

半参数 p -范极大似然回归

潘 雄^{1,2}, 孙海燕¹

(1. 武汉大学 测绘学院, 湖北 武汉 430079; 2. 武汉工业学院 数理系, 湖北 武汉 430023)

Semiparametric p -norm Maximum Likelihood Regression

PAN Xiong^{1,2}, SUN Haiyan¹

(1. School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China; 2. Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan 430023, China)

Abstract: In this paper, used the kernel weight function, we obtain the parameter estimation of p -norm distribution in semiparametric regression model, which is effective to decide the distribution of random errors. Under the assumption that the distribution of observations is unimodal and symmetrical, this method can give the estimates of X , S and σ . Finally, two simulated adjustment problems are constructed to explain this method. The new method presented in this paper shows an effective way of solving the problem, the estimated values are nearer to their theoretical ones than those by least squares adjustment.

Key words: p -norm distributions; semiparametric regression; kernel weight function; maximum likelihood adjustment

摘 要: 应用核权函数, 在观测为误差单峰、对称的情况下, 得到了一元 p -范分布的半参数模型的计算公式。详细推导了 p 已知时一元 p -范分布极大似然方程的解算公式, 将半参数回归模型应用到极大似然平差的参数估计理论中, 得到了一个比较好的算法。最后, 构造了两个模拟平差问题, 说明了此方法的优越性。

关键词: p -范分布; 半参数回归; 核函数; 极大似然平差

传统的测量平差问题在数学上归结为求在一组线性方程组约束下的残差二次型 ($V^T P V$) 的条件极小值及其极小值点, 这种方法称为最小二乘法。在测量数据处理理论中, 观测值中不可避免地含有误差, 按其性质通常分为粗差 Δ_g 、系统误差 Δ_s 和偶然误差 Δ_n 。即 $\Delta = \Delta_g + \Delta_s + \Delta_n$ 。在经典平差理论中^[1], 一般假定观测值中仅包含偶然误差, 即 $\Delta_s = 0$, $\Delta_g = 0$, $\Delta = \Delta_n$ 。对于 $\Delta_g \neq 0$ (即观测值中含有粗差的) 情况, 近代平差理论已作了比较充分的讨论。但是对于系统误差的处理, 目前讨论的还不多。在实际问题分析过程中,

常常会遇到模型的某些假定不能够完全满足的情形, 如: 反映变量与解释变量之间的具体依存关系不明确, 观测数据中含有系统误差、异常值的扰动等。理论研究和实践经验表明, 高斯-马尔可夫模型的最小二乘估计, 当有粗差发生时, 参数的最小二乘估计不可靠, 与其真值偏离太远。另外, 误差分布不是正态分布时, 最小二乘估计不是最优估计。统计学界在 20 世纪 80 年代提出了半参数回归模型, 放宽了线性模型中某一个解释变量的线性假定, 既含有参数分量又含有非参数分量, 使模型适应数据变化的能力更强, 用它描述实际问题

收稿日期: 2003-11-13; 修回日期: 2004-08-30

基金项目: 国家自然科学基金资助项目(40274005); 湖北省重点科研计划项目(2003X129); 湖北省自然科学基金资助项目(2004A BA032)

作者简介: 潘 雄, (1973), 男, 湖北兴山人, 博士, 主要研究方向为半参数回归模型在测量数据处理中的应用。

时,更能接近于真实情况。其次,文献[2]提出的 p -范分布和 p -范极大似然平差刚好能解决误差分布不是正态分布这个问题。本文试图将半参数回归分析的方法引入测量平差理论,建立适于测量数据处理的理论与方法,更好地确定模型中的参数和非参数,并将偶然误差与系统误差分离开来。

在 p 已知的前提下,引入核权函数,导出了半参数回归模型中 p -范极大似然平差,对相应的迭代计算进行分析,得到了相应的解算方法,作为特例,当 $p = 2$ 时,即为通常情况下的半参数模型;其次利用向后迭代(backfitting)方法得到了相应的解算步骤;在文章的最后,用两个模拟的平差问题($p = 1, p = 2$),分别用最小二乘法和半参数法平差求解,说明了此方法的优越性。

1 半参数平差模型及解算方法

间接平差的函数模型为^[2]

$$L = BX + \Delta \tag{1}$$

式中, L 为 n 维观测向量, X 为未知参数向量的真值, Δ 为 n 维误差观测向量, $B = (b_1, b_2, \dots, b_n)^T$ 为设计矩阵。在此模型中,通常假设 Δ 是期望为 0 的偶然误差。也就是说除去观测误差,观测值 L 完全表示为未知参数 X 的函数。如果模型不准确,或者观测值中有系统误差,式(1)并不能严格成立,而应改写为

$$L = BX + S + \Delta \tag{2}$$

式中, $S = (s(t_1), s(t_2), \dots, s(t_n))^T$ 是一个描述模型误差或系统误差,且与未知参数 X 有关的 n 维未知向量。在一般情况下,可以认为模型误差或观测值的系统误差的性态非常复杂,无法用简单的少数参数表示出来,也就是所谓的非参数分量。这种在观测方程中既有参数向量又有非参数分量的模型,称为半参数模型^[3]。

