

最小二乘法与协方差矩阵

Shuyong Chen

2023 年 8 月 9 日

—题记：知识要乱搭，美味要乱炖。

1 简介

把涉及最小二乘法的知识乱搭在一起，龙卷风来随风飘。

2 柯西-施瓦兹不等式的向量形式

2.1 定理证明

柯西-施瓦兹不等式的向量形式：若有向量 \mathbf{x} 和 \mathbf{y} ，则

$$|\mathbf{x}^T \mathbf{y}| \leq \|\mathbf{x}\| \|\mathbf{y}\|$$
$$-1 \leq \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \leq 1$$

当且仅当其中一个向量为 $\mathbf{0}$ ，或一个向量为另外一个向量的倍数时，等号成立。

证明：有向量 $\epsilon = \mathbf{x} - \lambda \mathbf{y}$ ，我们计算它的范数为：

$$\begin{aligned}\|\epsilon\|^2 &\geq 0 \\ (\mathbf{x} - \lambda \mathbf{y})^T (\mathbf{x} - \lambda \mathbf{y}) &\geq 0 \\ (\mathbf{x}^T - \lambda \mathbf{y}^T) (\mathbf{x} - \lambda \mathbf{y}) &\geq 0 \\ \mathbf{x}^T \mathbf{x} - \mathbf{x}^T \lambda \mathbf{y} - \lambda \mathbf{y}^T \mathbf{x} + \lambda^2 \mathbf{y}^T \mathbf{y} &\geq 0 \\ \mathbf{x}^T \mathbf{x} - \lambda (\mathbf{x}^T \mathbf{y} + \mathbf{y}^T \mathbf{x}) + \lambda^2 \mathbf{y}^T \mathbf{y} &\geq 0 \\ \text{Let } a = \mathbf{y}^T \mathbf{y}, b = (\mathbf{x}^T \mathbf{y} + \mathbf{y}^T \mathbf{x}), c = \mathbf{x}^T \mathbf{x} & \\ a\lambda^2 - b\lambda + c &\geq 0\end{aligned}$$

对下式求极值

$$a\lambda^2 - b\lambda + c = 0$$

这是开口向上的抛物线函数，有极值

$$\begin{aligned} 2a\lambda - b &= 0 \\ \lambda &= \frac{b}{2a} \end{aligned}$$

将 λ 代入上式得

$$\begin{aligned} a\left(\frac{b}{2a}\right)^2 - b\left(\frac{b}{2a}\right) + c &\geq 0 \\ \frac{b^2}{4a} - \frac{b^2}{2a} + c &\geq 0 \\ c &\geq \frac{b^2}{4a} \end{aligned}$$

根据向量的内积公式，我们有 $\mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x}$ ，并且向量 \mathbf{x} 的欧几里德范数为： $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$ 。将上述数值代回公式得

$$\begin{aligned} \mathbf{x}^T \mathbf{x} &\geq \frac{(\mathbf{x}^T \mathbf{y} + \mathbf{y}^T \mathbf{x})^2}{4\mathbf{y}^T \mathbf{y}} \\ \because \mathbf{x}^T \mathbf{y} &= \mathbf{y}^T \mathbf{x} \\ \mathbf{x}^T \mathbf{x} &\geq \frac{4(\mathbf{x}^T \mathbf{y})^2}{4\mathbf{y}^T \mathbf{y}} \\ (\mathbf{x}^T \mathbf{x})(\mathbf{y}^T \mathbf{y}) &\geq (\mathbf{x}^T \mathbf{y})^2 \\ \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 &\geq (\mathbf{x}^T \mathbf{y})^2 \\ \|\mathbf{x}\| \|\mathbf{y}\| &\geq \|\mathbf{x}^T \mathbf{y}\| \end{aligned}$$

我们还有

$$\begin{aligned} \lambda &= \frac{b}{2a} \\ &= (\mathbf{x}^T \mathbf{y})(\mathbf{y}^T \mathbf{y})^{-1} \\ &= (\mathbf{y}^T \mathbf{x})^T (\mathbf{y}^T \mathbf{y})^{-1} \end{aligned}$$

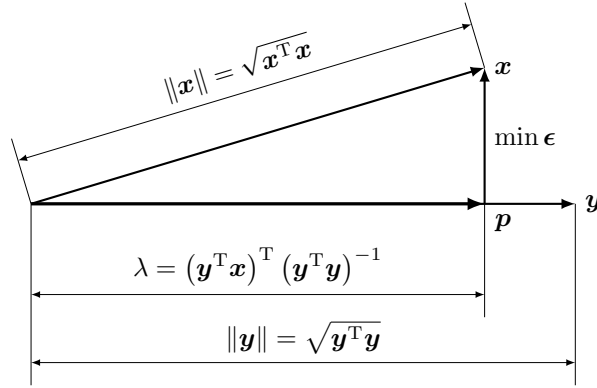
同时向量 \mathbf{x} 到 \mathbf{y} 的向量投影 (vector projection) \mathbf{p} 为

$$\mathbf{p} = \frac{(\mathbf{y}^T \mathbf{x})^T}{\mathbf{y}^T \mathbf{y}} \mathbf{y} = \lambda \mathbf{y}$$

