# MTH 511a - 2020: Lecture 18

Instructor: Dootika Vats

## 1 The EM algorithm

An important application of the MM algorithm is the Expectation-Maximization (EM) algorithm. However, the EM algorithm is an integral part of statistical algorithms, and hence we study it separately. We will first motivate the EM algorithm with an example.

### 1.1 Gaussian mixture likelihood

Suppose $X_1, X_2, \dots, X_n \sim F$, where $F$ is mixture of normal distribution so that the density is:

$$f(x|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*) = \pi^* f_1(x|\mu_1, \sigma_1^2) + (1 - \pi^*) f_2(x|\mu_2, \sigma_2^2)\,,$$

where $f_i(x|\mu_i, \sigma_i^2)$ is the density of $N(\mu_i, \sigma_i^2)$ distribution for $i = 1, 2$. Given the data, we will to find the maximum likelihood estimates of all 5 parameters: $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*)$. That is, we want to maximize:

$$\begin{aligned} l(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*|X) &= \sum_{i=1}^{n} \log f(x_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^*) \\ &= \sum_{i=1}^{n} \log \left[\pi^* f_1(x|\mu_1, \sigma_1^2) + (1 - \pi^*) f_2(x|\mu_2, \sigma_2^2)\right]\,. \end{aligned}$$

There is no analytical solution to the above optimization problem and we have to resort to numerical techniques. Instead of trying to use gradient-based tools, we use a common trick called the *latent variable* or *missing data* trick.

Recall is the data likelihood is a mixture of Gaussians. So with probability $\pi^*$, any observed $X_i$ is from $f_1$ and with probability $1 - \pi^*$ it is from $f_2$. Suppose we have the information that $x_i$ is coming from which of the two populations. Thus, suppose the *complete data* was of the form

$$(X_1, Z_1), (X_2, Z_2), \dots, (X_n, Z_n)\,,$$

where each $Z_i = k$ means that $X_i$ is from population $k$. If this complete data is available to us, then first note that the joint density is

$$f(x_i, z_i = k) = f(x_i | z_i = k) \Pr(Z_i = k)\,.$$

Suppose $\mathcal{D}_1 = \{i : 1 \leq i \leq n, z_i = 1\}$ and $\mathcal{D}_2 = \{i : 1 \leq i \leq n, z_i = 2\}$, with cardinality $d_1$ and $d_2$ respectively. The set $\mathcal{D}_1$ and $\mathcal{D}_\in$ have the indices of the data that belong to each component of the mixture. Then the likelihood from the full data is

$$\begin{aligned}
&L(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi^* | X) \\
&= \prod_{i=1}^{n} f(x_i, z_i) \\
&= \prod_{i \in \mathcal{D}_1} f(x_i, z_i = 1) \prod_{j \in \mathcal{D}_2} f(x_i, z_i = 2) \\
&= \prod_{i \in \mathcal{D}_1} f(x_i | z_i = 1) \Pr(Z_i = 1) \prod_{i \in \mathcal{D}_2} f(x_i | z_i = 2) \Pr(Z_i = 2) \\
&= \prod_{i \in \mathcal{D}_1} \left[\pi^* f_1(x_i | \mu_1, \sigma_1^2)\right] \prod_{j \in \mathcal{D}_2} \left[f_2(x_i | \mu_2, \sigma_2^2)(1 - \pi^*)\right] \\
&= (\pi^*)^{d_1} (1 - \pi^*)^{d_2} \prod_{i \in \mathcal{D}_1} \left[f_1(x_i | \mu_1, \sigma_1^2)\right] \prod_{i \in \mathcal{D}_2} \left[f_2(x_i | \mu_2, \sigma_2^2)\right]\,.
\end{aligned}$$

This means that the log likelihood is

$$\Rightarrow \log L = d_1 \log(\pi^*) + d_2 \log(1 - \pi^*) + \sum_{i \in \mathcal{D}_1} \log f_1(x_i | \mu_1, \sigma_1^2) + \sum_{i \in \mathcal{D}_2} \log f_2(x_i | \mu_2, \sigma_2^2)\,.$$

This log-likelihood is in a far nicer format, so that closed-form estimates are available.

