最小二乘法

耿修瑞

中国科学院空天信息创新研究院

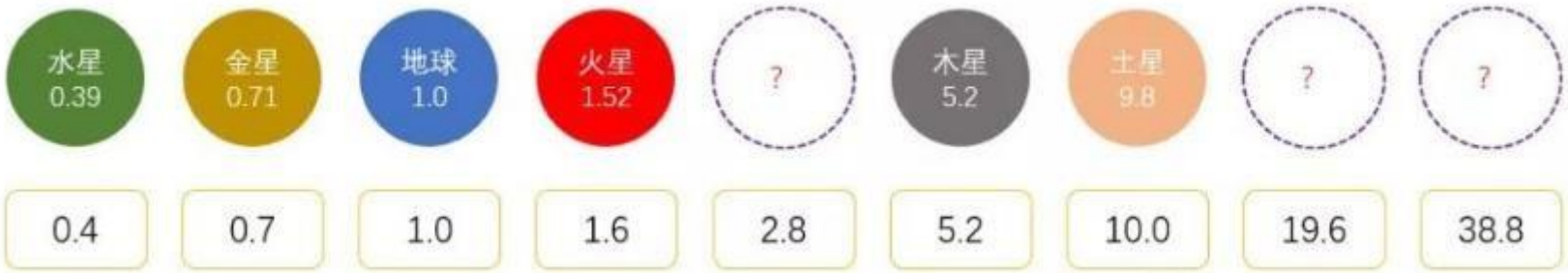
gengxr@sina.com.cn

2025.3

主要内容

- 问题背景
- 基本原理
- 变量问题
- 几何解释
- 概率解释
- 约束最小二乘
- 最小二乘法的应用

问题背景



1766年，德国有一位名叫约翰·提丢斯（Johann Daniel Titius）的中学教师，写下了下面的数列：

$$\frac{(3 \times 2^n + 4)}{10} \qquad n = -\infty, 0, 1, 2, 3, \cdots$$

令提丢斯惊奇的是，他发现这个数列的每一项与当时已知的六大行星（即水星、金星、地球、火星、木星、土星）到太阳的距离比例（地球到太阳的距离定为1个单位）有着一定的联系。

问题背景

提丢斯的朋友，天文学家波得深知这一发现的重要意义，就于1772年公布了提丢斯的这一发现，这串数从此引起了科学家的极大重视；并被称为提丢斯——波得定律即太阳系行星与太阳的平均距离．当时，人们还没有发现天王星、海王星，以为土星就是距太阳最远的行星。1781年，英籍德国人赫歇尔 在接近19.6的位置上（即数列中的第八项）发现了天王星，从此，人们就对这一定则深信不疑了。根据这一定则，在数列的第五项即2.8的位置上也应该对应一颗行星，只是还没有被发现。于是，许多天文学家和天文爱好者便以极大的热情，踏上了寻找这颗新行星的征程。

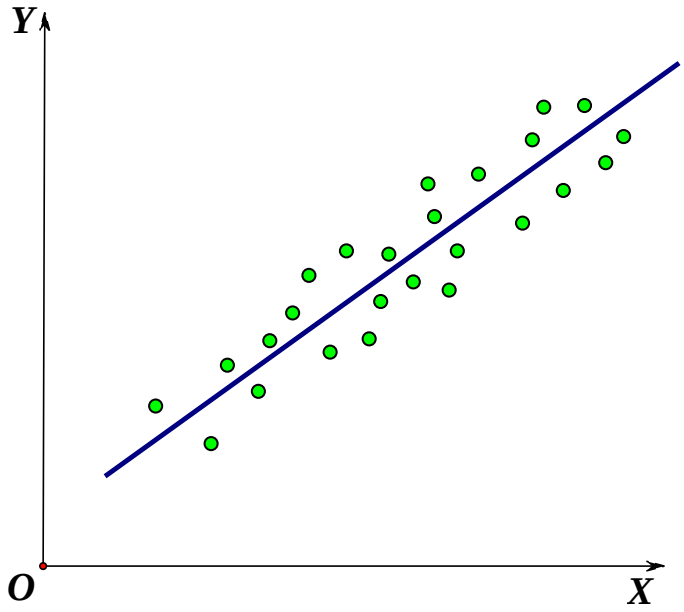
问题背景

1801年，意大利天文学家朱赛普·皮亚齐发现了第一颗小行星谷神星。经过40天的跟踪观测后，由于谷神星运行至太阳背后，使得皮亚齐失去了谷神星的位置。随后全世界的科学家利用皮亚齐的观测数据开始寻找谷神星，但是根据大多数人计算的结果来寻找谷神星都没有结果。时年24岁的高斯也采用了一种新方法（即为最小二乘法）计算了谷神星的轨道。奥地利天文学家海因里希·奥尔伯斯根据高斯计算出来的轨道重新发现了谷神星。高斯使用的最小二乘法的方法发表于1809年他的著作《天体运动论》中，而法国科学家勒让德于1806年独立发现最小二乘法，但因不为时人所知而默默无闻。两人曾为谁最早创立最小二乘法原理发生争执。最小二乘法自创立以来，在自然科学乃至社会科学的各个领域产生了广泛的应用。

高斯（Gauss, 1777~1855），德国数学家，最小二乘法创始人。

基本原理

最小二乘法通常用来研究两个变量或者多个变量之间的关系。以常用的直线回归为例，已知一组观测点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 分布在直角坐标系中（如图），如何用一条直线最佳的拟合这些散点？

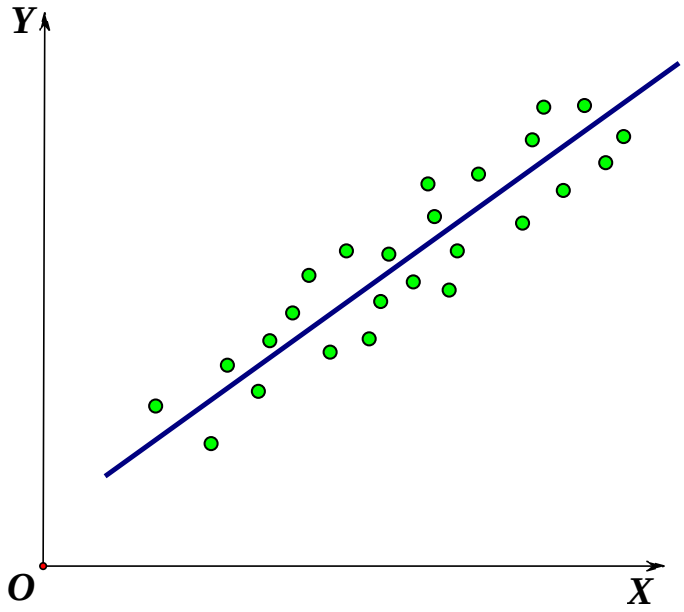


直线拟合示意图

$$y = ax + b$$

基本原理

最小二乘法通常用来研究两个变量或者多个变量之间的关系。以常用的直线回归为例，已知一组观测点 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 分布在直角坐标系中（如图），如何用一条直线最佳的拟合这些散点？



直线拟合示意图

$$y = ax + b$$

基本原理

最小二乘法的一般解法（行空间角度）

最小二乘法求解，相当于寻找a,b。使得如下公式尽量成立

$$\begin{cases} y_1 = ax_1 + b \\ y_2 = ax_2 + b \\ \vdots \\ y_n = ax_n + b \end{cases}$$

可建立优化模型如下：

$$\min_{a,b} f(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$$

基本原理

对自变量a,b求导，并令其为0：

$$\begin{cases} \dfrac{\partial f}{\partial a} = 2\sum_{i=1}^{n}(ax_i + b - y_i)x_i = 0 \\ \dfrac{\partial f}{\partial b} = 2\sum_{i=1}^{n}(ax_i + b - y_i) = 0 \end{cases}$$

化简可得解为：

$$\begin{cases} a = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ b = \dfrac{1}{n}\sum_{i=1}^{n} y_i - \dfrac{a}{n}\sum_{i=1}^{n} x_i \end{cases}$$

基本原理

最小二乘法的矩阵解法（列空间角度）

首先记：

$$\mathbf{x}=\begin{bmatrix}x_1 & x_2 & \cdots & x_n\end{bmatrix}^{\mathrm{T}},\ \mathbf{y}=\begin{bmatrix}y_1 & y_2 & \cdots & y_n\end{bmatrix}^{\mathrm{T}},\ \mathbf{1}=\begin{bmatrix}1 & 1 & \cdots & 1\end{bmatrix}^{\mathrm{T}}$$

则线性模型可以表示为：

$$\mathbf{y}=a\mathbf{x}+b\mathbf{l}$$

进一步，令

$$\mathbf{X}=\begin{bmatrix}\mathbf{x} & \mathbf{l}\end{bmatrix}\qquad \mathbf{c}=\begin{bmatrix}a & b\end{bmatrix}^{\mathrm{T}}$$

则线性模型可以进一步表示为：

$$\mathbf{y}=\mathbf{X}\mathbf{c}$$

基本原理

优化模型转化为：

$$\min_{\mathbf{c}} f(\mathbf{c}) = \left\| \mathbf{y} - \mathbf{X}\mathbf{c} \right\|^2$$

