PS 4.

1. (a).

x	$P[X=x]$	$P[Y=1 \mid X=x]$	$h(x)$	Y
0	0.5	0.3	0.3	1
1	0.5	0.3	0.3	1

$$P[Y=1 \mid h(X)=\overset{0.3}{\cancel{p}}] = \frac{P[Y=1 \text{ and } h(X)=0.3]}{\cancel{P[h(X)=0.3]}=1}$$

$$= P[Y=1 \text{ and } X=0] + P[Y=1 \text{ and } X=1]$$

$$= P[Y=1 \mid X=0] \cdot P[X=0] + P[Y=1 \mid X=1] \cdot P[X=1]$$

$$= 0.5 \cdot 0.3 + 0.5 \cdot 0.3$$

$$= 0.3 = h(X).$$

$\therefore$ that's a perfect calibration. but accuracy is:

$$P[\mathbb{I}[h(X) \geq 0.5] = Y] = 0 \quad (\text{not perfect})$$

~~the~~ the converse is true. use table 1 as example.

x	$P[X=x]$	$P[Y=1 \mid X=x]$	$h(x)$	Y
0	0.25	0.2	0.3	0
1	0.25	0.0	0.3	0
2	0.25	1.0	0.3	1
3	0.25	0.8	0.3	1

Accuracy is 100% but not perfect calibration.

1.(b) $MSE(h) = E[(Y-h(X))^2]$

$= E[(Y - T(X) + T(X) - h(X))^2]$

$= E[(Y-T(X))^2 + 2(Y-T(X))(T(X)-h(X)) + (T(X)-h(X))^2]$

$= E[(Y-T(X))^2] + 2E[(Y-T(X))(T(X)-h(X))] + \underbrace{E[(T(X)-h(X))^2]}_{\text{Calibration Error.}}$

$E[(Y-T(X))^2] = E[(Y-E(Y)+E(Y)-T(X))^2]$

$= \underbrace{E[(Y-E(Y))^2}_{Var[Y]} + 2(Y-E(Y))(E(Y)-T(X)) + (E(Y)-T(X))^2]$

$MSE(h) = Var[Y] + E[(T(X)-h(X))^2] + S$

where

$S = 2E[YT(X) - T(X)^2 - Yh(X) + T(X)h(X)]$

$+ 2E[YE(Y) - YT(X) - E(Y)^2 + E(Y)T(X)]$

$+ E[(E(Y))^2 - 2E(Y)(T(X)) + T(X)^2]$

$= 2\cancel{E[YT(X)]} - 2E(T(X)^2) - 2E(Yh(X)) + 2E[T(X)h(X)]$

$+ 2\cancel{E[YE(Y)]} - 2\cancel{E(YT(X))} - 2\cancel{E(Y)^2} + 2E\cancel{(Y)E(T(X))}$

$+ E(Y)^2 - 2E(Y)\cancel{E(T(X))} + E(T(X)^2)$

$= E(Y)^2 - E(T(X)^2) - 2E(Yh(X)) + 2E(T(X)h(X))$

$= E(Y)^2 - 2E(Yh(X)) + 2E(T(X)h(X)) - \underbrace{(E(T(X)^2) - E(T(X))^2)}_{Var[T(X)]}$

$- E(T(X))^2$

Let $Z = E(Y)^2 - 2E(Yh(X)) + 2E(T(X)h(X)) - E[T(X)]^2$

1.(b) cont.

$\therefore MSE = Var[Y] - Var[T(x)] + E[(T(x)-h(x))^2]$

$+ Z$.

$Z = E(y)^2 - E[T(x)]^2 + 2E(T(x)h(x)) - 2E(y\,h(x))$

$E(y) = E(E(y|h(x))) = E[T(x)]$

$\therefore E(y)^2 - E(T(x))^2 = 0$

$2E(T(x)h(x)) - 2E(y\,h(x))$

$= 2E(T(x)) \cdot E(h(x)) - 2E(y) \cdot E(h(x))$ y, h(x) are independent

T(x), h(x) are independant

$= 2E(h(x))\,(E(T(x)) - E(y)) = 0$

$\therefore Z = 0$

$\therefore MSE = var[Y] - var[T(x)] + E[(T(x)-h(x))^2]$

2.

for a datapoint $x^{(i)}$, let $\alpha^{(i)}$ ~~be the~~ satisfy that

$$f_u(x^{(i)}) = \underset{v \in \mathcal{V}}{\arg\min} \, \|x^{(i)} - v\|^2.$$

$$\therefore f_u(x^{(i)}) = \alpha^{(i)} \cdot u.$$

$$\cancel{\underset{u:u^Tu=1}{\arg\min} \sum_{i=1}^{n}} \|x^{(i)} - f_u(x)^{i}\|_2^2$$

$$= \|x^{(i)} - \alpha^{(i)} u\|_2^2$$

$$= (x^{(i)} - \alpha^{(i)} u)^T \cdot (x^{(i)} - \alpha^{(i)} u)$$

$$= x^{(i)T} \cdot x^{(i)} - 2\alpha^{(i)} x^{(i)T} u + \alpha^{(i)2} \cdot \underbrace{u^T u}_{= I}$$

$$\underset{u:u^Tu=1}{\arg\min} \sum_{i=1}^{n} \|x^{(i)} - f_u(x)^{i}\|_2^2$$

$$= \underset{u:u^Tu=1}{\arg\min} \sum_{i=1}^{n} x^{(i)T} \cdot x^{(i)} - 2\alpha^{(i)} \cancel{x}^{(i)T} u + \alpha^{(i)2} \underbrace{u^T u}_{\to I}$$

independant of u.

