欧氏平面上平方距离最小的直线拟合

BY JIN

2019-06-10

摘要

从平面上的一族点拟合一条直线是一个常见问题。然而统计上常用的针对$y=kx+b$ 的最小二乘优化$\{k=(n\sum xy-\sum x\sum y)\,/\,(n\sum x^2-(\sum x)^2)=\mathrm{cov}(x,y)\,/\,\mathrm{cov}(x,x), b=(\sum y-k\sum x)\,/\,n=\bar{y}-k\bar{x}\}$，最小化的是$y$方向上的误差$\sum\,(y_i-kx_i-b)^2$，并不是欧几里德平面上的距离最小。本文将针对点到直线的垂直距离，使用完全最小二乘法求最优，得出为更优雅的结果，并自多种角度对这一结果进行解释和验证。

1 问题描述

待拟合的直线用法向式表示为

$$x\cos\alpha+y\sin\alpha-r=0 \tag{1}$$

则任意一点(x,y)到该直线的距离为

$$d=|x\cos\alpha+y\sin\alpha-r| \tag{2}$$

则问题是对已知点集$\{(x,y)\}_n$，以最小化$F(\alpha,r)=\sum d^2$为目标，求最优解

$$\underset{\alpha,r}{\mathrm{ArgMin}}\sum_i\,(x_i\cos\alpha+y_i\sin\alpha-r)^2 \tag{3}$$

2 求解最优

记$\cos\alpha=c, \sin\alpha=s$，待优化的和式有如下展开

$$\begin{aligned} F(\alpha,r) &= \sum\,(x\cos\alpha+y\sin\alpha-r)^2 \\ &= \sum\,(c^2x^2+s^2y^2+2csxy-2rcx-2rsy+r^2) \\ &= c^2\sum\,x^2+s^2\sum\,y^2+2cs\sum\,xy-2rc\sum\,x-2rs\sum\,y+nr^2 \end{aligned} \tag{4}$$

求偏导，有

$$\begin{aligned} 0=\frac{\partial F}{\partial r} &= -2\big(c\sum\,x+s\sum\,y-nr\big) \\ \text{即}\quad r &= \big(c\sum\,x+s\sum\,y\big)/n=c\,\bar{x}+s\,\bar{y} \end{aligned} \tag{5}$$

以及

$$\begin{aligned} 0=\frac{\partial F}{\partial\alpha} &= -2\big(cs\big(\sum\,x^2-\sum\,y^2\big)-(c^2-s^2)\sum\,xy-r\,\big(s\sum\,x-c\sum\,y\big)\big) \\ \text{代入}r,\quad 0 &= cs(\bar{x^2}-\bar{y^2})-(c^2-s^2)\overline{xy}-(c\bar{x}+s\bar{y})\,(s\bar{x}-c\bar{y}) \\ &= cs(\bar{x^2}-\bar{y^2})-(c^2-s^2)\overline{xy}-cs\,(\bar{x}^2-\bar{y}^2)+(c^2-s^2)\bar{x}\bar{y} \\ &= cs\,(\mathrm{cov}(x,x)-\mathrm{cov}(y,y))-(c^2-s^2)\mathrm{cov}(x,y) \end{aligned} \tag{6}$$

整理，得

$$\frac{2\,\mathrm{cov}(x,y)}{\mathrm{cov}(x,x)-\mathrm{cov}(y,y)}=\frac{2\cos\alpha\sin\alpha}{\cos^2\alpha-\sin^2\alpha}=\tan2\alpha \tag{7}$$

由此解出

$$\alpha = \frac{1}{2}\arctan\frac{2\,\mathrm{cov}(x,y)}{\mathrm{cov}(x,x)-\mathrm{cov}(y,y)} + \frac{k\pi}{2} \tag{8}$$

包含相垂直的两个解，正好对应直线的方向向量θ和法向量方向α，也是误差函数的最大和最小值。

3 分析验根

统计角度上看，$\mathrm{cov}(x,\,x) - \mathrm{cov}(y,\,y)$的符号表示点集在x, y方向上的分散程度大小，而$\mathrm{cov}(x,\,y)$的符号表示了点集中x, y是正相关还是负相关：据此可以确定α的正确取值。