So, if the complete data was available to us, we can easily find the MLE of all the 5 parameters. Unfortunately, the $Z$s are usually not observed, and only the $X$s have been observed. The EM algorithm will solve this problem by estimating the unobserved $z_i$ corresponding to each $x_i$ in an iterative manner.

We will come back this Gaussian problem again.

## 1.2 The Expectation-Maximization Algorithm

Suppose, we have a vector of parameters $\theta$, and wee have observed the marginal data $X_1, \dots, X_n$ from the complete data $(X_i, Z_i)$. The objective function is to maximize is

$$l(\theta | X) = \log \int f(\mathbf{x}, \mathbf{z} | \theta) d\nu_z\,,$$

where the $\int \cdot \, d\nu_z$ denotes integral or summation based on whether $Z$ is continuous or discrete. The EM algorithm iterates through the following: Consider a starting value $\theta_0$. Then for any $k+1$ iteration

1. **E-Step**: Compute

$$q(\theta; \theta_{(k)}) = \mathrm{E}_{Z|x}\Big[\log f(\mathbf{x}, \mathbf{z}|\theta) \mid X = x, \theta_{(k)}\Big]$$

   where the expectation is computed with respect to the conditional distribution of $Z$ given $X = x$ for the current iterate $\theta_{(k)}$.

2. **M-Step**: Compute

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} q(\theta; \theta_k)\,.$$

**Theorem 1.** *The EM algorithm is an MM algorithm and thus has the ascent property.*

*Proof.* The objective function is $\log f(\mathbf{x}|\theta)$. We will construct a minorizing function $\tilde{f}(\theta|\theta_{(k)})$ on this objective function and show that

$$\tilde{f}(\theta|\theta_{(k)}) = q(\theta|\theta_{(k)}) \, + \text{constants}.$$

The, maximizing $\tilde{f}(\theta|\theta_{(k)})$ is equivalent to maximizing $q(\theta|\theta_{(k)})$. Let

$$\tilde{f}(\theta|\theta_{(k)}) = \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta)\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz + \log f(\mathbf{x}|\theta_{(k)}) - \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz\,.$$

(The proof technique is setup for continuous $Z$, but the same proof works for discrete $Z$ as well).

Naturally, we can see that at $\theta = \theta_{(k)}$, $\tilde{f}(\theta|\theta_{(k)}) = f(\theta_{(k)})$. We will now show that minorizing property.

$$\begin{aligned}
&\tilde{f}(\theta|\theta_{(k)})\\
&= \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta)\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz + \log f(\mathbf{x}|\theta_{(k)}) - \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz\\
&= \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta)\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz + \int_z \log f(\mathbf{x}|\theta_{(k)}) f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz - \int_z \log\{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) dz\\
&= \int_z \log\left\{\frac{f(\mathbf{x}, \mathbf{z}|\theta)\, f(\mathbf{x}|\theta_{(k)})}{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})}\right\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)})\\
&= \int_z \log\left\{\frac{f(\mathbf{x}, \mathbf{z}|\theta)\, f(\mathbf{x}|\theta_{(k)})}{f(\mathbf{x}, \mathbf{z}|\theta_{(k)})}\right\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) + \log f(\mathbf{x}|\theta) - \log f(\mathbf{x}|\theta)\\
&= \int_z \log\left\{\frac{f(\mathbf{x}, \mathbf{z}|\theta)\, f(\mathbf{x}|\theta_{(k)})}{f(\mathbf{x}, \mathbf{z}|\theta_{(k)}) f(\mathbf{x}|\theta)}\right\} f(\mathbf{z}|\mathbf{x}, \theta_{(k)}) + \log f(\mathbf{x}|\theta)
\end{aligned}$$

