Homework 1

HE Xuan

March 2016

1 Mathematics Basics

1.1 Optimization
Construct the Lagrange function:

$$L(x_1, x_2, \lambda_1, \lambda_2) = x_1^2 + x_2^2 - 1 - \lambda_1(x_1 + x_2 - 1) - \lambda_2(2x_1 - x_2)$$

Calculate derivatives of function L to $x_1, x_2, \lambda_1, \lambda_2$ respectively, and set them to zero and according to KTT conditions:

$$\begin{aligned}
&2x_1 - \lambda_1 - 2\lambda_2 = 0 \\
&2x_2 - \lambda_1 + \lambda_2 = 0 \\
&1 - x_1 - x_2 = 0 \\
&\lambda_2(2x_1 - x_2) = 0 \\
&2x_1 - x_2 \geq 0 \\
&\lambda_1 \geq 0, \lambda_2 \geq 0
\end{aligned}$$

Then solved $x_1 = x_2 = 1/2$, $\lambda_1 = 1$, $\lambda_2 = 0$
Target value is $x_1^2 + x_2^2 - 1 = 1/4 + 1/4 - 1 = -1/2$

1.2 Conjugate Prior
The prior pdf:

$$p(p|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

The likelihood function:

$$p(x|p) = p^x(1-p)^{1-x}$$

So the posterior pdf:

$$p(p|x) = \frac{p(p)p(x|p)}{p(x)}$$
$$\propto p(p|\alpha, \beta)p(x|p)$$
$$\propto p^{\alpha-1}(1-p)^{\beta-1}p^x(1-p)^{1-x}$$
$$\propto p^{\alpha+x-1}(1-p)^{\beta-x+1-1}$$

Then integrate this result to 1:

$$\int p(p|x) = 1 \Rightarrow p(p|x) = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+x)\Gamma(\beta-x+1)}p^{\alpha+x-1}(1-p)^{\beta-x+1-1}$$

So Beta distribution can serve as a conjugate prior to the Bernoulli distribution.

1.3 Parameter Estimation

1. The likelihood of μ and Σ is:

$$L(\mu, \Sigma|x_1, ..., x_N)$$
$$= \prod_{i=1}^{N} N(x_i|\mu, \Sigma)$$
$$= (2\pi)^{-\frac{NK}{2}}|\Sigma|^{-\frac{N}{2}}exp(-\frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))$$

2. From 1 we know the log likelihood of μ is:

$$log(L) = -\frac{NK}{2}log(2\pi) - \frac{N}{2}log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$
$$\frac{\partial log(L)}{\partial \mu} = \sum_{i=1}^{N}\Sigma^{-1}(x_i-\mu)$$

Set the partial differential to 0, then:

$$\sum_{i=1}^{N}\Sigma^{-1}(x_i-\mu) = 0 \Rightarrow \mu = \frac{1}{N}\sum_{i=1}^{N}x_i$$

So $\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N}x_i$

Because $E[\mu] = \frac{1}{N}\sum_{i=1}^{N}E[x_i] = \frac{1}{N} * N * \mu = \mu$

So the MLE above of μ is unbiased.

3. Posterior distribution of μ:

$$N(\mu|x, \Sigma) \propto N(x|\mu)N(\mu|\mu_0, \lambda^{-1}\Sigma)$$
$$\propto (2\pi)^{-K/2}|(N+\lambda)^{-1}\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu)^T(N+\lambda)\Sigma^{-1}(x-\mu))$$

so:

$$\mu_{MAP} = \frac{\lambda\mu_0 + N\overline{x}}{N+\lambda}$$

2 Mixture of Multinomials

2.1 MLE for multinomials

The log likelihood function of $\mu = (\mu_i)_{i=1}^{d}$:

$$\begin{aligned} \ln(L(\mu|\mathbf{x})) &= \ln P(\mathbf{x}|\mu) \\ &= \ln n! - \sum_i \ln x_i! + \sum_i x_i \ln \mu_i \end{aligned}$$