对自变量求导，并令其为0

$$f_{\mathbf{c}}^{'} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{c} - 2\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{0}$$

可得模型的解为

$$\mathbf{c} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad \mathbf{X}^{\#} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}$$

基本原理

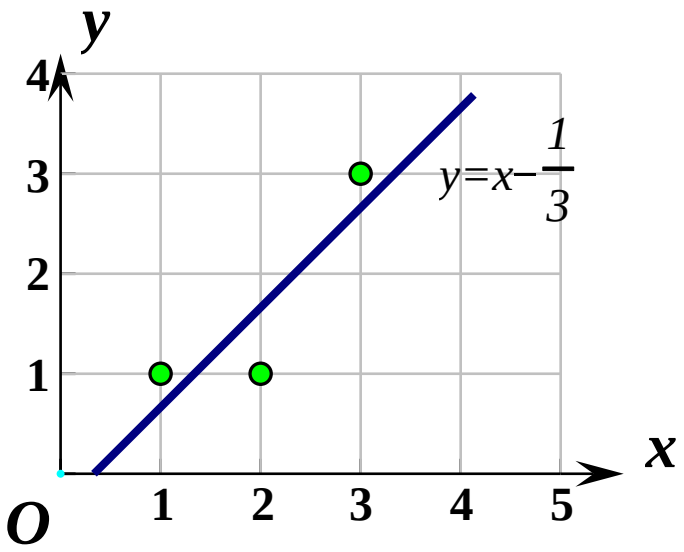
一个简单的例子：假设有三个观测点，分别为

$$(1,1),(2,1)和(3,3)$$

方法1：

$$a=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}=\frac{3(1*1+2*1+3*3)-(1+2+3)(1+1+3)}{3(1^2+2^2+3^2)-(1+2+3)^2}=1$$

$$b=\frac{1}{n}\sum_{i=1}^{n}y_i-\frac{a}{n}\sum_{i=1}^{n}x_i=\frac{1}{3}(1+1+3)-\frac{1}{3}(1+2+3)=-\frac{1}{3}$$



方法2：

$$\mathbf{c}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}=\left(\begin{bmatrix}1&2&3\\1&1&1\end{bmatrix}\begin{bmatrix}1&1\\2&1\\3&1\end{bmatrix}\right)^{-1}\begin{bmatrix}1&2&3\\1&1&1\end{bmatrix}\begin{bmatrix}1\\1\\3\end{bmatrix}=\begin{bmatrix}1\\-1/3\end{bmatrix}$$

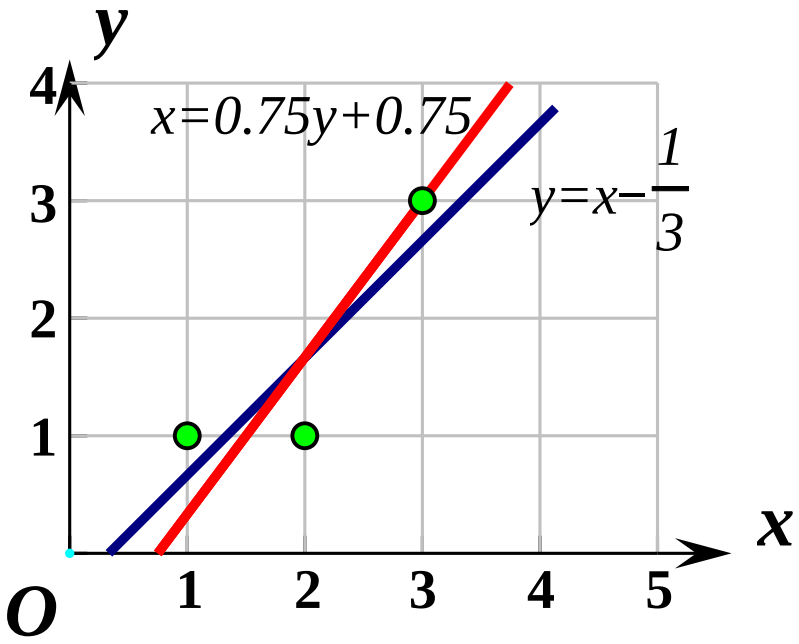
变量问题

最小二乘法中的变量问题

在前面，我们通过给出一组散点的最佳直线拟合阐述了最小二乘法的基本原理。对于同样的一组散点，我们用 $x = a'y + b'$ 来拟合这组散点是否可以得到同样的结果呢？

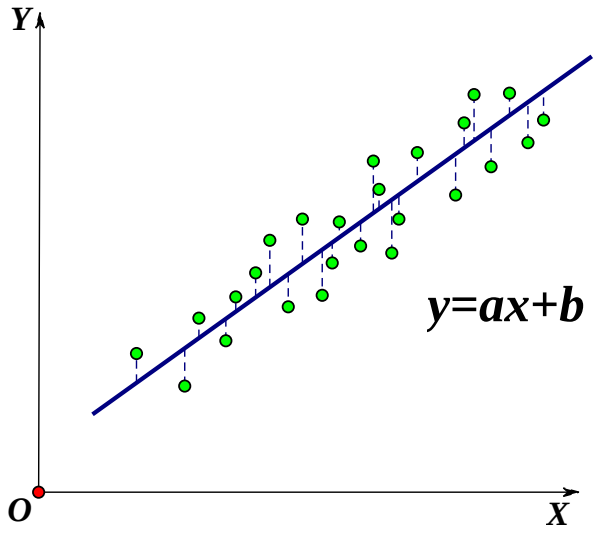
$$\begin{cases} x_1 = a'y_1 + b' \\ x_2 = a'y_2 + b' \\ \vdots \\ x_n = a'y_n + b' \end{cases}$$

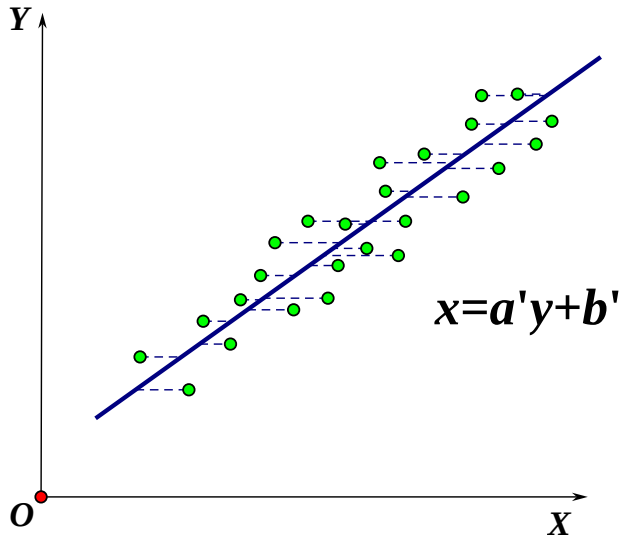
$$\mathbf{c}' = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{x} = \begin{bmatrix} 0.75 \\ 0.75 \end{bmatrix}$$



变量问题

最小二乘法中的变量问题

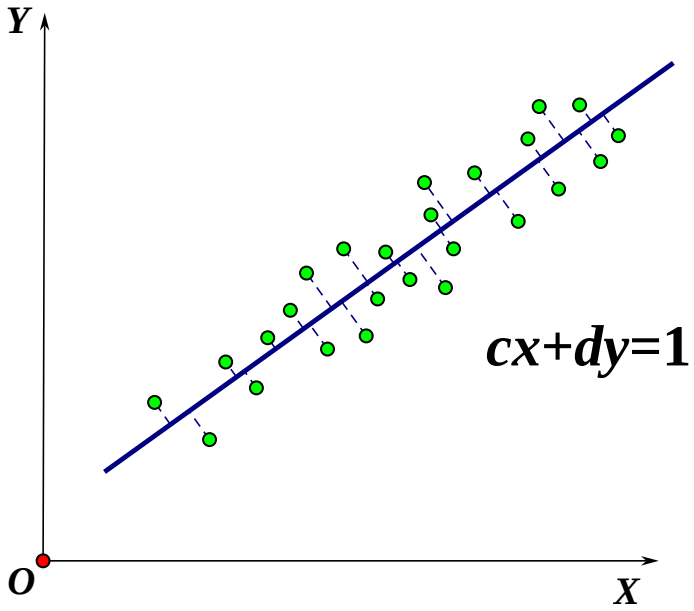




当x和y互为自变量和因变量直线拟合的直观解释

变量问题

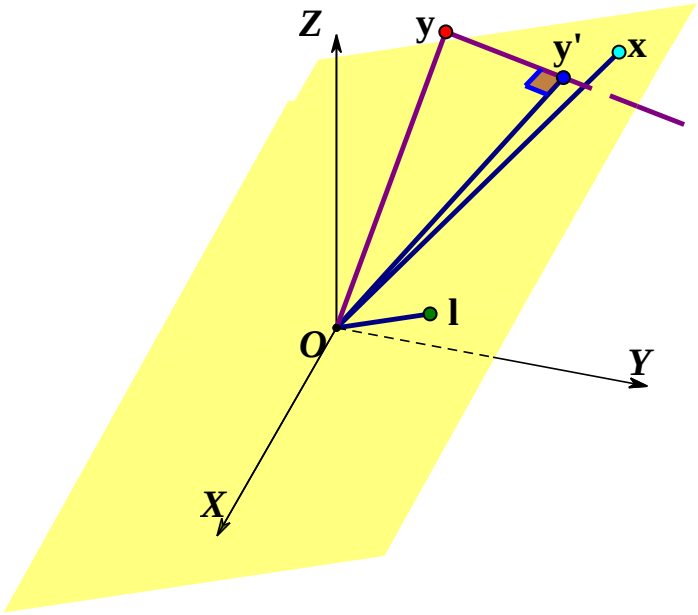
总体最小二乘法



总体最小二乘法的直观理解

几何解释

最小二乘法的几何解释（线性变换角度）



最小二乘法对应的样本空间的几何：其中 $\mathbf{x}=[1\ \ 2\ \ 3]^T$，$\mathbf{y}=[1\ \ 1\ \ 3]^T$，$\mathbf{1}=[1\ \ 1\ \ 1]^T$。最小二乘揭示的是三个散点之间的关系。当以$y=ax+b$ 来拟合散点时，对应到样本空间，相当于寻求向量**y**到**x**和**1**这两个向量所张成平面的投影