By Jensen's inequality,

$$
\begin{aligned}
&\leq \log\left[\int_z \left\{\frac{f(\mathbf{x},\mathbf{z}|\theta)\, f(\mathbf{x}|\theta_{(k)})}{f(\mathbf{x},\mathbf{z}|\theta_{(k)}) f(\mathbf{x}|\theta)}\right\} f(\mathbf{z}|\mathbf{x},\theta_{(k)})\right] + \log f(\mathbf{x}|\theta)\\
&= \log\left[\int_z \frac{f(\mathbf{z}|\mathbf{x},\theta)}{f(\mathbf{z}|\mathbf{x},\theta_{(k)})} f(\mathbf{z}|\mathbf{x},\theta_{(k)})\right] + \log f(\mathbf{x}|\theta)\\
&= \log \int_z f(\mathbf{z}|\mathbf{x},\theta) dz + \log f(\mathbf{x}|\theta)\\
&= \log f(\mathbf{x}|\theta)\,.
\end{aligned}
$$

Thus, $\tilde{f}(\theta|\theta_{(k)})$ is a minorizing function, and the next iterate is

$$\theta_{(k+1)} = \arg\max_{\theta} \tilde{f}(\theta|\theta_{(k)}) = \arg\max_{\theta} q(\theta|\theta_{(k)})$$

□

## 1.3 (Back to) Gaussian mixture likelihood

We will look at the general setup of $C$ groups, so that the density for $X_1, \dots, X_n$ is

$$f(x|\theta) = \sum_{j=1}^{C} \pi_j f_j(x \mid \mu_j, \sigma_j^2)\,,$$

where $\theta = (\mu_1, \dots, \mu_C, \sigma_1^2, \dots, \sigma_C^2, \pi_1, \dots, \pi_{C-1})$. The setup is the same as before, and suppose we the *complete data* $(X_i, Z_i)$ where $X_i \mid Z_i = c \sim N(\mu_c, \sigma_c^2)$ and $\Pr(Z_i = c) = \pi_c$.

To implement the EM algorithm for this example, we first need to find $q(\theta|\theta_{(k)})$. First recall the conditional distribution of $Z|X$ is

$$\Pr(Z = c|X = x_i) = \frac{f(x_i|Z = c)\Pr(Z = c)}{f(x_i)} \frac{f_c(x_i|\mu_c, \sigma_c^2)\pi_c}{\sum_{j=1}^{C} f_j(x_i|\mu_j, \sigma_j^2)\pi_j} := \gamma_{i,c}\,.$$

So for any $k$th iterate with current step $\theta_{(k)} = (\mu_{1,k}, \mu_{2,k}, \sigma_{1,k}^2, \sigma_{2,k}^2, \pi_k^*)$, we have

$$\Pr(Z = c|X = x, \theta_{(k)}) = \frac{f_c(x|\mu_{c,k}, \sigma_{c,k}^2)\pi_{c,k}}{\sum_{j=1,2} f_j(x|\mu_{j,k}, \sigma_{j,k}^2)\pi_{j,k}} := \gamma_{i,c,k}\,.$$

**NOTE:** $\gamma_{i,c}$ are itself quantities of interest since they tell us the probability of the $i$th observation being in class $c$. This helps in classifying the observed data.

Next,

$$q(\theta|\theta_{(k)}) = \mathrm{E}_{Z|x}\left[\log f(x, z|\theta) \mid X = x, \theta_{(k)}\right]$$

$$
\begin{aligned}
&= \mathrm{E}_{Z|x}\left[\sum_{i=1}^{n} \log f(x_i, z_i|\theta) \mid X = x, \theta_{(k)}\right] \\
&= \sum_{i=1}^{n} \mathrm{E}_{Z_i|x_i}\left[\log f(x_i, z_i|\theta) \mid X = x_i, \theta_{(k)}\right] \\
&= \sum_{i=1}^{n}\sum_{c=1}^{C} \log\left\{f_c(x_i|\mu_c, \sigma_c^2)\pi_c\right\} \; \frac{f_c(x_i|\mu_{c,k}, \sigma_{c,k}^2)\pi_{c,k}}{\sum_{j=1,2} f_j(x_i|\mu_{j,k}, \sigma_{j,k}^2)\pi_{j,k}} \,.
\end{aligned}
$$