Because $\ln n!$ and $\sum_i \ln x_i!$ are constant to μ. So the MLE question equals to:

$$\begin{aligned} &\underset{\mu}{argmax} \sum_i x_i \ln \mu_i \\ &s.t.: \sum_i x_i = n \\ &\sum_i \mu_i = 1 \end{aligned}$$

Construct Lagrange function:

$$L(x_i, \mu_i, \lambda_1, \lambda_2) = \sum_i x_i \ln \mu_i - \lambda_1(n - \sum_i x_i) - \lambda_2(\sum_i \mu_i - 1)$$

Calculate partial differentials and set them to 0:

$$\begin{aligned} &\ln \mu_i + \lambda_1 = 0 \\ &\tfrac{x_i}{\mu_i} - \lambda_2 = 0 \\ &\sum_i x_i = n \\ &\sum_i \mu_i = 1 \\ &\lambda_1 \geq 0; \lambda_2 \geq 0 \end{aligned}$$

Solved the equations above, the result is:

$$\mu_i = \tfrac{x_i}{n}$$

So $\mu_{MLE} = \frac{1}{n}\mathbf{x}$

2.2 EM for mixture of multinomials

The multinomial prior:

$$P(c_d = k) = \pi_k, k = 1, 2, ..., K$$

The multinomial likelihood:

$$P(d|c_d = k) = \frac{n_d!}{\prod_w T_{dw}!} \prod_w \mu_{wk}^{T_{dw}}, n_d = \sum_w T_{dw}$$

The marginal likelihood of d:

$$P(d) = \sum_{k=1}^{K} P(d|c_d = k)P(c_d = k) = \frac{n_d!}{\prod_w T_{dw}!} \sum_{k=1}^{K} \pi_k \prod_w \mu_{wk}^{T_{dw}}$$

So the posterior of c_d is:

$$\gamma(z_{dk}) = P(c_d = k|d) = \frac{P(d|c_d=k)P(c_d=k)}{P(d)} = \frac{\prod_w \pi_k \mu_{wk}^{T_{dw}}}{\sum_{j=1}^{K} \prod_w \pi_j \mu_{wj}^{T_{dw}}}$$

So in the E-step, calculate:

$$\gamma(z_{dk}) = \frac{P(d|c_d=k)P(c_d=k)}{P(d)} = \frac{\pi_k \prod_w \mu_{wk}^{T_{dw}}}{\sum_{j=1}^{K} \pi_j \prod_w \mu_{wj}^{T_{dw}}}$$

for every d in {1,...,D} and k in {1,...,K}
Then calculate the log expectation of posterior:

$$E[\gamma(z_{dk})] = \sum_{d=1}^{D} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln n_d! - \sum_w \ln T_{dw}! + \sum_w T_{dw} \ln \mu_{wk}\}$$

and the constraint conditions:

$$\sum_{k=1}^{K} \pi_k = 1$$

$$\sum_{w}^{W} \mu_{wk} = 1$$

where k = 1,2,...,K
Construct Lagrange function:

$$L(\pi, \mu, \lambda_0, ..., \lambda_K) =$$
$$\sum_{d=1}^{D} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln n_d! - \sum_w \ln T_{dw}! + \sum_w T_{dw} \ln \mu_{wk}\} + \lambda_0(1 - \sum_{k=1}^{K} \pi_k) + \sum_{k=1}^{K} \lambda_k(1 - \sum_{w}^{W} \mu_{wk})$$

where $\lambda_j \geq 0$ for every j in {0,1,...,K}
Maximize this function get the results:

$$\pi_k = \frac{\sum_{d=1}^{D} \gamma(z_{dk})}{\sum_{k=1}^{K} \sum_{d=1}^{D} \gamma(z_{dk})} = \frac{\sum_{d=1}^{D} \gamma(z_{dk})}{D}$$

$$\mu_{wk} = \frac{\sum_{d=1}^{D} T_{dw} \gamma(z_{dk})}{\sum_{w=1}^{W} \sum_{d=1}^{D} T_{dw} \gamma(z_{dk})}$$

So in the M-step, just calculate the values of π_k and μ_{wk} for every k and w according to the equations above.
Implement and result:
The algorithm was implement by python numpy:
Initiation step:

π and μ were all initiate according to dirichlet distribution

E-step:

Calculate the posterior.

