Lecture 19 — Expectation-Maximization.

Alex Schwing and Matus Telgarsky

April 3, 2018

Announcements.

- Midterms available in **TA** office hours.
- 10 more days for regrade requests.
- Ongoing questions:
  - Solutions available ...?
  - Delay last homework...?
  - Videos...?
- Any other course/midterm concerns?

Schedule for today.

- k-means and GMM E-M review.
- E-M in general.

Lloyd's method ("k-means algorithm").

1. Choose initial centers $(\mu_1, \ldots, \mu_k)$.
2. Alternate the following two steps until convergence:

   2.1 **(Reassignment.)** Hold centers fixed, optimally update hard assignments $A \in \{0,1\}^{n \times k}$, $A\mathbf{1}_k = \mathbf{1}_n$: for every $i \in \{1, \ldots, n\}$,

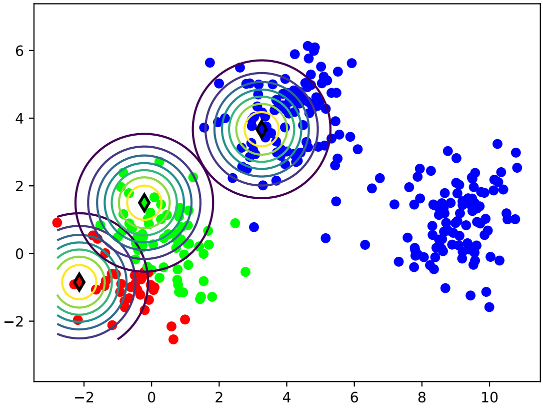
   $$A_{ij} = \mathbb{1}\left[\mu(x_i) = \mu_j\right].$$

   2.2 **(Recentering.)** Hold assignments fixed, optimally update centers: for every $j \in \{1, \ldots, k\}$,

   $$\mu_j := \frac{\sum_{i=1}^n A_{ij} x_i}{\sum_i A_{ij}}.$$

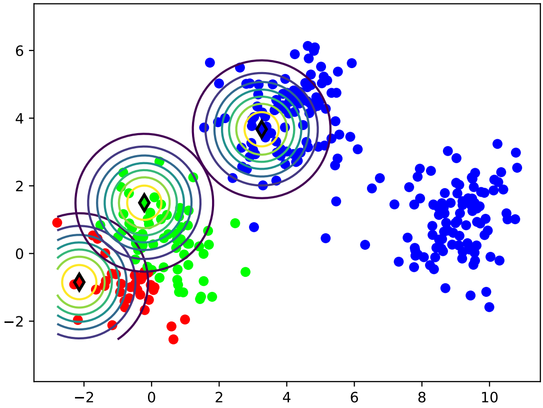
Lloyd's method ("k-means algorithm).

Works well on spherical data.



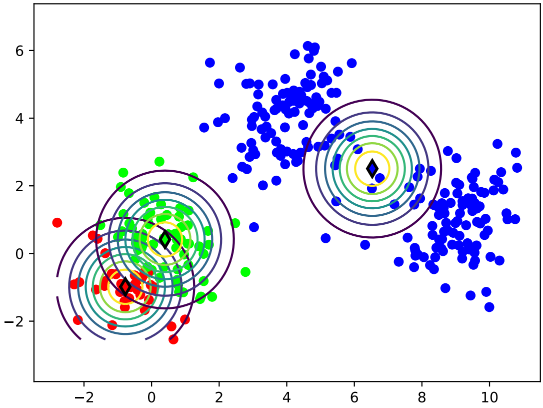
Lloyd's method ("*k*-means algorithm).

Works well on spherical data.



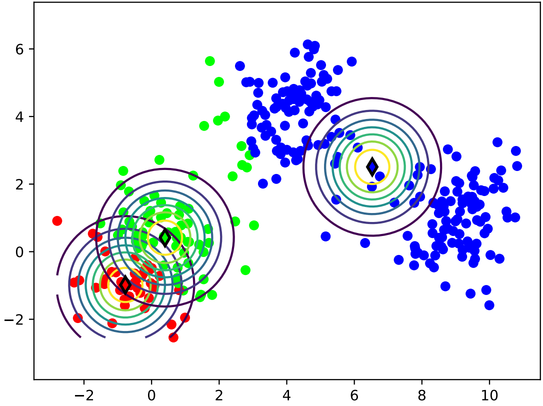
Lloyd's method ("k-means algorithm).

Works well on spherical data.



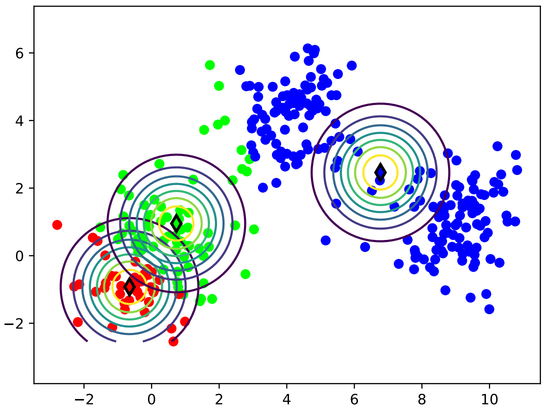
Lloyd's method ("*k*-means algorithm).

Works well on spherical data.



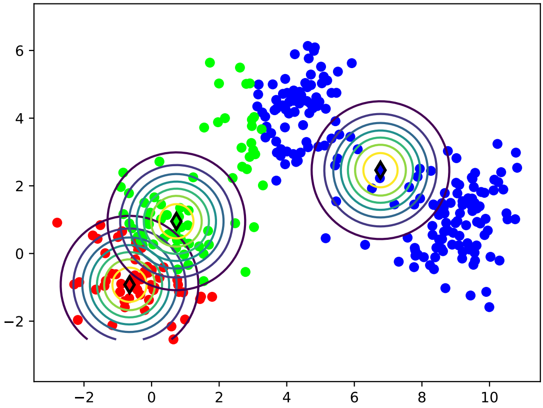
Lloyd's method ("k-means algorithm).

Works well on spherical data.



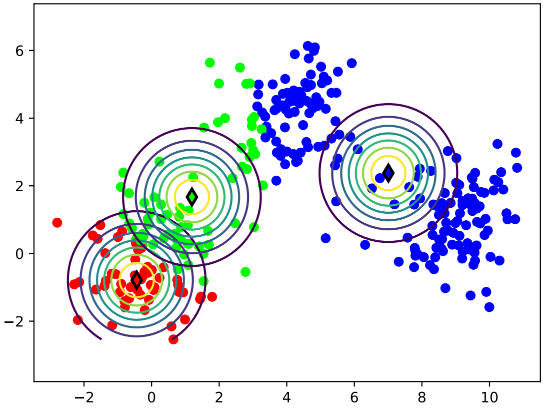
Lloyd's method ("*k*-means algorithm).

Works well on spherical data.



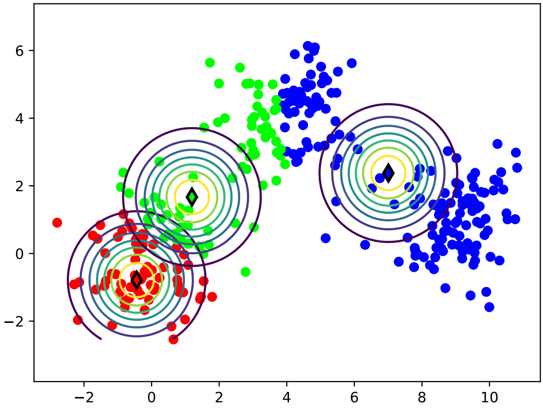
Lloyd's method ("k-means algorithm).

Works well on spherical data.



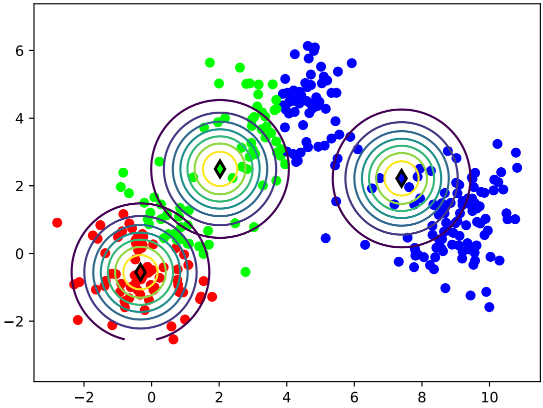
Lloyd's method ("*k*-means algorithm).

Works well on spherical data.



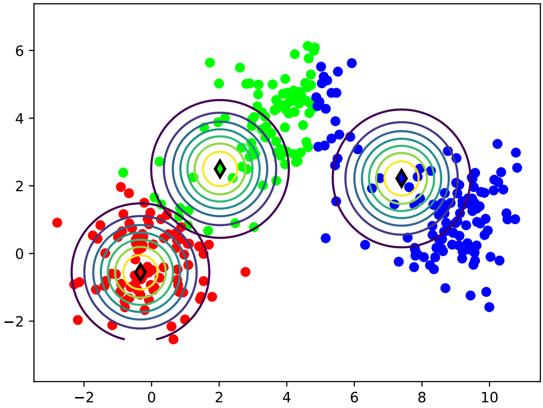
Lloyd's method ("*k*-means algorithm).

Works well on spherical data.



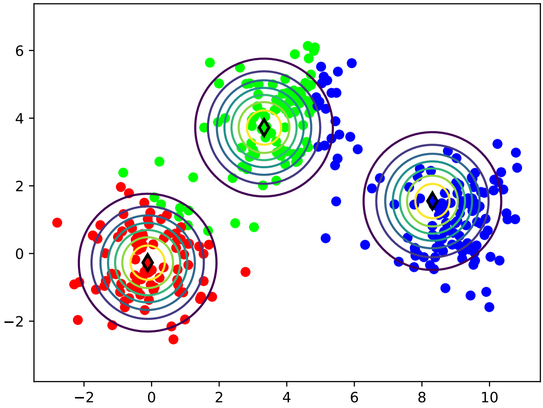
Lloyd's method ("*k*-means algorithm).

Works well on spherical data.



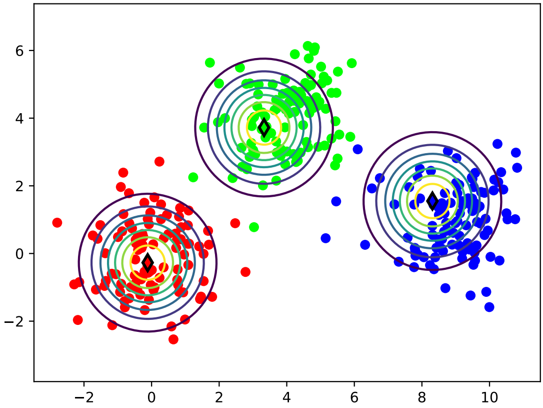
Lloyd's method ("k-means algorithm).

Works well on spherical data.



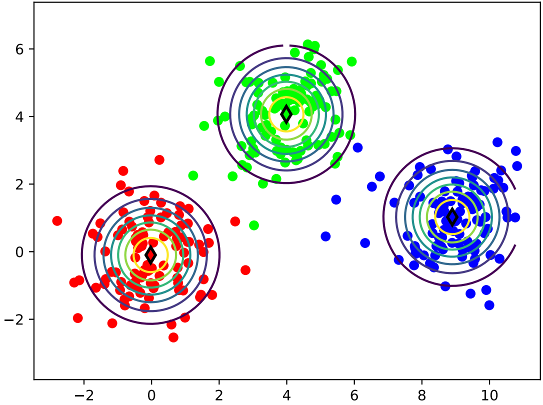
Lloyd's method ("k-means algorithm).

Works well on spherical data.



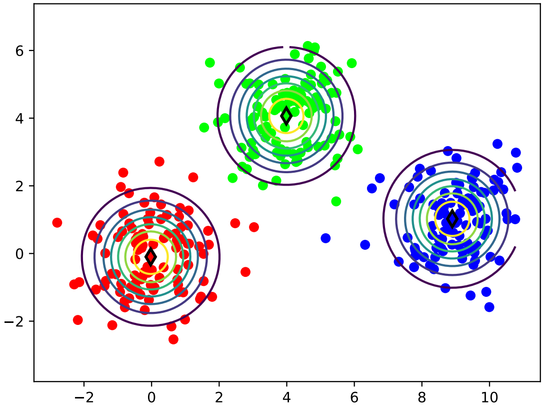
Lloyd's method ("k-means algorithm).

Works well on spherical data.



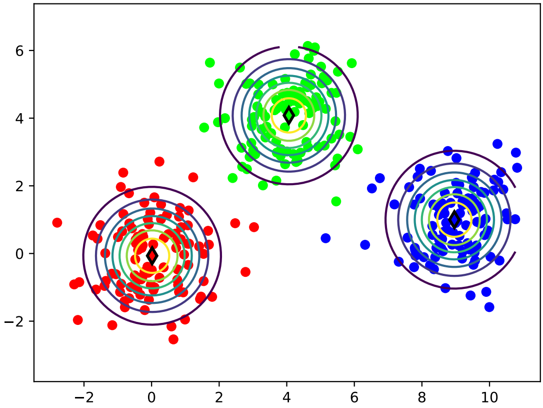
Lloyd's method ("k-means algorithm).

Works well on spherical data.



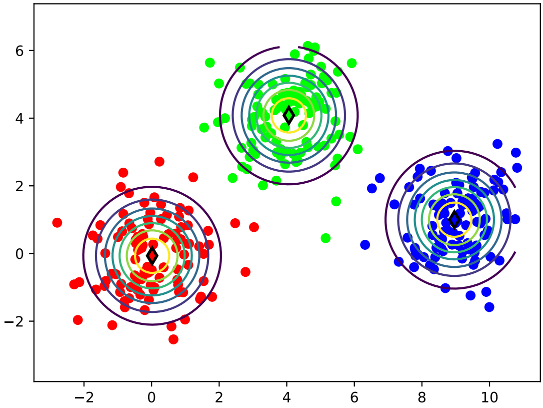
Lloyd's method ("k-means algorithm).