1. 当$\mathrm{cov}(x,x)-\mathrm{cov}(y,y)>0, \mathrm{cov}(x,y)>0$，对应$0<\theta<\frac{\pi}{4}, -\frac{\pi}{2}<\alpha<-\frac{\pi}{4}$，因此

$$\begin{aligned} \theta &= \frac{1}{2}\mathrm{arctan2}(2\,\mathrm{cov}(x,y), \mathrm{cov}(x,x)-\mathrm{cov}(y,y)) \\ \alpha = \theta \pm \frac{\pi}{2} &= \frac{1}{2}\mathrm{arctan2}(-2\,\mathrm{cov}(x,y), \mathrm{cov}(y,y)-\mathrm{cov}(x,x)) \end{aligned}$$

2. 当$\mathrm{cov}(x,x)-\mathrm{cov}(y,y)<0, \mathrm{cov}(x,y)>0$，对应$\frac{\pi}{4}<\theta<\frac{\pi}{2}, -\frac{\pi}{4}<\alpha<0$，因此仍有

$$\begin{aligned} \theta &= \frac{1}{2}\mathrm{arctan2}(2\,\mathrm{cov}(x,y), \mathrm{cov}(x,x)-\mathrm{cov}(y,y)) \\ \alpha = \theta \pm \frac{\pi}{2} &= \frac{1}{2}\mathrm{arctan2}(-2\,\mathrm{cov}(x,y), \mathrm{cov}(y,y)-\mathrm{cov}(x,x)) \end{aligned}$$

3. 同理（或由连续性）可证$\mathrm{cov}(x,x)-\mathrm{cov}(y,y)<0, \mathrm{cov}(x,y)<0, -\frac{\pi}{2}<\theta<-\frac{\pi}{4}, 0<\alpha<\frac{\pi}{4}$的情况，以及$\mathrm{cov}(x,x)-\mathrm{cov}(y,y)>0, \mathrm{cov}(x,y)<0, -\frac{\pi}{4}<\theta<0, \frac{\pi}{4}<\alpha<\frac{\pi}{2}$的情况也有同样的结论

因此统一地，有结论

$$\alpha = \frac{1}{2}\mathrm{arctan2}(-2\,\mathrm{cov}(x,y), \mathrm{cov}(y,y)-\mathrm{cov}(x,x)) \tag{9}$$

4 统计角度

按$r=\bar{x}\cos\alpha+\bar{y}\sin\alpha$，若对点集$X=\begin{bmatrix} \vdots & \vdots \\ x_i & y_i \\ \vdots & \vdots \end{bmatrix}_{n\times 2}$作平移变换$X'=X-\bar{X}$，可将直线方程写成齐次形式：

$$x'\cos\alpha + y'\sin\alpha = 0 \tag{10}$$

再进行旋转变换$X''=X'R,\quad R\,R^T=I$，此时新的协方差矩阵

$$\begin{aligned} \Sigma(X'') = E((X'R-0)^T(X'R-0)) &= R^T E(X'^T X')\,R \\ &= R^{-1}E((X-\bar{X})^T(X-\bar{X}))\,R \\ &= R^{-1}\Sigma(X)\,R \end{aligned} \tag{11}$$

为原协方差矩阵的一个相似变换。

由X''为对称正定矩阵，故必存在恰当的正交变换R使其对角化，即

$$C(X'') = \begin{bmatrix} \sigma_{x''}^2 & 0 \\ 0 & \sigma_{y''}^2 \end{bmatrix} \tag{12}$$

如此可将点集的误差分解到彼此独立且正交的两个方向上。设$\sigma_{x''}^2 > \sigma_{y''}^2$，则$\sigma_{x''}^2$对应了直线方向，$\sigma_{y''}^2$对应了直线的法线方向。由此引出算法：

1. 求点集$\{(x,y)\}_n$的协方差矩阵$C = \left[\begin{smallmatrix} \mathrm{cov}(x,x) & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \mathrm{cov}(y,y) \end{smallmatrix}\right]$

