CSC2515 — Assignment #3 Answers

Behdad Esfahbod
993505827

December 1, 2004

1 Mixture of Base Rates (7%)

- The complete (joint) log likelihood:

$$\begin{aligned}
\log p(\mathbf{x}, k) &= \log p(c = k) + \log p(\mathbf{x}|c = k) \\
&= \log \alpha_k + \sum_i \log p(x_i|c = k) \\
&= \log \alpha_k + \sum_i \log \prod_{j \in V_i} a_{ijk}^{[x_i = j]} \\
&= \log \alpha_k + \sum_i \sum_{j \in V_i} [x_i = j] \log a_{ijk}
\end{aligned}$$

- The E-step: The posterior probability:

$$\begin{aligned}
p(k|\mathbf{x}^n) &= \frac{p(\mathbf{x}^n|k)p(k)}{p(\mathbf{x}^n)} \\
&= \frac{\alpha_k \prod_i \prod_{j \in V_i} a_{ijk}^{[x_i^n = j]}}{\sum_k \alpha_k \prod_i \prod_{j \in V_i} a_{ijk}^{[x_i^n = j]}}
\end{aligned}$$

- The expected complete log likelihood:

$$\begin{aligned}
\ell &= \log p(\{\mathbf{x}\}) = \sum_n \sum_k p(k|\mathbf{x}^n) \log p(\mathbf{x}^n, k) \\
&= \sum_n \sum_k p(k|\mathbf{x}^n) \left\{ \log \alpha_k + \sum_i \sum_{j \in V_i} [x_i^n = j] \log a_{ijk} \right\}
\end{aligned}$$

- The M-step: a_{ijk}'s:

$$\begin{aligned}
c_{ik}(\mathbf{a}) &= 1 - \sum_j a_{ijk} \\
L(\mathbf{a}, \lambda) &= \ell + \sum_i \sum_k \lambda_{ik} c(\mathbf{a})
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial L}{\partial a_{ijk}} &= \sum_n p(k|\mathbf{x}^n)[x_i^n = j]\frac{1}{a_{ijk}} - \lambda_{ik} = \frac{1}{a_{ijk}}\sum_n p(k|\mathbf{x}^n)[x_i^n = j] - \lambda_{ik} = 0 \\
a_{ijk} &= \frac{1}{\lambda_{ik}}\sum_n [x_i^n = j]p(k|\mathbf{x}^n) \\
c_{ik}(\mathbf{a}) &= 1 - \sum_j a_{ijk} = 1 - \frac{1}{\lambda_{ik}}\sum_j\sum_n [x_i^n = j]p(k|\mathbf{x}^n) = 1 - \frac{1}{\lambda_{ik}}\sum_n p(k|\mathbf{x}^n) = 0 \\
\lambda_{ik} &= \sum_n p(k|\mathbf{x}^n) = 0 \\
a_{ijk} &= \frac{\sum_n [x_i^n = j]p(k|\mathbf{x}^n)}{\sum_n p(k|\mathbf{x}^n)}
\end{aligned}
$$

- The M-step: α_k's:

$$
\begin{aligned}
c(\alpha) &= 1 - \sum_k \alpha_k \\
L(\alpha, \lambda) &= \ell + \lambda c(\alpha) \\
\frac{\partial L}{\partial \alpha_k} &= \sum_n p(k|\mathbf{x}^n)\frac{1}{\alpha_k} - \lambda = \frac{1}{\alpha_k}\sum_n p(k|\mathbf{x}^n) - \lambda = 0 \\
\alpha_k &= \frac{1}{\lambda}\sum_n p(k|\mathbf{x}^n) \\
c(\alpha) &= 1 - \sum_k \alpha_k = 1 - \frac{1}{\lambda}\sum_k\sum_n p(k|\mathbf{x}^n) = 1 - \frac{1}{\lambda}\sum_n\sum_k p(k|\mathbf{x}^n) \\
&= 1 - \frac{1}{\lambda}\sum_n 1 = 1 - \frac{n}{\lambda} = 0 \\
\lambda &= n \\
\alpha_k &= \frac{1}{n}\sum_n p(k|\mathbf{x}^n)
\end{aligned}
$$

- The given q_k is simply $\log(p(x|k)p(k))$. So, starting with e^{q_k}, all we need is to divide it by $p(x)$ which is the sum over k of e^{q_k} itself. Putting together:

$$p(k|x) = \frac{e^{q_k}}{\sum_k e^{q_k}}$$

so, taking log and using logsum instead of summation, the log of the answer is:

$$\log p(k|x) = q_k - \text{logsum}_k(q_k)$$

which is stable. Now we can achieve the non-log'ed answer by exponentiating this, getting:

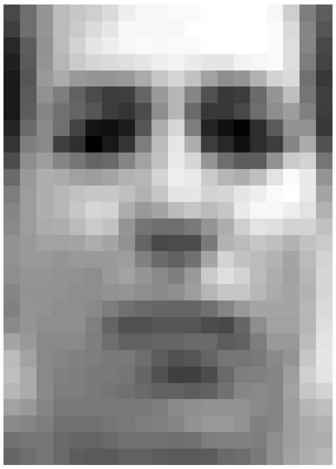
$$p(k|x) = \exp\{q_k - \text{logsum}_k(q_k)\}.$$

- "Mixture of Gaussians with diagonal covariances"

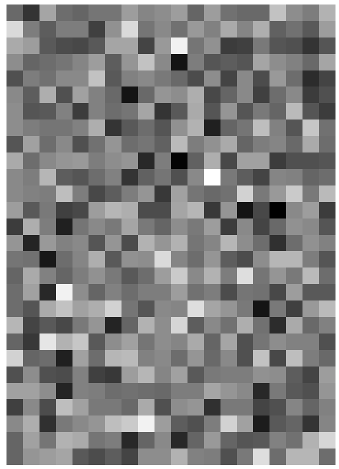
2 EM Algorithm for Factor Analysis

Six Faces

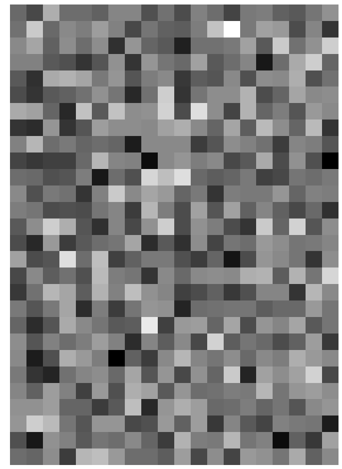
μ
black=57
white=224

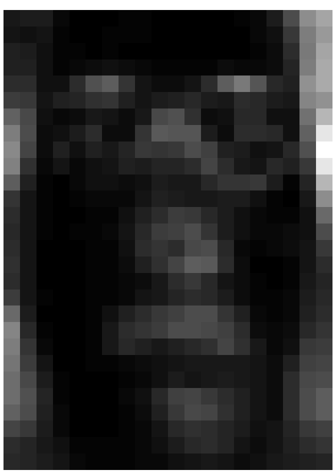