可以看到， λ 是向量的投影系数。

2.2 几何解释

从几何上, 有向量 x 和 y , λ 代表比例因子, $\epsilon = x - \lambda y$ 则代表从 λy 指向 x 的向量。柯西-施瓦兹不等式的含义是调整 λ 的大小, 找到向量 $x - \lambda y$ 的最小范数, 就是向量 λy 到向量 x 的最短距离。当向量 x 和向量 y 共线时, 该距离为 0。而 $\lambda = (y^T x)^T (y^T y)^{-1}$, 则意味着当 λy 为向量 x 到 y 的向量投影时 ($\epsilon \perp y$), ϵ 有最小范数。



根据向量的内积公式, 有 $x^T y = y^T x$, 我们将公式 $(x^T x)(y^T y) \geq (x^T y)^2$ 做变形, 以待以后的比较:

$$\begin{aligned} x^T x &\geq \frac{(x^T y)^2}{y^T y} \\ x^T x &\geq \frac{(x^T y)(y^T x)}{y^T y} \\ x^T x &\geq (y^T x)^T (y^T y)^{-1} (y^T x) \end{aligned}$$

3 柯西-施瓦兹不等式的矩阵形式

3.1 定理证明

矩阵施瓦兹不等式 (Matrix Schwarz inequality): 若 P 和 Q 分别是 $m \times n$ 和 $m \times l$ 矩阵, $P^T P$ 非奇异, 则

$$Q^T Q \geq (P^T Q)^T (P^T P)^{-1} (P^T Q)$$

此外, 对于某些 $n \times l$ 矩阵 S , 当且仅当 $Q = PS$ 时, 上式的等号成立。该定理来自于参考文献 [3]。

证明: 有矩阵 $Q - PS$, 则 $(Q - PS)^T (Q - PS)$ 非负定, 所以有

$$\begin{aligned}
(Q - PS)^T (Q - PS) &\geq 0 \\
(Q^T - S^T P^T) (Q - PS) &\geq 0 \\
Q^T Q - Q^T PS - S^T P^T Q + S^T P^T PS &\geq 0 \\
Q^T Q &\geq S^T (P^T Q) + (P^T Q)^T S - S^T (P^T P) S \\
Q^T Q &\geq S^T ((P^T Q) - (P^T P) S) + (P^T Q)^T S
\end{aligned}$$

令

$$S = (P^T P)^{-1} (P^T Q)$$

则

$$\begin{aligned}
Q^T Q &\geq S^T ((P^T Q) - (P^T P) (P^T P)^{-1} (P^T Q)) + (P^T Q)^T (P^T P)^{-1} (P^T Q) \\
Q^T Q &\geq (P^T Q)^T (P^T P)^{-1} (P^T Q)
\end{aligned}$$

3.2 几何解释

该不等式粗略看起来很不直观。我们可以从 Q 的列向量的角度考虑，也就是 Q 的列向量之间的内积，要大于某一个向它投影的矩阵 P 的列向量和它的列向量的内积。

其实 $Q^T Q$ 是一个半正定的实对称矩阵，有十分广泛的应用，如协方差矩阵，权重矩阵（增益矩阵）等等。在后面该定理的在最小二乘的应用中，就有更直观的含义。

4 普通最小二乘法

4.1 定义

一个超定方程组含有的方程多于变量数。这种方程组通常是不相容的。因此，给定一个 $m \times n$ 的方程组 $Ax = b$ ，其中 $m > n$ ，一般我们不能期望找到一个向量 $x \in \mathbf{R}^n$ ，使得 $Ax = b$ 。事实上，可以寻找一个向量 x ，使得 Ax “最接近” b 。正如所期望的，正交性在求 x 的过程中扮演了重要的角色。

给定一个超定方程组 $Ax = b$ ，其中 A 为一个 $m \times n$ ($m > n$) 矩阵，并且 $b \in \mathbf{R}^m$ ，则对每一个 $x \in \mathbf{R}^n$ ，可以构造一个残差 (residual)

$$r(x) = b - Ax$$

则 b 和 Ax 之间的距离为

$$\|b - Ax\| = \|r(x)\|$$

我们希望寻找一个向量 $x \in \mathbf{R}^n$ ，使得 $\|r(x)\|$ 最小。最小化 $\|r(x)\|$ 等价于最小化 $\|r(x)\|^2$ 。达到最小值的向量 \hat{x} 称为方程组 $Ax = b$ 的最小二乘 (least squares) 解，这是一个近似解。