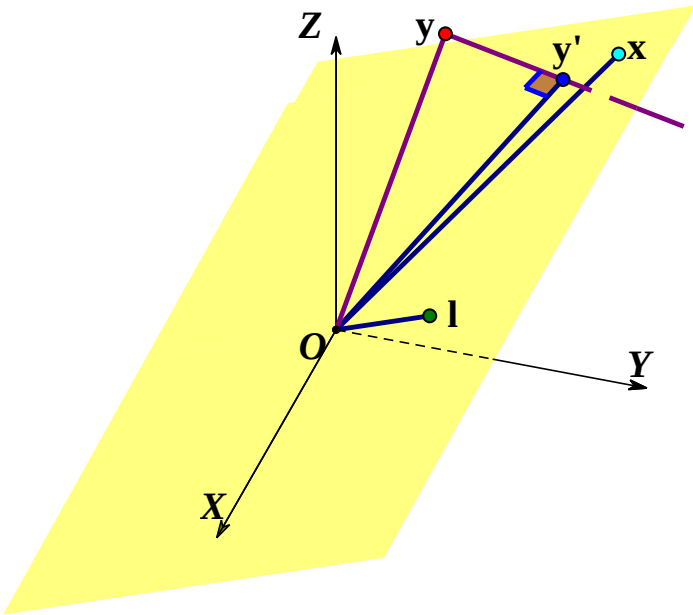
几何解释

最小二乘法的几何解释（线性变换角度）

$$\min_{\mathbf{c}} f(\mathbf{c}) = \left\| \mathbf{y} - \mathbf{X}\mathbf{c} \right\|^2$$

$$\mathbf{c} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

$$\mathbf{y}' = \mathbf{X}\mathbf{c} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}}\mathbf{y}$$



矩阵的广义逆：$\mathbf{X}^{\#} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}}$

投影矩阵 $\mathbf{P}_{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} \right)^{-1} \mathbf{X}^{\mathrm{T}} = \mathbf{X}\mathbf{X}^{\#}$

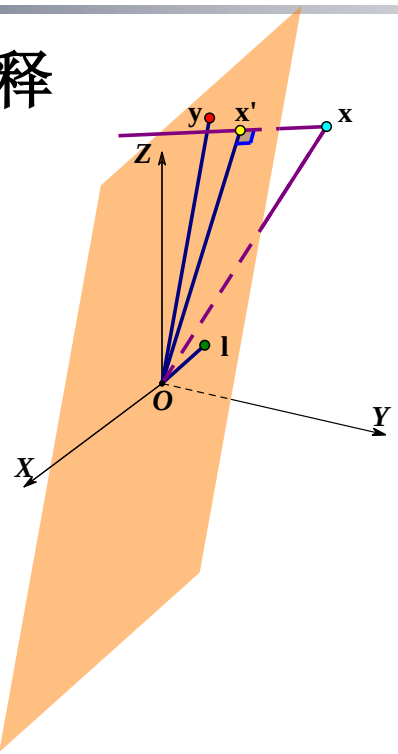
$$\mathbf{y}' = \mathbf{P}_{\mathbf{X}}\mathbf{y}$$

正交补投影算子 $\mathbf{P}_{\mathbf{X}}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{X}} = \mathbf{I} - \mathbf{X}\mathbf{X}^{\#}$

讨论：
投影矩阵的性质？

几何解释

最小二乘法的几何解释



最小二乘法对应的样本空间的几何：其中 $\mathbf{x}=[1\ \ 2\ \ 3]^T$，$\mathbf{y}=[1\ \ 1\ \ 3]^T$，$\mathbf{1}=[1\ \ 1\ \ 1]^T$。最小二乘揭示的是三个散点之间的关系。当以$x=a'y+b'$来拟合散点时，对应到样本空间，相当于寻求 $\mathbf{x}$ 到 $\mathbf{y}$ 和 $\mathbf{1}$ 这两个向量所张成平面的投影

概率解释

最小二乘法的概率解释

在前面的讨论中，因变量的总体观测误差用

$$f(a,b)=\sum_{i=1}^{n}(ax_i+b-y_i)^2$$

来表示。即对于每个观测点的因变量的观测误差，都选用模型解与观测值的差的平方来衡量。这正是相应的方法命名为最小二乘法而不是最小一乘或者最小三乘的原因所在。

Why？

概率解释

最小二乘法的概率解释

记

$$\begin{cases} y_1 = \mathbf{c}^{\mathrm{T}}\mathbf{x}_1 + \varepsilon_1 \\ y_2 = \mathbf{c}^{\mathrm{T}}\mathbf{x}_2 + \varepsilon_2 \\ \vdots \\ y_n = \mathbf{c}^{\mathrm{T}}\mathbf{x}_n + \varepsilon_n \end{cases}$$

$$\mathbf{c} = \begin{bmatrix} a & b \end{bmatrix}^{\mathrm{T}}$$

$$\mathbf{x}_i = \begin{bmatrix} x_i & 1 \end{bmatrix}^{\mathrm{T}}$$

其中 ε_i 对应第i个因变量y_i的观测误差，它对应着不能被线性模型刻画的因素。假设ε_i服从正态分布，

$$\varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$$

则其概率密度函数为，

$$f\left(\varepsilon_i\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

概率解释

这意味着，在给定 $\mathbf{x}_i$ 和参数**c**的情况下，因变量 y_i 也服从正态分布，即

$$f\left(y_i|\mathbf{x}_i;\mathbf{c}\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\mathbf{c}^{\mathrm{T}}\mathbf{x}_i\right)^2}{2\sigma^2}\right)$$

$$\mathbf{c}=\begin{bmatrix}a & b\end{bmatrix}^{\mathrm{T}}$$

$$\mathbf{x}_i=\begin{bmatrix}x_i & 1\end{bmatrix}^{\mathrm{T}}$$

假设所有的 ε_i 独立同分布，我们可以定义所有观测数据关于参数**c**的似然函数如下：

$$L(\mathbf{c})=f\left(\mathbf{y}|\mathbf{X};\mathbf{c}\right)=\prod_{i=1}^{n}f\left(y_i|\mathbf{x}_i;\mathbf{c}\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i-\mathbf{c}^{\mathrm{T}}\mathbf{x}_i\right)^2}{2\sigma^2}\right)$$

其中

$$\mathbf{y}=\begin{bmatrix}y_1 & y_2 & \cdots & y_n\end{bmatrix}^{\mathrm{T}},\ \mathbf{X}=\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix}^{\mathrm{T}}$$

概率解释

为了便于求解，定义 $l(\mathbf{c}) = \ln(L(\mathbf{c}))$

$$
\begin{aligned}
l(\mathbf{c}) &= \ln\left(L\left(\mathbf{c}\right)\right) = \ln\left(\prod_{i=1}^{n} f\left(y_i \middle| \mathbf{x}_i; \mathbf{c}\right)\right) \\
&= \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - \mathbf{c}^{\mathrm{T}}\mathbf{x}_i\right)^2}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - \mathbf{c}^{\mathrm{T}}\mathbf{x}_i\right)^2}{2\sigma^2}\right)\right) \\
&= n\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{n} \frac{\left(y_i - \mathbf{c}^{\mathrm{T}}\mathbf{x}_i\right)^2}{2\sigma^2}
\end{aligned}
$$

因此，最大化似然函数，相当于最小化

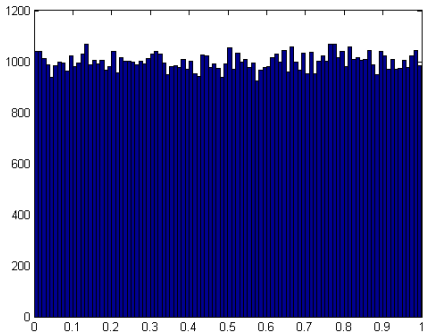
$$
\sum_{i=1}^{n}\left(y_i - \mathbf{c}^{\mathrm{T}}\mathbf{x}_i\right)^2 = \sum_{i=1}^{n}(ax_i + b - y_i)^2 = f(a,b)
$$