init Λ_1
black=-3
white=3

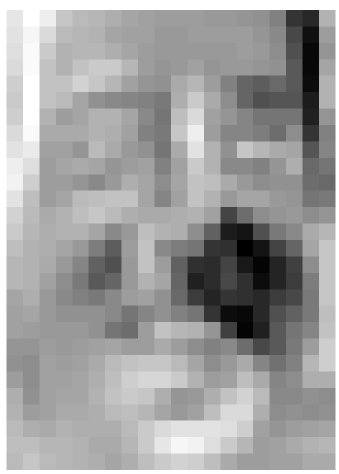


init Λ_2
black=-3
white=3

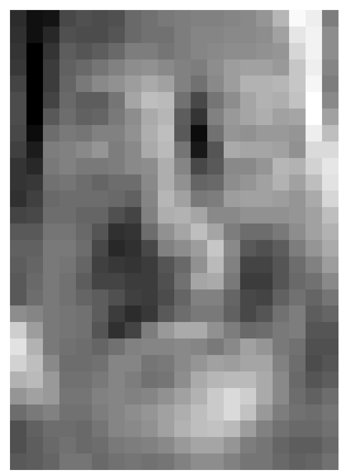




diag(Ψ)
black=11
white=3906



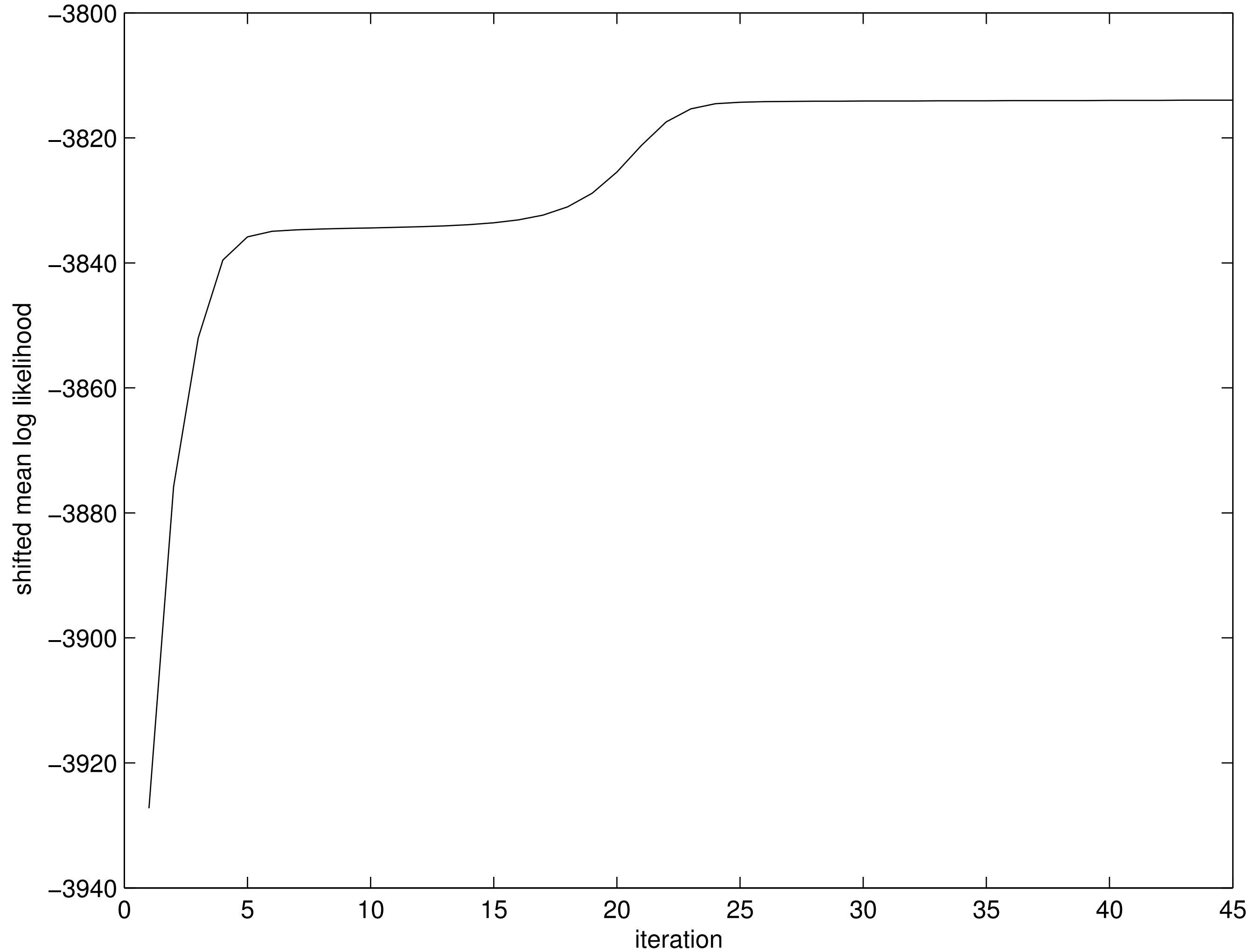
Λ_1
black=-26
white=16



Λ_2
black=-26
white=21

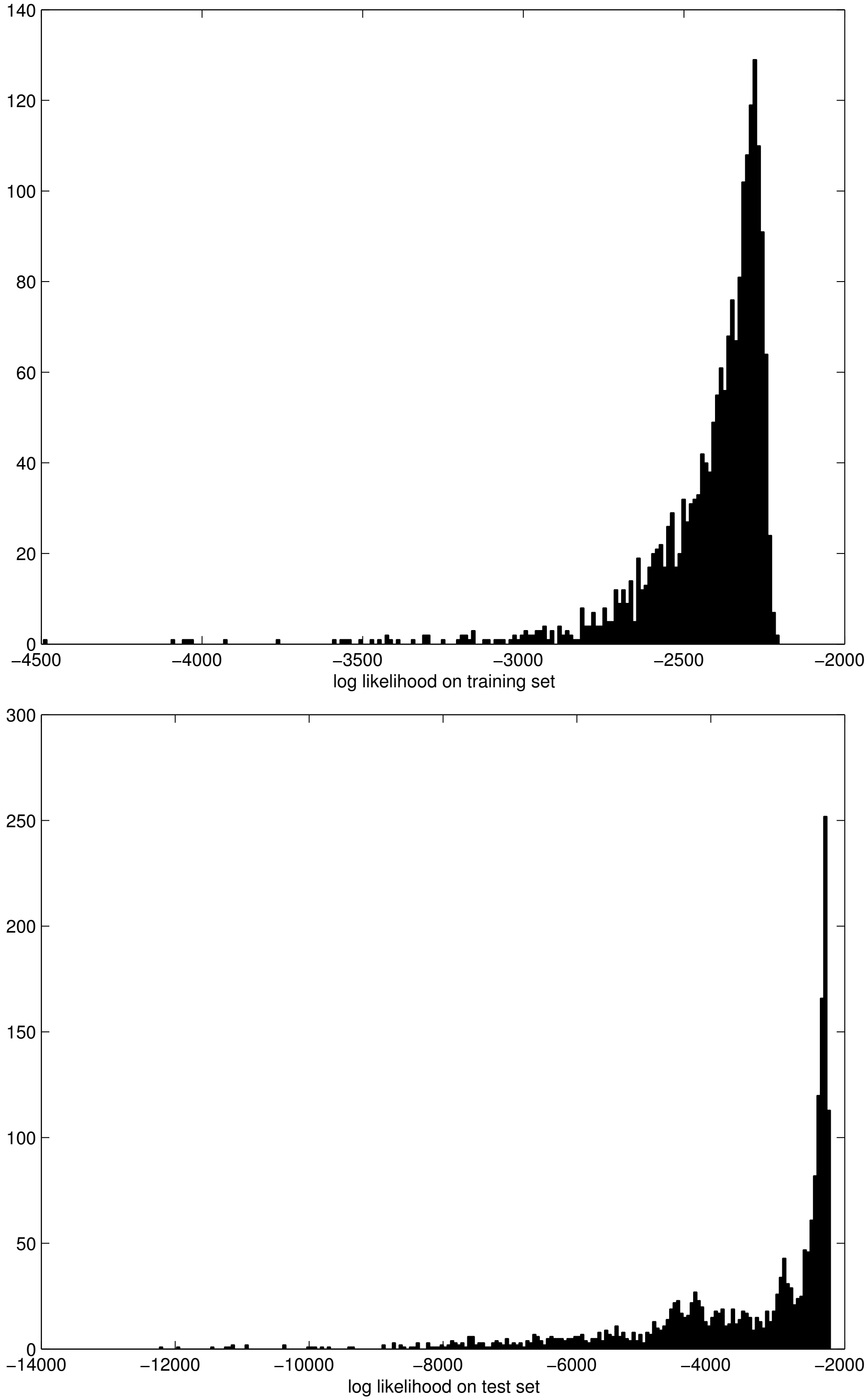
All black and white values have been rounded to nearest integer. Grayscales are linear between the black and white values. Ψ has been initialized to a diagonal matrix, with each entry uniformly chosen between zero and twice the same entry in the covariance matrix of the data. Λ's have been initialized using standard normal distribution.

