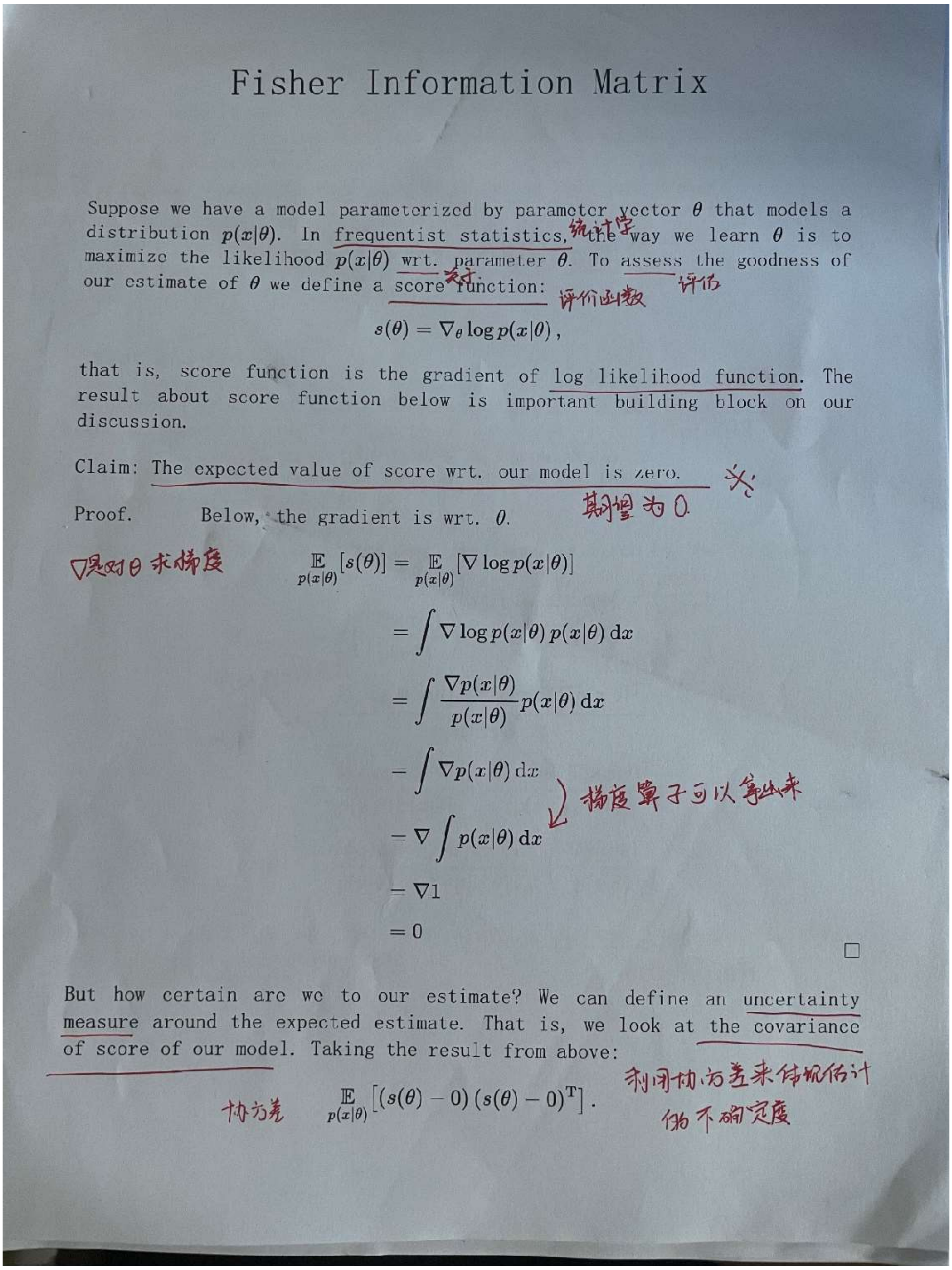
2、证明信息矩阵与协方差矩阵的逆之间的关系

搜集了一些资料，以下这个证明过程是我认为最完整，也容易理解的

这个证明过程是可以理解的，我现在还有一点疑问是对于这个信息矩阵的定义，如何体现极大似然估计的不确定度，我大概理解是：协方差矩阵可以体现随机变量的不确定度，然后信息矩阵是对似然函数的对数的梯度求协方差矩阵，和似然函数相关肯定是能体现一些性质。这个需要更加详细的推导，以后再查，这里的证明已经完全说明了信息矩阵和协方差矩阵的逆之间的关系。



Fisher Information Matrix

Suppose we have a model parameterized by parameter vector θ that models a distribution $p(x|\theta)$. In frequentist statistics, 统计学 the way we learn θ is to maximize the likelihood $p(x|\theta)$ wrt. 对 parameter θ. To assess the goodness of our estimate of θ we define a score function: 评价函数 评估

$$s(\theta) = \nabla_\theta \log p(x|\theta),$$

that is, score function is the gradient of log likelihood function. The result about score function below is important building block on our discussion.