Works well on spherical data.



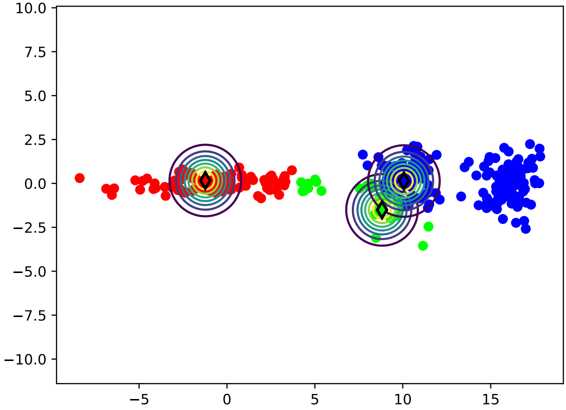
Lloyd's method ("k-means algorithm).

Works well on spherical data.



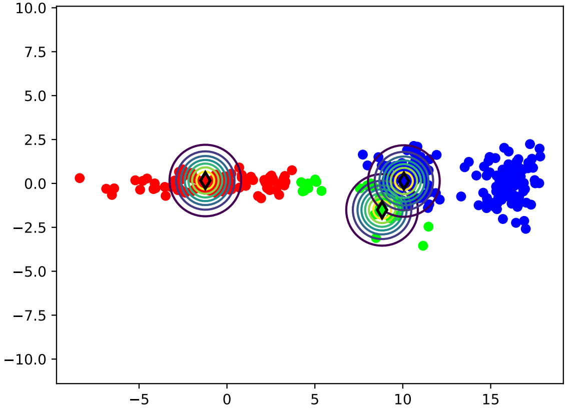
Lloyd's method ("*k*-means algorithm).

On general ellipses has trouble.



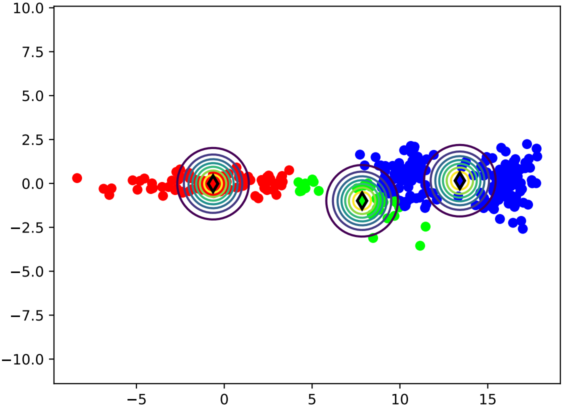
Lloyd's method ("*k*-means algorithm).

On general ellipses has trouble.



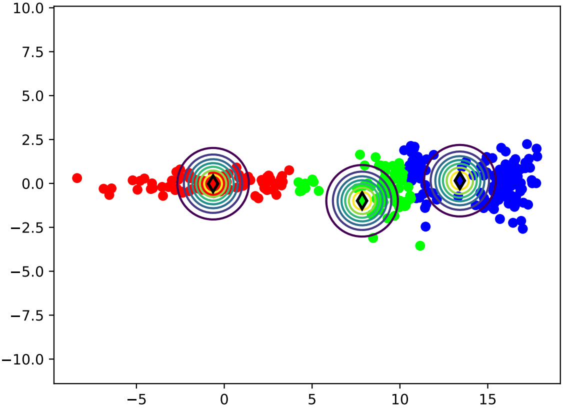
Lloyd's method ("*k*-means algorithm).

On general ellipses has trouble.



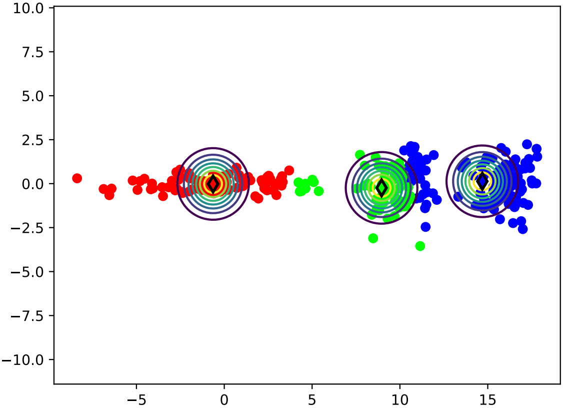
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



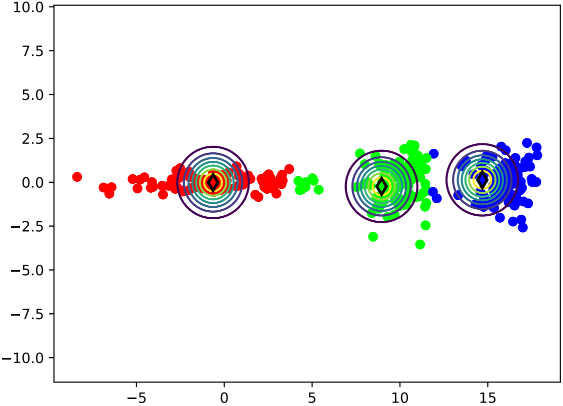
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



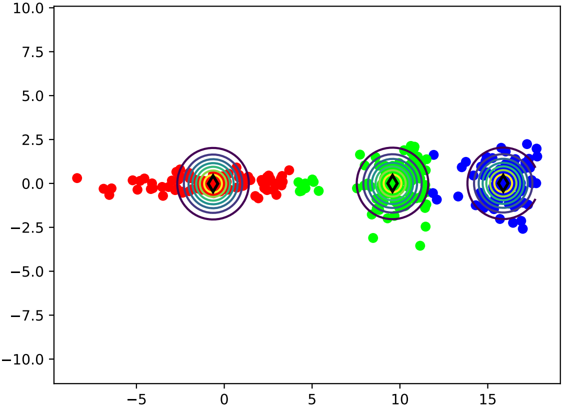
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



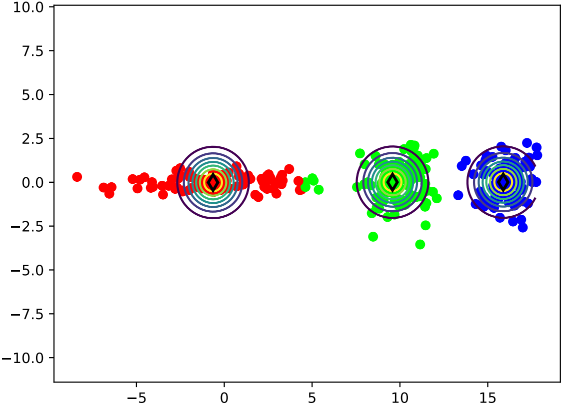
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



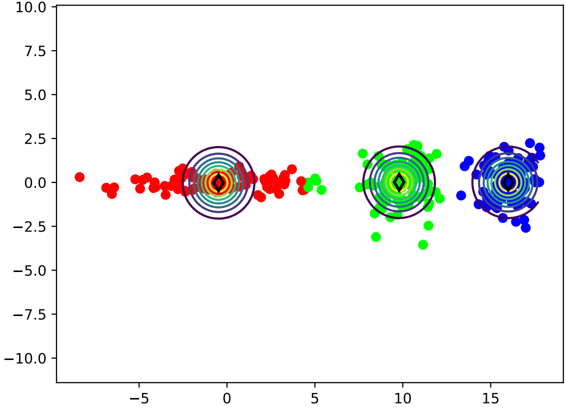
Lloyd's method ("*k*-means algorithm).

On general ellipses has trouble.



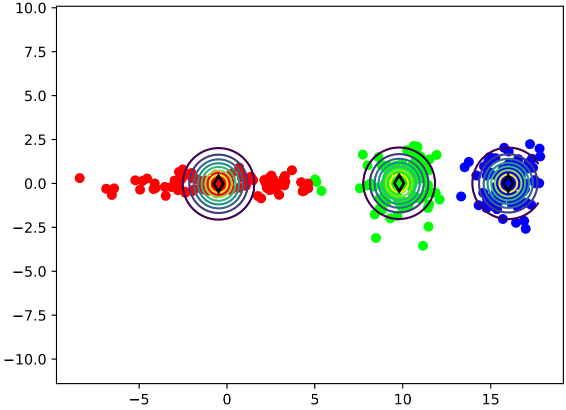
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



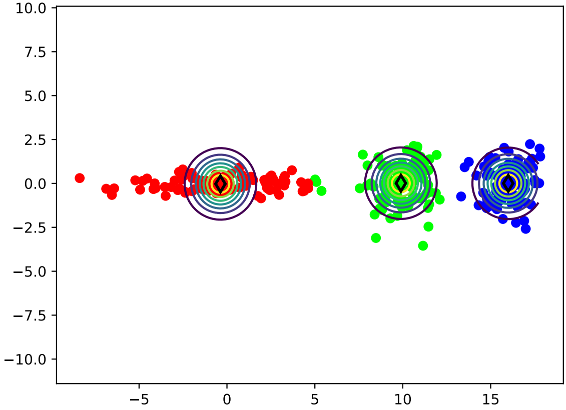
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



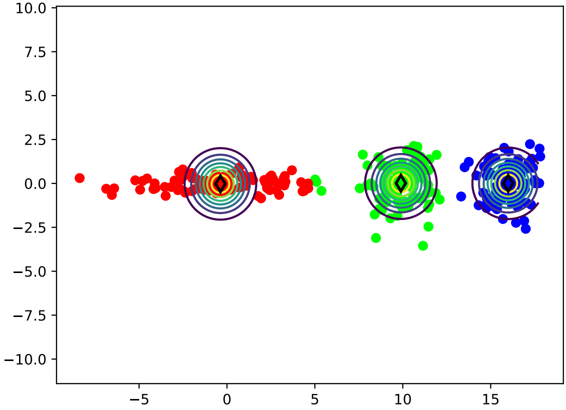
Lloyd's method ("k-means algorithm).

On general ellipses has trouble.



Lloyd's method ("k-means algorithm).

On general ellipses has trouble.

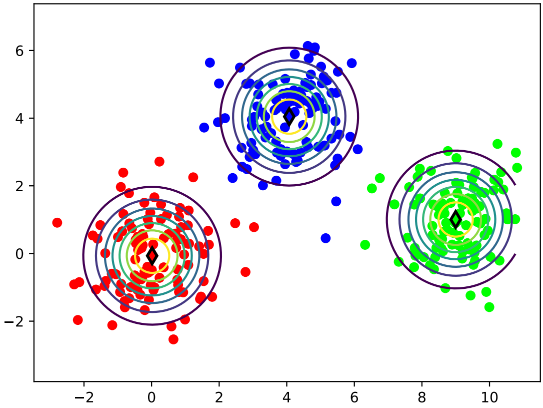


E-M for Gaussian Mixture Model (GMM).

1. Choose initial parameters $\theta = ((\pi_1, \mu_1, \Sigma_1), \ldots, (\pi_k, \mu_k, \Sigma_k))$.
2. Alternate the following two steps until convergence:
   2.1 **(E step)** (Reassignment). Hold parameters fixed, optimally update soft assignments $A \in [0,1]^{n \times k}$, $A\mathbf{1}_k = \mathbf{1}_n$: for every $i \in \{1, \ldots, n\}$,
   $$A_{ij} \propto \pi_j p_{\theta_j}(x_i),$$
   where p_{θ_j} is the gaussian density with $\theta_j = (\mu_j, \Sigma_j)$,
   $$\Big((2\pi)^d \det(\Sigma_j)\Big)^{-1/2} \exp(-\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j)).$$
   2.2 **(M step).** Hold assignments fixed, optimally update parameters: for every $j \in \{1, \ldots, k\}$,
   $$\pi_j := \frac{\sum_i A_{ij}}{n},$$
   $$\mu_j := \frac{\sum_i A_{ij} x_i}{n\pi_j},$$
   $$\Sigma_j := \frac{\sum_i A_{ij}(x_i - \mu_j)(x_i - \mu_j)^\top}{n\pi_j}.$$

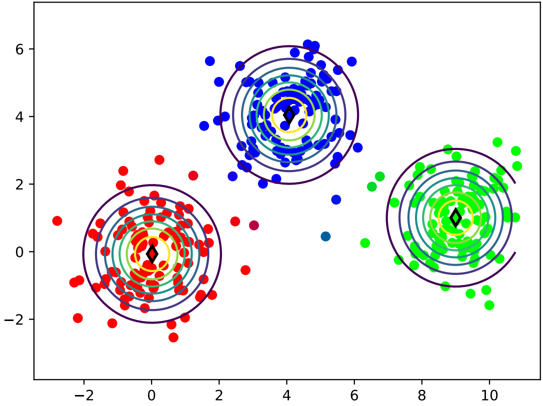
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



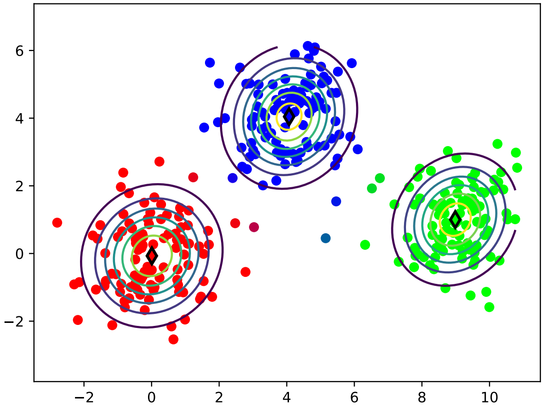
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



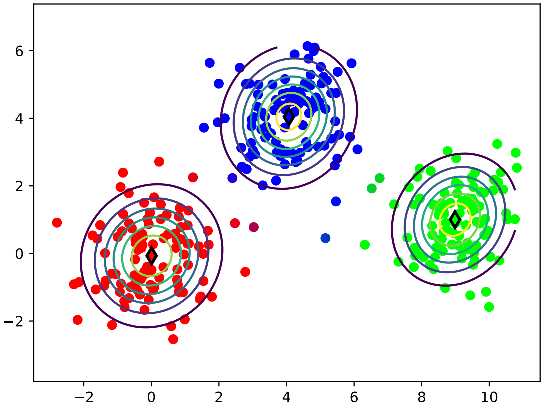
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data. . .



