Supplemental material for "Bayesian Face Revisited: A Joint Formulation"

Dong Chen[1], Xudong Cao[3], Liwei Wang[2], Fang Wen[3], and Jian Sun[3]

[1] University of Science and Technology of China
chendong@mail.ustc.edu.cn
[2] The Chinese University of Hong Kong
lwwang@cse.cuhk.edu.hk
[3] Microsoft Research Asia, Beijing, China
{xudongca,fangwen,jiansun}@microsoft.com

1 Efficient Computation for Block-wise Matrix

In Section 2.3, we need to compute the following formula,

$$E(\mathbf{h}|\mathbf{x}) = \Sigma_h \mathbf{P}^T \Sigma_x^{-1} \mathbf{x}. \tag{1}$$

Directly computing Eqn.(1) leads to expensive memory and computational complexity. Let d represents the dimension of feature and m represents the number of images of the subject, the computational complexity is $O\left(d^3m^3\right)$ and the memory complexity is $O\left(d^2m^2\right)$ for naive implementation. However, by taking the advantage of block-wise structure of the matrix, the complexity can be reduced to $O\left(d^3 + md^2\right)$ in computation and $O\left(d^2\right)$ in memory. In this part, we describe the details of the efficient implementation.

1.1 Efficient Inverse

In Eqn.(1), we need the calculate the inverse of Σ_x. We will show in the rest of part how to compute Σ_x^{-1} in $O\left(d^3\right)$ complexity. As discussed in Section 2.3, Σ_x is the covariance matrix of m observations(features of face images) for one subject, which can be derived as,

$$\Sigma_x = \begin{bmatrix} S_\mu + S_\varepsilon & S_\mu & \cdots & S_\mu \\ S_\mu & S_\mu + S_\varepsilon & \cdots & S_\mu \\ \vdots & \vdots & \ddots & \vdots \\ S_\mu & S_\mu & \cdots & S_\mu + S_\varepsilon \end{bmatrix}.$$

Suppose its inverse satisfy the following form:

$$\Sigma_x^{-1} = \begin{bmatrix} F + G & G & \cdots & G \\ G & F + G & \cdots & G \\ \vdots & \vdots & \ddots & \vdots \\ G & G & \cdots & F + G \end{bmatrix} \tag{2}$$

Using $\Sigma_x \Sigma_x^{-1} = I$, for the elements on the diagonal, we have:

$$(S_\mu + S_\varepsilon)(F + G) + mS_\mu G = \mathbf{I} \tag{3}$$

Where m is the image number for each subject. For other elements, we have:

$$(S_\mu + S_\varepsilon)G + S_\mu F + mS_\mu G = \mathbf{0} \tag{4}$$

Eqn. (3) - Eqn. (4):

$$S_\varepsilon F = \mathbf{I}$$

$$F = S_\varepsilon^{-1} \tag{5}$$

Put Eqn. (5) into Eqn. (4), we have:

$$G = -((m+1)S_\mu + S_\varepsilon)^{-1} S_\mu S_\varepsilon^{-1} \tag{6}$$

By plugging into the results of F and G into Eqn.(2), we have the inverse matrix of Σ_x. Since the computational complexity for computing F and G are $O(d^3)$, the computational complexity for matrix Σ_x is $O(d^3)$.

1.2 Efficient Matrix Multiplication

We rewrite the Eqn.(1) here for the convenience of reading,

$$E(\mathbf{h}|\mathbf{x}) = \Sigma_h \mathbf{P}^T \Sigma_x^{-1} \mathbf{x},$$

Where $\mathbf{h} = [\mu; \varepsilon_1; \cdots; \varepsilon_m]$ are the latent variables and $\mathbf{x} = [x_1; \cdots; x_m]$ are the observations.

We can see that directly computing Eqn.(1) is still expensive, even the inverse of matrix Σ_x is known. The complexities in both computation and memory are $O(m^2 d^2)$. In this part, we show that the complexity can be reduced to $O(md^2)$ and $O(d^2)$ in computation and memory respectively.

