Brief Article

The Author

April 14, 2015

1 12.1

See the reference manual. It's conventional convex optimization. Note that we know $\eta_j = 0$ because $\boldsymbol{u}_i, i = 1, \ldots, M$ are linearly independent.

2 12.2

In (12.93), the second term is just the sum of all Lagrangian terms. we have $(D - M)(D - M)$ constraints, for every pair of vectors (maybe somewhat redundant, but it's still fine), and we have $(D - M)(D - M)$ Lagrangian terms in H. Look at the solution manual. Note that if there exists a H that minimizes this equation, we can always set it to be a symmetric one. (Note that $I - \hat{U}^T\hat{U}$ is symmetric, just as $x_i x_j = x_j x_i$ in Problem 1.14).

The derivation in the reference manual just says that 1) $H = L$, where L is a diagonal matrix having all eigenvalues of S is OK, but 2) if we allow H to be a symmetric matrix, then to satisfy the equation, the eigenvalues of H must be L (eigenvalues of S), and any transformation Ψ will be equally good.

Note that by solving H and $\hat{U}$, we only get stationary points. Any $D - M$ eigenvalues of S will do, or, any rotation of $D - M$ eigenvectors will do.

Basically, we can rotate freely in the remaining $D - M$ dimensions, and I think we can certainly do so in the first M dimensions, as long as orthogonality is preserved.

3 12.3

Trivial. Just pre-multiply (12.30) by $\boldsymbol{u}_i^T$.

4 12.4

Trivial. The first equation in solution manual is wrong. Σ^{-1} should be Σ.

5 12.5

Well, just look at the solution manual. I don't know some of the solutions. But generally speaking, I think singular distributions are not useful, and they can be approximated by giving a very small covariance matrix, without any difference in practice.

6 12.6

Trivial.

7 12.7

Trivial.

8 12.8

Trivial, just by rote.

9 12.9

Should be trivial.

10 12.10

Trivial, using properties of convex / concave.

11 12.11

By doing SVD on W, we are done. I don't like solution in the manual much. Actually, that solution tells us, if the rotation matrix is identity matrix, then mean of z is the whitened projection in PCA. I like Byron Yu's $\tilde{\boldsymbol{z}}$ in his notes. By plugging in the general solution for W (12.45), I can show that $\tilde{\boldsymbol{z}}$ is the projection made by conventional PCA.

12 12.12

Don't understand what's in solution manual. I want to derive Byron's result myself.

Using (12.45), (12.48), and formula for $\tilde{\boldsymbol{z}}$, we have

$$
\begin{aligned}
\tilde{\boldsymbol{z}} &= (L-\sigma^2 I)^{1/2} R E(\boldsymbol{z} \mid \boldsymbol{x}) && (1)\\
&= (L-\sigma^2 I)^{1/2} R M^{-1} W^T(\boldsymbol{x}-\overline{\boldsymbol{x}}) && (2)\\
&= (L-\sigma^2 I)^{1/2} R [W^T W+\sigma^2 I]^{-1} W^T(\boldsymbol{x}-\overline{\boldsymbol{x}}) && (3)\\
&= (L-\sigma^2 I)^{1/2} R [R^T(L-\sigma^2 I)^{1/2} U^T U (L-\sigma^2 I)^{1/2} R+\sigma^2 I]^{-1} W^T(\boldsymbol{x}-\overline{\boldsymbol{x}}) && (4)\\
&= (L-\sigma^2 I)^{1/2} R L^{-1} W^T(\boldsymbol{x}-\overline{\boldsymbol{x}}) && (5)\\
&= (L-\sigma^2 I)^{1/2} R L^{-1} R^T (L-\sigma^2 I)^{1/2} U^T(\boldsymbol{x}-\overline{\boldsymbol{x}}) && (6)\\
&= (I-\sigma^2 L^{-1}) U^T(\boldsymbol{x}-\overline{\boldsymbol{x}}). && (7)
\end{aligned}
$$

This is just what Byron showed in the notes.

13 12.13

Trivial. Plug in (12.48) into the expression, and use (12.45). Well, this question is asking, given W solved by PPCA, what's the best way to project that point to the space of W, so that square error is minimized... Well, it's just writing orthogonal projection in terms of $E(z|x)$!

14 12.14

See the reference solution... This solution just says that the number of independent parameters match, but doesn't say if the solution spaces match.

15 12.18

Easy to derive, but don't know why (why the rotation matrix in total subtracts $M(M-1)$ degrees of freedom) ...

16 12.19

Trivial...

17 12.25

Look at the reference manual. I like the first part on how to prove the transformation of ML solutions are new ML solutions: basically, you do some algebra to make the new log likelihood function have the same form of old ones.

Well, I think the solution for first part is completely wrong. They should do the variable substitution as follows: $\mu_A = A^{-1}\mu, W_A = A^{-1}W$, and $\Psi_A = A^{-1}\Psi A$, so that everything takes the same form, except that everything has a subscript. But we need to remember, the end goal is to solve μ, W, Ψ. we have $\mu_A = \mu_{ML} = A^{-1}\mu$, and get $\mu = A\mu_{ML}$. Similarly for others.

18 Kernel PCA (pp. 588 of book)

I think the original paper of KPCA and the presentation here in the book is redundant, and not logically correct.

I hate that $K^2 = K\alpha$ stuff.

Well let's denote that

$$X = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \cdots \end{bmatrix} \tag{8}$$

$$K = XX^T \tag{9}$$

$$C = \frac{1}{N}X^TX. \tag{10}$$

where C being covariance matrix.

First I show that $\frac{1}{N}K$ and C have the same set of positive eigenvalues.

$$
\begin{aligned}
&Cv_i = \lambda_i v_i && (11)\\
\rightarrow &\frac{1}{N} X^T X v_i = \lambda_i v_i && (12)\\
\rightarrow &\frac{1}{N} X X^T X v_i = \lambda_i X v_i && (13)\\
\rightarrow &\frac{1}{N} X X^T (X v_i) = \lambda_i (X v_i) && (14)\\
\rightarrow &\frac{1}{N} K a_i = \lambda_i a_i && (15)
\end{aligned}
$$

Notice that $a_i = Xv_i$ must be non zero, otherwise $X^T X v_i$ would be zero, and $\lambda_i v_i$ be zero.

What if we have duplicate positive eigenvalues? In that case, we can show that all $a_i = Xv_i, a_i' = Xv_i', a_i'' = Xv_i'', \ldots$ are linear independent as well. (because $X^T a_i, X^T a_i', X^T a_i'', \ldots$ are independent. So $X^T(x_1 a_i + x_2 a' + x_3 a_i'') = 0$ have only zero solution, so does $(x_1 a_i + x_2 a' + x_3 a_i'') = 0$ only when $x_1 = x_2 = x_3 = 0$, otherwise we have contradiction)

We can prove the other direction by changing the role of C and $\frac{1}{N}K$, so we have

$$
\begin{aligned}
&\frac{1}{N} K a_i = \lambda_i a_i && (16)\\
\rightarrow &C(X^T a_i) = \lambda_i (X^T a_i) && (17)\\
\rightarrow &C v_i = \lambda_i v_i && (18)
\end{aligned}
$$

So we know that all solutions to (12.80) in the book correspond to some v_i for (12.74), and we also established (12.76).

So, since $a_i = Xv_i$, for a existing data point, its projection to the KPCA space (upto some multiplication) can be read out directly by eigenvectors of K. Check `test_KPCA.m` to see this!!! By the book, the projection should be Ka, which should be just some multiplication of a, since they are eigenvectors!