$$= \underset{u:u^Tu=1}{\arg\max} \sum \alpha^{(i)} x^{(i)T} u \quad = \arg\max \sum \alpha^{(i)} \cdot u^T u \cdot x^{(i)T} u$$

$$= \arg\max \sum u^T \cdot (\alpha^{(i)} u \cdot x^{(i)T}) u \quad = \arg\max \sum u^T \cdot (f_u(x^{(i)}) \cdot x^{(i)T}) u$$

propotional to $x^{(i)}$

$$= \underset{u:u^Tu=1}{\arg\max} \sum_{i=1}^{n} u^T (x^{(i)} \cdot x^{(i)T}) u$$

$$= \underset{u:u^Tu=1}{\arg\max} \, u^T \left(\sum_{i=1}^{n} x^{(i)} \cdot x^{(i)T} \right) u.$$

Original large image

Updated large image

3. (a) · compress image.

(b) Instead of storing a [r, g, b] color per pixel, we store 1 int per pixel ~~to~~ ~~to~~ as the assigned centroid

so we compressed the image by 3 approximately

~~with~~ ~~but~~ because we do need an extra ~~E~~ 16 × 3 to store the centroids

4.(a)

$$D(P\|Q) = \sum_{x\in X} p(x)\log\frac{p(x)}{Q(x)}$$

$$= -\sum_{x\in X} p(x)\log\frac{Q(x)}{p(x)} \quad \text{by Jensen's inequality}$$

$$\geq -\log\sum_{x\in X} p(x)\cdot\frac{Q(x)}{p(x)}$$

$$= -\log\underline{\sum_{x\in X} Q(x)} \quad \text{by definition} = 1$$

$$= -\log 1$$

$$= 0$$

$\therefore D(P\|Q) \geq 0$

when $P(P\|Q) = 0 \Leftrightarrow \sum_{x\in X} P(x)\log\frac{Q(x)}{P(x)} = 0 = \log\sum_{x\in X} P(x)\cdot\frac{Q(x)}{P(x)}$

$\Rightarrow X$ is aconstant

$\therefore \frac{Q(x)}{P(x)} =$ constant c.

$$\log\sum_{x\in X} P(x)\cdot c = 0$$

$$\Rightarrow \log c\cdot\sum_{x\in X} P(x) = 0$$

$$\Rightarrow \log c\cdot 1 = 0$$

$$\Rightarrow c = 1$$

$\therefore \frac{Q(x)}{P(x)} = 1 \Rightarrow Q = P.$

4(b)

$$D(P(X,Y)\|Q(X,Y)) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

$$= \sum_x \sum_y P(x)P(y|x) \cdot \log \frac{P(x)P(y|x)}{Q(x)\,Q(y|x)}$$

$$= \underbrace{\sum_x \sum_y P(x)P(y|x) \cdot \log \frac{P(x)}{Q(x)}}_{\parallel} + \sum_x \sum_y P(x)P(y|x) \log \frac{P(y|x)}{Q(y|x)}$$

$$= \sum_x P(x) \cdot \log \frac{P(x)}{Q(x)} + \sum_x P(x) \left(\sum_y P(y|x) \cdot \log \frac{P(y|x)}{Q(y|x)} \right)$$

$$= D(P(X)\|Q(X)) + D(P(Y|X)\|Q(Y|X))$$

4.(c)

$$\cancel{\arg\min_\theta}\, D(\hat{P}\|P_\theta) = \sum_{x \in X} \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)}$$

$\hat{P}(x) = \frac{1}{n}$ as uniform distribution

$$D = \sum_{x \in X} \frac{1}{n} \cdot \log \frac{1}{n \cdot P_\theta(x)}$$

$$= -\frac{1}{n} \left(\sum_{x \in X} \left[\log n + \log P_\theta(x) \right] \right)$$

$$\arg\min_\theta D(\hat{P}\|P_\theta) = \arg\min_\theta \underbrace{-\frac{1}{n}}_{\text{constant}} \left(\sum_{x \in X} \left(\underbrace{\log n}_{\text{independant of } \theta} + \log P_\theta(x) \right) \right)$$

$$= \arg\max_\theta \sum_{x \in X} \log P_\theta(x)$$

$$= \arg\max_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)})$$

5.(a) from class, we know $l_{unsup}(\theta^{(t+1)}) \geq l_{unsup}(\theta^{(t)})$

$$l_{sup}(\theta^{(t+1)}) = \sum_{i=1}^{\tilde{n}} \log(P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})$$