若 \hat{x} 为方程组 $Ax = b$ 的最小二乘解，且 $p = A\hat{x}$ ，则 p 就是 A 的列空间中和 b 最接近的向量（投影）。

4.2 证明

简单而快速证明如下：

- 最小化残差平方的范数，

$$\|r(\mathbf{x})\|^2 = \mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{b}^T A \mathbf{x} + \mathbf{b}^T \mathbf{b}$$

把这看成是开口向上的抛物线函数，可以有极小值。

- 相对于 \mathbf{x} 将梯度设置为 0：

$$\nabla_{\mathbf{x}} \|r(\mathbf{x})\|^2 = 2A^T A \mathbf{x} - 2A^T \mathbf{b} = 0$$

- 得出正规方程：

$$A^T A \mathbf{x} = A^T \mathbf{b}$$

- 上述方程表示的 $n \times n$ 线性方程组称为正规方程组 (normal equations)。它有唯一解

$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$$

且 $\hat{\mathbf{x}}$ 为方程组 $A\mathbf{x} = \mathbf{b}$ 唯一的最小二乘解。

- 在上式中，

$$A^\dagger = (A^T A)^{-1} A^T$$

称为 A 的伪逆矩阵，并且 A^\dagger 是矩阵 A 的 (满秩，瘦型) 左逆：

$$A^\dagger A = (A^T A)^{-1} A^T A = I$$

- 投影向量 \mathbf{p}

$$\mathbf{p} = A\hat{\mathbf{x}} = A(A^T A)^{-1} A^T \mathbf{b}$$

为 A 的列空间中 $C(A)$ 中的元素，并在最小二乘意义下最接近 \mathbf{b} ，也就是 \mathbf{p} 是 \mathbf{b} 在 A 的列空间中的投影， $\boldsymbol{\epsilon} = \mathbf{b} - \mathbf{p}$ 为残差 (误差) 向量。矩阵 $P = A(A^T A)^{-1} A^T$ 称为投影矩阵 (projection matrix)。

- 投影矩阵 P 为对称矩阵，并且有性质： $P^2 = P$ 。

4.3 几何解释

我们可以基于矩阵的 QR 分解去直观地理解最小二乘法：

- 矩阵 $A \in \mathbf{R}^{m \times n}$ ， $m > n$ ，是满秩，瘦型矩阵。
- 将其进行 QR 分解， $A = QR$ ，其中 $Q^T Q = I$ ，并且 $R \in \mathbf{R}^{n \times n}$ 为上三角矩阵且可逆。

- 伪逆矩阵为

$$(A^T A)^{-1} A^T = (R^T Q^T Q R)^{-1} R^T Q^T = R^{-1} Q^T$$

所以

$$\hat{\mathbf{x}} = R^{-1} Q^T \mathbf{b}$$

- 投影到 A 的列空间 $C(A)$ 中的矩阵为

$$A (A^T A)^{-1} A^T = A R^{-1} Q^T = Q Q^T$$

- 矩阵 A 的完全 QR 因子分解为:

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

其中, $\begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \in \mathbf{R}^{m \times m}$ 且为正交矩阵, $R_1 \in \mathbf{R}^{n \times n}$ 为上三角矩阵且可逆。

- 与正交矩阵相乘不改变矩阵范数, 所以

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|^2 &= \left\| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \mathbf{x} - \mathbf{b} \right\|^2 \\ &= \left\| \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^T \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \mathbf{x} - \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}^T \mathbf{b} \right\|^2 \\ &= \left\| \begin{bmatrix} R_1 \mathbf{x} - Q_1^T \mathbf{b} \\ -Q_2^T \mathbf{b} \end{bmatrix} \right\|^2 \\ &= \|R_1 \mathbf{x} - Q_1^T \mathbf{b}\|^2 + \|Q_2^T \mathbf{b}\|^2 \end{aligned}$$

- 这显然可以通过选择最小化 $\hat{\mathbf{x}}$, 即上式左边第一项为 0:

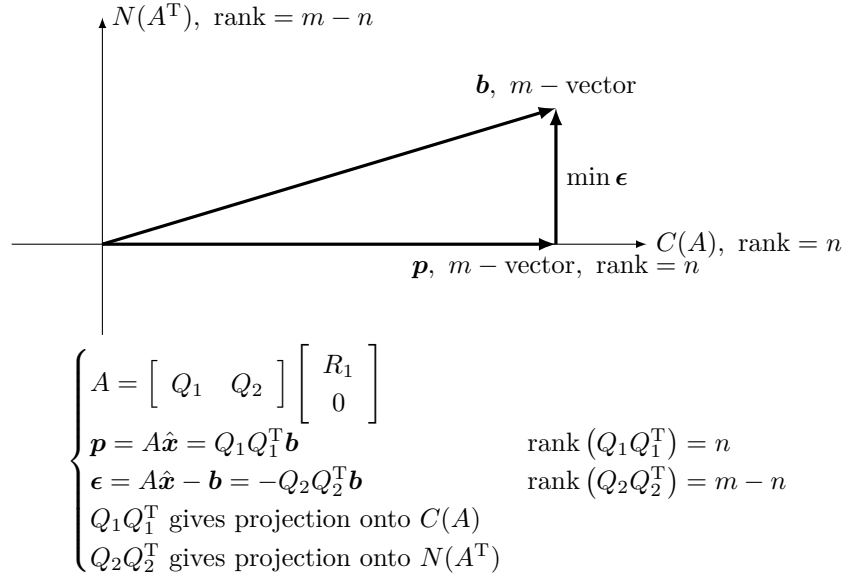
$$\hat{\mathbf{x}} = R_1^{-1} Q_1^T \mathbf{b}$$