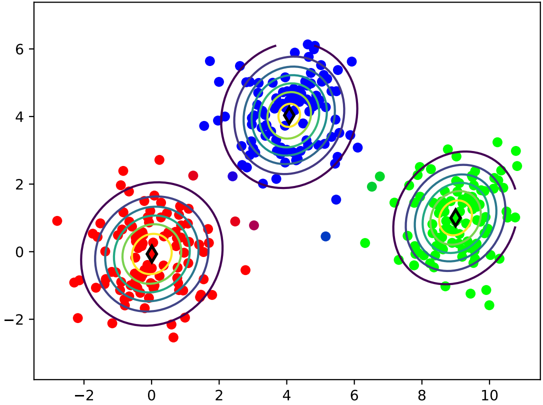
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



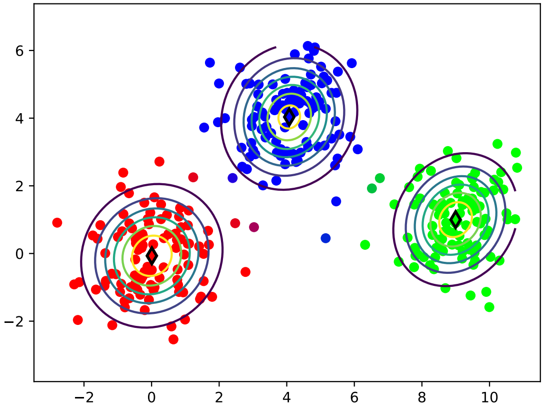
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



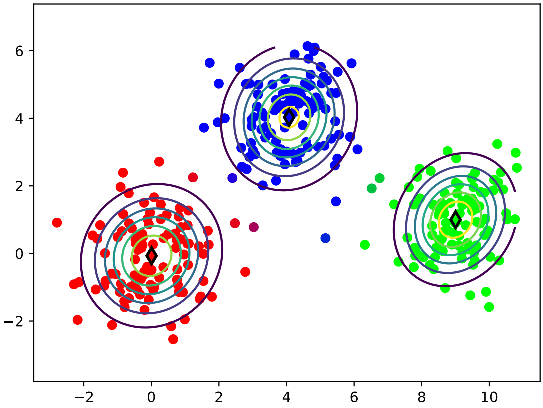
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



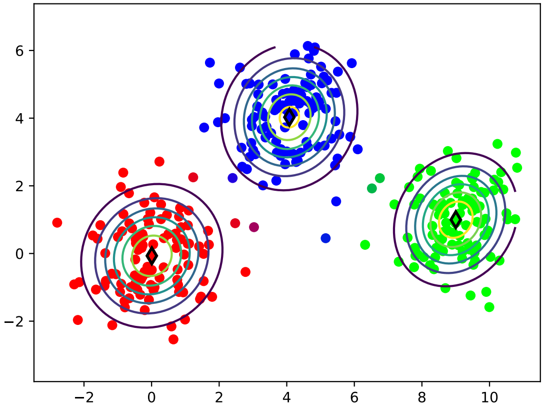
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



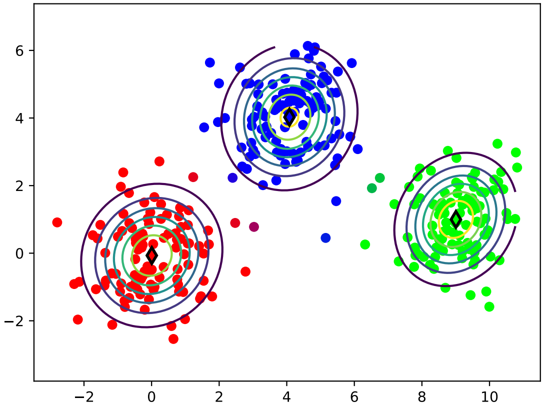
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data. . .



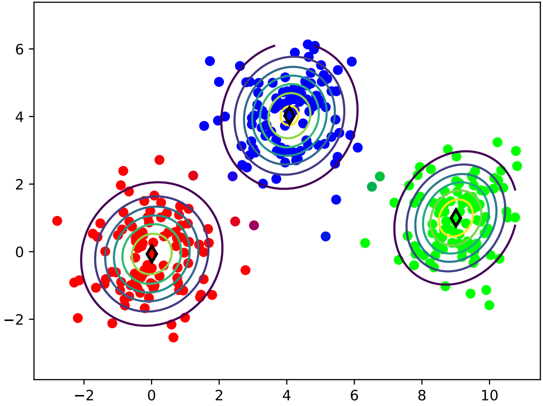
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



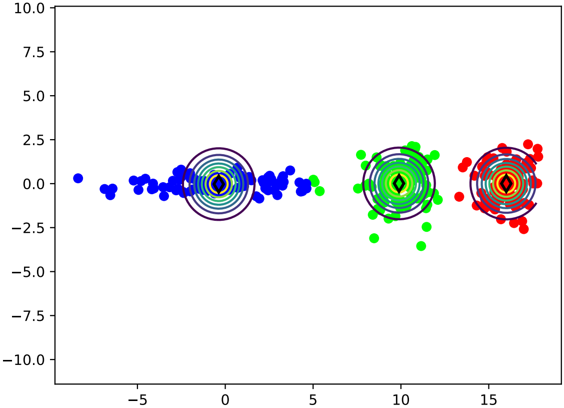
E-M for Gaussian Mixture Model (GMM).

Fine with spherical data...



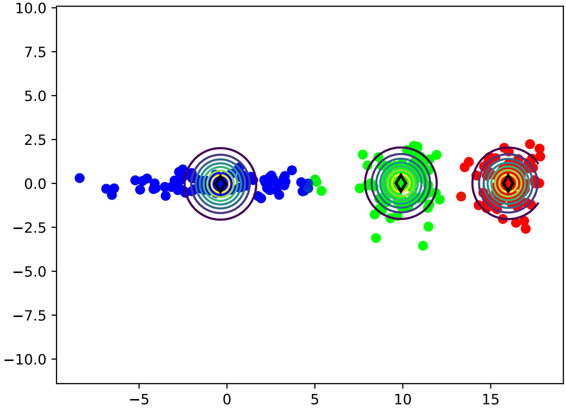
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



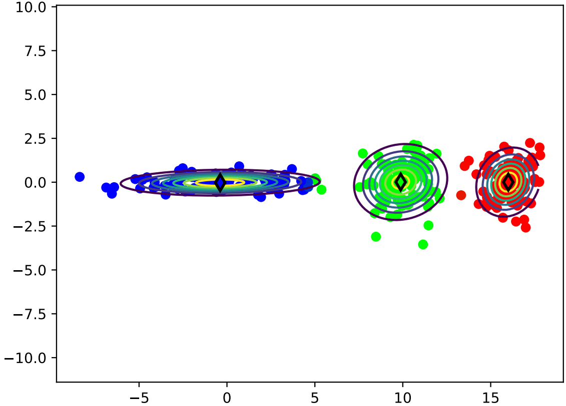
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



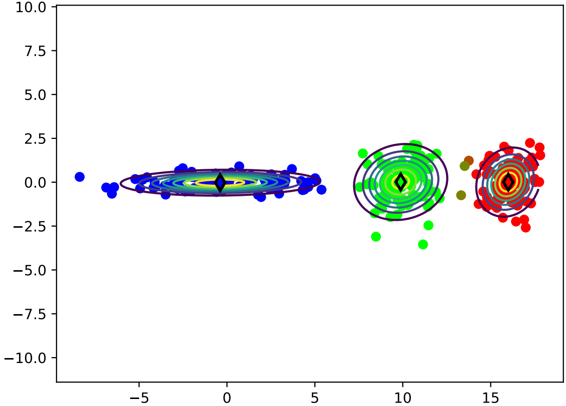
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



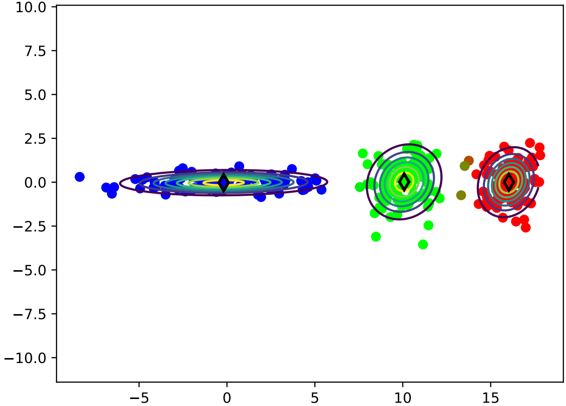
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



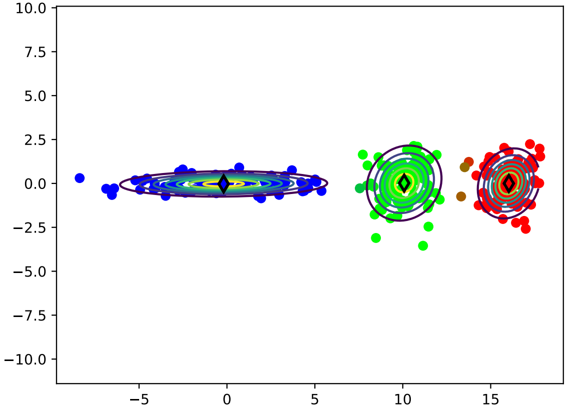
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



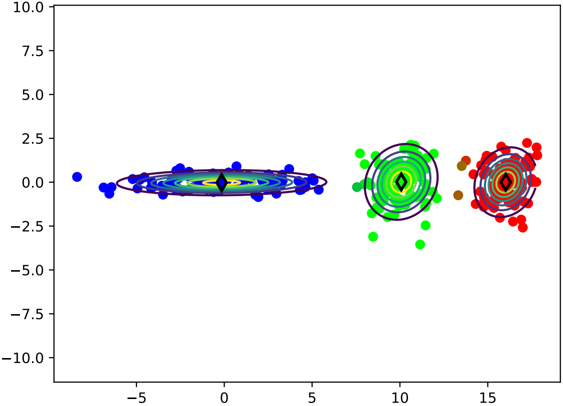
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



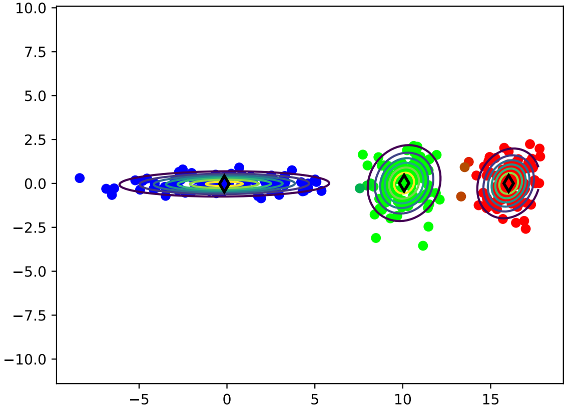
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



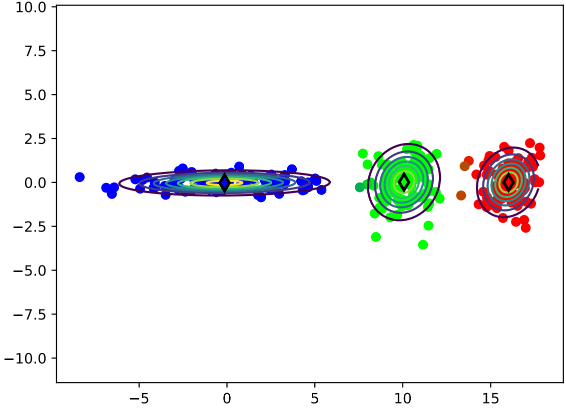
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



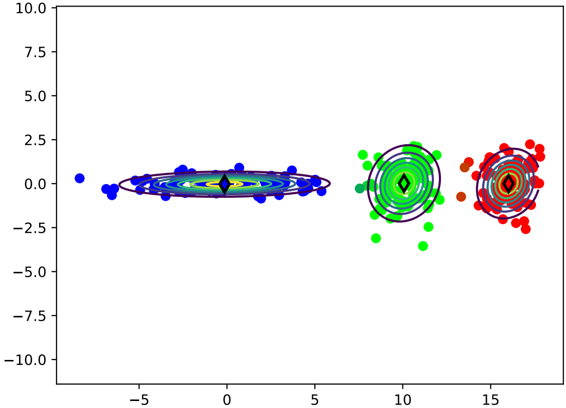
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



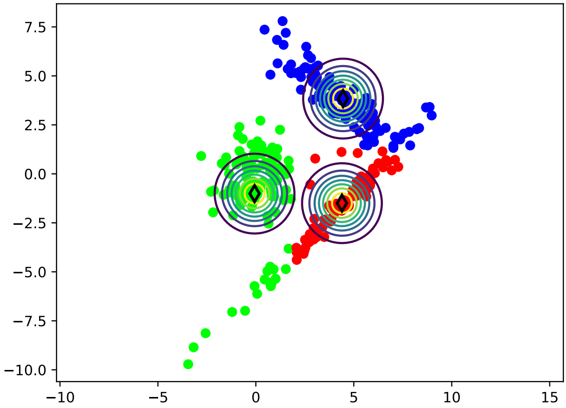
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



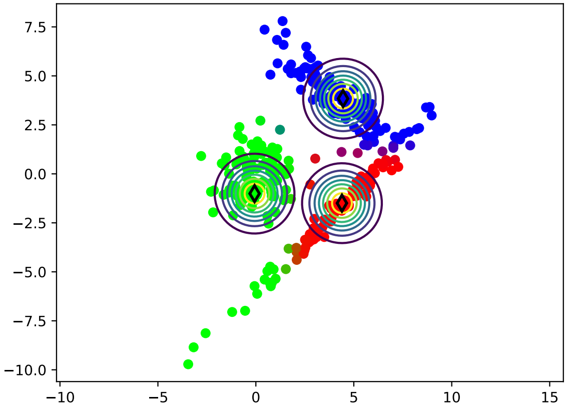
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