$$= \underbrace{\sum_{i=1}^{\tilde{n}} \log(P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})}_{l_{sup}(\theta^{t})} + \sum_{i=1}^{\tilde{n}} \log \frac{P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})}{P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})}$$

$$= \cancel{\log} \prod \underbrace{P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{t+1}) \geq P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})}_{\text{by M-Step definition}}$$

$$\therefore \sum_{i=1}^{\tilde{n}} \log \frac{P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t+1)})}{P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta^{(t)})} \geq \sum_{i=1}^{\hat{n}} \log 1 = 0$$

$$\therefore l_{sup}(\theta^{(t+1)}) \geq l_{sup}(\theta^{(t)})$$

$$\therefore l_{semi}(\theta^{(t+1)}) = l_{unsup}(\theta^{(t+1)}) + \alpha\, l_{sup}(\theta^{(t+1)})$$

$$\geq l_{usup}(\theta^{(t)}) + \alpha\, l_{sup}(\theta^{(t)})$$

$$= l_{semi}(\theta^{(t)})$$

5(b) only $w_j^{(i)}$ needs to be re-estimated.

$$w_j^{(i)} = P(z^{(i)}=j \mid x^{(i)}; \phi, \mu, \Sigma)$$
$$= Q_i(z^{(i)}=j)$$

5.(c)

let
$$S = \sum_{i=1}^{n}\left(\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}\right) + \alpha\left(\sum_{i=1}^{\hat{n}} \log P(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \cdot \Sigma_j^{-1}(x^{(i)}-\mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

$$+ \alpha \sum_{i=1}^{\hat{n}} \log \frac{1}{(2\pi)^{d/2}|\Sigma_{\tilde{z}^{(i)}}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})^T \cdot \Sigma_{\tilde{z}^{(i)}}^{-1}(\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})\right)\phi_{\tilde{z}^{(i)}}$$

$$\nabla_{\mu_l} S = \sum_{i=1}^{n} w_l^{(i)}(\Sigma_l^{-1}x^{(i)} - \Sigma_l^{-1}\mu_l) +$$
$$-\frac{\alpha}{2}\nabla_{\mu_l}\sum_{i=1}^{\hat{n}} 1\{\tilde{z}^{(i)}=l\}(\tilde{x}^{(i)}-\mu_l)^T \cdot \Sigma_l^{-1}(\tilde{x}^{(i)}-\mu_l)$$
$$= \sum_{i=1}^{n} w_l^{(i)}(\Sigma_l^{-1}x^{(i)} - \Sigma_l^{-1}\mu_l) + \alpha\sum_{i=1}^{\hat{n}} 1\{\tilde{z}^{(i)}=l\} \cdot (\Sigma_l^{-1}\tilde{x}^{(i)} - \Sigma_l^{-1}\mu_l)$$

set $\nabla_{\mu_l} = 0$

$$\mu_l = \frac{\sum_{i=1}^{n} w_l^{(i)} \cdot x^{(i)} + \alpha\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)}=l\} \cdot \tilde{x}^{(i)}}{\left(\sum_{i=1}^{n} w_l^{(i)} + \alpha\sum_{i=1}^{\hat{n}} 1\{\hat{z}^{(i)}=l\}\right)}$$

$$\nabla_{\phi_j} S = \nabla\left[\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log\phi_j + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)}=j\} \cdot \log\phi_j\right]$$

$$= \nabla\left[\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \log\phi_j + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)}=j\} \log\phi_j + \beta\left(\sum_{j=1}^{k}\phi_j - 1\right)\right]$$

$$= \sum_{i=1}^{n} \frac{w_j^{(i)}}{\phi_j} + \alpha \cdot \sum_{i=1}^{\tilde{n}} \frac{1\{\tilde{z}^{(i)}=j\}}{\phi_j} + \beta$$

set it to 0

$$\phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)}=j\}}{-\beta}$$

$$\sum_j \phi_j = 1$$

$$-\beta = n$$

$$= \frac{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)}=j\}}{n}$$

$\cancel{\nabla_{\Sigma_j} S =}$

terms in S related to Σ_j:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} w_j^{(i)} \left[-\log(w_j^{(i)}) - \log(2\pi^{d/2}) - \frac{1}{2}\log|\Sigma_j| + \left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right) + \log\phi_j\right]$$

$$+ \alpha \sum_{i=1}^{\tilde{n}} \left[-\log(2\pi^{d/2}) - \frac{1}{2}\log|\Sigma_{\tilde{z}^{(i)}}| + \left(-\frac{1}{2}(\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})^T \Sigma_{\tilde{z}^{(i)}}^{-1} (\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})\right) + \log\phi_j\right]$$

$$\cancel{\alpha \sum_{i=1}^{\hat{n}} \left(-\left(\log(2\pi)^{1/2} + \tfrac{1}{2}\log|\Sigma_j| + \left(-\tfrac{1}{2}\right.\right.\right.}$$

$$\nabla_{\Sigma_j} S = \nabla \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \left[-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j) \right]$$

$$+ \nabla \alpha \sum_{i=1}^{\tilde{n}} \left[-\frac{1}{2}\log|\Sigma_{\tilde{z}^{(i)}}| - \frac{1}{2}(\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}})^T \Sigma_{\tilde{z}^{(i)}}^{-1} (\tilde{x}^{(i)}-\mu_{\tilde{z}^{(i)}}) \right]$$

$$= -\frac{1}{2}\left[\sum_{i=1}^{n} \left(w_j^{(i)} \Sigma_j^{-1} - w_j^{(i)} (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T \Sigma_j^{-2} \right) \right]$$