概率解释

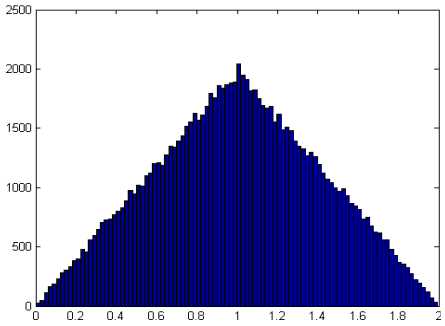
从上面的推导可以看出，各个观测点的模型误差满足独立同分布的高斯分布是最小二乘法能够行之有效的概率机制前提。如果此条件不能得到满足，用最小二乘法拟合数据将不能得到最优解。而根据**中心极限定理**，这个条件一般情况下是近似成立的。

中心极限定理：对于多个独立的随机变量，它们和的平均，满足

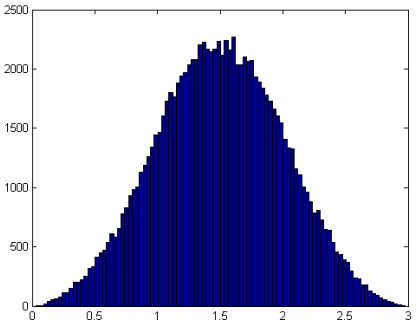
$$\lim_{n\to\infty} \frac{x_1 + x_2 + \cdots + x_n}{n} \rightarrow \text{高斯分布}$$



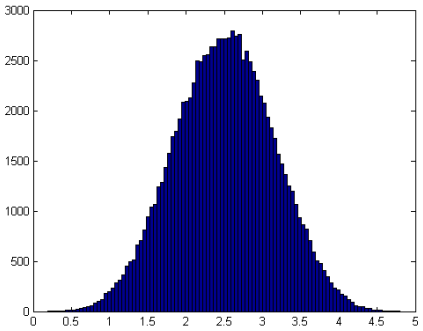
一个均匀分布



两个均匀分布之和



三个均匀分布之和



五个均匀分布之和

约束最小二乘

- 讨论：四种模型所对应的几何投影

$$\min_{\mathbf{c}} \|\mathbf{y}-\mathbf{X}\mathbf{c}\|^2$$ 无约束

$$\begin{cases} \min\limits_{\mathbf{c}} \|\mathbf{y}-\mathbf{X}\mathbf{c}\|^2 \\ \mathbf{c}^{\mathrm{T}}\mathbf{1}=1 \end{cases}$$ 等式约束

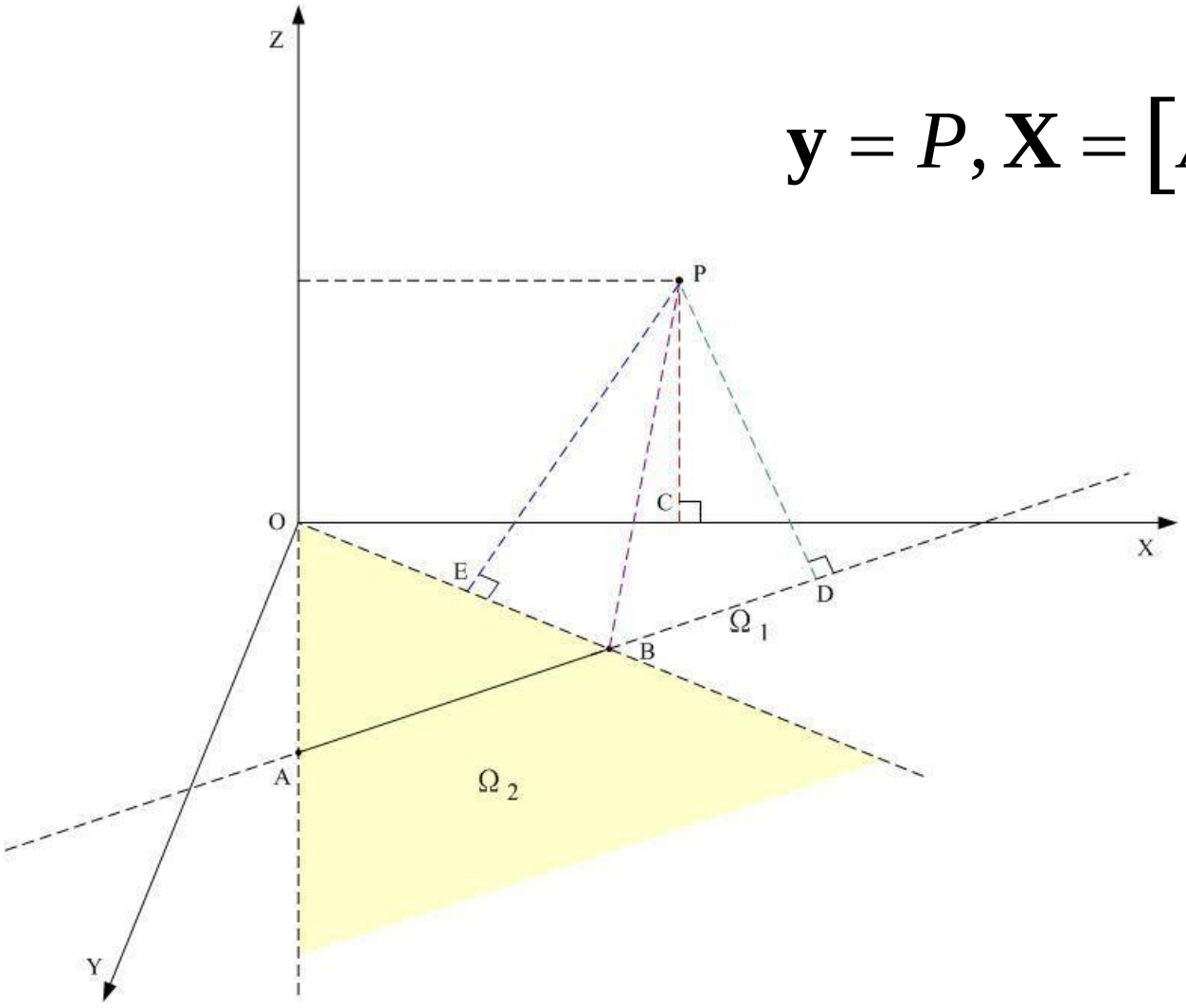
$$\begin{cases} \min\limits_{\mathbf{c}} \|\mathbf{y}-\mathbf{X}\mathbf{c}\|^2 \\ \mathbf{c}\geq\mathbf{0} \end{cases}$$ 不等式约束

$$\begin{cases} \min\limits_{\mathbf{c}} \|\mathbf{y}-\mathbf{X}\mathbf{c}\|^2 \\ \mathbf{c}^{\mathrm{T}}\mathbf{1}=1 \\ \mathbf{c}\geq\mathbf{0} \end{cases}$$ 全约束

其中**1**为所有元素都为1的列向量，**0**为所有元素都为0的列向量

约束最小二乘

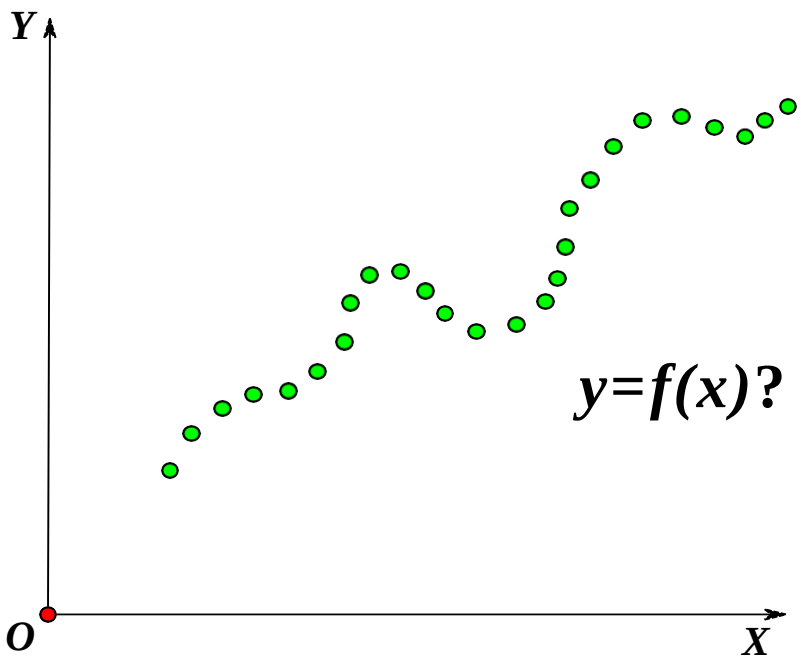
$$\mathbf{y} = P, \mathbf{X} = \begin{bmatrix} A & B \end{bmatrix}$$