- 则 \mathbf{x} 的最优化的残差为

$$\boldsymbol{\epsilon} = A\hat{\mathbf{x}} - \mathbf{b} = -Q_2 Q_2^T \mathbf{b}$$

- 矩阵 $Q_1 Q_1^T$ 给出了到 A 的列空间 $C(A)$ 中的投影。
- 矩阵 $Q_2 Q_2^T$ 给出了到列空间的正交空间 $C(A)^\perp$, 即 A^T 的零空间 $N(A^T)$ 的投影。
- 列空间 $C(A)$ 的满秩为 n , 则零空间 $N(A^T)$ 的秩为 $m - n$, 两者的秩之和为 m 。

上述关系的示意图如下:



5 基于权重的最小二乘法

令 $\{\underline{\xi}_k\}$ 为一系列的随机向量,称为**随机序列** (random sequence)。符号表示 $E(\underline{\xi}_k) = \underline{\mu}_k$, $\text{Cov}(\underline{\xi}_k, \underline{\xi}_j) = R_{kj}$, 则 $\text{Var}(\underline{\xi}_k) = R_{kk} := R_k$ 。如果随机向量有协方差矩阵 $\text{Cov}(\underline{\xi}_k, \underline{\xi}_j) = R_{kj} = R_k \delta_{kj}$, 其中当 $k = j$ 时, $\delta_{kj} = 1$, 当 $k \neq j$ 时, $\delta_{kj} = 0$, 则该随机向量 $\{\underline{\xi}_k\}$ 称为**白噪声序列** (white noise sequence)。如果每个 $\underline{\xi}_k$ 都是白色且正态的, 则 $\{\underline{\xi}_k\}$ 称为**高斯或正态白噪声序列** (sequence of Gaussian or normal white noise)。

考虑观测数据受噪声污染的线性系统的观测方程, 即:

$$\mathbf{v}_k = C_k \mathbf{x}_k + D_k \mathbf{u}_k + \underline{\xi}_k,$$

其中, $\{\mathbf{x}_k\}$ 通常是状态序列, $\{\mathbf{u}_k\}$ 是控制序列, $\{\mathbf{v}_k\}$ 是量测数据序列。我们假设, 对于每个 k , $q \times n$ 常数矩阵 C_k 、即量测矩阵, $q \times p$ 常数矩阵 D_k 、即控制矩阵, 以及确定性控制的 p 维控制向量 \mathbf{u}_k 是给定的。通常, 系统噪声 $\{\underline{\xi}_k\}$ 未知, 但将其假定为一个零均值高斯白噪声序列, 即: 对于 $k, j = 1, 2, \dots$, $E(\underline{\xi}_k) = 0$ 并且 $E(\underline{\xi}_k \underline{\xi}_j^T) = R_{kj} \delta_{kj}$, 其中系统噪声协方差矩阵 R_k 是对称且正定的矩阵。

我们的目标是从量测数据 $\{\mathbf{v}_k\}$ 的信息中获得状态向量 \mathbf{x}_k 的最优估计 $\hat{\mathbf{y}}_k$ 。如果没有噪音, 那么很明显 $\mathbf{z}_k - C_k \hat{\mathbf{y}}_k = 0$, 其中 $\mathbf{z}_k := \mathbf{v}_k - D_k \mathbf{u}_k$, 每当线性系统有解时; 否则, 一些测量误差 $\mathbf{z}_k - C_k \mathbf{y}_k$ 必须在所有 \mathbf{y}_k 中最小化。一般来说, 当数据被噪声污染时, 我们将在所有 n 维向量 \mathbf{y}_k 上, 最小化测量误差 (残差) 的平方范数的数量:

$$F(\mathbf{y}_k, W_k) = E(\mathbf{z}_k - C_k \mathbf{y}_k)^T W_k (\mathbf{z}_k - C_k \mathbf{y}_k)$$

其中 W_k 是正定对称 $q \times q$ 矩阵, 称为**权重矩阵** (weight matrix), 也称**增益矩阵** (gain matrix) G_k 。也