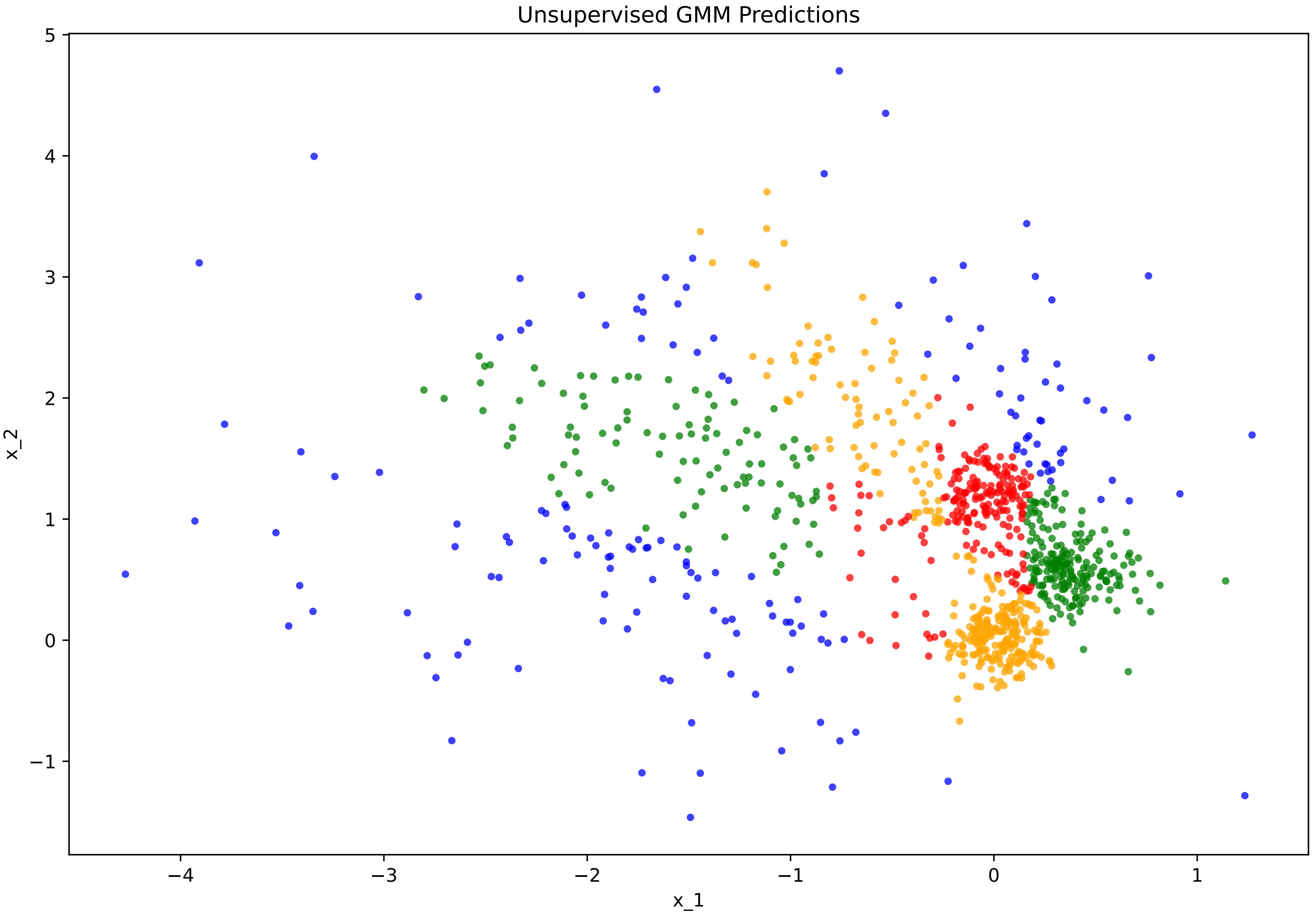
$$- \frac{\alpha}{2}\left[\sum_{i=1}^{\tilde{n}} 1\{\tilde{z}^{(i)}=j\} \left(\Sigma_j^{-1} - (\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T \Sigma_j^{-2} \right) \right]$$

set it to 0.