讨论：

- 除了可以拟合直线，最小二乘可以拟合任意次数的多项式，乃至任意已知表达式的曲线。那么如何用最小二乘法拟合任意形状的未知表达式的曲线呢？

$$y = f(x) = x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n$$

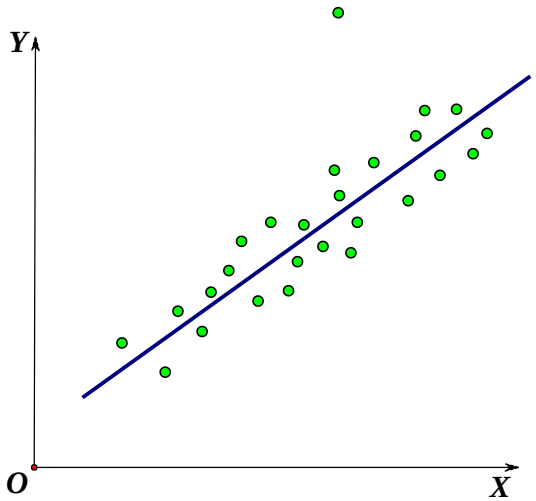


思考题（每题1分，共3分）：

- 病态情形：当$\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 不可逆时，如何处理？

$$\mathbf{c}=\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- 异常情形：当数据中存在异常点时，如何处理？



思考题（每题1分，共3分）：

- 求解如下等式约束的最小二乘解

$$\begin{cases} \min\limits_{\mathbf{c}} \left\| \mathbf{y} - \mathbf{X}\mathbf{c} \right\|^2 \\ s.t.\ \ \mathbf{c}^{\mathrm{T}}\mathbf{1} = 1 \end{cases}$$