M-step:

Maximize the expectation of posterior.

Stop condition:
When the Euclidean distance between new μ, π and old μ, π were less than the threshold which was set small enough. Here $threshold = 1e-20$
Results:

K	**Most-frequent words in each topic**
5	network network network network model
10	network model learning network network network model network network model
20	point network model cell network network function model cell network network model learning model input network model network network network
30	unit speaker weight network network model input network model model learning network network model network distance model network circuit network network network network network network method network learning model parameter

Table 1: EM Results

From the table we know, as k increasing from 5 to 30, the entropy of result topic words set is increased, this means that the topics are classified better, so in this case when k = 30, the result is best.

3 PCA

3.1 Minimum Error Formulation

Introduce a set of complete orthonormal basis:

$$\{\mu_i\}, i = 1, ..., p$$

$$\mu_i^T \mu_j = \delta_{ij}$$

Represent each data point by the basis vectors linear combination:

$$\mathbf{x_n} = \sum_{i=1}^{D} \alpha_{ni}\mu_\mathbf{i}$$

then $\alpha_{ni} = \mathbf{x_n^T}\mu_\mathbf{j}$,so:

$$\mathbf{x_n} = \sum_{i=1}^{D} (\mathbf{x_n^T}\mu_\mathbf{i})\mu_\mathbf{i}$$

Assume that a M-dimensional linear subspace can approximate the original space, where M $\leq$ D, then:

$$\widetilde{\mathbf{x}}_n = \sum_{i=1}^{M} z_{ni}\mu_\mathbf{i} + \sum_{i=M+1}^{D} b_i\mu_\mathbf{i}$$

So the Error of choose the M-subspace can represent as:

$$J = \frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x_n} - \widetilde{\mathbf{x}}_\mathbf{n}||^2$$

To minimize this error:

$$min_{\{z_{ni}\},\{b_i\}} J$$

$$\frac{\partial J}{\partial z_{ni}} = -\frac{2}{N}(\mathbf{x}_n - z_{ni}\mu_\mathbf{i}){\mu_\mathbf{i}}^T = 0$$

$$\frac{\partial J}{\partial b_i} = -\frac{2}{N}(\mathbf{x}_n - (D-M)b_i\mu_i){\mu_\mathbf{i}}^T = 0$$

$$\Rightarrow z_{ni} = \mathbf{x}_n{\mu_\mathbf{i}}^T = \mathbf{x}_n^T\mu_\mathbf{i}$$

$$\Rightarrow b_i = \frac{\mathbf{x}_n}{D-M}{\mu_\mathbf{i}}^T = \overline{\mathbf{x}}_n^T\mu_\mathbf{i}$$

So

$$\mathbf{x}_n - \widetilde{\mathbf{x}}_\mathbf{n} = \sum_{i=M+1}^{D} \{(\mathbf{x}_n - \overline{\mathbf{x}}_n)^T\mu_\mathbf{i}\}\mu_\mathbf{i}$$

So the error can be expressed as:

$$J = \sum_{i=M+1}^{D} {\mu_\mathbf{i}}^T\mathbf{S}\mu_\mathbf{i}$$

So the original question equals to:

$$\underset{\mu_i}{argmin}\ J$$

$$s.t.\ \mu_\mathbf{i}^\mathbf{T}\mu_\mathbf{i} = 1$$

Construct Lagrange function:

$$L = {\mu_\mathbf{i}}^T\mathbf{S}\mu_\mathbf{i} + \lambda_i(1 - \mu_\mathbf{i}^\mathbf{T}\mu_\mathbf{i})$$