就是说, 我们希望找到一个 $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_k(W_k)$, 使得

$$F(\hat{\mathbf{y}}_k, W_k) = \min_{\mathbf{y}_k} F(\mathbf{y}_k, W_k).$$

另外, 我们希望确定**最优权重** (*optimal weight*) \hat{W}_k 。为了找到 $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_k(W_k)$, 假设 $(C_k^\top W_k C_k)$ 是非奇异的, 我们重写方程

$$\begin{aligned} F(\mathbf{y}_k, W_k) &= E(\mathbf{z}_k - C_k \mathbf{y}_k)^\top W_k (\mathbf{z}_k - C_k \mathbf{y}_k) \\ &= E[(C_k^\top W_k C_k) \mathbf{y}_k - C_k^\top W_k \mathbf{z}_k]^\top (C_k^\top W_k C_k)^{-1} [(C_k^\top W_k C_k) \mathbf{y}_k - C_k^\top W_k \mathbf{z}_k] \\ &\quad + E\left(\mathbf{z}_k^\top \left[I - W_k C_k (C_k^\top W_k C_k)^{-1} C_k^\top\right] W_k \mathbf{z}_k\right), \end{aligned}$$

其中右边的第一项为非负定项。为了最小化 $F(\mathbf{y}_k, W_k)$, 右边的第一项必须消失, 则

$$\hat{\mathbf{y}}_k = (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k.$$

注意, 如果 $(C_k^\top W_k C_k)$ 是奇异的, 则 $\hat{\mathbf{y}}_k$ 不是唯一的。为了找到最优的权重 \hat{W}_k , 让我们考虑

$$F(\hat{\mathbf{y}}_k, W_k) = E(\mathbf{z}_k - C_k \hat{\mathbf{y}}_k)^\top W_k (\mathbf{z}_k - C_k \hat{\mathbf{y}}_k).$$

很明显, 在正定权重 W_k 下, 该数量不会达到最小值, 因为 $W_k = 0$ 会导致该最小值。因此, 我们需要另一个测量量来确定最优 \hat{W}_k 。注意到原始问题是用 $\hat{\mathbf{y}}_k(W_k)$ 估计状态向量 \mathbf{x}_k , 考虑误差 $(\mathbf{x}_k - \hat{\mathbf{y}}_k(W_k))$ 的测量量是自然的。但是, 由于对 \mathbf{x}_k 的所知不多, 并且只能测量带噪声的数据, 因此应通过误差的方差来确定该测量量。也就是说, 在所有正定对称矩阵 W_k 之上, 我们将最小化 $\text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k(W_k))$ 。我们简记 $\hat{\mathbf{y}}_k = \hat{\mathbf{y}}_k(W_k)$ 并有

$$\begin{aligned} \mathbf{x}_k - \hat{\mathbf{y}}_k &= (C_k^\top W_k C_k)^{-1} (C_k^\top W_k C_k) \mathbf{x}_k - (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k \\ &= (C_k^\top W_k C_k)^{-1} C_k^\top W_k (C_k \mathbf{x}_k - \mathbf{z}_k) \\ &= - (C_k^\top W_k C_k)^{-1} C_k^\top W_k \underline{\boldsymbol{\xi}}_k. \end{aligned}$$

因此, 通过期望运算的线性特性, 我们有

$$\begin{aligned} \text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k) &= (C_k^\top W_k C_k)^{-1} C_k^\top W_k E(\underline{\boldsymbol{\xi}}_k \underline{\boldsymbol{\xi}}_k^\top) W_k C_k (C_k^\top W_k C_k)^{-1} \\ &= (C_k^\top W_k C_k)^{-1} C_k^\top W_k R_k W_k C_k (C_k^\top W_k C_k)^{-1}. \end{aligned}$$

这就是要最小化的数量, 即误差的方差与系统噪声的方差正相关。为了把它写成一个完全平方方程, 我们需要正定对称矩阵 R_k 的**正平方根** (*positive square root*), 定义如下: 设系统噪声的协方差矩阵 R_k 的特征值为 $\lambda_1, \dots, \lambda_n$, 它们都是正值, 并且写为 $R_k = U^\top \text{diag}[\lambda_1, \dots, \lambda_n] U$, 其中 U 是酉矩阵 (由 $\lambda_i, i = 1, \dots, n$ 对应的正规化特征向量组成)。然后我们定义 $R_k^{1/2} = U^\top \text{diag}[\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}] U$, 其给出 $(R_k^{1/2}) (R_k^{1/2})^\top = R_k$ 。因此

$$\text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k) = Q^\top Q,$$

其中 $Q = \left(R_k^{1/2}\right)^\top W_k C_k (C_k^\top W_k C_k)^{-1}$ 。根据**矩阵施瓦兹不等式** (*matrix Schwarz inequality*)，参见第 3 章，在假设 P 是 $q \times n$ 矩阵且 $P^\top P$ 非奇异的情况下，我们有

$$Q^\top Q \geq (P^\top Q)^\top (P^\top P)^{-1} (P^\top Q).$$

因此，如果 $(C_k^\top R_k^{-1} C_k)$ 是非奇异的，我们可以选择 $P = \left(R_k^{1/2}\right)^{-1} C_k$ ，以便

$$P^\top P = C_k^\top \left(\left(R_k^{1/2}\right)^\top \right)^{-1} \left(R_k^{1/2}\right)^{-1} C_k = C_k^\top R_k^{-1} C_k$$

