(c_i)) = \exp\left(\log \frac{1}{K}\right) = \frac{1}{\mathrm{K}}$ is a constant and independent of $\mathrm{c_i}$, $\boldsymbol{\mu}$, $\mathbf{s}^2$, and thus can be ignored tempo

4.2.2 We use this to compute the expected log probability,

$$\mathbb{E}[\log p(x_i \mid c_i, \boldsymbol{\mu})] = \sum_k c_{ik} \mathbb{E}\left[\log p(x_i \mid \mu_k); m_k, s_k^2\right] \quad (23)$$

$$= \sum_k c_{ik} \mathbb{E}\left[-(x_i - \mu_k)^2/2; m_k, s_k^2\right] + \text{const.} \quad (24)$$

$$= \sum_k c_{ik} \left(\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\right) + \text{const.} \quad (25)$$

Thus the variational update for the i^{th} cluster assignment is:

$$\varphi_{ik} \propto \exp\left\{\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\right\}$$

From statistics class:

if w~normal(m_k, S_k),

then: $E[\boldsymbol{w}] = \boldsymbol{m_k}$

$E[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{m_k}\, \boldsymbol{m_k^T} + \boldsymbol{S_k}$

$E[\boldsymbol{w}^T\boldsymbol{w}] = \boldsymbol{m_k^T m_k} + \boldsymbol{Tr}(\boldsymbol{S_k})$

$$\mathbb{E}[\mu_k; m_k, s_k^2] = m_k$$

$$\mathbb{E}[\mu_k^2; m_k, s_k^2] = m_k^2 + s_k^2$$

$$\varphi_{ik} \propto exp\left(m_k x_i - \frac{m_k^2 + s_k^2}{2}\right)$$

1D: $\varphi_{ik} \propto \exp\left(m_k\ X_i\ - \frac{m_k^2 + Tr(S_k)}{2}\right)$

2D: $\varphi_{ik} \propto \exp\left(\boldsymbol{m_k^T}\ \boldsymbol{X_i}\ - \frac{\boldsymbol{m_k^T m_k} + \boldsymbol{Tr}(\boldsymbol{S_k})}{2}\right)$

4.2.3 Trick: due to mean-field assumption, S_k is a diagram with identical non-zero elements.

$\mathrm{Tr}(\mathrm{S_k}) = \mathrm{M} \times \mathbf{S_k}[1,1]$

4.3 The variational density of the mixture-component means

4.3.1 Mean-field VI for $q(\mu_k)$

$$q(\mu_k) \propto \exp\left\{\log p(\mu_k) + \textstyle\sum_{i=1}^n \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2\right]\right\}. \quad (27)$$

$$
\begin{aligned}
\log q(\mu_k) &= \log p(\mu_k) + \textstyle\sum_i \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}^2_{-k}\right] + \text{const.} && (28)\\
&= \log p(\mu_k) + \textstyle\sum_i \mathbb{E}[c_{ik} \log p(x_i \mid \mu_k); \varphi_i] + \text{const.} && (29)\\
&= -\mu_k^2/2\sigma^2 + \textstyle\sum_i \mathbb{E}[c_{ik}; \varphi_i] \log p(x_i \mid \mu_k) + \text{const.} && (30)\\
&= -\mu_k^2/2\sigma^2 + \textstyle\sum_i \varphi_{ik}\left(-(x_i - \mu_k)^2/2\right) + \text{const.} && (31)\\
&= -\mu_k^2/2\sigma^2 + \textstyle\sum_i \varphi_{ik} x_i \mu_k - \varphi_{ik}\mu_k^2/2 + \text{const.} && (32)\\
&= \left(\textstyle\sum_i \varphi_{ik} x_i\right)\mu_k - \left(1/2\sigma^2 + \textstyle\sum_i \varphi_{ik}/2\right)\mu_k^2 + \text{const.} && (33)
\end{aligned}
$$

where $\mathrm{p}(\mu) = \mathrm{Normal}(\mu|0, \sigma^2) = \frac{1}{\square\sqrt{2\pi\,\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)$ is the prior distribution.

4.3.2 For 1D GMM:

这里做一个小小的修正, 我习惯把括号写完:

$$\log \mathrm{q}(\mu_k) = -\frac{\mu_k^2}{2\sigma^2} + \sum\nolimits_i^n (\varphi_{ik} x_i \mu_k \ - \frac{\varphi_{ik}\mu_k^2}{2}) + const$$

$$\Rightarrow q(\mu_k; m_k, s_k^2) = normal(\mu_k | m_k, s_k^2)$$

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}, \qquad s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}. \qquad (34)$$