其中为**1**所有元素都为1的列向量

最小二乘法的应用

最小二乘法的遥感应用：

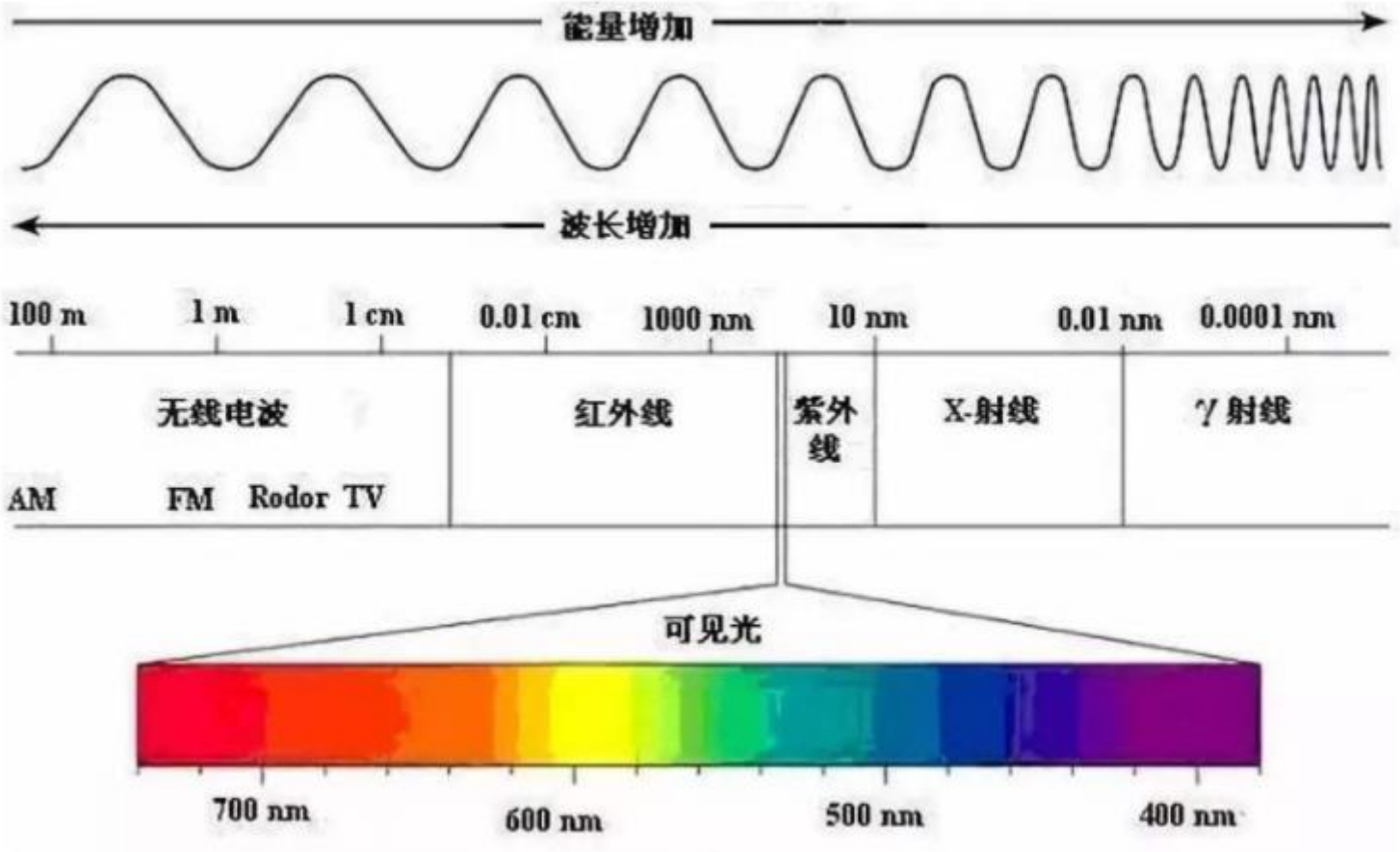
遥感图像简介

遥感图像混合像元分析

遥感图像条带消除

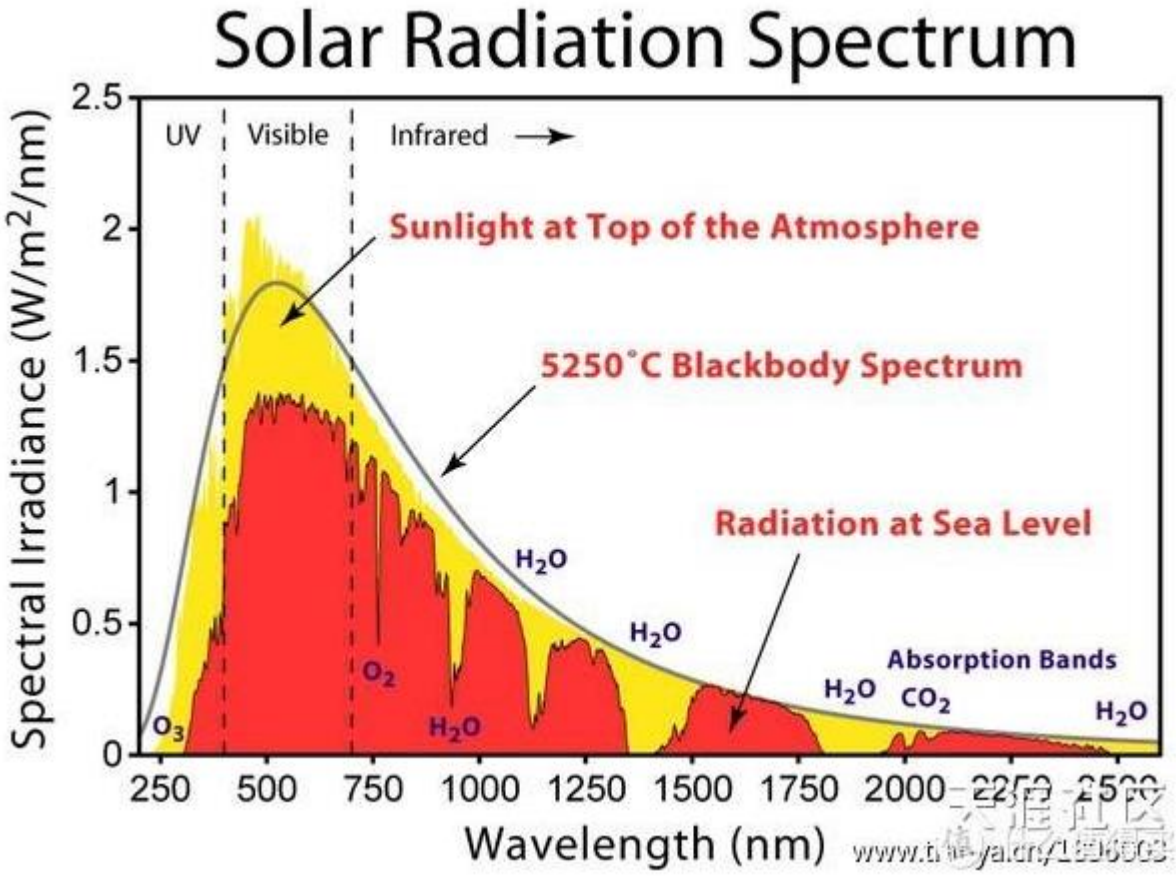
最小二乘法的应用

遥感图像、电磁波谱（讨论不同波段范围的用途？）



最小二乘法的应用

太阳辐射光谱



最小二乘法的应用

遥感图像发展趋势：

空间分辨率

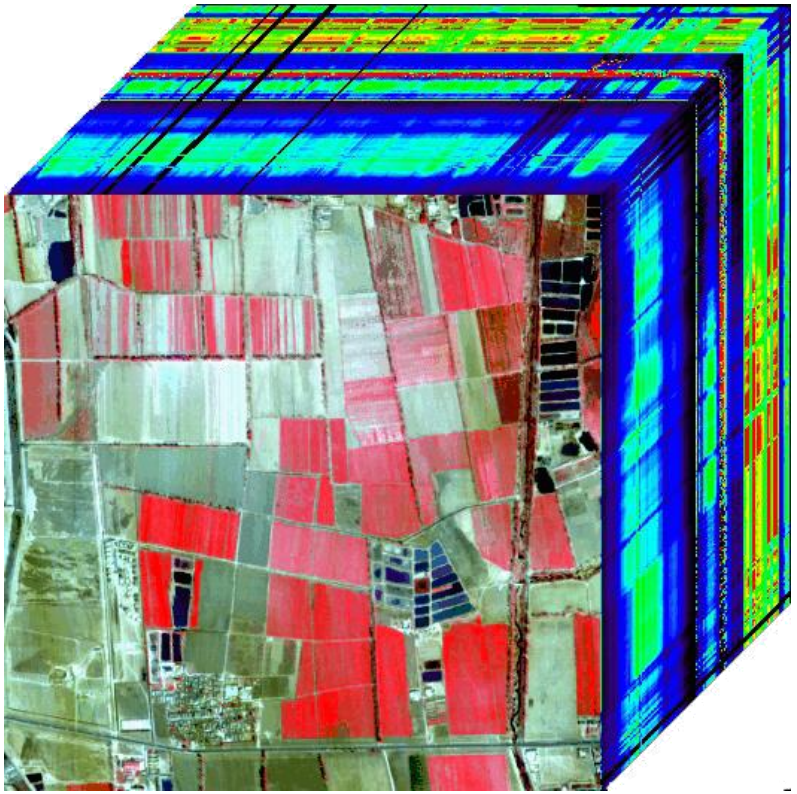
光谱分辨率

时间分辨率

波谱范围

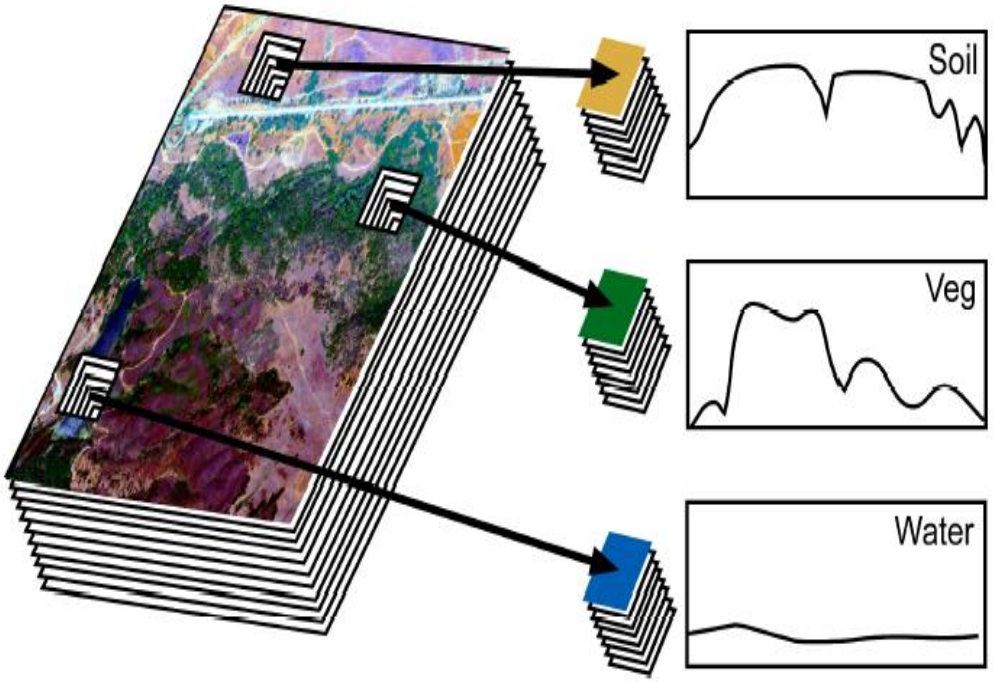
最小二乘法的应用

全色-(彩色)-多光谱-高光谱



最小二乘法的应用

高光谱遥感的概念（ENVI演示）

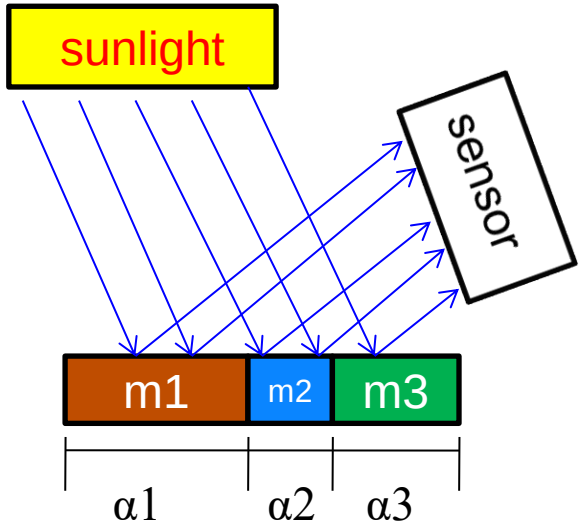


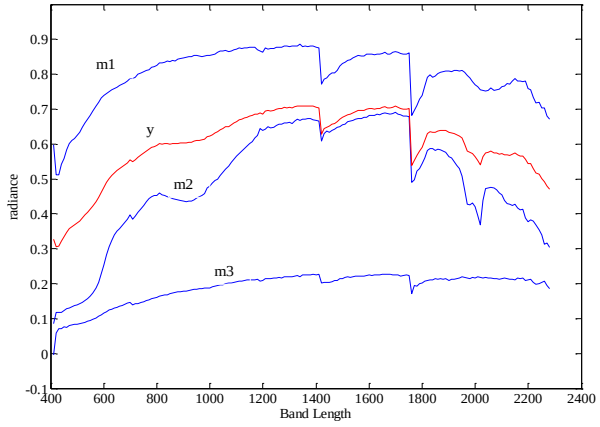
从每个象元均可提取一条连续的光谱曲线

最小二乘法的应用

混合像元分析

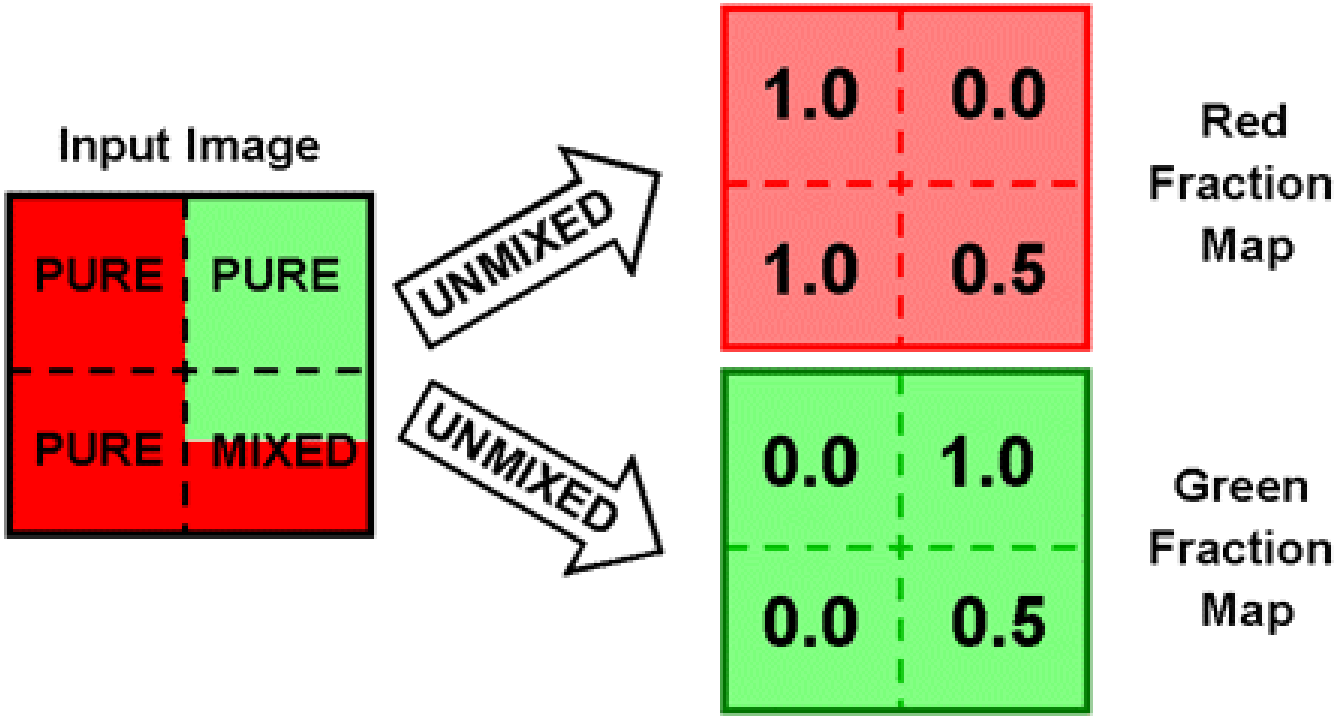
遥感图象中每个象元所对应的地表，往往包含不同的覆盖类型，他们有着不同的光谱响应特征。由于传感器分辨率的限制，混合像元普遍存在于遥感图像中。





