# Assignment 4

## Problem 1

Suppose we observe a set of data points $x_{1..4} = \{-1, 0.5, 1, 100\}$ and we would like to use Gaussian Mixture Model (GMM) for clustering $x_{1..4}$. Consider the following mixture of two Gaussians: $\mathcal{N}(0, 10)$ and $\mathcal{N}(1, 2)$ and we believe the data points are drawn from $\mathcal{N}(0, 10)$ 40% of the time and $\mathcal{N}(1, 2)$ 60% of the time.

1. Calculate the responsibility of each component for each data point in $x_{1..4}$.
2. For $x_4 = 100$ which component has higher responsibility? Briefly explain why that is the case.

## Problem 2

Suppose we want to model traffic on a given road. At each hour $i$, we measure the number of cars $y_i$. We think the number of cars is well modeled by a Gaussian distribution. However, we know that some times are rush hour, so we decide to use a mixture of Gaussian distributions with two components (normal and rush hour traffic respectively.) For simplicity, imagine that traffic at each hour is independent of one another.

Draw a graphical model that represents this mixture model. You should use plate notation and include variables $\mathbf{z} = (z_1, ..., z_n)$ and $\mathbf{y} = (y_1, ..., y_n)$ as well as the parameters of the distributions $\pi$, $\mu_1$, $\mu_2$, $\sigma_1$ and $\sigma_2$ in your graph.

## Problem 3

Suppose we have two sensors with known (and different) variances $v_1$ and $v_2$, but unknown (and the same) mean $\mu$. Suppose we observe $n_1$ observations $y_i^{(1)} \sim \mathcal{N}(\mu, v_1)$ from the first sensor and $n_2$ observations $y_i^{(2)} \sim \mathcal{N}(\mu, v_2)$ from the second sensor. (For example, suppose $\mu$ is the true temperature outside, and sensor 1 is a precise (low variance) digital thermosensing device, and sensor 2 is an imprecise (high variance) mercury thermometer.)

Let D represent all the data from both sensors. What is the posterior $p(\mu|D)$, assuming a non-informative prior for $\mu$ (which we can simulate using a Gaussian with a variance of $\infty$)? Give an explicit expression for the posterior mean and variance.