4.3.2 For Multi-Dimension GMM:

$$\log \mathrm{q}(\mu_k) = -\frac{\boldsymbol{\mu}_\mathbf{k} \mathbf{V}^{-1} \boldsymbol{\mu}_\mathbf{k}^\mathbf{T}}{2} + \quad \sum\nolimits_i^n (\varphi_{ik} \mathbf{x_i} \boldsymbol{\mu}_\mathbf{k}^\mathbf{T} \ - \frac{\varphi_{ik} \boldsymbol{\mu}_\mathbf{k} \mathbf{V}_0^{-1} \boldsymbol{\mu}_\mathbf{k}^\mathbf{T}}{2}) + const$$

$$\Rightarrow q(\boldsymbol{\mu}_{\boldsymbol{k}}; \boldsymbol{m}_{\boldsymbol{k}}, \boldsymbol{V}_{\boldsymbol{k}}) = normal(\mu_k | m_k, V_k)$$

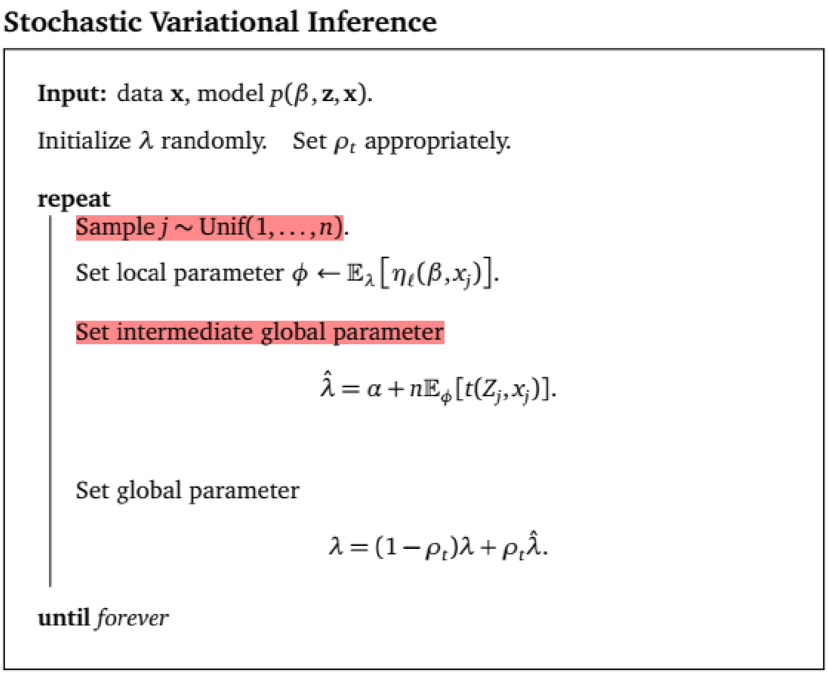
$$\boldsymbol{m}_{\boldsymbol{k}} = \{\sum\nolimits_\mathrm{i}^\mathrm{n} \varphi_{ik} \mathrm{x_i}\ \} \left(\mathbf{V}^{-1} + \mathbf{V}_0^{-1} \sum\nolimits_i^n \varphi_{ik}\right)^{-1}$$

$$\mathbf{V_k} = \left(\mathbf{V}^{-1} + \mathbf{V}_0^{-1} \sum\nolimits_i^n \varphi_{ik}\right)^{-1}$$

where $\mathbf{V}_0 = \begin{bmatrix} I & \square \\ \square & I \end{bmatrix}$, $\mathbf{V} = \begin{bmatrix} I & \square \\ \square & I \end{bmatrix}$

4.4 Stochastic Variational Inference for GMM

4.4.1 Algorithm for Stochastic Variational Inference



Stochastic Variational Inference

```
Input: data $\mathbf{x}$, model $p(\beta, \mathbf{z}, \mathbf{x})$.
Initialize $\lambda$ randomly.   Set $\rho_t$ appropriately.
repeat
    Sample $j \sim \mathrm{Unif}(1, \ldots, n)$.
    Set local parameter $\phi \leftarrow \mathbb{E}_\lambda[\eta_\ell(\beta, x_j)]$.
    Set intermediate global parameter
        $\hat{\lambda} = \alpha + n\mathbb{E}_\phi[t(Z_j, x_j)]$.
    Set global parameter
        $\lambda = (1 - \rho_t)\lambda + \rho_t\hat{\lambda}$.
until forever
```

$q(\boldsymbol{c}; \varphi_{ik})$ 不变， $q(\boldsymbol{\mu_k})$ 采用随机优化（参考随机梯度下降）

Deriving from Mean-field Variational Inference:

$$q(\mu_k) \propto \exp\left\{\log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2\right]\right\}. \qquad (27)$$

In Stochastic Variational Inference we may update for global variables λ

Randomly sample $j \sim Unif(1,2,3, \ldots, n)$.