Set the derivatives with respect to μ_i to 0, then:

$$\mathbf{S}\mu_{\mathbf{i}} = \lambda_i \mu_i$$

for $1 \le i \le D$
So the Error:

$$J = \sum_{i=M+1}^{D} \lambda_i$$

This means we choose largest M eigenvalues and let the last be small, this is truth of PCA.
3.2 PCA implement and results:
figure 2 - 9 are substracting the sample mean and figure 10-17 are not substracting the sample mean:

Figure 1: Original figure

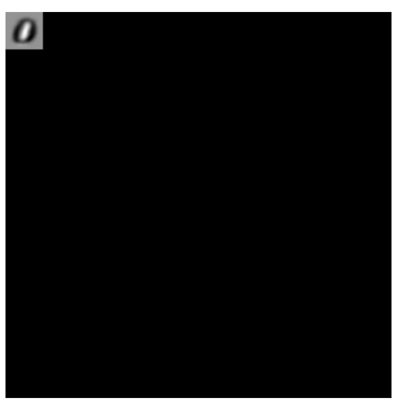

Figure 2: First component

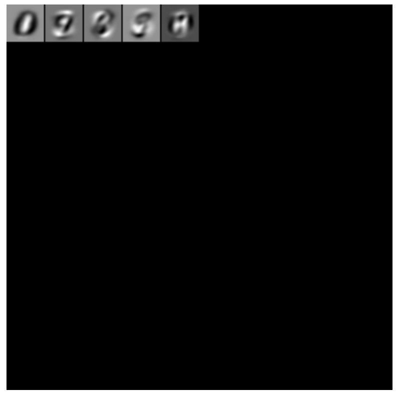

Figure 3: First 5 components

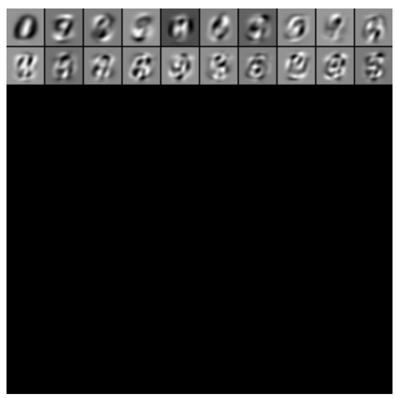

Figure 4: First 20 components

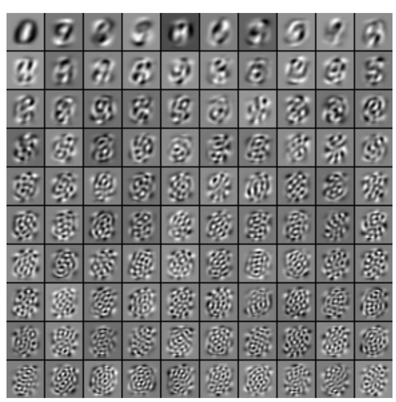

Figure 5: First 100 components

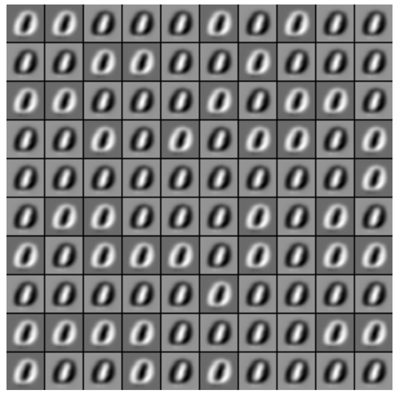

Figure 6: Reconstructed figure based on first component