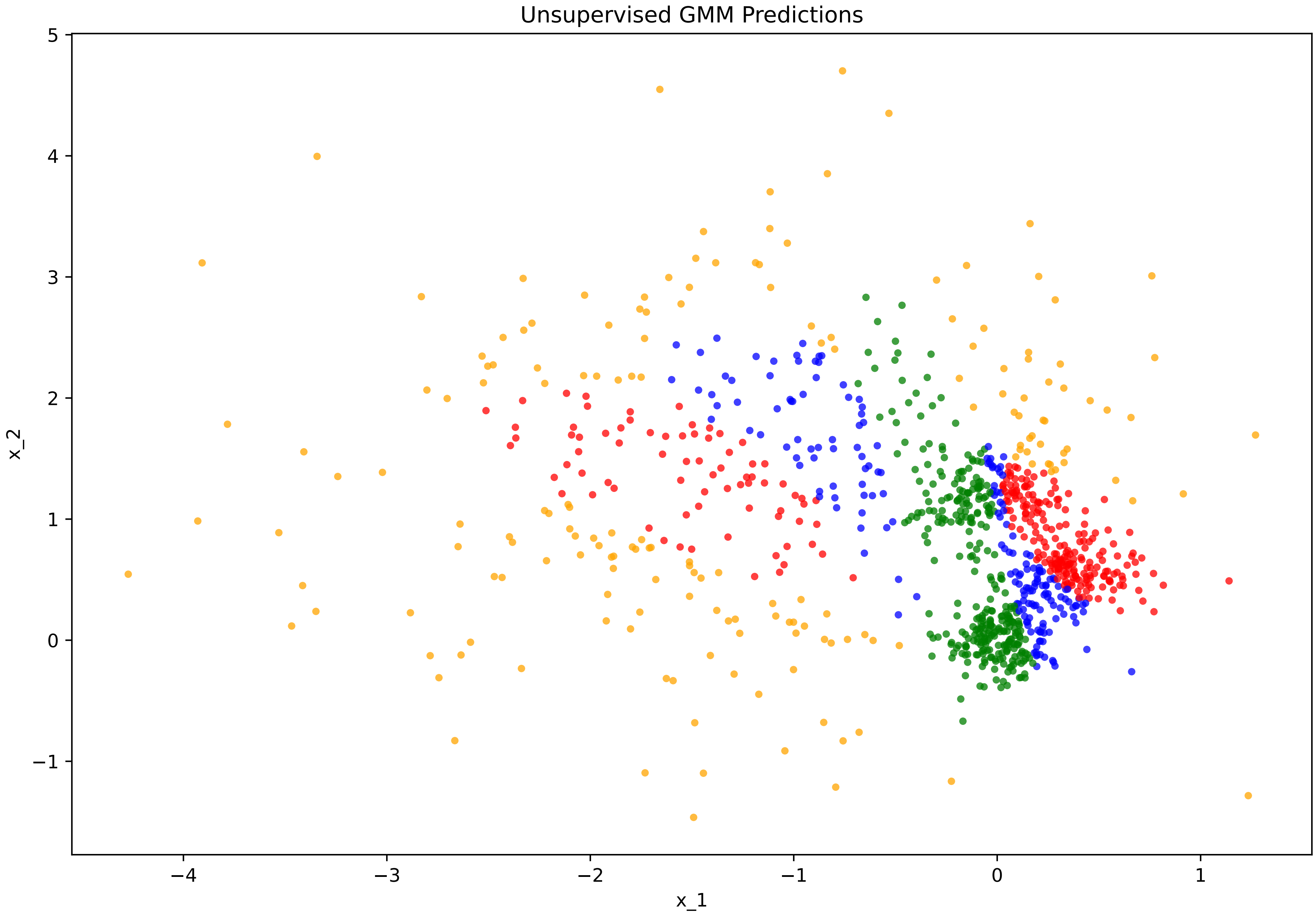
$$\sum_{i=1}^{n} \left(w_j^{(i)} \Sigma_j - w_j^{(i)} (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T \right)$$

$$+ \alpha \left(\sum_{i=1}^{\tilde{n}} 1\{z^{(i)}=j\} \left(\Sigma_j - (\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T \right) \right) = 0$$