EM Average Training Log Likelihood Trajectory



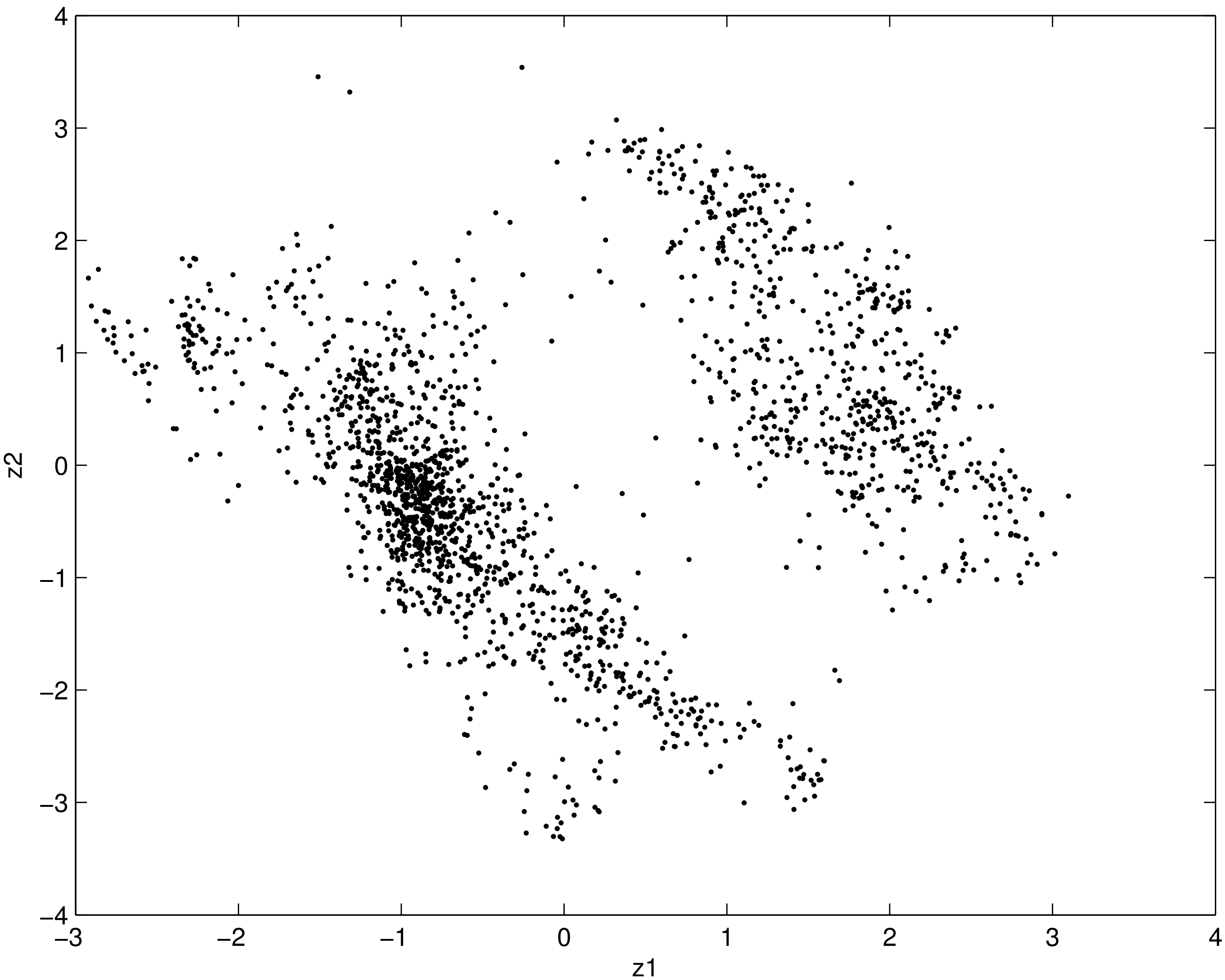
For the convergence condition, I used (`newlik-oldlik`)/`newlik`<2e-5 and ran EM 10 times and got the best solution. Then continued the EM algorithm on this best solution until the tighter convergence of (`newlik-oldlik`)/`newlik`<2e-6 was achieved. The step shape above shows that it took some time until the best rotation of the hidden parameters was found. The computed value of the mean log likelihood ignores some constant additive terms, so not to be compared to the log likelihood values that follow.

Training and Test Sets Log Likelihood



There are 200 bins per histogram. Means, an average of ten data points per bin.

Hidden Variables



2D scatter plot of z_1 and z_2. Apparently data is mostly divided into two clusters, mainly identified by the z_1 value.

And Their Interpretations

By *animating* the images in a sequence, sorted one on z_1 and another time on z_2, it can be seen that:

- z_1 models whether the face is smiling or not. The negative values mostly correspond to non-smiling faces, while positive values mostly correspond to smiling faces.
- z_2 models the direction that face is facing to. The negative values mostly correspond to faces facing to the right of the page, while positive values mostly correspond to faces facing to the left.

These observations can be verified by looking at the images in the next page.

Six More Faces

The three images in a vertical row, from bottom to top, correspond to the images with smallest, median, and largest posterior values on z_2. The three remaining images, from left to right, correspond to the images with smallest, median, and largest posterior values on z_1.

Classification

By using the minimum training log likelihood as the threshold, I got -4494 as the threshold and 29.1% test error. Since this in not the best possible choice of threshold, because it is quite unstable and depends on the *worst* training case, I used another approach to build a better classifier.

First I divided training points into 196 bins based on their log likelihood (200 was again chosen such that average number of data points per bin is ten), then I found the bin with least log likelihood center that had a number of data points in it that was not less than 5% of the maximum number of data points in all bins. The the edge of this bin and the previous bin was taken as the threshold. This way, threshold was set to -2823, training error was 3.6% and test error of 3.97%.