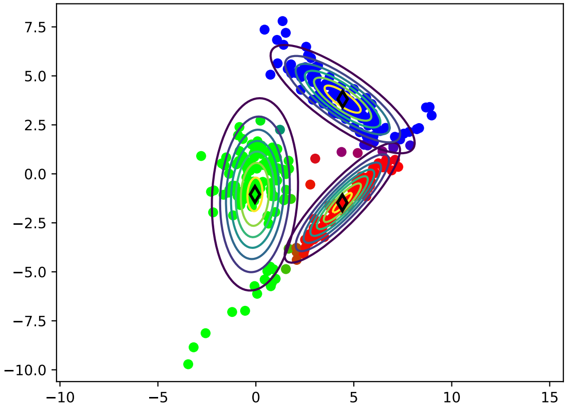
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



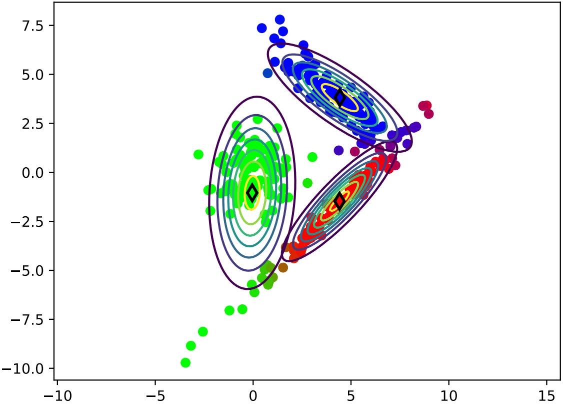
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



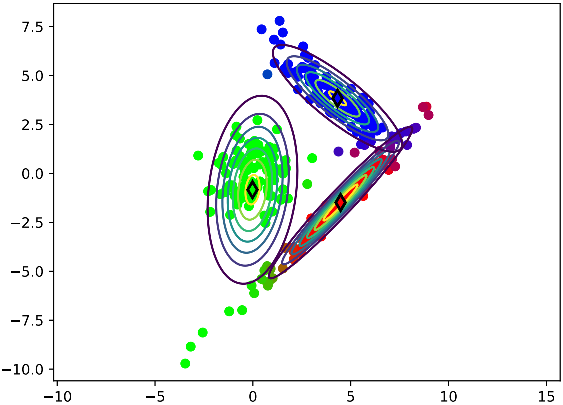
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



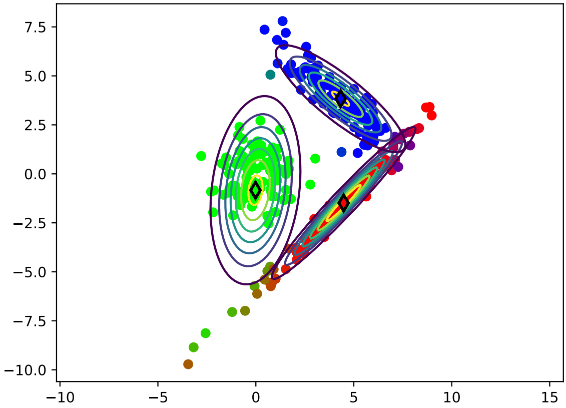
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



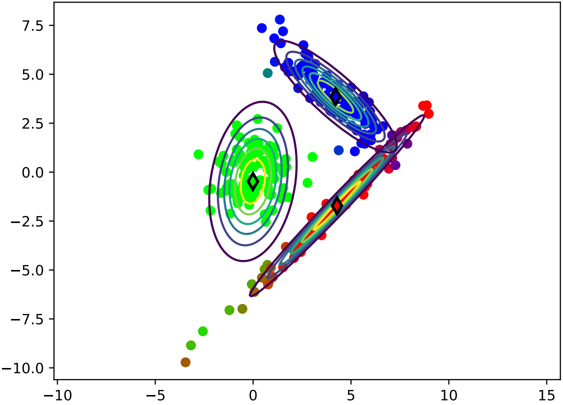
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



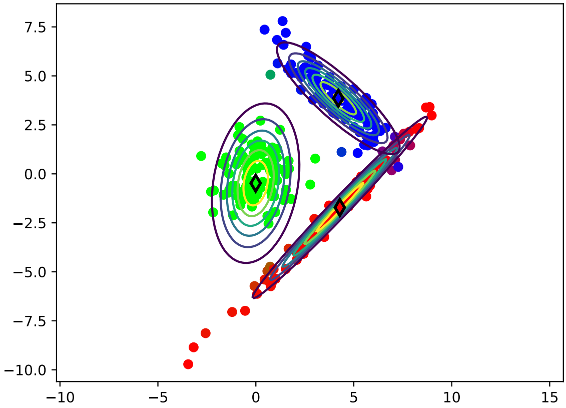
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



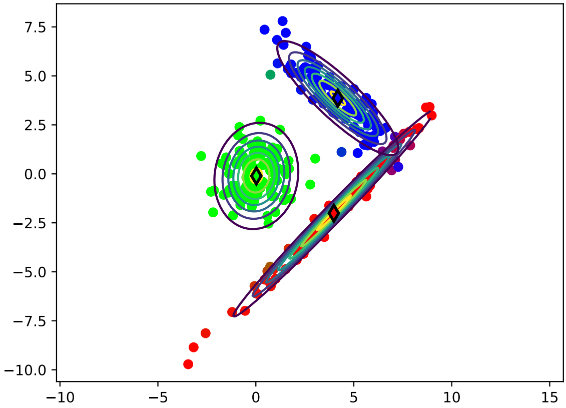
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



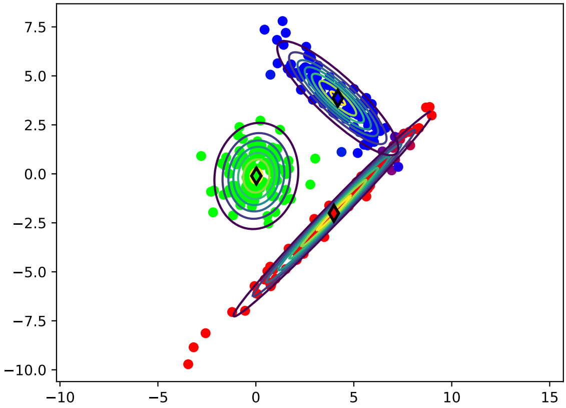
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



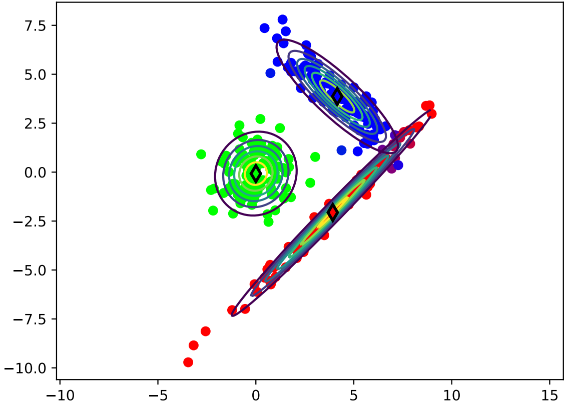
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



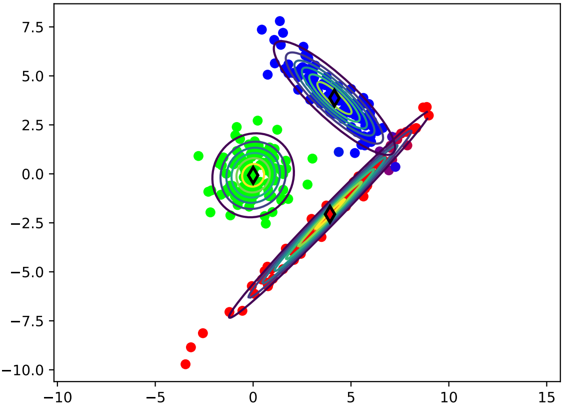
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



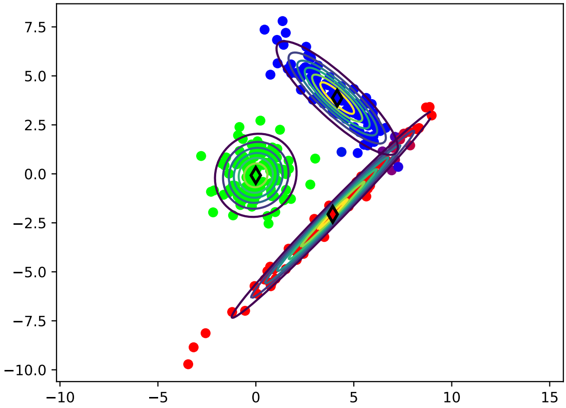
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



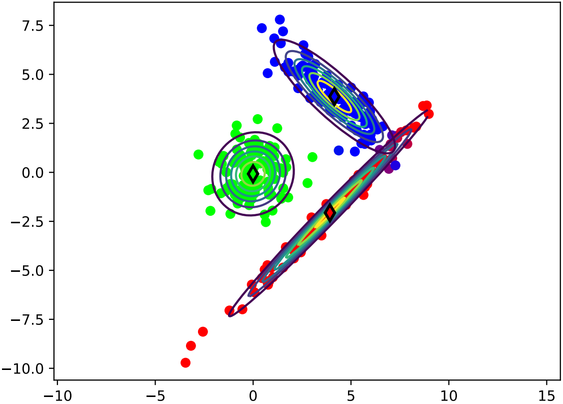
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



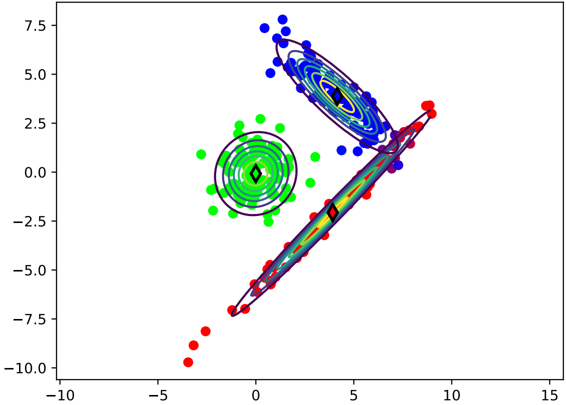
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



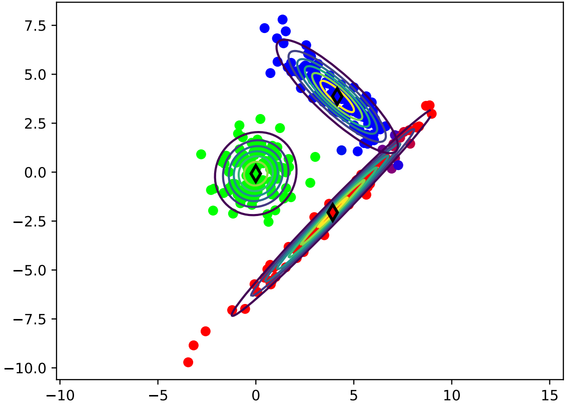
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



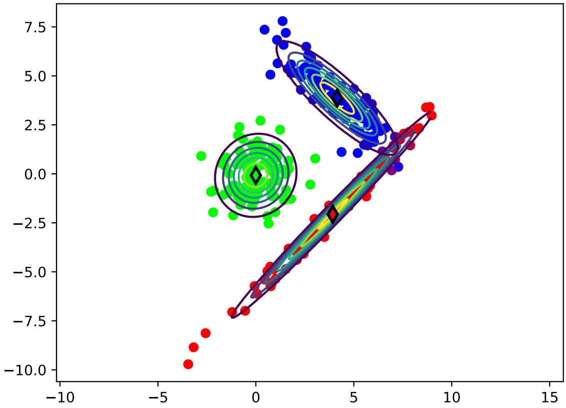
E-M for Gaussian Mixture Model (GMM).

...and elliptical data



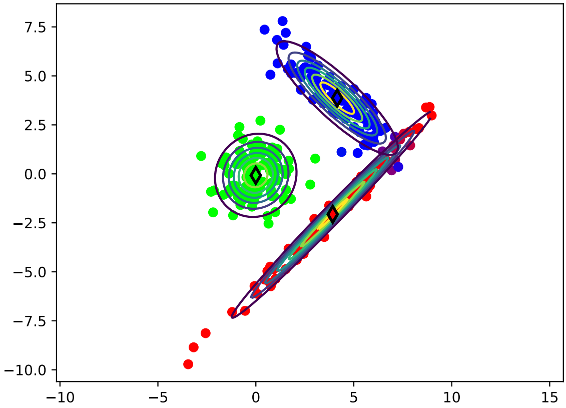
E-M for Gaussian Mixture Model (GMM).

. . . and elliptical data



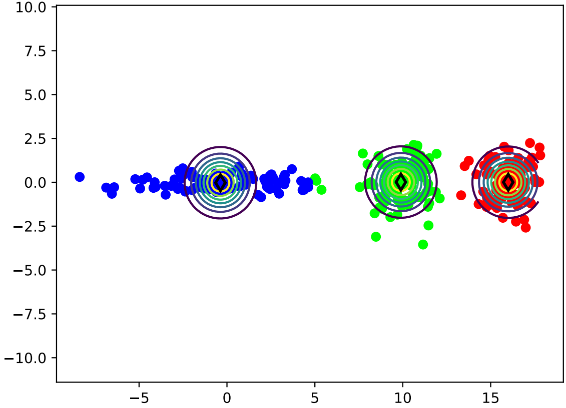
E-M for Gaussian Mixture Model (GMM).