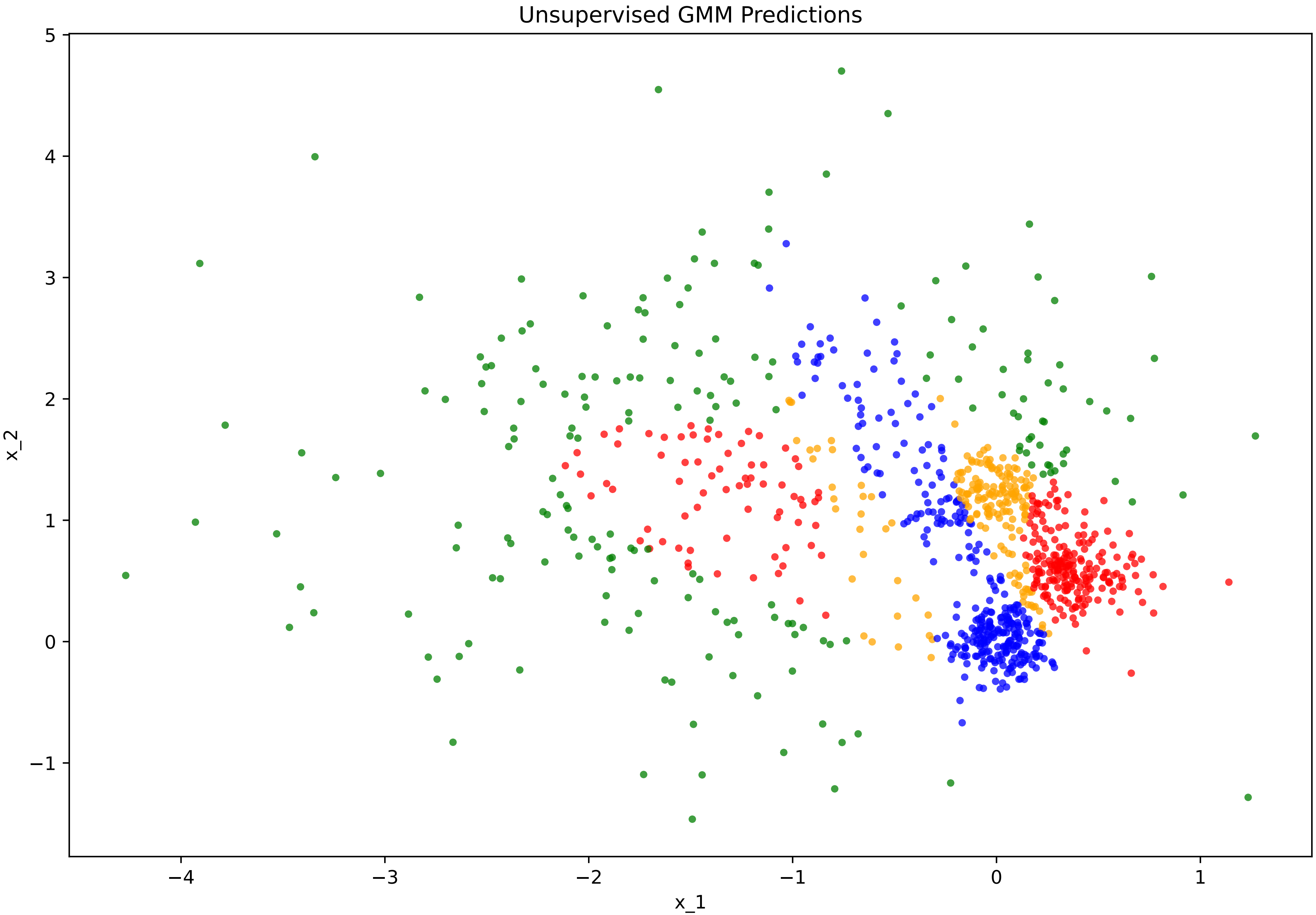
$$\Sigma_j = \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T + \alpha \sum_{i=1}^{\tilde{n}} 1\{z^{(i)}=j\} (\tilde{x}^{(i)}-\mu_j)(\tilde{x}^{(i)}-\mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} 1\{z^{(i)}=j\}}$$



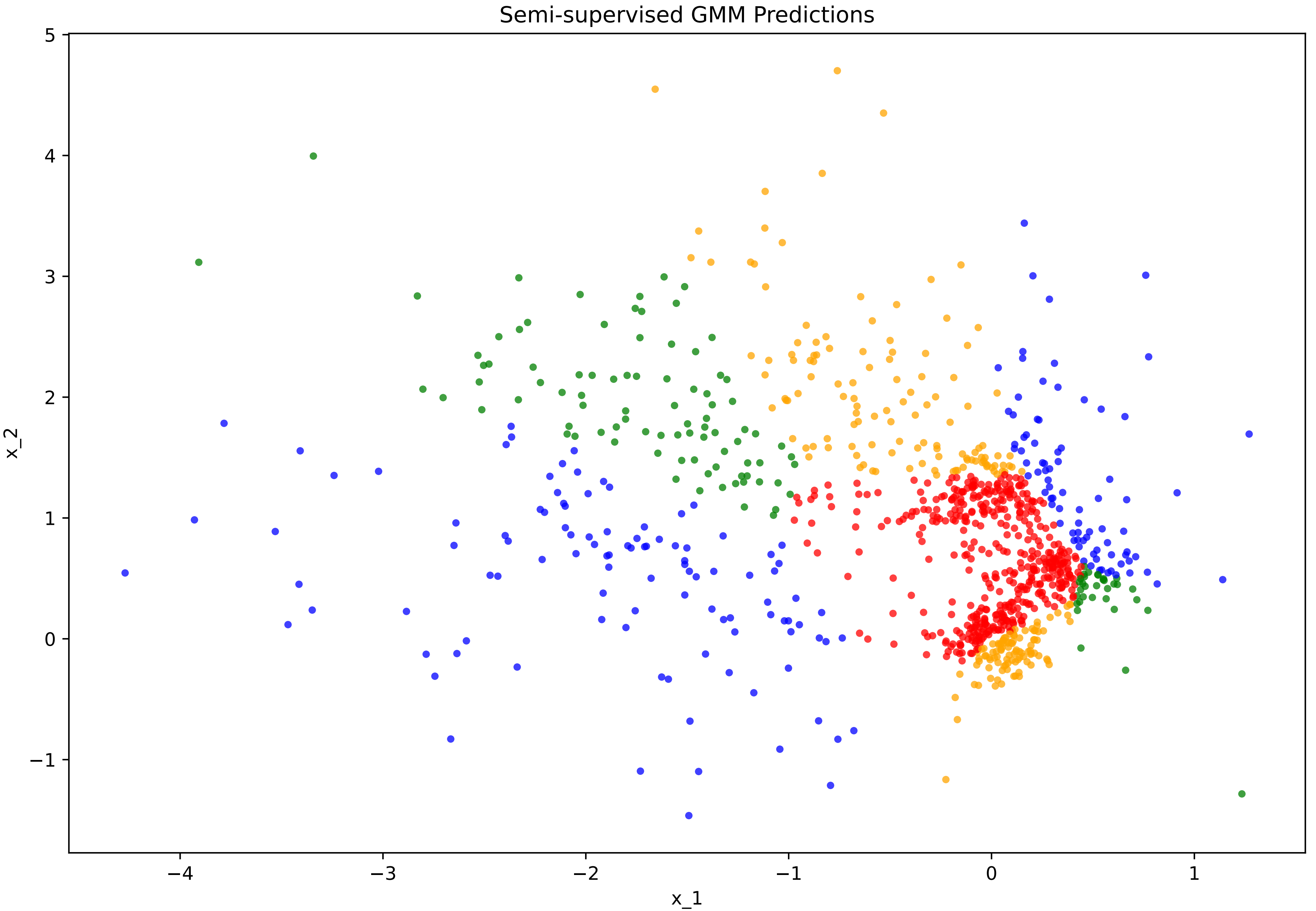
Unsupervised GMM Predictions
5
4
3
2
1
0
−1
x_2
−4
−3
−2
−1
0
1
x_1