2. 对C作特征分解$C = T\Sigma' T^{-1}$, 其中$C' = \left[\begin{smallmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{smallmatrix}\right], \sigma_1^2 > \sigma_2^2, T = \left[\begin{smallmatrix} u & v \\ -v & u \end{smallmatrix}\right] = \left[\begin{smallmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{smallmatrix}\right]$

3. 该特征分解对应了点集（围绕其均值）的一个旋转变换$[\, x'', y'' \,] = [\, x', y' \,]\, T^{-1}$，该变换将直线的方向向量旋转到了0度。而反过来$\left[\begin{smallmatrix} x' \\ y' \end{smallmatrix}\right] = \left[\begin{smallmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{smallmatrix}\right]\left[\begin{smallmatrix} x'' \\ y'' \end{smallmatrix}\right]$意味着$X''$旋转$-\beta$即可到达$X'$的角度，因此直线方向为$\theta = -\beta$， 法向$\alpha = \pi - \beta$， 即$\left[\begin{smallmatrix} \cos\alpha \\ \sin\alpha \end{smallmatrix}\right] = \left[\begin{smallmatrix} -\sin\theta \\ \cos\theta \end{smallmatrix}\right] = \left[\begin{smallmatrix} \sin\beta \\ \cos\beta \end{smallmatrix}\right] = \left[\begin{smallmatrix} v \\ u \end{smallmatrix}\right]$.

实践中，没有必要对$X^T X$做特征分解，完全可以等效地对X做奇异值分解：

5 完全最小二乘(SVD分解)

同样按(10)的方式齐次化直线方程。令$X' = \left[\begin{smallmatrix} \vdots & \vdots \\ x_i' & y_i' \\ \vdots & \vdots \end{smallmatrix}\right]_{n\times 2}, A = \left[\begin{smallmatrix} \cos\alpha \\ \sin\alpha \end{smallmatrix}\right]$，此时优化目标可以写成

$$E(X) = (XA)^T (XA) = A^T X^T X A, \tag{13}$$

求$A = \mathrm{Argmin}_A\, E(X)$。对X作SVD分解：$X = U_{n\times 2}\, \Sigma_{2\times 2}\, V_{2\times 2}^T$

$$E(X) = A^T V \Sigma^T U^T U \Sigma V^T A = A^T V \Sigma^T \Sigma V^T A \tag{14}$$

设$V = \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right), \Sigma = \left(\begin{smallmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{smallmatrix}\right), \sigma_1 \geqslant \sigma_2 \geqslant 0$，可直接求出

$$E(X) = \frac{1}{2}(\sigma_1^2 + \sigma_2^2 + (\sigma_1^2 - \sigma_2^2)\cos(2\alpha - 2\theta)) \tag{15}$$

当$2(\alpha - \theta) = \pi + 2k\pi$即$\alpha = \theta + \frac{\pi}{2} + k\pi$时取最小值。考虑到$\alpha + \pi$仅仅是$\alpha$的$180°$反方向，故只取一个解$\alpha = \theta + \frac{\pi}{2}$即可。此时

$$A = \begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix} = V[\![2]\!] \tag{16}$$

正好与特征分解算法的结果一致。

另一方面，按Total Least Squares求解，也能有：

$$A = \mathrm{Argmin}_A \|\tilde{X}\|_F, \quad (X + \tilde{X})A = 0 \tag{17}$$

SVD分解

$$\begin{aligned} X &= [\, U_1 \;\; U_2 \,]\begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}[\, V_1 \;\; V_2 \,]^T \\ X + \tilde{X} &= [\, U_1 \;\; U_2 \,]\begin{bmatrix} \sigma_1 & 0 \\ 0 & 0 \end{bmatrix}[\, V_1 \;\; V_2 \,]^T = U_1 \sigma_1 V_1^T \\ \text{两式相减} \quad \tilde{X} &= -U_2 \sigma_2 V_2^T \end{aligned} \tag{18}$$

由此也有，在$X + \tilde{X}$的补空间中取单位向量

$$A = \pm V_2$$

可满足 $(X + \tilde{X})A = U_1 \sigma_1 V_1^T \cdot V_2 = U_1 \sigma_1 (V_1^T V_2) = 0$。

Q.E.D.