Although we won't be able to write the full expectation in closed form, we don't need to since all we need in the next step is to maximize the expectation. So to implement the E-step we need to update

$$
\gamma_{i,c,k} = \frac{f_c(x_i|\mu_{c,k}, \sigma_{c,k}^2)\pi_{c,k}}{\sum_{j=1,2} f_j(x_i|\mu_{j,k}, \sigma_{j,k}^2)\pi_{j,k}} \,.
$$

This completes the E-step. We move on to the M-step. To complete the M-step

$$
\theta_{(k+1)} = \arg\max q(\theta|\theta_{(k)}) \,.
$$

$$
\begin{aligned}
q(\theta|\theta_{(k)}) &= \sum_{i=1}^{n}\sum_{c=1}^{C} \left\{\log f_c(x_i|\mu_c, \sigma_c^2) + \log \pi_c\right\} \; \gamma_{i,c,k} \\
&= \sum_{i=1}^{n}\sum_{c=1}^{2} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma_c^2 - \frac{(X_i - \mu_c)^2}{2\sigma_c^2} + \log\pi_c\right]\gamma_{i,c,k} \\
&= \text{const} - \frac{1}{2}\sum_{i=1}^{n}\sum_{c=1}^{C} \log\sigma_c^2 \gamma_{i,c,k} - \sum_{i=1}^{n}\sum_{c=1}^{C} \frac{(X_i - \mu_c)^2}{2\sigma_c^2}\gamma_{i,c,k} + \sum_{i=1}^{n}\sum_{c=1}^{C} \log\pi_c \, \gamma_{i,c,k} \,.
\end{aligned}
$$

Taking derivatives and setting to 0, we get For any $c$,

$$
\frac{\partial q}{\partial \mu_c} = \sum_{i=1}^{n} \frac{(x_i - \mu_c)\,\gamma_{i,c,k}}{\sigma_c^2} \overset{\text{set}}{=} 0 \;\Rightarrow\; \mu_{c,(k+1)} = \frac{\sum_{i=1}^{n}\gamma_{i,c,k}\; x_i}{\sum_{i=1}^{n}\gamma_{i,c,k}} \tag{1}
$$

$$
\frac{\partial L}{\partial \sigma_c^2} = -\frac{1}{2}\sum_{i=1}^{n}\frac{\gamma_{i,c,k}}{\sigma_c^2} + \sum_{i=1}^{n}\frac{(X_i - \mu_c)^2}{2\sigma_c^4}\gamma_{i,c,k} \overset{\text{set}}{=} 0 \;\Rightarrow\; \sigma_{c,(k+1)}^2 = \frac{\sum_{i=1}^{n}\gamma_{i,c,k}(x_i - \mu_{c,(k+1)}^2)}{\sum_{i=1}^{n}\gamma_{i,c,k}} \tag{2}
$$

For $\pi_c$ note that the optimization requires a constraint, since $\sum_c \pi_c = 1$. So we will use Lagrange multipliers. The objective function is

$$
\tilde{q}(\theta|\theta_{(k)}) = q(\theta|\theta_{(k)}) - \lambda\left(\sum_{c=1}^{C}\pi_c - 1\right) \,.
$$

Taking derivative

$$\begin{aligned}
&\Rightarrow \frac{\partial \tilde{q}}{\partial \pi_c} = \sum_{i=1}^{n} \frac{\gamma_{i,c,k}}{\pi_c} - \lambda \overset{\text{set}}{=} 0 \\
&\Rightarrow \pi_c = \sum_{i=1}^{n} \frac{\gamma_{i,c,k}}{\lambda} \\
&\Rightarrow \sum_{c=1}^{C} \pi_c = \sum_{c=1}^{C} \sum_{i=1}^{n} \frac{\gamma_{i,c,k}}{\lambda} \\
&\Rightarrow 1 = \frac{1}{\lambda} \sum_{i=1}^{n} 1 \\
&\Rightarrow \lambda = n \\
&\Rightarrow \pi_{c,(k+1)} = \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,c,k} \,. \qquad (3)
\end{aligned}$$

Thus equations (1) and (3) provide the iterative updates for the parameters.