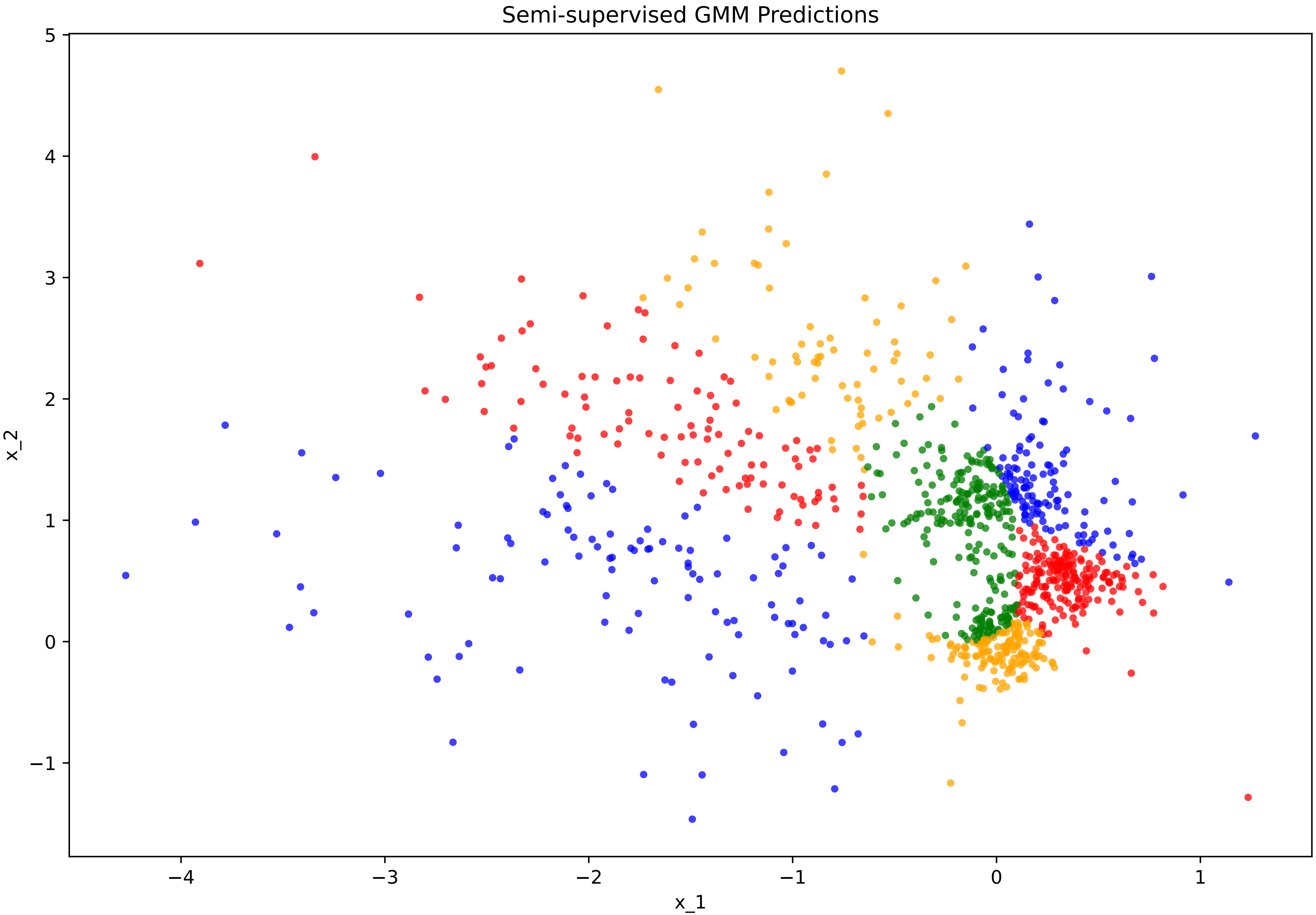


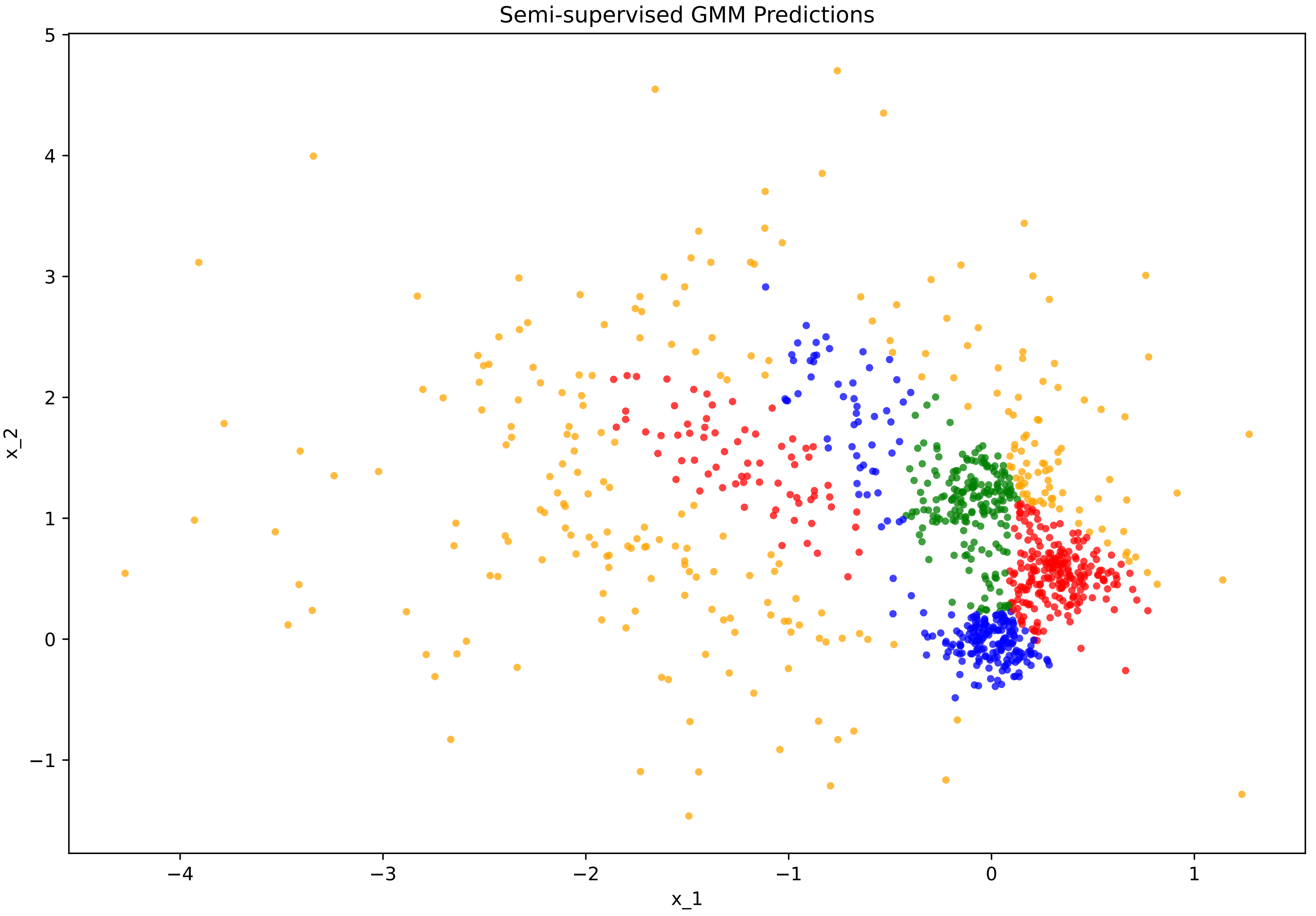
Unsupervised GMM Predictions
x_2
x_1
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1



Unsupervised GMM Predictions
5
4
3
2
1
0
−1
x_2
−4
−3
−2
−1
0
1
x_1


Semi-supervised GMM Predictions
x_1
x_2
−4
−3
−2
−1
0
1
−1
0
1
2
3
4
5


Semi-supervised GMM Predictions
5
4
3
2
1
0
−1
−4
−3
−2
−1
0
1
x_1
x_2


Semi-supervised GMM Predictions
x_1
x_2
−4
−3
−2
−1
0
1
5
4
3
2
1
0
−1

5.(e)

5.(f),

	unsupervised :	semi-supervised
iteration :	~40	<10
Stability :	low ~~changes~~	high.
overall quality :	high	high.

intuitively, semi-supervised is like regulation.
it converges and has high stability