最小二乘法的应用

混合像元分析



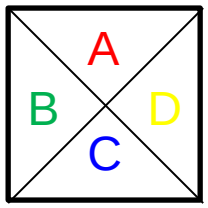
最小二乘法的应用

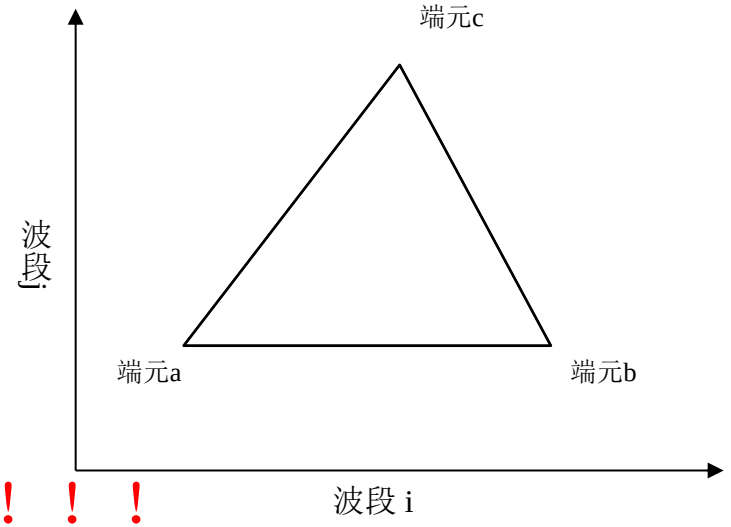
混合像元分析

$$\mathbf{p} = \sum_{i=1}^{N} c_i \mathbf{e}_i + \mathbf{n} = \mathbf{E}\mathbf{c} + \mathbf{n}$$

$$\sum_{i=1}^{N} c_i = 1$$

$$0 \le c_i \le 1$$





注意，这里有约束！！！

图象上的线性混合模型

光谱空间的单形体结构

最小二乘法的应用

- 线性混合模型（灰度图像情形）

$$p = \sum_{i=1}^{N} c_i e_i + n$$

一个方程，N个未知量，方程有无穷解，但解都没意义

- 线性混合模型（多光谱图像情形，波段数M）

$$\mathbf{p} = \sum_{i=1}^{N} c_i \mathbf{e}_i + \mathbf{n} = \mathbf{E}\mathbf{c} + \mathbf{n}$$

M个方程，N个未知量，方程组解的情况取决于M和N的大小

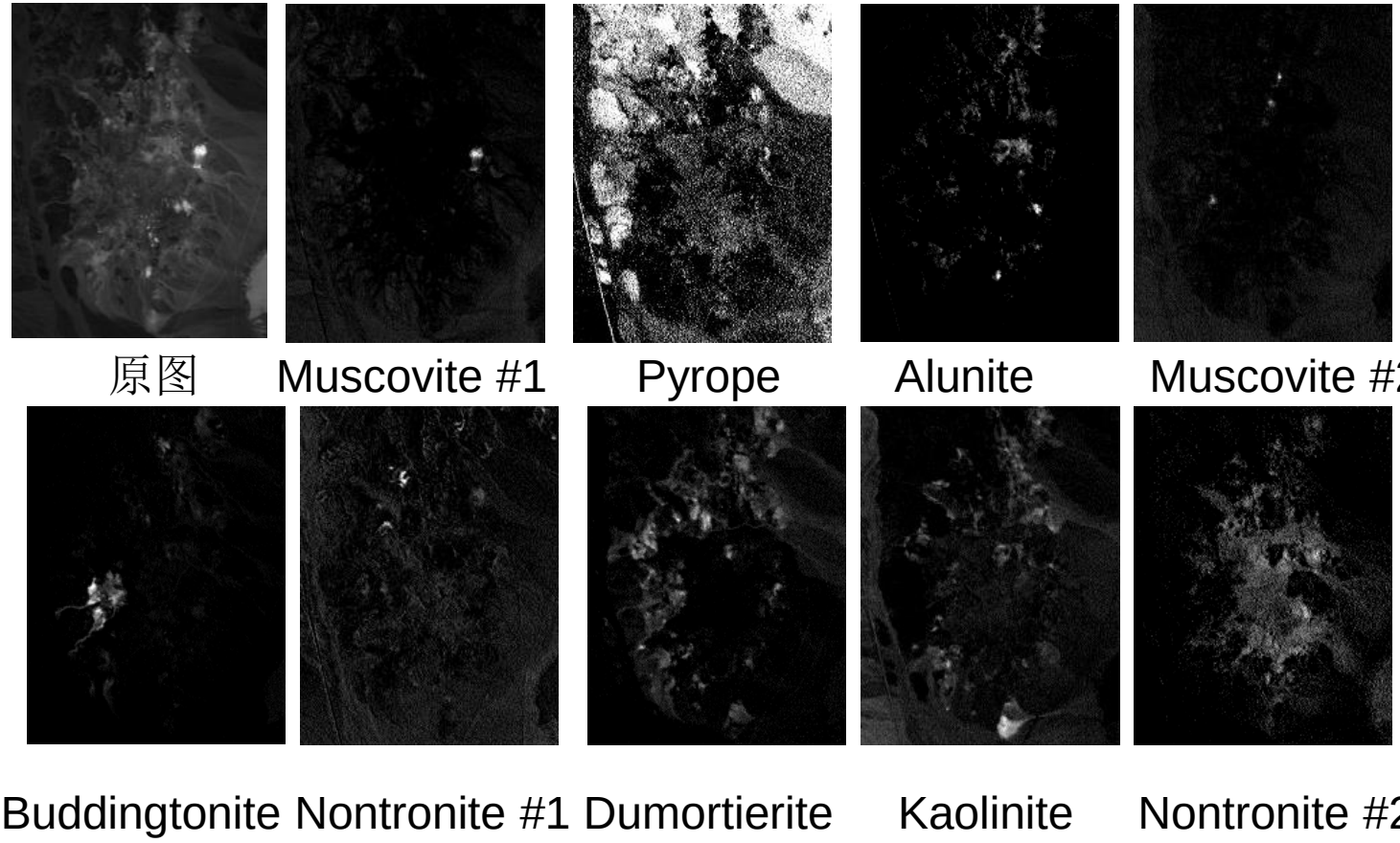
最小二乘法的应用

- 线性混合模型（高光谱图像情形，波段数M）

$$\mathbf{p}=\sum_{i=1}^{N}c_i\mathbf{e}_i+\mathbf{n}=\mathbf{E}\mathbf{c}+\mathbf{n}$$