是非奇异的，并且不等式的等号成立，则

$$\begin{aligned} (P^\top Q)^\top (P^\top P)^{-1} (P^\top Q) &= \left[C_k^\top \left(\left(R_k^{1/2}\right)^{-1} \right)^\top \left(R_k^{1/2}\right)^\top W_k C_k (C_k^\top W_k C_k)^{-1} \right]^\top (C_k^\top R_k^{-1} C_k)^{-1} \\ &\quad \cdot \left[C_k^\top \left(\left(R_k^{1/2}\right)^{-1} \right)^\top \left(R_k^{1/2}\right)^\top W_k C_k (C_k^\top W_k C_k)^{-1} \right] \\ &= (C_k^\top R_k^{-1} C_k)^{-1} \\ &= \text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k(R_k^{-1})). \end{aligned}$$

因此，对于所有正定对称权重矩阵 W_k ，有 $\text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k(W_k)) \geq \text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k(R_k^{-1}))$ 。因此，最优权重矩阵为 $\hat{W}_k = R_k^{-1}$ ，即协方差矩阵 R_k 的逆矩阵，使用该最优权重的状态向量 \mathbf{x}_k 的最优估计为

$$\hat{\mathbf{x}}_k := \hat{\mathbf{y}}_k(R_k^{-1}) = (C_k^\top R_k^{-1} C_k)^{-1} C_k^\top R_k^{-1} (\mathbf{v}_k - D_k \mathbf{u}_k)$$

我们称 $\hat{\mathbf{x}}_k$ 为 \mathbf{x}_k 的**最小二乘最优估计** (*least-squares optimal estimate*)。注意， $\hat{\mathbf{x}}_k$ 是 \mathbf{x}_k 的**线性估计** (*linear estimate*)。作为数据 $\mathbf{v}_k - D_k \mathbf{u}_k$ 的线性变换的图像，它给出了在 $E\hat{\mathbf{x}}_k = E\mathbf{x}_k$ 意义上的 \mathbf{x}_k 的**无偏估计** (*unbiased estimate*)，即

$$\begin{aligned} E\hat{\mathbf{x}}_k &= (C_k^\top R_k^{-1} C_k)^{-1} C_k^\top R_k^{-1} E(\mathbf{v}_k - D_k \mathbf{u}_k) \\ &= (C_k^\top R_k^{-1} C_k)^{-1} C_k^\top R_k^{-1} E(C_k \mathbf{x}_k - \boldsymbol{\eta}_k) \\ &= E\mathbf{x}_k \end{aligned}$$

其中 $\boldsymbol{\eta}_k$ 为测量噪声向量，并且它还给出了 \mathbf{x}_k 的**最小方差估计** (*minimum variance estimate*)，因为

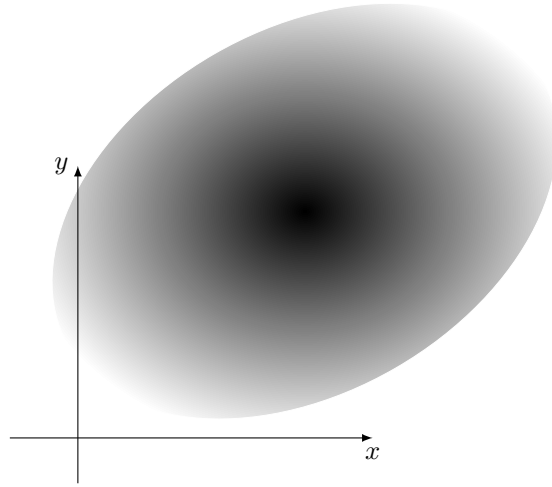
$$\text{Var}(\mathbf{x}_k - \hat{\mathbf{x}}_k) \leq \text{Var}(\mathbf{x}_k - \hat{\mathbf{y}}_k(W_k))$$

对于所有正定对称权重矩阵 W_k 成立。

6 几何解释

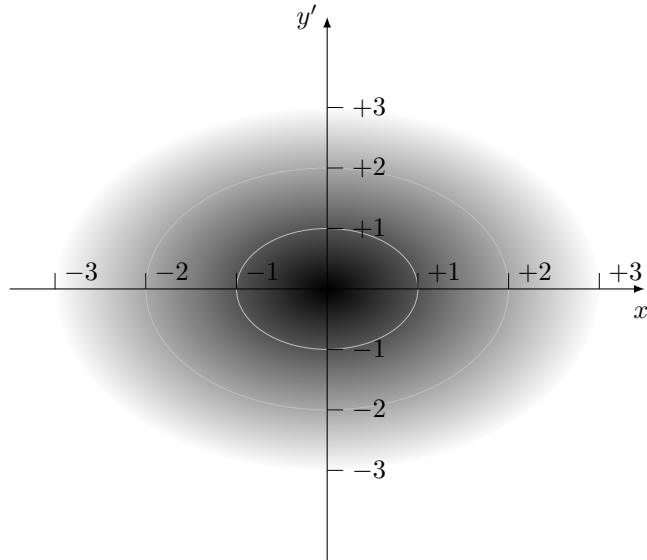
基于权重的最小二乘法的算法相当复杂，不容易理解。但是如果从协方差矩阵的角度去理解，就相当明了。