Note: for each update it need a new j, where j = 1,2,3,...,n

$$\hat{\lambda} = \log p(\mu_k) + n \times E\left[\log p\left(\boldsymbol{x_j} \middle| c_j, \boldsymbol{\mu}\right); \ \varphi_j, \boldsymbol{m}_{-k}, \boldsymbol{s}_{-k}^2\right]$$

$$\lambda = (1 - \rho_t)\lambda + \rho_t\hat{\lambda}$$

$\Rightarrow$

4.4.2 For 1D GMM:

Randomly sample $j \sim Unif(1,2,3, \ldots, n)$.

$$\log q(\mu_k) = -\frac{\mu_k^2}{2\sigma^2} + n \times (\varphi_{jk} x_j \mu_k - \frac{\varphi_{jk}\mu_k^2}{2}) + const$$

$$\Rightarrow q(\mu_k; m_k, s_k^2) = normal(\mu_k | m_k, s_k^2)$$

$$m_k = \frac{n \times \varphi_{jk} x_j}{\frac{1}{\sigma^2} + n \times \varphi_{jk}}, \qquad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + n \times \varphi_{jk}}$$

4.4.3 For Multi-Dimension GMM:

Randomly sample $j \sim Unif(1,2,3, \ldots, n)$.

$$\log q(\boldsymbol{\mu_k}) = -\frac{\boldsymbol{\mu_k} V^{-1} \boldsymbol{\mu_k}^T}{2} + n \times (\varphi_{jk} \boldsymbol{x_j} \boldsymbol{\mu_k}^T - \frac{\varphi_{jk}\boldsymbol{\mu_k} V_0^{-1} \boldsymbol{\mu_k}^T}{2}) + const$$

$$\Rightarrow q(\boldsymbol{\mu_k}; \boldsymbol{m_k}, \boldsymbol{V_k}) = normal(\boldsymbol{\mu_k}|\boldsymbol{m_k}, \boldsymbol{V_k})$$

$$\boldsymbol{m_k} = (n \times \varphi_{jk}\mathbf{x_j})\left(\mathbf{V^{-1}} + \mathbf{V_0^{-1}}\, n \times \varphi_{jk}\right)^{-1}$$

$$\mathbf{V_k} = \left(\mathbf{V^{-1}} + \mathbf{V_0^{-1}}\ \mathrm{n} \times \varphi_{jk}\right)^{-1}$$

where n is the quantity of all data,

$\mathbf{V}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is from the prior distribution for $p(\boldsymbol{\mu}) = \mathrm{normal}(\boldsymbol{\mu}|\mathbf{0}, \boldsymbol{V_0})$

$\mathbf{V} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ is the observation for $\mathrm{p}(\mathbf{x_i}|\mathbf{c_{i;}}\ \boldsymbol{\mu})$

5.Comparison of Stochastic VI and Mean-field VI

5.1 Classical VI is inefficient:

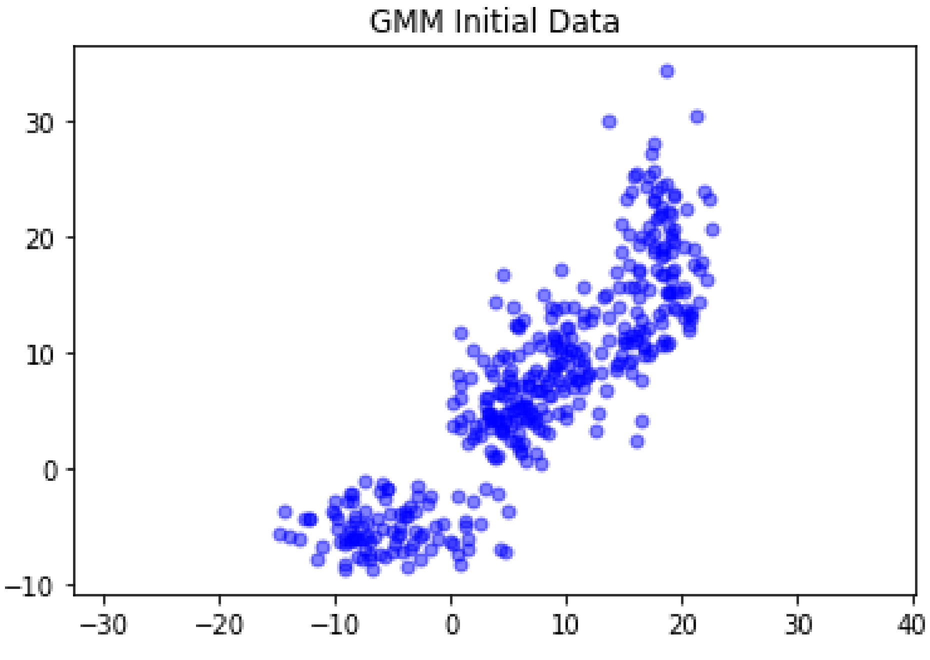
- Do some local computation *for each data point.*
- Aggregate these computations to re-estimate global structure.
- Repeat.
- This cannot handle massive data.