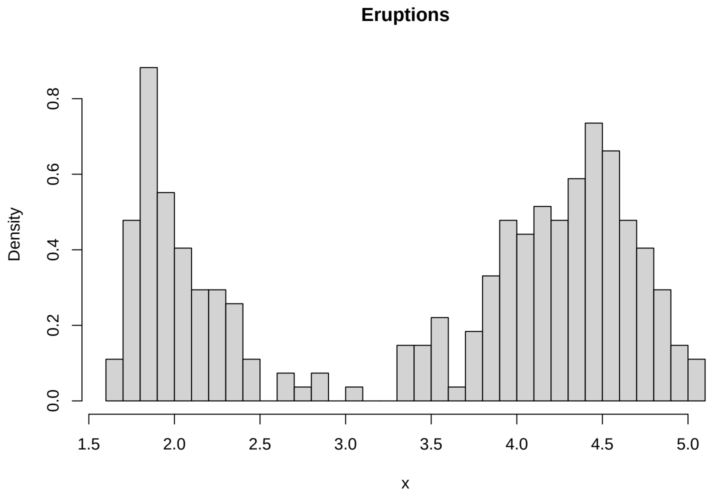
### 1.3.1 Old Geyser Eruptions

In Yellowstone National Park, Wyoming, USA, there is geyser that erupts quite often. The eruptions are either long or short. Below in data `faithful`, there are eruption times of this geyser.

```
#################################################
## Old Faithful Geyser data
#################################################
data(faithful)

head(faithful)
#  eruptions waiting
#1    3.600     79
#2    1.800     54
#3    3.333     74
#4    2.283     62
#5    4.533     85
#6    2.883     55

x <- faithful$eruptions
hist(x, breaks = 30, main = "Eruptions")
```



From the looks of it, the eruption time can be is bimodal. There's a mode at around 2 seconds and another at around 4.5 seconds. We would like to model the eruption with a Gaussian mixture model and estimate the following quantities (although notice that the mode on the left is slightly asymmetric so Gaussianity is not a great assumption there):

- What is the probability that any given eruption is a short eruption. ($\pi_1$)
- If it is a short eruption, what is the average time? What is the average time for a long eruption. ($\mu_1, \mu_2$)
- How much variability is there in these eruption times ($\sigma_1^2, \sigma_2^2$)?

We will use EM algorithm to obtain the MLE for these parameters.

```
#################################################
## EM Algorithm for the Old Faithful Geyser data
#################################################
# (pi_1, mu_1, mu_2, sigma^2_1, sigma^2_2)
theta <- c(.6, 1,5, 1, 1) # starting values
current <- theta
diff <- 100
tol <- 1e-5
iter <- 0
store <- current

while(diff > tol)
{
  iter <- iter + 1

  # E step: find gamma_{i,c,k} for just c = 1, since for c = 2 is just 1-Ep
  Ep <- current[1]*dnorm(x, current[2], sqrt(current[4]))/
```

```
    (current[1]*dnorm(x, current[2], sqrt(current[4])) + (1 -
       current[1])*dnorm(x, current[3], sqrt(current[5])))

  # M-step
  theta[1] <- mean(Ep)
  theta[2] <- sum(Ep*x) / sum(Ep)
  theta[3] <- sum((1-Ep)*x) / sum(1-Ep)
  theta[4] <- sum(Ep*(x - theta[2])^2) / sum(Ep)
  theta[5] <- sum((1-Ep)*(x - theta[3])^2) / sum(1-Ep)

  diff <- max( abs(theta - current)) # choosing absolute difference here
  current <- theta
  store <- rbind(store, theta)
}

current # final estimates
# [1] 0.34840894 2.01861785 4.27335295 0.05552515 0.19101167
```

So the above are the final estimates: about 35% of the eruptions are short. Short eruptions have mean 2.02 and variance .0555 and long eruptions have mean 4.27 and variance .19.