我们测量得到这样的数据集合



我们知道一些这个系统的状态变量和变换方程。但是因为各种原因，系统存在误差，因此这是一个超定方程，没有解析解。但是我们可以用最小二乘法计算得出近似解。

我们假定数据按照高斯分布，并且系统噪声为零均值的白高斯噪声。我们知道，是变量自身的方差在坐标轴上拉伸了图像，是变量之间的协方差旋转了图像，但是该旋转并没有改变各数据点到数据中心的距离，因此我们可以在算法中直接忽略变量之间的协方差，并在数据中心建立新的坐标系。



注意，因为各轴的方差不一样，因此各轴的数据变形程度也不一样，体现在数据刻度不一样。这时直接在这样的数据集合上进行估计并不是最优估计。因此我们需要引入权重矩阵 W 以校正这种变形。经过推导，数据的误差的方差和系统噪声的方差正相关。根据矩阵施瓦兹不等式，我们有

$$Q^T Q \geq (P^T Q)^T (P^T P)^{-1} (P^T Q)$$

其中, $Q^\top Q$ 为数据的误差协方差矩阵, 因此 Q 为误差的标准差矩阵。矩阵 P 为原始量测矩阵 C 经过权重矩阵 W 校正过后新的量测矩阵, 以抵消系统噪声的协方差矩阵 R 的影响。

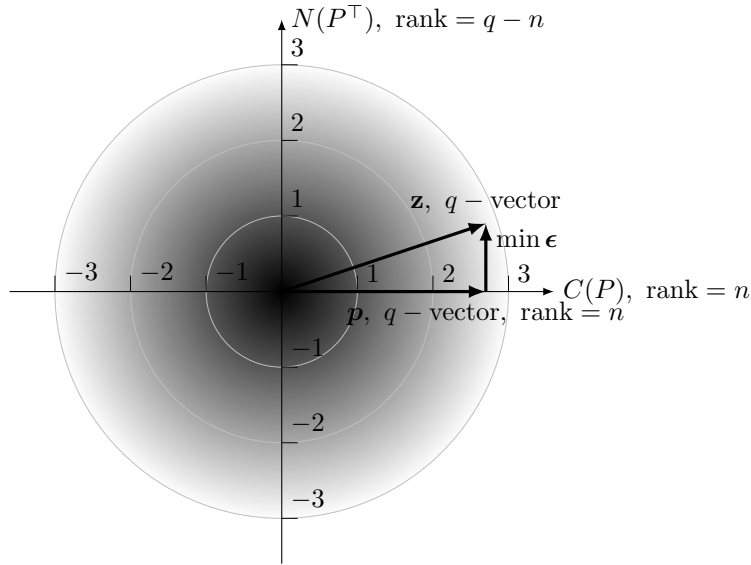
其它的推导就是复杂且累人的矩阵代数运算, 根据实对称矩阵的一些特性以及一些配元的技巧, 最终得到的最优权重矩阵为 $\hat{W} = R^{-1}$, 因此对冲了系统噪声的影响, 这很符合直觉。这时上述不等式的等号成立, 数据的误差最小, $Q^\top Q = (P^\top P)^{-1}$ 。此时 $P = (R^{1/2})^{-1} C$, 其中标准差的倒数 $(R^{1/2})^{-1}$ 又称精度矩阵, 由此左乘校正了量测矩阵 C , 因此

$$\Sigma = (P^\top P)^{-1} = (C^\top R^{-1} C)^{-1}$$

被称为最小二乘解的协方差, 并且这种二次型矩阵被称为 Fisher 信息矩阵或 Hessian 矩阵。并且 $(C^\top R^{-1} C)^{-1} C^\top R^{-1}$ 为矩阵 C 的最优权重伪逆矩阵, 因为

$$\begin{aligned} (C^\top R^{-1} C)^{-1} C^\top R^{-1} C &= \left(\left((R^{1/2})^{-1} C \right)^\top \left((R^{1/2})^{-1} C \right) \right)^{-1} \left((R^{1/2})^{-1} C \right)^\top (R^{1/2})^{-1} C \\ &= (P^\top P)^{-1} P^\top P \\ &= I \end{aligned}$$