Claim: The expected value of score wrt. our model is zero. 关

期望为0

Proof. Below, the gradient is wrt. θ.

∇是对θ求梯度

$$\begin{aligned}
\mathop{\mathbb{E}}_{p(x|\theta)}[s(\theta)] &= \mathop{\mathbb{E}}_{p(x|\theta)}[\nabla \log p(x|\theta)] \\
&= \int \nabla \log p(x|\theta)\, p(x|\theta)\, \mathrm{d}x \\
&= \int \frac{\nabla p(x|\theta)}{p(x|\theta)} p(x|\theta)\, \mathrm{d}x \\
&= \int \nabla p(x|\theta)\, \mathrm{d}x \\
&= \nabla \int p(x|\theta)\, \mathrm{d}x \\
&= \nabla 1 \\
&= 0
\end{aligned}$$

梯度算子可以拿出来

□

But how certain are we to our estimate? We can define an uncertainty measure around the expected estimate. That is, we look at the covariance of score of our model. Taking the result from above:

协方差

$$\mathop{\mathbb{E}}_{p(x|\theta)}\left[(s(\theta) - 0)\,(s(\theta) - 0)^{\mathrm{T}}\right].$$

利用协方差来体现估计的不确定度

We can then see it as an information. The covariance of score function above is the definition of Fisher Information. As we assume $\boldsymbol{\theta}$ is a vector, the Fisher Information is in a matrix form, called Fisher Information Matrix:

信息矩阵的定义
$$\mathbf{F} = \mathop{\mathbb{E}}_{p(x|\theta)}\left[\nabla \log p(x|\theta)\, \nabla \log p(x|\theta)^{\mathrm{T}}\right].$$

However, usually our likelihood function is complicated and computing the expectation is intractable. 难的 We can approximate 估算 the expectation in $\mathbf{F}$ using empirical distribution $\hat{q}(x)$, which is given by our training data $X = \{x_1, x_2, \cdots, x_N\}$. In this form, $\mathbf{F}$ is called Empirical Fisher:

经验主义的，以实验为依据的

$$\mathrm{F} = \frac{1}{N}\sum_{i=1}^{N} \nabla \log p(x_i|\theta)\, \nabla \log p(x_i|\theta)^{\mathrm{T}}.$$

Fisher and Hessian

理解，解释

One property 性质 of $\mathbf{F}$ that is not obvious is that it has the interpretation of being the negative expected Hessian of our model' s log likelihood.

负的似然函数的对数的Hessian矩阵.

Claim: The negative expected Hessian of log likelihood is equal to the Fisher Information Matrix F. ※

Proof.   The Hessian of the log likelihood is given by the Jacobian of its gradient:

$$\nabla \log p(x|\theta) = \frac{\nabla p(x|\theta)}{p(x|\theta)}$$

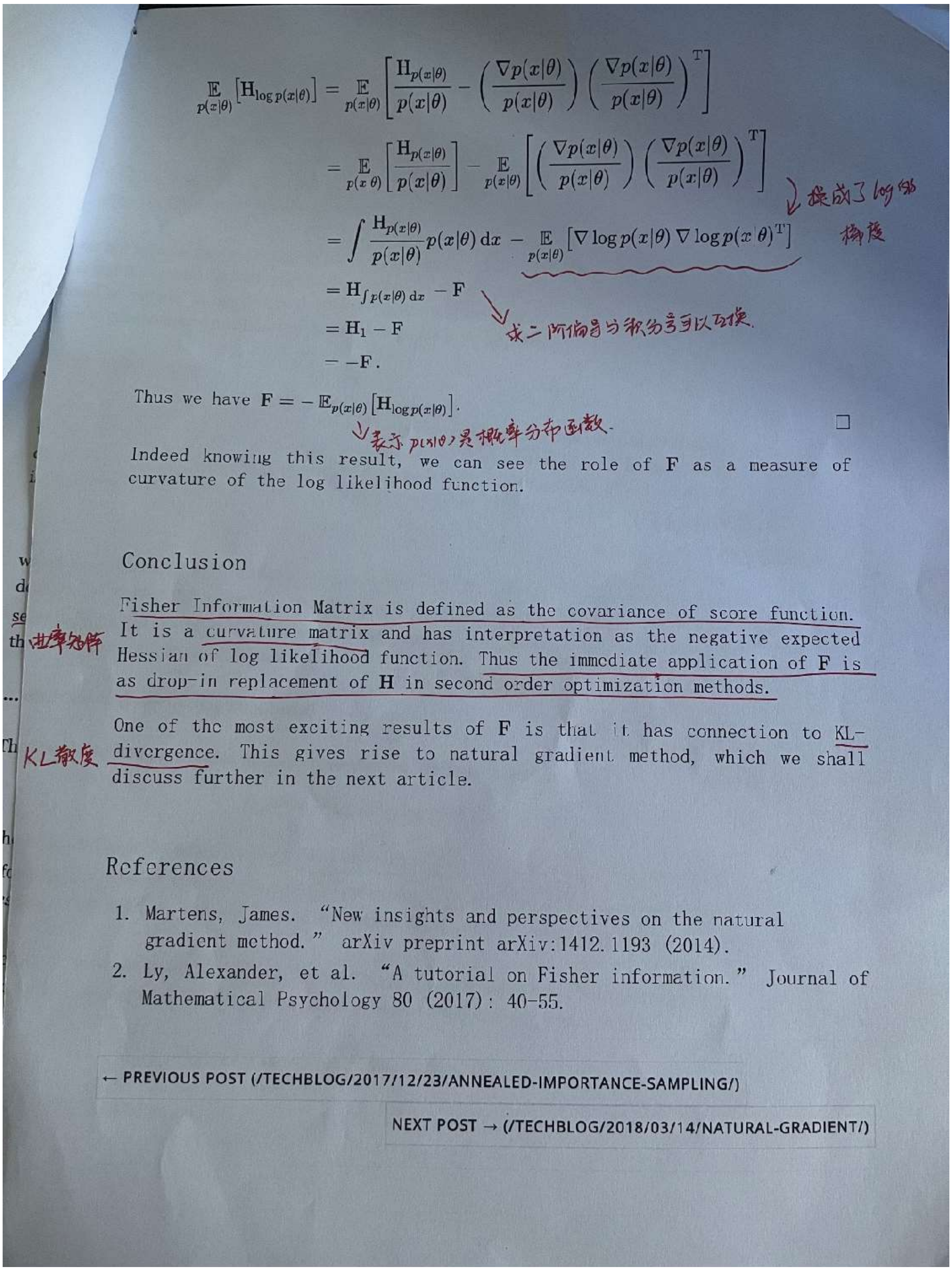
$$\mathrm{H}_{\log p(x|\theta)} = \mathrm{J}\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)$$

整体对θ求梯度

$$\begin{aligned}
&= \frac{\mathrm{H}_{p(x|\theta)}\, p(x|\theta) - \nabla p(x|\theta)\, \nabla p(x|\theta)^{\mathrm{T}}}{p(x|\theta)\, p(x|\theta)} \\
&= \frac{\mathrm{H}_{p(x|\theta)}\, p(x|\theta)}{p(x|\theta)\, p(x|\theta)} - \frac{\nabla p(x|\theta)\, \nabla p(x|\theta)^{\mathrm{T}}}{p(x|\theta)\, p(x|\theta)} \\
&= \frac{\mathrm{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}},
\end{aligned}$$

where the second line is a result of applying quotient rule of derivative. Taking expectation wrt. our model, we have:

商的求导法则



$$\begin{aligned}
\underset{p(x|\theta)}{\mathbb{E}}\left[\mathbf{H}_{\log p(x|\theta)}\right] &= \underset{p(x|\theta)}{\mathbb{E}}\left[\frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} - \left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}}\right] \\
&= \underset{p(x|\theta)}{\mathbb{E}}\left[\frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)}\right] - \underset{p(x|\theta)}{\mathbb{E}}\left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)\left(\frac{\nabla p(x|\theta)}{p(x|\theta)}\right)^{\mathrm{T}}\right] \\
&= \int \frac{\mathbf{H}_{p(x|\theta)}}{p(x|\theta)} p(x|\theta)\,\mathrm{d}x - \underbrace{\underset{p(x|\theta)}{\mathbb{E}}\left[\nabla \log p(x|\theta)\, \nabla \log p(x|\theta)^{\mathrm{T}}\right]} \\
&= \mathbf{H}_{\int p(x|\theta)\,\mathrm{d}x} - \mathbf{F} \\
&= \mathbf{H}_1 - \mathbf{F} \\
&= -\mathbf{F}\,.
\end{aligned}$$

换成了log的梯度

求二阶偏导与积分可以互换.