... and elliptical data



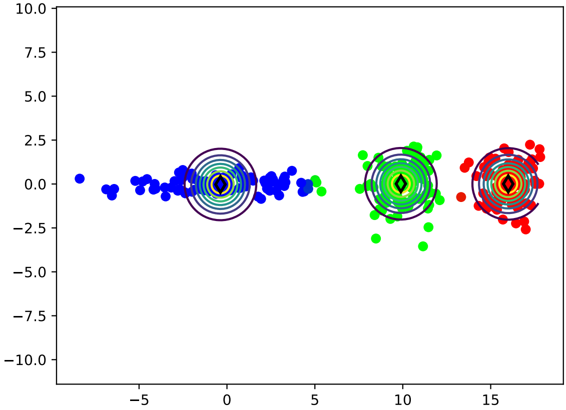
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



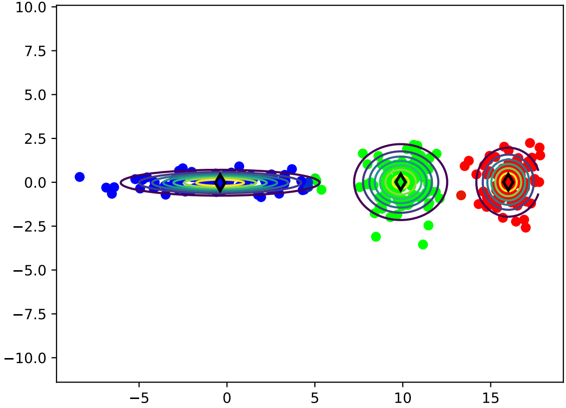
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



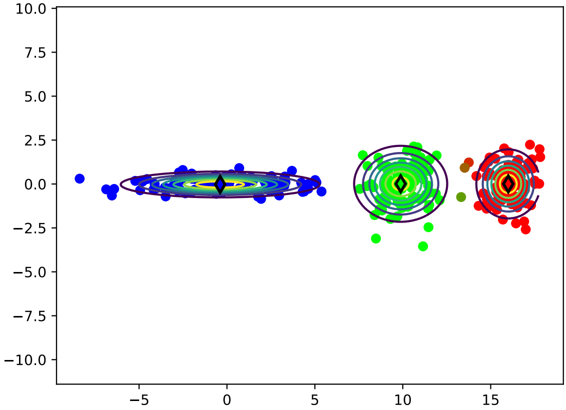
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



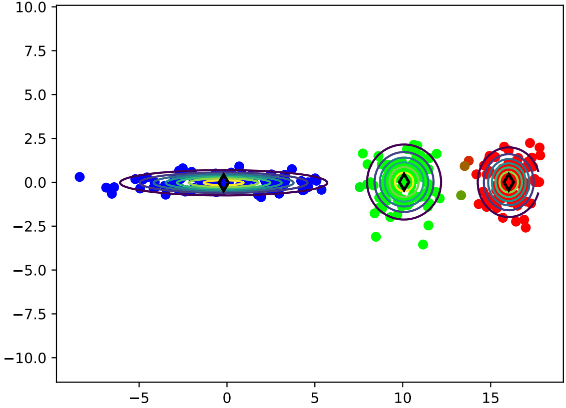
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



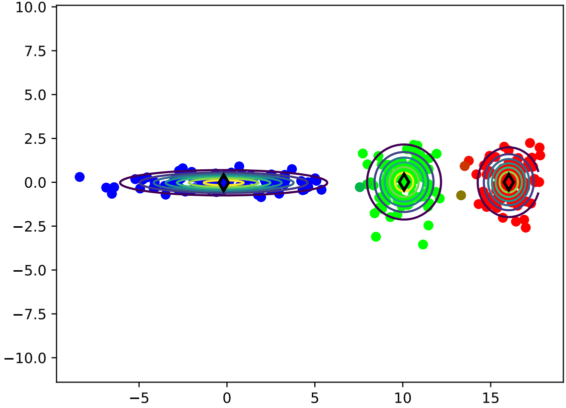
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



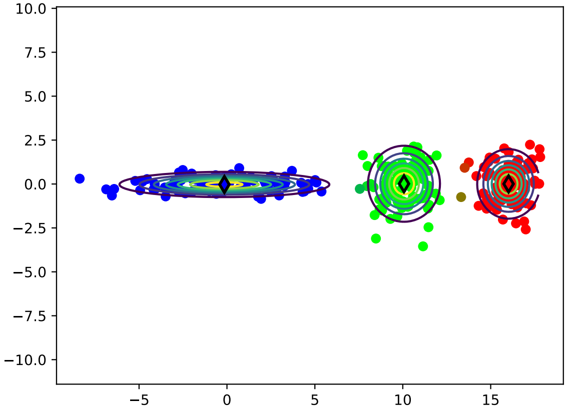
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



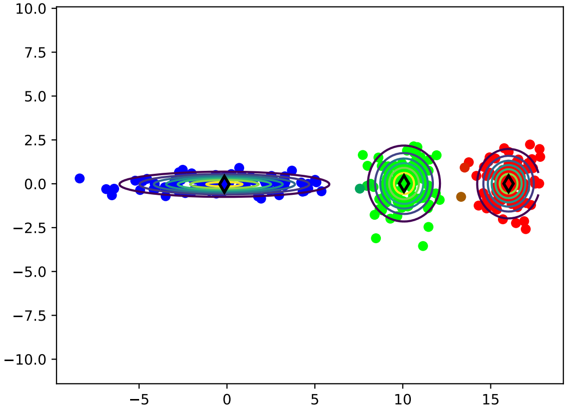
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



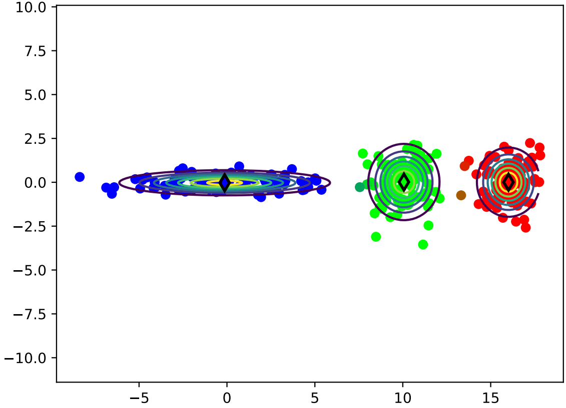
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



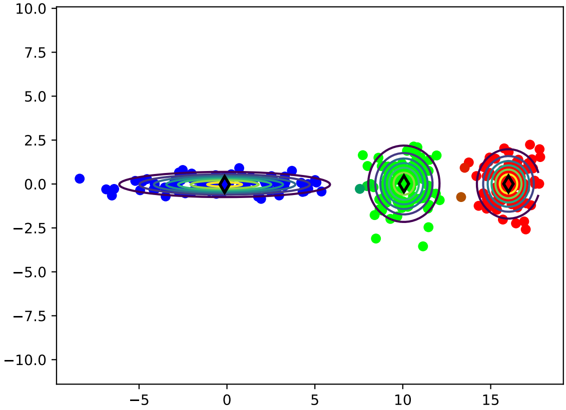
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



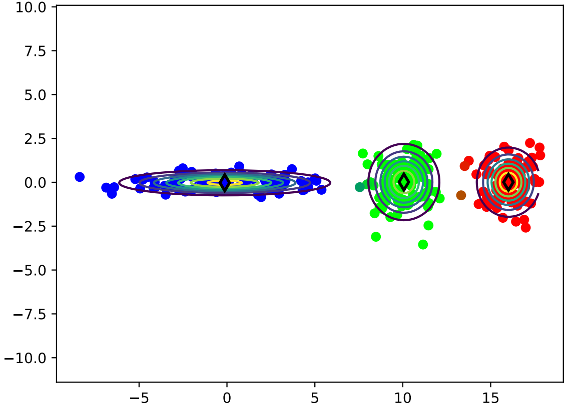
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



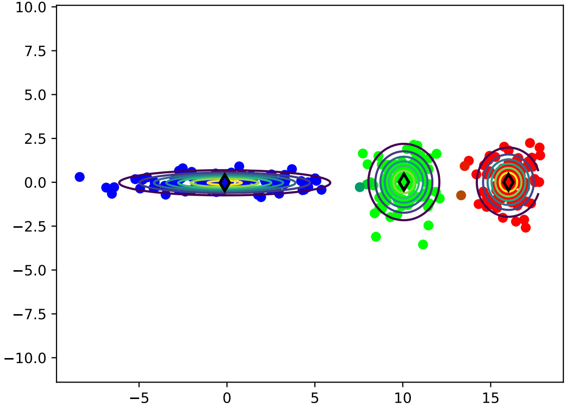
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



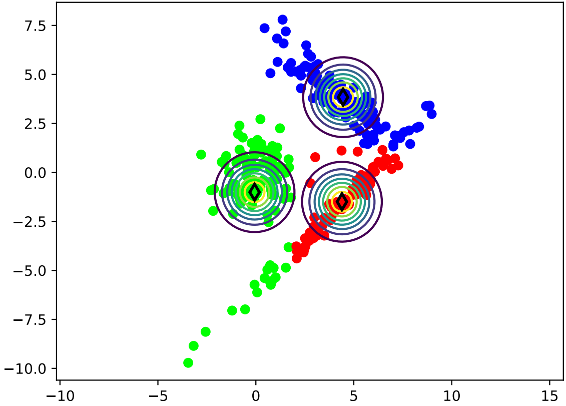
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



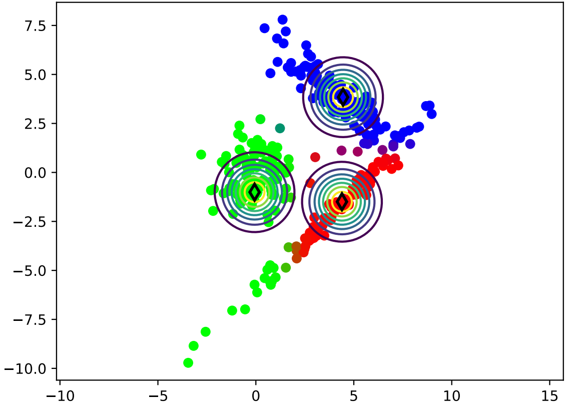
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



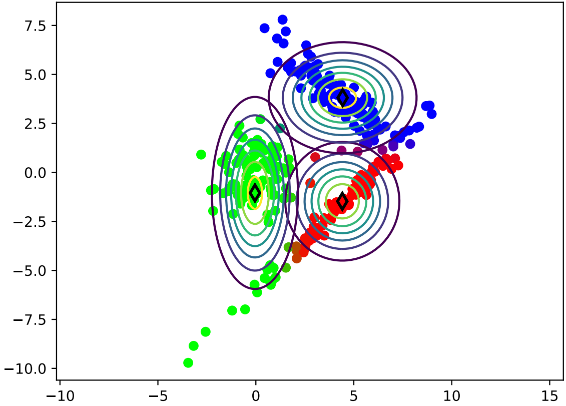
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



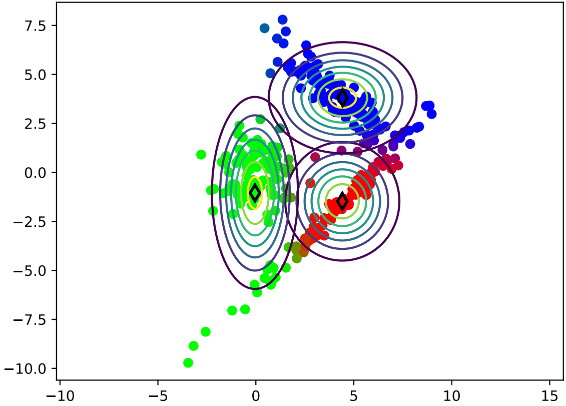
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



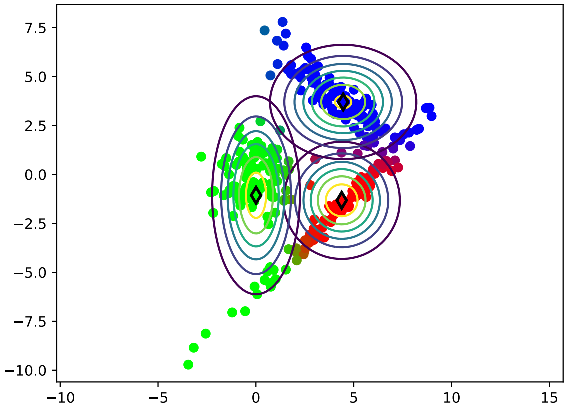
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



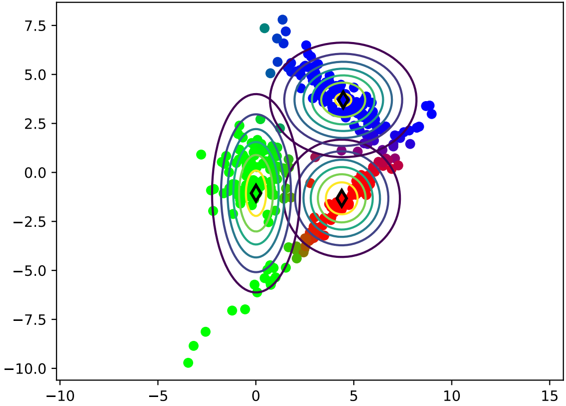
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



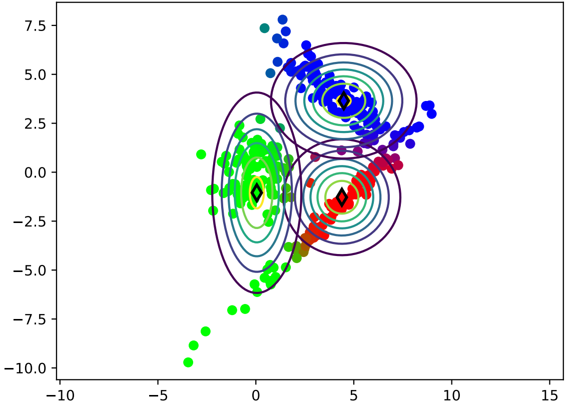
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