对于高光谱图像，一般M远大于N,方程有唯一的最小二乘解

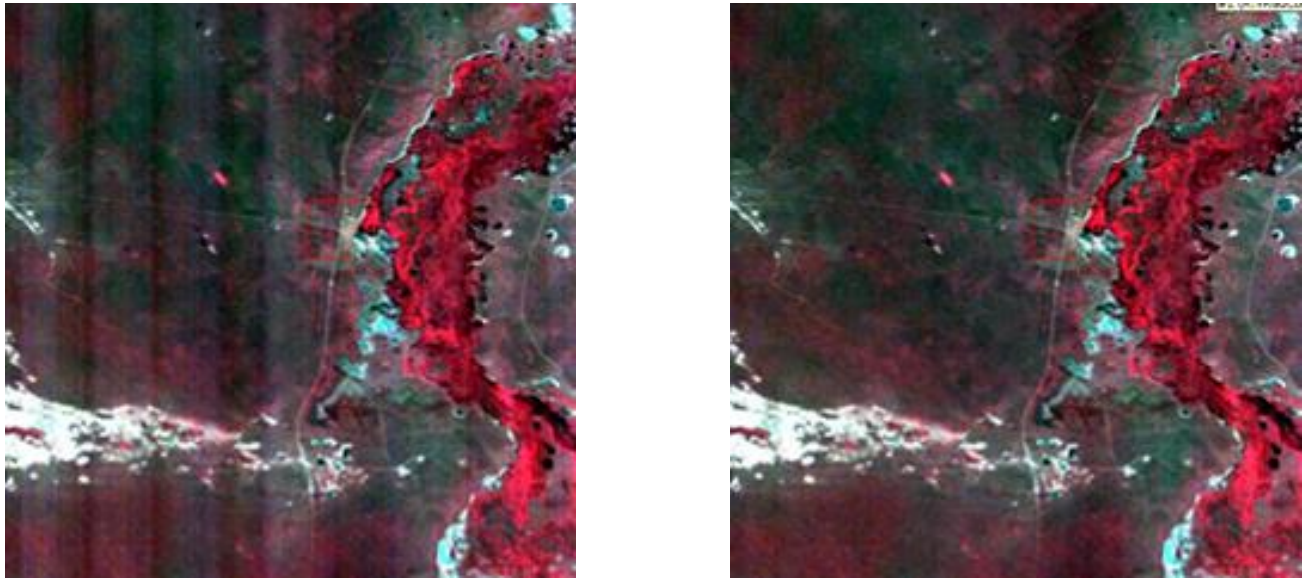
最小二乘法的应用



美国cuprite区域Aviris高光谱数据混合像元分解部分结果图

最小二乘法的应用

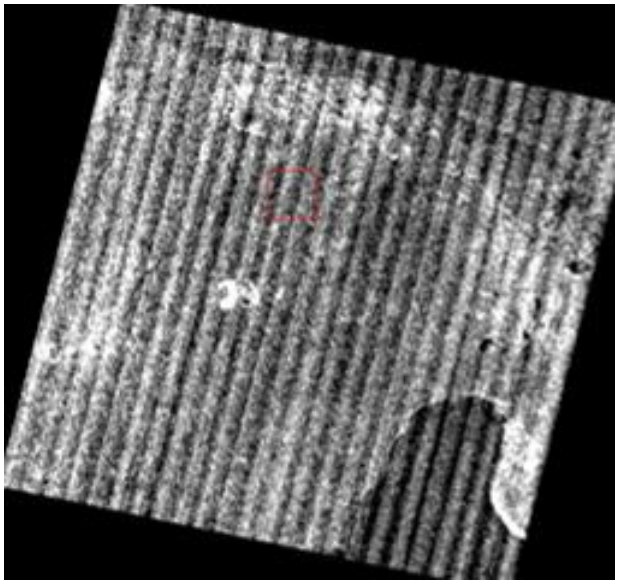
条带消除

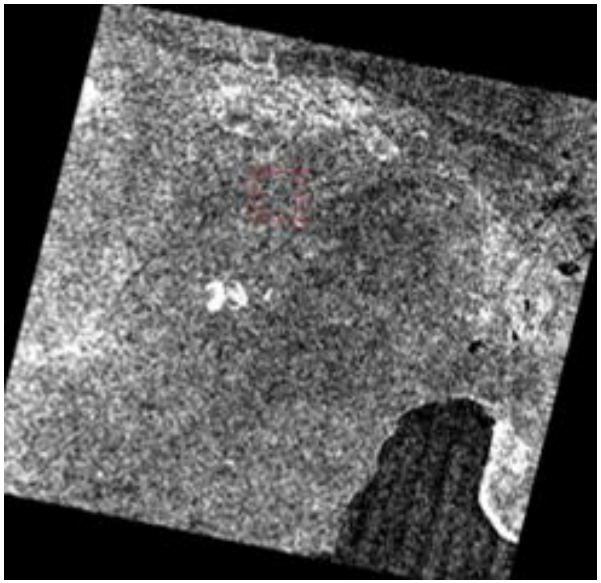


VRadCor（赵永超）批处理HJ-1A已经地面处理系统相对辐射校正后的高光谱单景图像中残存条纹的前(左)后(右)效果对比

最小二乘法的应用

条带消除

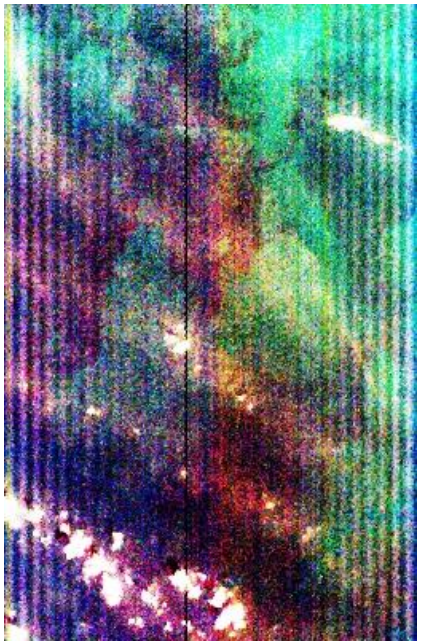


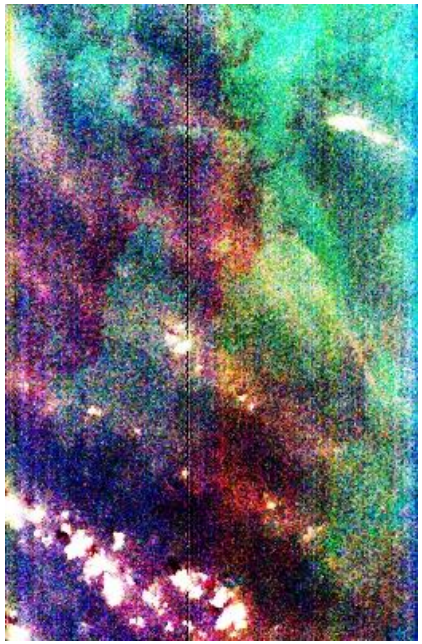


VRadCor（赵永超）批处理HJ-1A已经几何纠正的高光谱单景图像时的前(左)后(右)效果对比，只显示了其中一个波段

最小二乘法的应用

条带消除



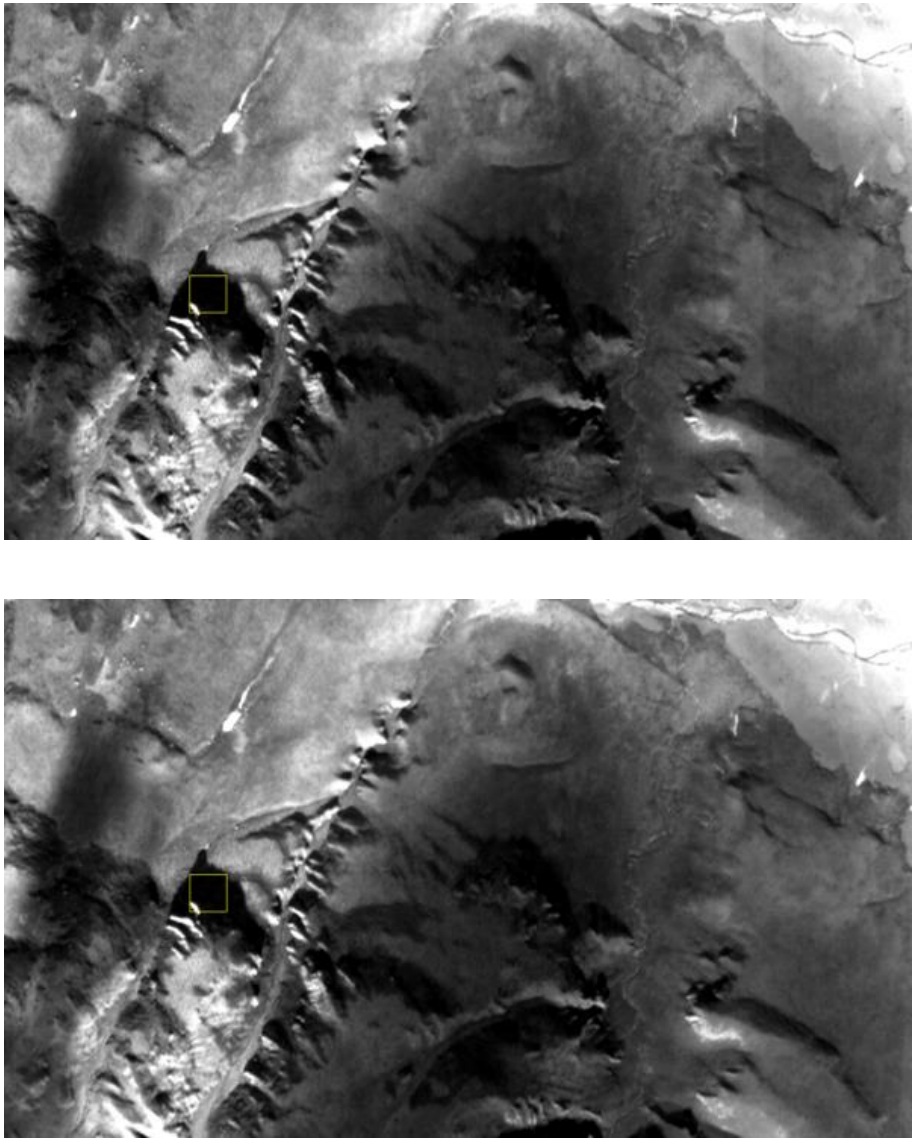


南佛罗里达大学和NOAA在墨西哥湾的海草遥感监测决策系统中采用VRadCor对有限海域HYPERION图像中残存的高频条纹进行修正的效果（赵永超）

最小二乘法的应用

条带消除

快舟1号利用VRadCor计算的定标系数处理前(上)后(下)的陆域效果，消除了其中的低频条带条纹（赵永超）



总结

1. 从行空间角度，线性方程组的解为各个方程所对应的超平面的交集
2. 从列空间角度，线性方程组的常数项向量为系数矩阵的各个列向量的线性表出，且表出系数即为待求的解
3. 最小二乘法是求解矛盾方程组的常用方法
4. 代数上，线性方程组的最小二乘解对应着矩阵的广义逆操作
5. 几何上，线性方程组的最小二乘解等价于线性方程组的常数项向量在系数矩阵的列空间的正交投影
6. 概率上，最小二乘法基于线性方程组模型误差的高斯分布。

谢　谢

耿修瑞

中国科学院空天信息创新研究院

gengxr@sina.com.cn