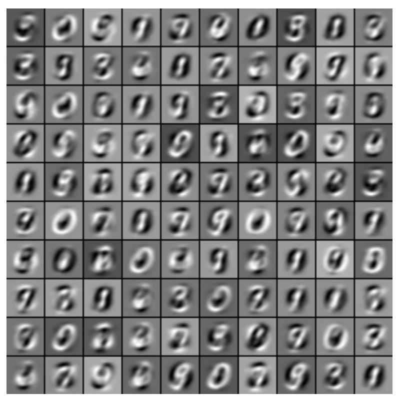

Figure 7: Reconstructed figure based on first 5 components

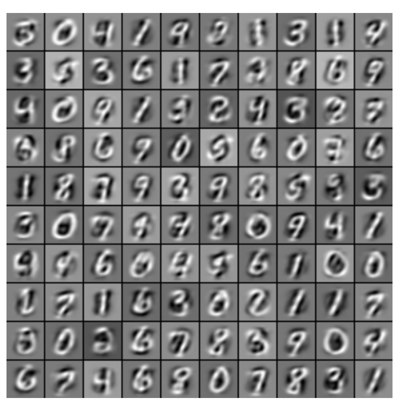

Figure 8: Reconstructed figure based on first 20 components

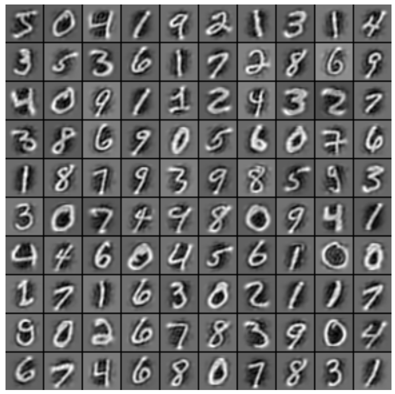

Figure 9: Reconstructed figure based on first 100 components

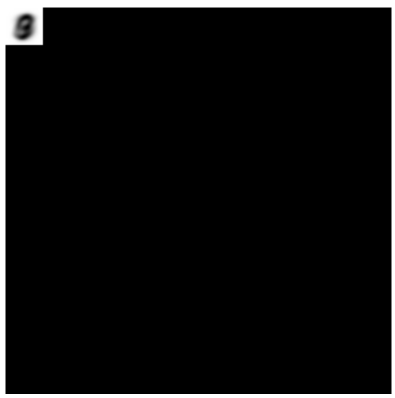

Figure 10: First component

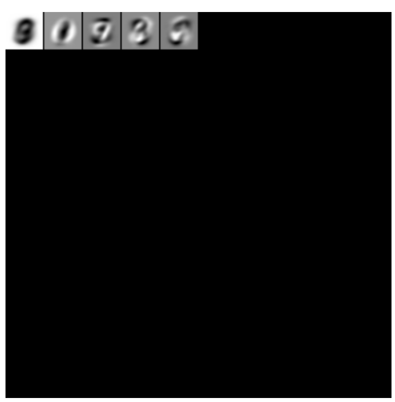

Figure 11: First 5 components

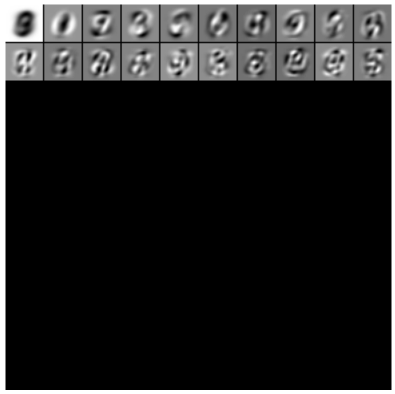

Figure 12: First 20 components

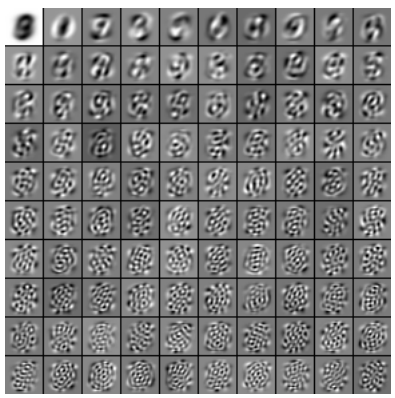

Figure 13: First 100 components

Figure 14: Reconstructed figure based on first component

Figure 15: Reconstructed figure based on first 5 components

Figure 16: Reconstructed figure based on first 20 components

Figure 17: Reconstructed figure based on first 100 components