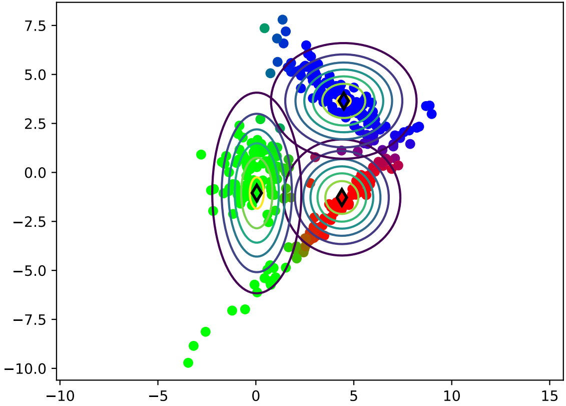
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



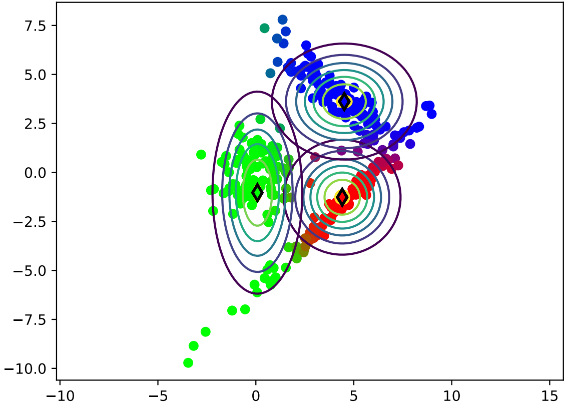
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



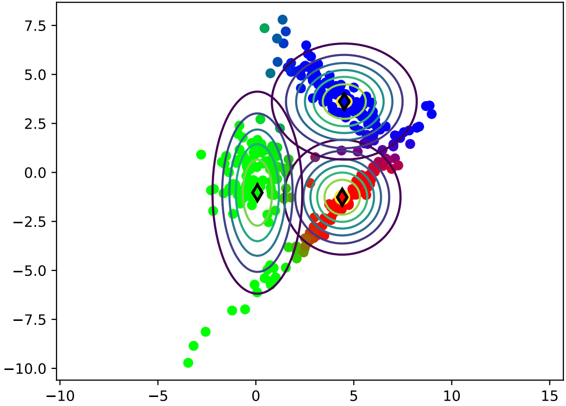
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



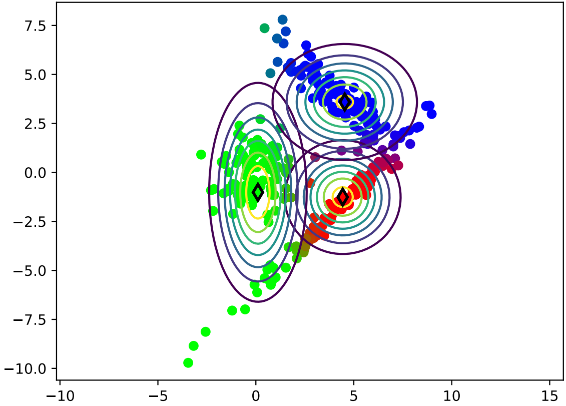
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



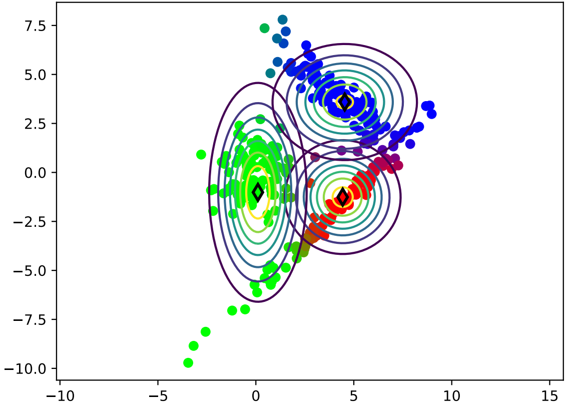
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



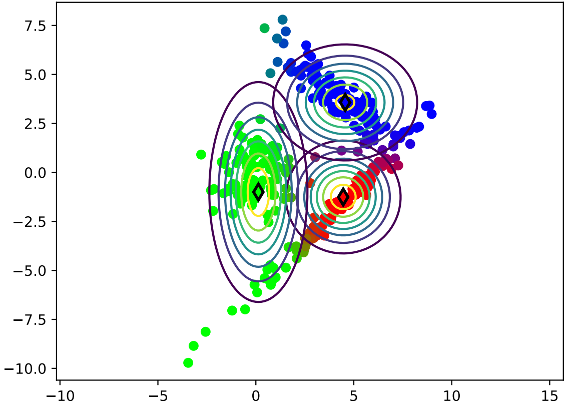
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



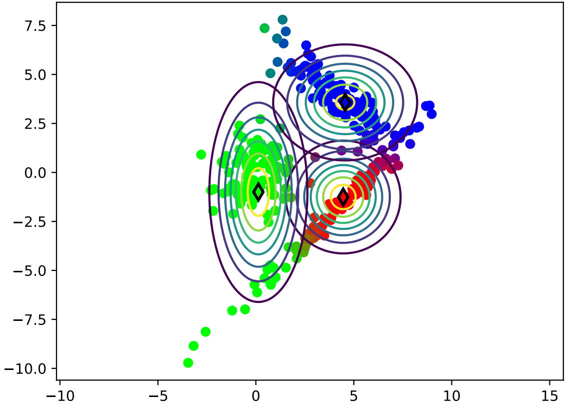
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



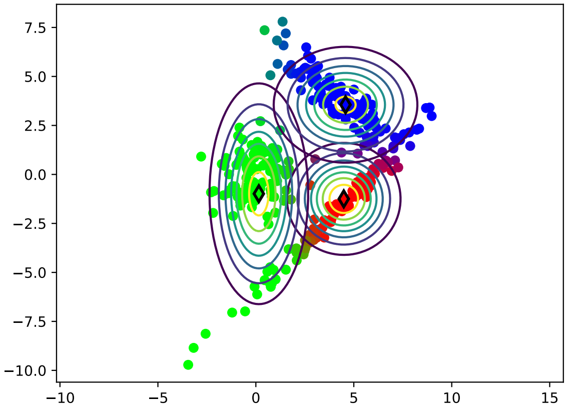
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



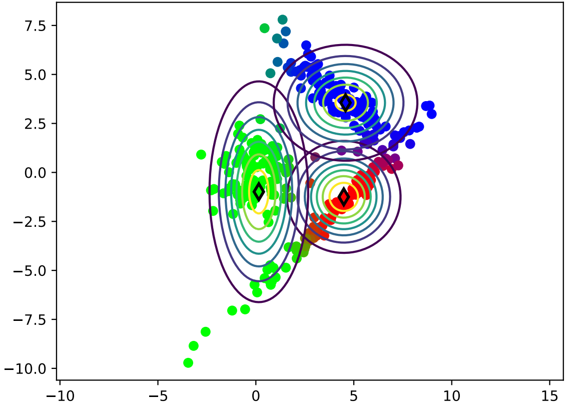
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



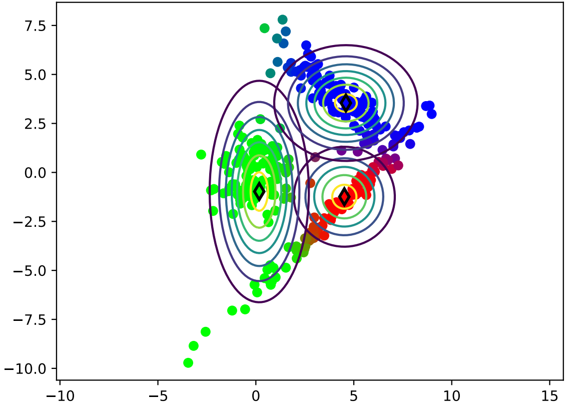
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



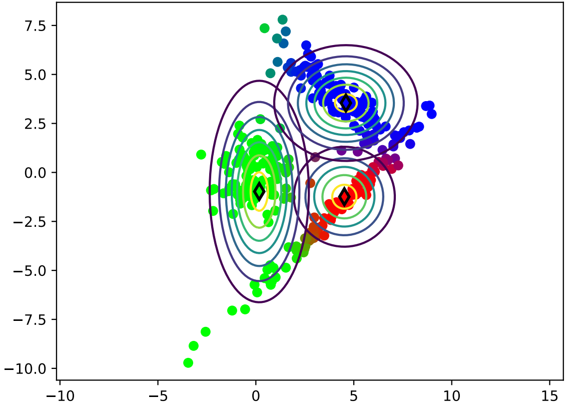
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



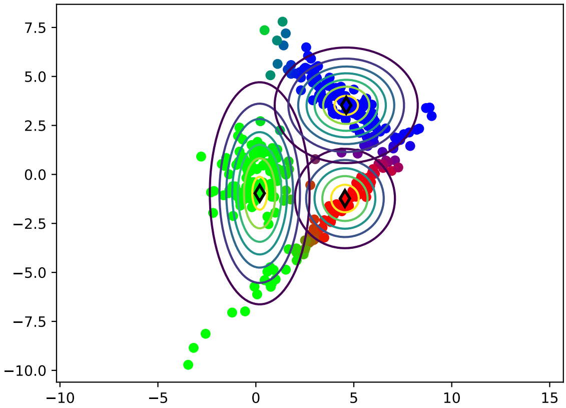
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



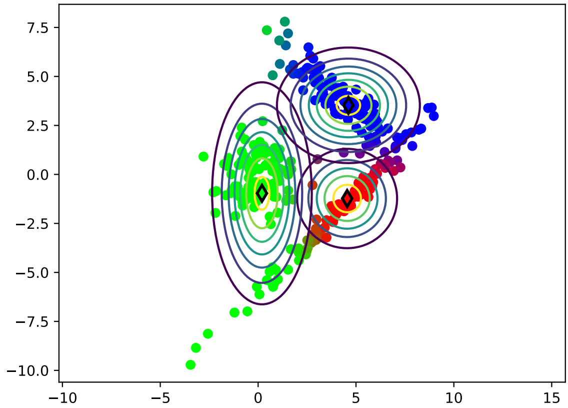
Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices.
(Why?)



Gaussian Mixture Model with diagonal covariances.

Note: GMMs often simplified via diagonal covariance matrices. (Why?)



Interpolating between k-means and GMMS.

Start with GMM but unweighted and with identity covariance.

1. Choose initial parameters $\theta = ((1/k, \mu_1, cI), \ldots, (1/k, \mu_k, cI))$.
2. Alternate the following two steps until convergence:
   2.1 **(E step)** (Reassignment). Hold parameters fixed, optimally update soft assignments $A \in [0,1]^{n \times k}$, $A\mathbf{1}_k = \mathbf{1}_n$: for every $i \in \{1, \ldots, n\}$,
   $$A_{ij} \propto \pi_j p_{\theta_j}(x_i),$$
   where p_{θ_j} is the gaussian density with $\theta_j = (\mu_j, \Sigma_j)$,
   $$\left((2\pi)^d \det(\Sigma_j)\right)^{-1/2} \exp(-\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j)).$$
   2.2 **(M step).** Hold assignments fixed, optimally update parameters: for every $j \in \{1, \ldots, k\}$,
   $$\pi_j := \frac{1}{k}$$
   $$\mu_j := \frac{\sum_i A_{ij} x_i}{\sum_i A_{ij}}$$
   $$\Sigma_j := cI.$$

Interpolating between k-means and GMMS.

Consider the E-step, fix $i \in \{1, \ldots, n\}$, define $r_{ij} := \frac{1}{2}\|x_i - \mu_j\|^2$. Then

$$A_{ij} = \frac{\pi_j p_{\theta_j}(x_i)}{\sum_l \pi_l p_{\theta_l}(x_i)} = \frac{e^{-r_{ij}/c}}{\sum_l e^{-r_{il}/c}} = \frac{1}{1 + \sum_{l \neq j} e^{(r_{ij} - r_{il})/c}}$$

Suppose $m_i := \arg\min_j r_{ij}$ unique. Then

$$\lim_{c \downarrow 0} \sum_{l \neq j} e^{\frac{r_{ij} - r_{il}}{c}} = \infty \cdot \mathbb{1}[j \neq m_i].$$

That is, A_{ij} becomes **hard assignment** as $c \downarrow 0$..

Interpolating between k-means and GMMS.

Consider the E-step, fix $i \in \{1, \ldots, n\}$, define $r_{ij} := \frac{1}{2}\|x_i - \mu_j\|^2$.
Then

$$A_{ij} = \frac{\pi_j p_{\theta_j}(x_i)}{\sum_l \pi_l p_{\theta_l}(x_i)} = \frac{e^{-r_{ij}/c}}{\sum_l e^{-r_{il}/c}} = \frac{1}{1 + \sum_{l \neq j} e^{(r_{ij} - r_{il})/c}}$$

Suppose $m_i := \arg\min_j r_{ij}$ unique. Then

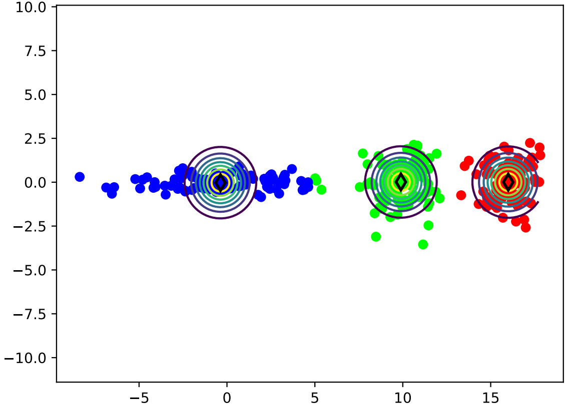
$$\lim_{c \downarrow 0} \sum_{l \neq j} e^{\frac{r_{ij} - r_{il}}{c}} = \infty \cdot \mathbb{1}[j \neq m_i].$$

That is, A_{ij} becomes **hard assignment** as $c \downarrow 0$..