Recall that, as a bonus, we also get estimates of the probability that each observed data point is in which class ($\gamma_{i,c}$)

```
# Final estimates of the probability
# that each observation is in Class C.
Prob.Z <- current[1]*dnorm(x, current[2], sqrt(current[4]))/
  (current[1]*dnorm(x, current[2], sqrt(current[4])) + (1 -
     current[1])*dnorm(x, current[3], sqrt(current[5])))

head(round(Prob.Z, 4))
```

**WARNING:** The EM algorithm often does *label switching*. That is, it doesn't know which class we are calling 1 and which class we call 2. So from one starting point to the other, individual probabilities will change, although probabilities within a same group will remain same

Below is a plot of the what the iterations of the algorithm fit and how it converged. I have overlayed the histogram with fitted mixture densities at each iteration $k$:

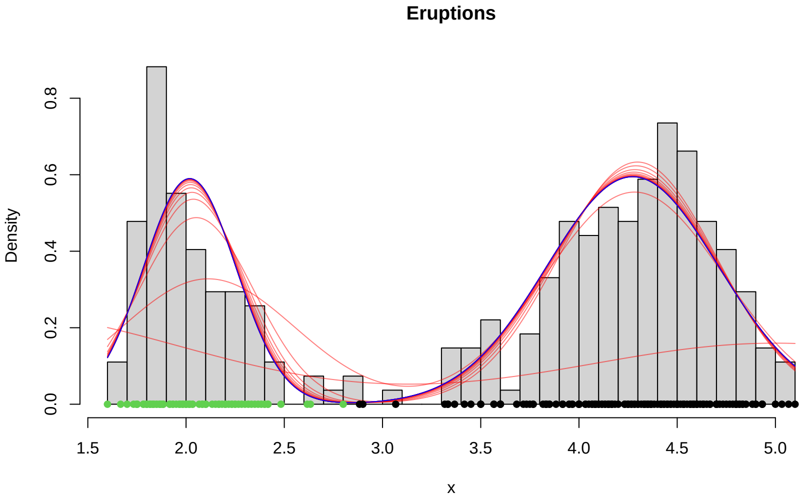
$$\pi^*_{1,(k)} f(x|\mu_{1,(k)}, \sigma^2_{1,(k)}) + (1 - \pi^*_{1,(k)}) f(x|\mu_{2,(k)}, \sigma^2_{2,(k)})\,.$$

I also add points on the $x$ axis indicating the classification obtained by $\gamma_{i,c}$. If estimated $\hat{\gamma}_{i,c} < .5$ I assign them color black and if estimated $\hat{\gamma}_{i,c} > .5$, I assign it color green.

```
# Make plot of iterative model fits
for(i in 1:dim(store)[1])
```

```
{
  test.x <- seq(min(x), max(x), length = 1000)
  test.y <- store[i,1]* dnorm(test.x, mean = store[i,2], sd =
      sqrt(store[i,4])) + (1-store[i,1]) *dnorm(test.x, mean = store[i,3],
      sd = sqrt(store[i,5]))
  lines(test.x, test.y, col = rgb(1,0,0, alpha = .5))
}
lines(test.x, test.y, col = rgb(0,0,1, alpha = 1))

# add color
color <- 1*(Ep < .5) + 3*(Ep >= .5)
points(x, rep(0, length(x)), pch = 16, col = color)
```



You can see how our model fitted progressed from bad fit in the beginning to slowly converging to reasonable estimates. Note: the left mode is still a little skewed, and that is because the assumption of normality on that mode is not ideal.

## 2 Questions to think about

- The `faithful` dataset has waiting times between eruptions as well. You can do a similar model on the waiting time.
- What happens when you change the starting values drastically?
- Can you setup the EM algorithm for multivariate normal distributions?