最终, 我们是在校正后的白化数据集上进行最小二乘估计, 并且此时的估计为无偏估计。



该算法如果应用在动态递归的估计系统中, 就成了卡尔曼滤波器的核心部分。

7 参考文献

1. 《线性代数》Steven J. Leon — 第 5 章正交性
2. Lecture 5 — Least-squares
3. 《卡尔曼滤波及其实时应用》— 第 1.3 节最小二乘初步

8 草稿纸

整理并验证方程相等

$$\begin{aligned}
 & (\mathbf{z}_k - C_k \mathbf{y}_k)^\top W_k (\mathbf{z}_k - C_k \mathbf{y}_k) \\
 &= [(C_k^\top W_k C_k) \mathbf{y}_k - C_k^\top W_k \mathbf{z}_k]^\top (C_k^\top W_k C_k)^{-1} [(C_k^\top W_k C_k) \mathbf{y}_k - C_k^\top W_k \mathbf{z}_k] \\
 &+ \left(\mathbf{z}_k^\top \left[I - W_k C_k (C_k^\top W_k C_k)^{-1} C_k^\top \right] W_k \mathbf{z}_k \right)
 \end{aligned}$$

展开等式左边

$$\begin{aligned}
 (\mathbf{z}_k - C_k \mathbf{y}_k)^\top W_k (\mathbf{z}_k - C_k \mathbf{y}_k) &= \mathbf{z}_k^\top W_k \mathbf{z}_k \\
 &- (C_k \mathbf{y}_k)^\top W_k \mathbf{z}_k \\
 &- \mathbf{z}_k^\top W_k (C_k \mathbf{y}_k) \\
 &+ (C_k \mathbf{y}_k)^\top W_k (C_k \mathbf{y}_k)
 \end{aligned}$$

展开等式右边

$$\begin{aligned}
 & [(C_k^\top W_k C_k) \mathbf{y}_k - C_k^\top W_k \mathbf{z}_k]^\top (C_k^\top W_k C_k)^{-1} [(C_k^\top W_k C_k) \mathbf{y}_k - C_k^\top W_k \mathbf{z}_k] \\
 &= ((C_k^\top W_k C_k) \mathbf{y}_k)^\top (C_k^\top W_k C_k)^{-1} (C_k^\top W_k C_k) \mathbf{y}_k \\
 &- ((C_k^\top W_k C_k) \mathbf{y}_k)^\top (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k \\
 &- (C_k^\top W_k \mathbf{z}_k)^\top (C_k^\top W_k C_k)^{-1} (C_k^\top W_k C_k) \mathbf{y}_k \\
 &+ (C_k^\top W_k \mathbf{z}_k)^\top (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k
 \end{aligned}$$

并有余项

$$\mathbf{z}_k^\top \left[I - W_k C_k (C_k^\top W_k C_k)^{-1} C_k^\top \right] W_k \mathbf{z}_k$$

利用对称矩阵 $W_k = W_k^\top$ 展开各项并整理

1. 右边第 1 项

$$\begin{aligned}
 & ((C_k^\top W_k C_k) \mathbf{y}_k)^\top (C_k^\top W_k C_k)^{-1} (C_k^\top W_k C_k) \mathbf{y}_k \\
 &= ((C_k^\top W_k C_k) \mathbf{y}_k)^\top \mathbf{y}_k \\
 &= (C_k \mathbf{y}_k)^\top W_k (C_k \mathbf{y}_k)
 \end{aligned}$$

与左边第 4 项相等。

2. 右边第 2 项

$$\begin{aligned}
 & - ((C_k^\top W_k C_k) \mathbf{y}_k)^\top (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k \\
 &= - (C_k \mathbf{y}_k)^\top W_k C_k (W_k C_k)^{-1} (C_k^\top)^{-1} C_k^\top W_k \mathbf{z}_k \\
 &= - (C_k \mathbf{y}_k)^\top W_k \mathbf{z}_k
 \end{aligned}$$

与左边第 2 项相等。

3. 右边第 3 项

$$\begin{aligned}
 & - (C_k^\top W_k \mathbf{z}_k)^\top (C_k^\top W_k C_k)^{-1} (C_k^\top W_k C_k) \mathbf{y}_k \\
 &= - (C_k^\top W_k \mathbf{z}_k)^\top \mathbf{y}_k \\
 &= - \mathbf{z}_k^\top W_k (C_k \mathbf{y}_k)
 \end{aligned}$$

与左边第 3 项相等。

4. 剩下各项凑成余项

$$\begin{aligned}
 & \mathbf{z}_k^\top W_k \mathbf{z}_k - (C_k^\top W_k \mathbf{z}_k)^\top (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k \\
 &= \mathbf{z}_k^\top W_k \mathbf{z}_k - \mathbf{z}_k^\top W_k C_k (C_k^\top W_k C_k)^{-1} C_k^\top W_k \mathbf{z}_k \\
 &= \mathbf{z}_k^\top \left(I - W_k C_k (C_k^\top W_k C_k)^{-1} C_k^\top \right) W_k \mathbf{z}_k
 \end{aligned}$$

所以两式相等。