In summary, k-means is obtained from E-M on GMMs via: uniform mixture weights, diagonal covariances cI with $c \downarrow 0$.

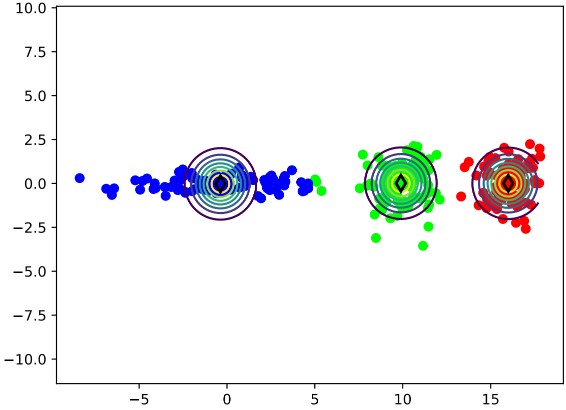
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



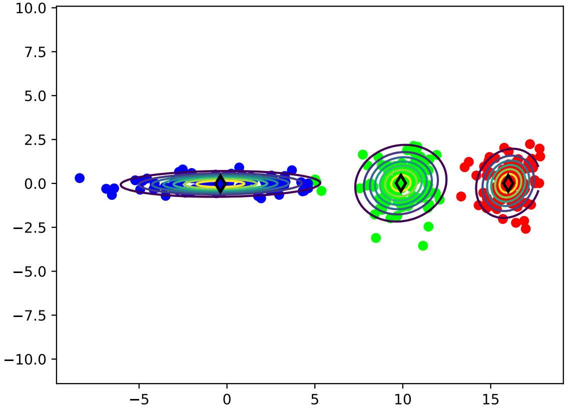
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



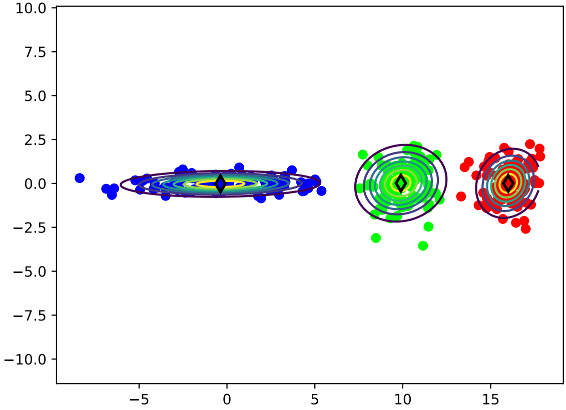
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



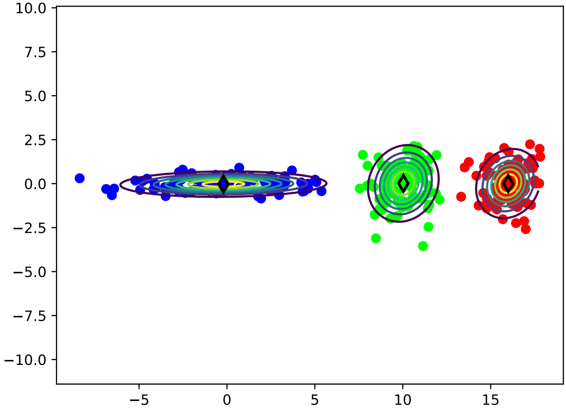
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



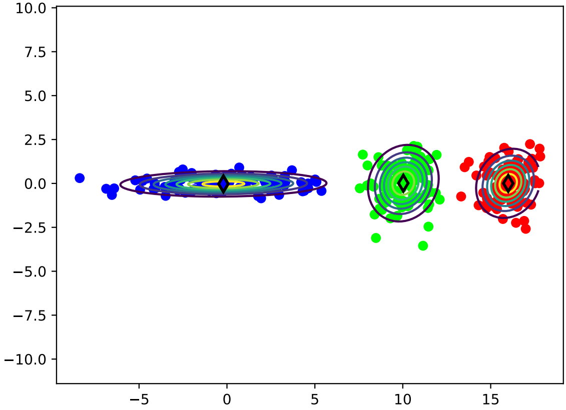
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



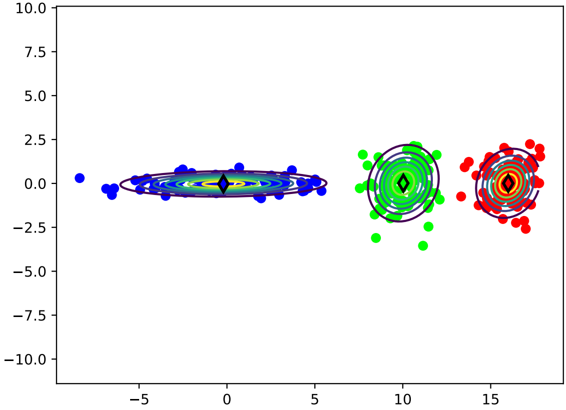
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



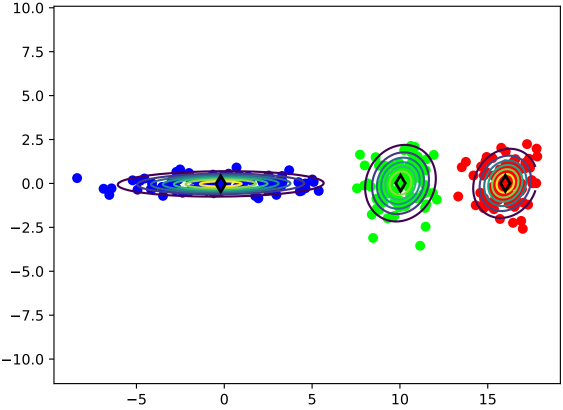
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



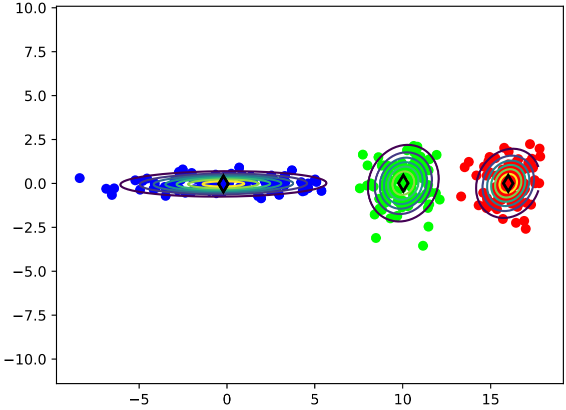
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



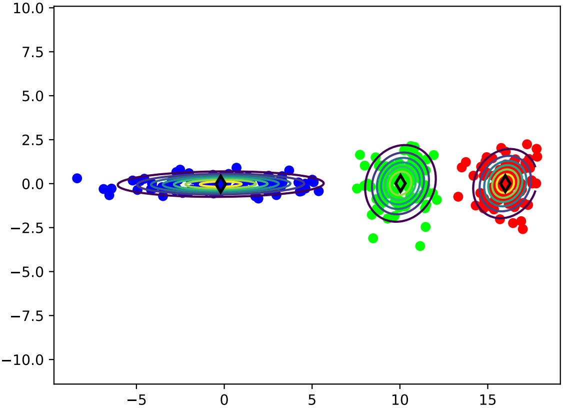
k-means with elliptical clusters.

Can use the same updates to derive elliptical *k*-means.



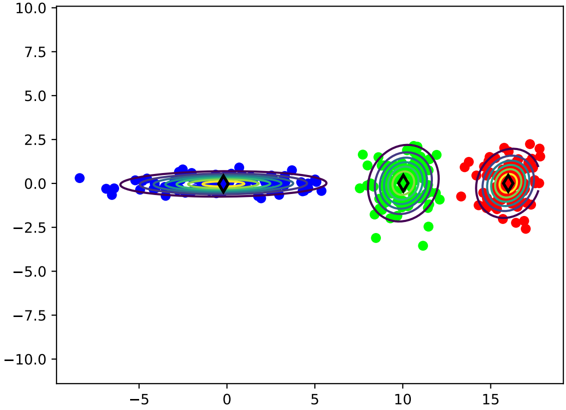
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



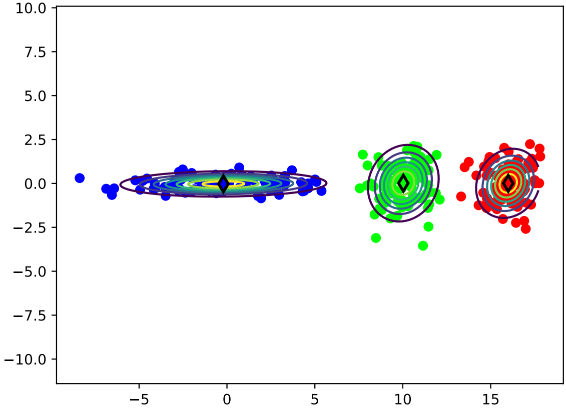
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



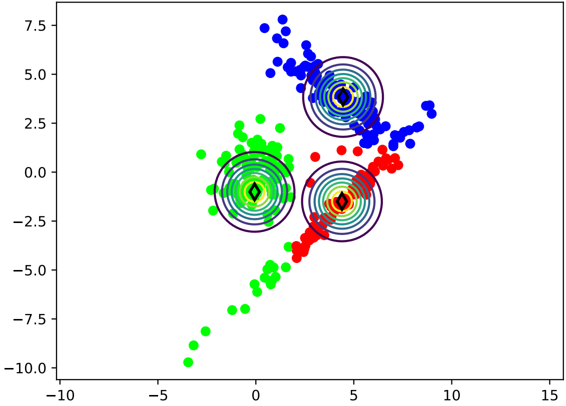
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



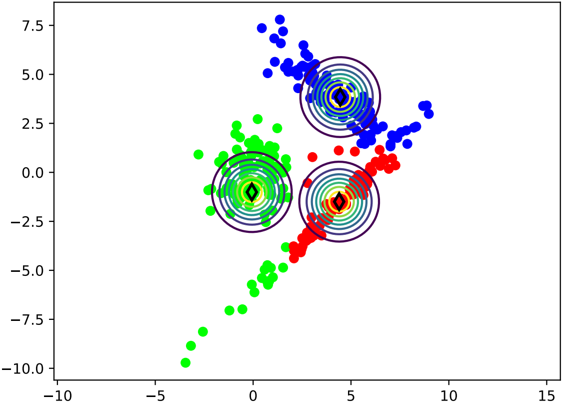
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



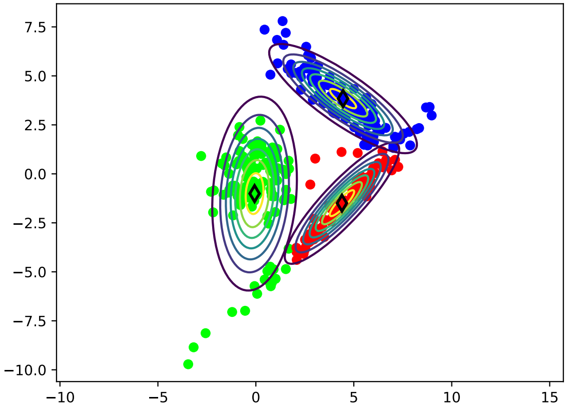
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



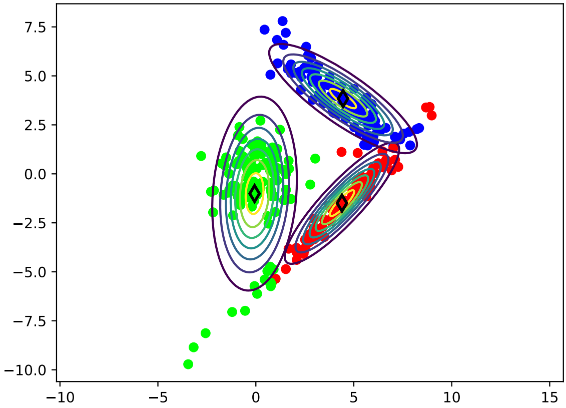
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



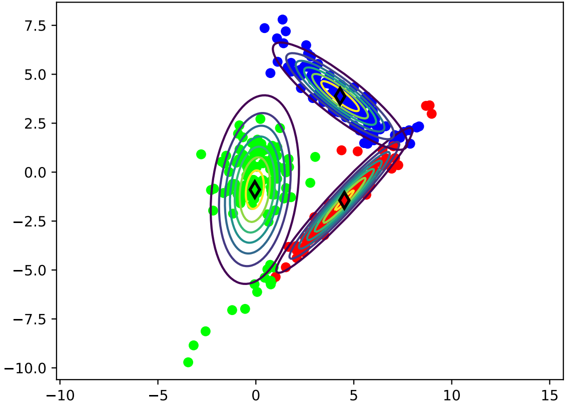
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



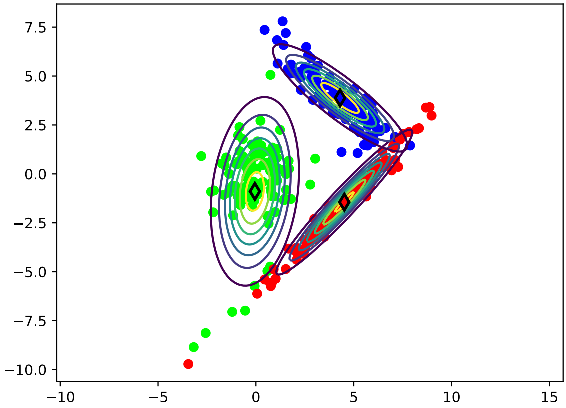
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