By putting Eqn.(2) into Eqn.(1), we can get:

$$\mu = \sum_{i=1}^{m} S_\mu (F + (m+1)G) x_i \tag{7}$$

$$\varepsilon_j = S_\varepsilon x_j + \sum_{i=1}^{m} S_\varepsilon G x_i \tag{8}$$

From Eqn.(7) and Eqn.(8), we can tell that the computational complexity is $O(md^2)$ and the complexity in memory is $O(d^2)$.

1.3 Summary

Here we present the summary of complexity analysis for our algorithm.

1. The complexities of naive implementation are $O\left(d^3m^3\right)$ in computation and $O\left(d^2m^2\right)$ in memory.
2. The complexities of our method are $O\left(d^3 + md^2\right)$ in computation and $O\left(d^2\right)$ in memory.
3. Usually the dimension of feature(around 1000) is larger than the number of images for one subject(around 100), the complexities of our method can be simplified to $O\left(d^3\right)$ in computation and $O\left(d^2\right)$ in memory.

2 Negative Definite

In this part, we prove the negative definiteness of matrix A and G described in Section 2.2. In the beginning, we introduce three lemmas which will be used in the following proof.

Lemma 1. (Schur's complement theory) Let $A \succ \mathbf{0}$,

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \succ \mathbf{0} \Leftrightarrow C - B^T A^{-1} B \succ \mathbf{0}$$

Lemma 2. Let $A \succ \mathbf{0}$ and $B \succ \mathbf{0}$,

$$A - B \succ \mathbf{0} \Leftrightarrow \lambda_{min}\left(B^{-1/2} A B^{-1/2}\right) > 1$$

where $\lambda_{min}(\cdot)$ represents the minimum eigen value of the matrix.

Lemma 3. Block matrix inversion:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

2.1 *G*

It is easy to prove that matrix G is negative definite. Referring Eqn.(6), matrix G can be represented as,

$$G = -(2S_\mu + S_\varepsilon)^{-1} S_\mu S_\varepsilon^{-1}$$

As S_μ might not be invertible, we introduce a new variable:

$$P = -(2(S_\mu + \lambda I) + S_\varepsilon)^{-1}(S_\mu + \lambda I) S_\varepsilon^{-1}$$

we prove $(-P)^{-1}$ is positive definite by reformulating its form,

$$(-P)^{-1} = S_\varepsilon (S_\mu + \lambda I)^{-1}(2(S_\mu + \lambda I) + S_\varepsilon) = 2S_\varepsilon + S_\varepsilon (S_\mu + \lambda I)^{-1} S_\varepsilon$$

Both $2S_\varepsilon$ and $S_\varepsilon(S_\mu + \lambda I)^{-1} S_\varepsilon$ are positive definite. The summation of them are also positive definite. Since $(-P)^{-1}$ is positive definite, matrix P is negative definite. So

$$G = \lim_{\lambda \to 0} P$$

is also negative definite.

2.2 *A*

Using the block matrix inversion in lemma 1, the matrix A can be reformulated:

$$A = (S_\mu + S_\varepsilon)^{-1} - ((S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu)^{-1} \tag{9}$$

As the covariance matrix of the distribution $P(x_1, x_2|H_I)$:

$$\Sigma_I = \begin{bmatrix} S_\mu + S_\varepsilon & S_\mu \\ S_\mu & S_\mu + S_\varepsilon \end{bmatrix}.$$

is positive definite. Using Schur's complement theory in lemma 2, the formulation:

$$N = (S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu$$

is also positive definite. Let,

$$\begin{aligned} M &= (S_\mu + S_\varepsilon)^{-1/2}N(S_\mu + S_\varepsilon)^{-1/2} \\ &= (S_\mu + S_\varepsilon)^{-1/2}((S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu)(S_\mu + S_\varepsilon)^{-1/2} \\ &= \mathbf{I} - (S_\mu + S_\varepsilon)^{-1/2}S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu(S_\mu + S_\varepsilon)^{-1/2} \end{aligned}$$