5.2 Stochastic VI more efficient:

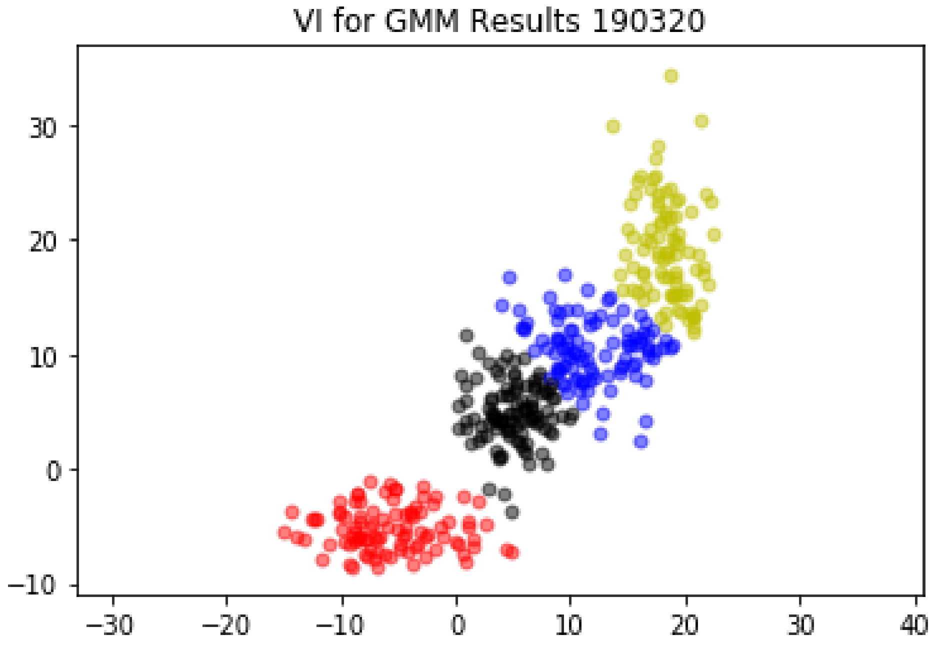
- Stochastic variational inference (SVI) scales VI to massive data.

6. simulation results

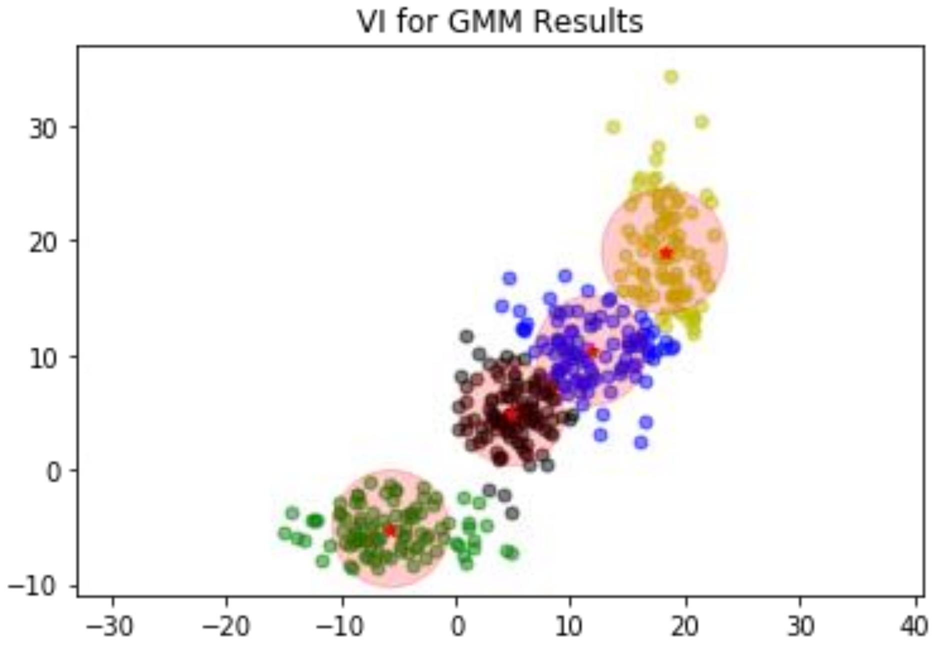
6.1 Initial data:



6.2 Visualization of results 1



6.3 Visualization of results 2



7. references

1. Variational Inference: A Review for Statisticians, https://arxiv.org/abs/1601.00670
2. VARIATIONAL INFERENCE: FOUNDATIONS AND INNOVATIONS http://www.cs.columbia.edu/~blei/talks/Blei_VI_tutorial.pdf

→