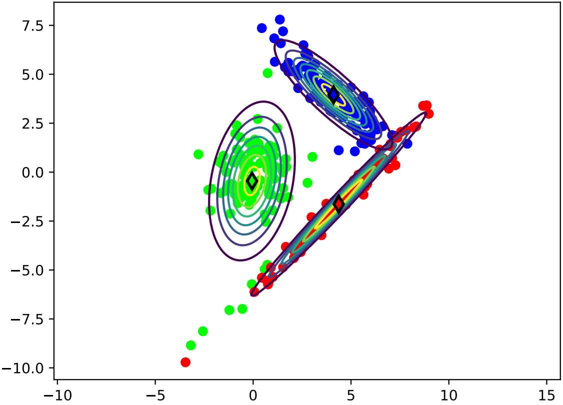
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



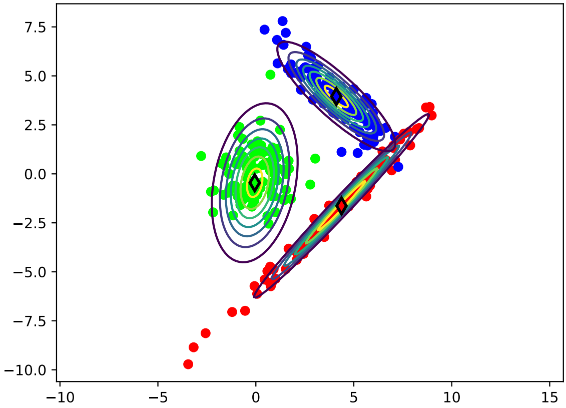
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



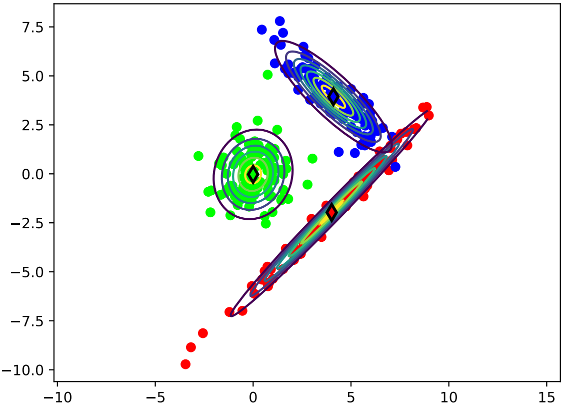
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



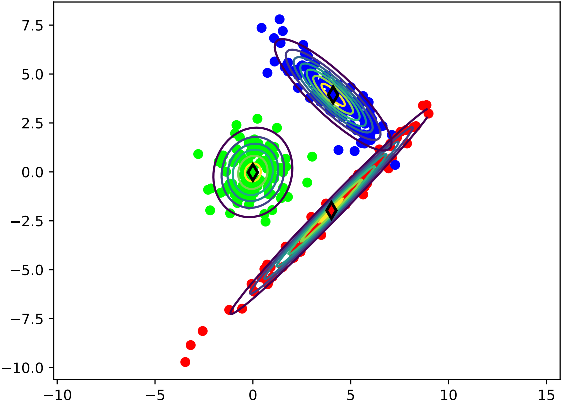
k-means with elliptical clusters.

Can use the same updates to derive elliptical *k*-means.



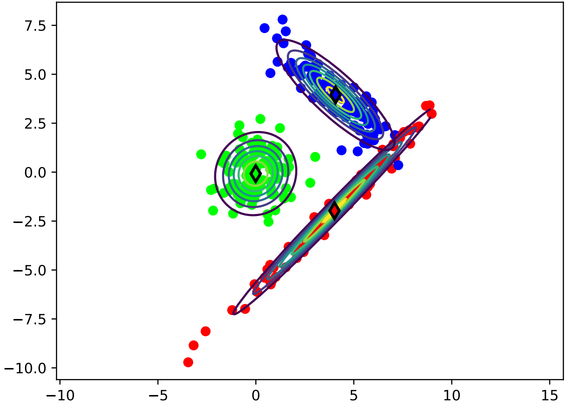
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



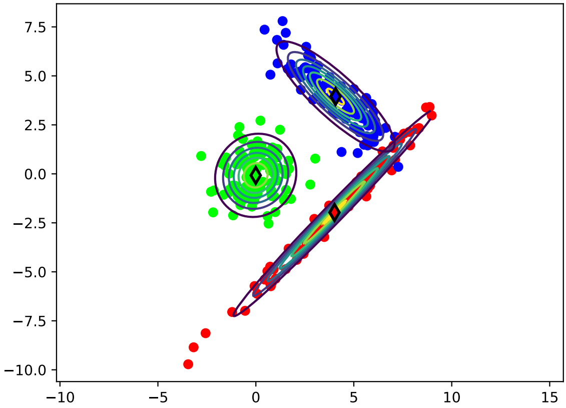
k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



k-means with elliptical clusters.

Can use the same updates to derive elliptical k-means.



Expectation-Maximization (E-M) in general.

Expectation-Maximization (E-M) in general.

The goal in E-M for GMMs was to find parameters $\theta = (\theta_1, \ldots, \theta_k)$ to maximize the log-likelihood of the data:

$$\ell(\theta) := \sum_{i=1}^{n} \ln p_\theta(x_i) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j p_{\theta_j}(x_i) \right).$$

Question: why didn't we just solve this directly?

Expectation-Maximization (E-M) in general.

The goal in E-M for GMMs was to find parameters $\theta = (\theta_1, \ldots, \theta_k)$ to maximize the log-likelihood of the data:

$$\ell(\theta) := \sum_{i=1}^{n} \ln p_\theta(x_i) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{k} \pi_j p_{\theta_j}(x_i) \right).$$

Question: why didn't we just solve this directly?

Answer: the summation inside the ln is annoying (unless $k = 1 \ldots$).
We can run gradient ascent, but we'll see E-M also increases ℓ.

What did E-M do?

Define **latent** or **hidden** variables $(y_1, \ldots, y_n)$ identifying which Gaussian generated x_i. If we knew them, estimating parameters θ would be easy.

E-M approach:
maintain $A \in \mathbb{R}^{n \times k}$;
with A given, estimating parameters was easy,
and then $A_{ij} \propto p_\theta(x_i, y_j) \ldots$

What did E-M do?

Define **latent** or **hidden** variables $(y_1, \ldots, y_n)$ identifying which Gaussian generated x_i. If we knew them, estimating parameters θ would be easy.

E-M approach:
maintain $A \in \mathbb{R}^{n \times k}$;
with A given, estimating parameters was easy,
and then $A_{ij} \propto p_\theta(x_i, y_j) \ldots$ **Why?**

What did E-M do?

Define **latent** or **hidden** variables $(y_1, \ldots, y_n)$ identifying which Gaussian generated x_i. If we knew them, estimating parameters θ would be easy.

E-M approach:
maintain $A \in \mathbb{R}^{n \times k}$;
with A given, estimating parameters was easy,
and then $A_{ij} \propto p_\theta(x_i, y_j)\ldots$ **Why?**

Some adjusted notation:
$p_\theta(y_j) = \pi_j$, whereas $p_\theta(x_i | y_j)$ replaces old $p_{\theta_j}(x_i)$.
(I'll unify this tonight.)

Some notation.

Define likelihood $\ell(\theta)$ and a helper $\underline{\ell}(A, \theta)$:

$$\ell(\theta) := \sum_{i=1}^{n} \ln p_\theta(x_i),$$

$$\underline{\ell}(A, \theta) := \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \ln \frac{p_\theta(x_i, y_j)}{A_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \left(\ln p_\theta(x_i, y_j) - \ln A_{ij} \right).$$

- The term $\sum_{i,j} A_{ij} \ln A_{ij}$ does not affect $\arg\max \underline{\ell}(A, \theta)$.

Some notation.

Define likelihood $\ell(\theta)$ and a helper $\underline{\ell}(A,\theta)$:

$$\ell(\theta) := \sum_{i=1}^{n} \ln p_\theta(x_i),$$

$$\underline{\ell}(A,\theta) := \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \ln \frac{p_\theta(x_i, y_j)}{A_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \left(\ln p_\theta(x_i, y_j) - \ln A_{ij} \right).$$

- The term $\sum_{i,j} A_{ij} \ln A_{ij}$ does not affect $\arg\max \underline{\ell}(A,\theta)$.
- Moreover, given Lagrangian (with Lagrange multipliers $\alpha \in \mathbb{R}^n$)

$$\sum_{i=1}^{n} \alpha_i (\sum_j A_{ij} - 1) + \underline{\ell}(A,\theta),$$

applying $\frac{\mathrm{d}}{\mathrm{d}A_{ij}}$ and setting to 0 gives

$$\alpha_i + \ln p_\theta(x_i, y_j) - \ln A_{ij} - 1 = 0 \qquad \Longrightarrow \qquad A_{ij} = p_\theta(x_i, y_j) e^{\alpha_i - 1},$$

Some notation.

Define likelihood $\ell(\theta)$ and a helper $\underline{\ell}(A, \theta)$:

$$\ell(\theta) := \sum_{i=1}^{n} \ln p_\theta(x_i),$$

$$\underline{\ell}(A, \theta) := \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \ln \frac{p_\theta(x_i, y_j)}{A_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{k} A_{ij} \left(\ln p_\theta(x_i, y_j) - \ln A_{ij} \right).$$

- The term $\sum_{i,j} A_{ij} \ln A_{ij}$ does not affect $\arg\max \underline{\ell}(A, \theta)$.
- Moreover, given Lagrangian (with Lagrange multipliers $\alpha \in \mathbb{R}^n$)

$$\sum_{i=1}^{n} \alpha_i (\sum_j A_{ij} - 1) + \underline{\ell}(A, \theta),$$

applying $\frac{\mathrm{d}}{\mathrm{d}A_{ij}}$ and setting to 0 gives

$$\alpha_i + \ln p_\theta(x_i, y_j) - \ln A_{ij} - 1 = 0 \qquad \Longrightarrow \qquad A_{ij} = p_\theta(x_i, y_j) e^{\alpha_i - 1},$$

- Thus M and E steps maximize $\underline{\ell}$! **Is this useful?**

E-M theorem.

Theorem. Define

$$\begin{aligned}
\ell(\theta) &:= \sum_{i=1}^{n} \ln p_\theta(x_i), \\
\underline{\ell}(A,\theta) &:= \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij} \ln \frac{p_\theta(x_i, y_j)}{A_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij} \left(\ln p_\theta(x_i, y_j) - \ln A_{ij}\right), \\
(A_\theta)_{ij} &\propto p_\theta(x_i, y_j), \\
\theta' &:= \arg\max_{\theta} \underline{\ell}(A,\theta).
\end{aligned}$$

E-M theorem.

Theorem. Define

$$\begin{aligned}
\ell(\theta) &:= \sum_{i=1}^{n} \ln p_\theta(x_i), \\
\underline{\ell}(A,\theta) &:= \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij} \ln \frac{p_\theta(x_i,y_j)}{A_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij} \left(\ln p_\theta(x_i,y_j) - \ln A_{ij}\right), \\
(A_\theta)_{ij} &\propto p_\theta(x_i,y_j), \\
\theta' &:= \arg\max_\theta \underline{\ell}(A,\theta).
\end{aligned}$$

Then $\underline{\ell}(A,\theta) \le \ell(\theta)$, and

$$\ell(\theta) = \underline{\ell}(A_\theta,\theta) \le \underline{\ell}(A_\theta,\theta') \le \underline{\ell}(A_{\theta'},\theta') = \underline{\ell}(\theta')$$

and in particular

$$\ell(\theta_1) \le \ell(\theta_2) \le \ell(\theta_3) \le \cdots .$$

Proof 1/2.

By Jensen,

$$\begin{aligned}
\underline{\ell}(A,\theta) &= \sum_{i=1}^{n}\sum_{j=1}^{k} A_{ij}\ln\left(\frac{p_\theta(x_i,y_j)}{A_{ij}}\right) \\
&\le \sum_{i=1}^{n}\ln\left(\sum_{j=1}^{k} A_{ij}\frac{p_\theta(x_i,y_j)}{A_{ij}}\right) \\
&= \sum_{i=1}^{n}\ln\left(\sum_{j=1}^{k} p_\theta(x_i,y_j)\right) \\
&= \ell(\theta).
\end{aligned}$$

Proof 2/2.

On the other hand,

$$\begin{aligned}
\underline{\ell}(A_\theta, \theta) &= \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{p_\theta(x_i, y_j)}{p_\theta(x_i)} \ln \left(p_\theta(x_i, y_j) \left(\frac{p_\theta(x_i)}{p_\theta(x_i, y_j)} \right) \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{p_\theta(x_i, y_j)}{p_\theta(x_i)} \ln p_\theta(x_i) \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{k} \frac{p_\theta(x_i, y_j)}{p_\theta(x_i)} \right) \ln p_\theta(x_i) \\
&= \ell(\theta).
\end{aligned}$$

Soft vs Hard assignment.

E-M increases $\underline{\ell}$ *and* ℓ.
Hard assignment increases $\underline{\ell}$; not clear for ℓ.

Key points.

- E-M can be derived as alternating maximization of $\underline{\ell}$.
- E-M can be shown to given non-decreasing likelihood ℓ.

Remarks.

Remarks.

- Easy to run E-M on more complicated latent variable models; write down $\underline{\ell}$ and do alternating maximization.
  Example: mixtures of other distributions.

Remarks.

- Easy to run E-M on more complicated latent variable models; write down $\underline{\ell}$ and do alternating maximization.
  Example: mixtures of other distributions.

- Latent variables are useful/magical;
  ℓ was not tractable, but $\underline{\ell}$ was tractable.

Remarks.

- Easy to run E-M on more complicated latent variable models;
  write down $\underline{\ell}$ and do alternating maximization.
  Example: mixtures of other distributions.

- Latent variables are useful/magical;
  ℓ was not tractable, but $\underline{\ell}$ was tractable.

- Not perfect: "singularities", local optima and slow convergence, sensitivity to initialization, ...

Summary.

Things to know.

- k-means objective, Lloyd's method, hard assignment.
- Log-likelihood of GMM, E-M for GMM, soft assignment.
- E-M is alternating maximization of $\underline{\ell}$, increases ℓ.