On one hand, since M is positive definite, the eigen value of matrix M is larger than 0. On the other hand, as the second term in the equation above is positive definite, the eigen value of matrix M is smaller than 0, i.e.

$$1 > \lambda(M) > 0$$

Therefore, all the eigen values of M^{-1} are larger than 1.

$$\lambda(M^{-1}) > 1$$

As

$$M^{-1} = (S_\mu + S_\varepsilon)^{1/2}((S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu)^{-1}(S_\mu + S_\varepsilon)^{1/2}$$

According to lemma 3, we have

$$-A = ((S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu)^{-1} - (S_\mu + S_\varepsilon)^{-1} \succ \mathbf{0}$$

Therefore A is negative definite.

3 Invariance to Full Rank Linear Transform

From the Eqn.(9) and Eqn.(9) we have,

$$\begin{aligned} G &= -(2S_\varepsilon + S_\varepsilon S_\mu^{-1} S_\varepsilon)^{-1} \\ A &= (S_\mu + S_\varepsilon)^{-1} - ((S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1}S_\mu)^{-1} \end{aligned} \tag{10}$$

Let the full rank linear transform be,

$$y = Wx$$

where x is the original feature and y is transformed feature. W is the full rank linear transform matrix. The covariance matrixes of identity and intra-person variation for the new feature y can be derived as,

$$\begin{aligned} \tilde{S}_\mu &= W S_\mu W^T \\ \tilde{S}_\epsilon &= W S_\epsilon W^T \end{aligned} \tag{11}$$

By plugging Eqn.(11) into Eqn.(10) and simplifying the equation, we have

$$\begin{aligned} \tilde{G} &= -(W^T)^{-1}(2S_\varepsilon + S_\varepsilon S_\mu^{-1} S_\varepsilon)^{-1}(W)^{-1} \\ \tilde{A} &= (W^T)^{-1}((S_\mu + S_\varepsilon)^{-1} - ((S_\mu + S_\varepsilon) - S_\mu(S_\mu + S_\varepsilon)^{-1} S_\mu)^{-1})(W^T) \end{aligned}$$

W is inheritable, because it is full rank linear transform. Combining the above equations, we can get

$$\begin{aligned} r(y_1, y_2) =& {y_1}^T \tilde{A} y_1 + {y_2}^T \tilde{A} y_2 - 2{y_1}^T \tilde{G} y_2 \\ =& {x_1}^T A x_1 + {x_2}^T A x_2 - 2{x_1}^T G x_2 \\ =& r(x_1, x_2) \end{aligned} \tag{12}$$

Therefore the similarity based on joint Bayesian is invariant to full rank linear transform.

4 Connection with Reference Based Method

In this part, we show that the metric derived from the angle of reference based method is equivalent to the metric derived from joint Bayesian. The metric derived from reference based method can be written as,

$$\mathrm{Log}\left(\frac{\int P(x_1|\mu)\, P(x_2|\mu)\, P(\mu) d\mu}{\int P(x_1|\mu)\, P(\mu) d\mu \int P(x_2|\mu)\, P(\mu) d\mu}\right). \tag{13}$$

Let

$$\begin{aligned} \mu \sim& N(0, S_\mu) \\ x \sim& N(\mu, S_\epsilon) \end{aligned}$$

We have

$$\begin{aligned} \int P(x|\mu) P(\mu) d\mu =& P(x) = N(0, S_\mu + S_\epsilon) \\ \int P(x_1|\mu)\, P(x_2|\mu)\, P(\mu) d\mu =& P(x_1, x_2) = N(0, \Sigma) \\ \Sigma =& \begin{pmatrix} S_\mu + S_\epsilon & S_\mu \\ S_\mu & S_\mu + S_\epsilon \end{pmatrix} \end{aligned} \tag{14}$$

It is obvious from Eqn.(14) that the molecule of Eqn.(13) is equal to $P(x_1, x_2|H_I)$ and the denominator of Eqn.(13) is equal to $P(x_1, x_2|H_E)$. Therefore the two metrics are equivalent.