Thus we have $\mathbf{F} = -\mathbb{E}_{p(x|\theta)}\left[\mathbf{H}_{\log p(x|\theta)}\right]$. □

表示p(x|θ)是概率分布函数.

Indeed knowing this result, we can see the role of F as a measure of curvature of the log likelihood function.

Conclusion

Fisher Information Matrix is defined as the covariance of score function. It is a curvature matrix and has interpretation as the negative expected Hessian of log likelihood function. Thus the immediate application of F is as drop-in replacement of H in second order optimization methods.

曲率矩阵

One of the most exciting results of F is that it has connection to KL-divergence. This gives rise to natural gradient method, which we shall discuss further in the next article.

KL散度

References

1. Martens, James. "New insights and perspectives on the natural gradient method." arXiv preprint arXiv:1412.1193 (2014).
2. Ly, Alexander, et al. "A tutorial on Fisher information." Journal of Mathematical Psychology 80 (2017): 40-55.

← PREVIOUS POST (/TECHBLOG/2017/12/23/ANNEALED-IMPORTANCE-SAMPLING/)

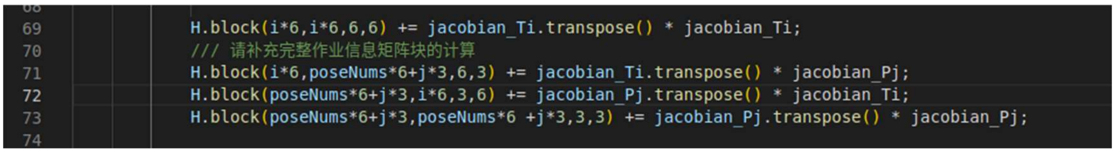
NEXT POST → (/TECHBLOG/2018/03/14/NATURAL-GRADIENT/)

3、补充作业代码中单目 BA 信息矩阵的计算，并输出正确的结果。正确的结果为：奇异值的最后 7 维接近于 0，表明零空间的维度为 7。

代码中，jacobian_Pj 是重投影误差对对特征点/路标点的雅可比矩阵（表示在世界坐标系下），jacobian_Ti 表示重投影误差对相机位姿的雅可比矩阵，具体计算代码中已经给出，推导在《十四讲》中有，我还没有推导过。

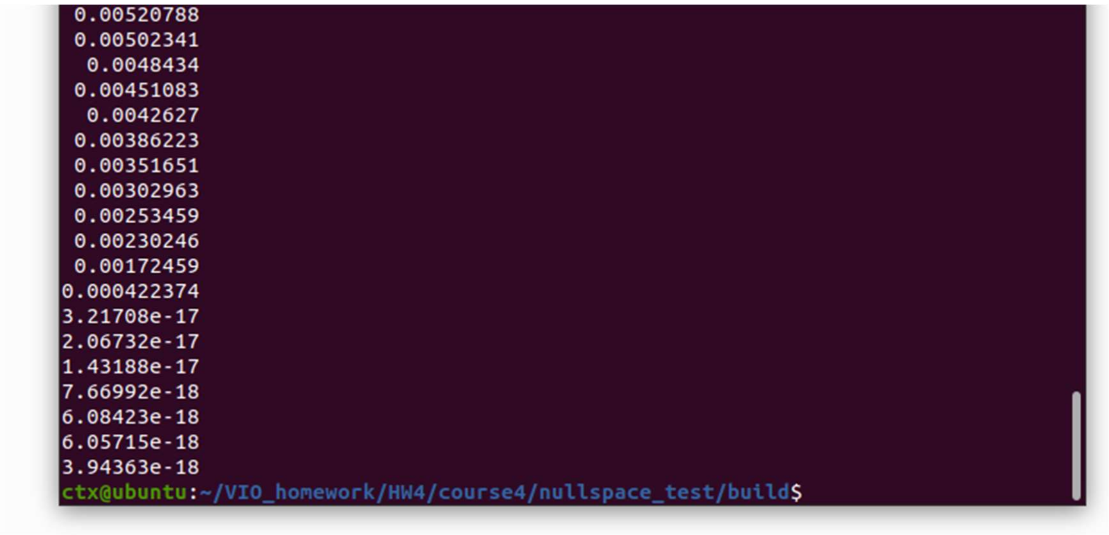
相机 Pose 是六维数据，路标点坐标是三维数据，代码中实际计算的 H 矩阵的横纵坐标对应的是按照先把所有 Pose 排列，再把所有路标点排列，进而求对应位置的 H 矩阵。

H 矩阵在相应位置上的子矩阵快就是 J^T * J，J 分别要对应上各自的雅可比矩阵，是对相机位姿求导还是对路标点求导。



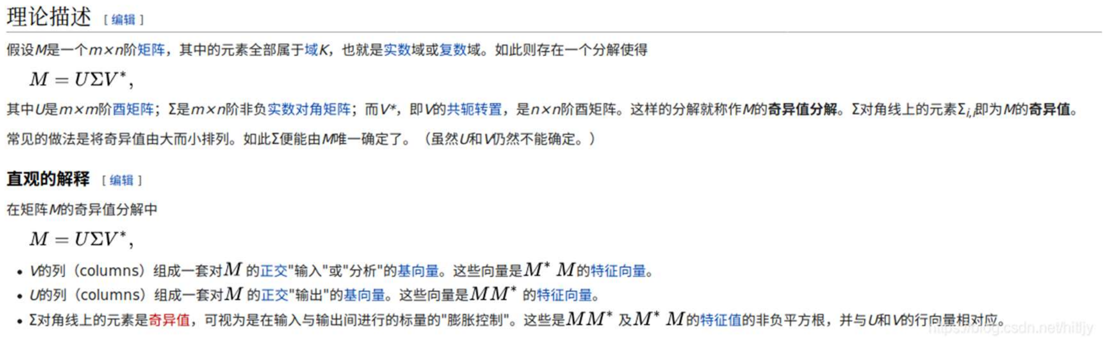
```
            H.block(i*6,i*6,6,6) += jacobian_Ti.transpose() * jacobian_Ti;
            /// 请补充完整作业信息矩阵块的计算
            H.block(i*6,poseNums*6+j*3,6,3) += jacobian_Ti.transpose() * jacobian_Pj;
            H.block(poseNums*6+j*3,i*6,3,6) += jacobian_Pj.transpose() * jacobian_Ti;
            H.block(poseNums*6+j*3,poseNums*6 +j*3,3,3) += jacobian_Pj.transpose() * jacobian_Pj;
```

最后输出的是 H 矩阵的奇异值，从大到小排列，奇异值的最后 7 维接近于 0，表明零空间的维度为 7。



```
 0.00520788
 0.00502341
  0.0048434
 0.00451083
  0.0042627
 0.00386223
 0.00351651
 0.00302963
 0.00253459
 0.00230246
 0.00172459
0.000422374
3.21708e-17
2.06732e-17
1.43188e-17
7.66992e-18
6.08423e-18
6.05715e-18
3.94363e-18
ctx@ubuntu:~/VIO_homework/HW4/course4/nullspace_test/build$
```

关于奇异值的意义：



理论描述 [编辑]

假设M是一个$m \times n$阶矩阵，其中的元素全部属于域K，也就是实数域或复数域。如此则存在一个分解使得

$$M = U\Sigma V^*,$$

其中U是$m \times m$阶酉矩阵；Σ是$m \times n$阶非负实数对角矩阵；而V^*，即V的共轭转置，是$n \times n$阶酉矩阵。这样的分解就称作M的**奇异值分解**。Σ对角线上的元素$\Sigma_{i,i}$即为M的**奇异值**。

常见的做法是将奇异值由大而小排列。如此Σ便能由M唯一确定了。（虽然U和V仍然不能确定。）

直观的解释 [编辑]

在矩阵M的奇异值分解中

$$M = U\Sigma V^*,$$

- V的列（columns）组成一套对M的正交"输入"或"分析"的基向量。这些向量是$M^* M$的特征向量。
- U的列（columns）组成一套对M的正交"输出"的基向量。这些向量是MM^*的特征向量。
- Σ对角线上的元素是奇异值，可视为是在输入与输出间进行的标量的"膨胀控制"。这些是MM^*及$M^* M$的特征值的非负平方根，并与U和V的行向量相对应。

对角矩阵的非零对角元素的个数对应于矩阵 M 的秩。与零奇异值对应的右奇异向量生成矩阵 M 的零空间，与非零奇异值对应的左奇异向量则生成矩阵 M 的列空间。在线性代数数值计算中奇异值分解一般用于确定矩阵的有效秩，这是因为，由于舍入误差，秩亏矩阵的零奇异值可能会表现为很接近零的非零值。

还没有完全理解，也还需要再参考矩阵分析的知识，这部分还没有做完，要赶路回家，在第二次交作业会给出答案。