设观测值 $L_i (i = 1, 2, \dots, n)$ 是一组独立观测量。 L_i 的先验方差为

$$\sigma_i^2 = \sigma_0^2 q_i \tag{3}$$

式中, σ_0^2 为单位权方差因子, q_i 为 L_i 的先验协因数。令 $\tilde{\omega}_i = \sigma_0 / \sigma_i = 1 / \sqrt{q_i}$, 这里 $\tilde{\omega}_i$ 与通常的权 p_i 有如下关系

$$\tilde{\omega}_i = \sqrt{p_i} \tag{4}$$

设 L_i 服从一元 p -范分布^[1], 其概率密度为

$$f(L_i) = \frac{p \lambda}{2 \alpha \Gamma(\frac{1}{p})} \exp \left\{ - \left[\frac{\lambda |L_i - \mu_{i,X}|}{\alpha} \right]^p \right\} =$$

$$\frac{p \tilde{\omega}_i \lambda}{2 \sigma_0 \Gamma(\frac{1}{p})} \exp \left\{ - \left[\frac{\tilde{\omega}_i |L_i - \mu_{i,X}|}{\sigma_0} \right]^p \right\} \tag{5}$$

式中, $\lambda = \left[\Gamma(\frac{3}{p}) / \Gamma(\frac{1}{p}) \right]^{\frac{1}{2}}$, $\mu_{i,X} = b_i X + s(t_i)$, $\Gamma(x)$ 为伽玛函数。用平滑似然法可求得参数 X, S, σ_0 的估值 $\hat{X}, \hat{S}, \hat{\sigma}_0$ 。

对式(5)取对数求和可得

$$\begin{aligned} \Phi = & \sum_{i=1}^n \Phi_i(\mu_{i,X}) = \sum_{i=1}^n \ln(L_i | \mu_{i,X}, \sigma_0) = \\ & n \left\{ \ln p - \ln 2 - \ln \sigma_0 - \ln \Gamma\left(\frac{1}{p}\right) \right\} + \\ & \ln \left[\prod_{i=1}^n \tilde{\omega}_i \right] + \frac{n}{2} \left\{ \ln \Gamma\left(\frac{3}{p}\right) - \ln \Gamma\left(\frac{1}{p}\right) \right\} - \\ & \frac{1}{\sigma_0^p} \left[\Gamma\left(\frac{3}{p}\right) / \Gamma\left(\frac{1}{p}\right) \right]^{\frac{p}{2}} \cdot \\ & \sum_{i=1}^n \tilde{\omega}_i^p |L_i - b_i X - s(t_i)|^p \end{aligned}$$

为了得到 X 和 S 的估计值,采用如下局部似然函数(local likelihood)^[4]:

$$\Phi^b(S(t)) = \sum_{i=1}^n K_h(t - t_i) \ln(L_i | \mu_{i,S(t)}, \sigma_0) \tag{6}$$

这里 $K_h(\cdot)$ 为核权函数, h 为带宽, $K_h(\cdot) = K(\cdot/h)/h$, $\mu_{i,S(t)} = b_i X + \hat{s}(t)$ 为模型在 t 时刻的最优估计。

令 $\frac{\partial \Phi^b}{\partial S} = 0$, 可得

$$\sum_{i=1}^n K_h(t - t_i) \Phi' (b_i X + \hat{s}(t)) = 0 \tag{7}$$

令 $\frac{\partial \Phi}{\partial \sigma_0} = 0$, 可得

$$\sigma_0^p = \frac{p}{n} \left[\Gamma\left(\frac{3}{p}\right) / \Gamma\left(\frac{1}{p}\right) \right]^{\frac{p}{2}} \cdot \sum_{i=1}^n \tilde{\omega}_i^p \cdot |L_i - \mu|^p \tag{8}$$

令 $\frac{\partial \Phi}{\partial X} = 0$, 可得

$$\sum_{i=1}^n \Phi' (b_i X + \hat{s}(t_i)) \cdot (b_i + \hat{s}'_i) = 0 \tag{9}$$

式(7)对 X 再求导,化简可得下式

$$\hat{s}'(t) = - \frac{\sum_{i=1}^n K_h(t - t_i) \Phi'' (b_i X + \hat{s}(t)) b_i}{\sum_{i=1}^n K_h(t - t_i) \Phi'' (b_i X + \hat{s}(t))} \tag{10}$$

Nelder 和 Wedderburn^[5] 提出 Fisher scoring 作为广义线性模型^[4] 中估计参数 β 的方法。即给定一个 β , 通过下式

$$\beta^{new} = \beta + \left\{ E \left[- \frac{\partial^2 \Phi}{\partial \beta \partial \beta^T} \right] \right\}^{-1} \cdot \frac{\partial \Phi}{\partial \beta} \quad (11)$$

来迭代, 从而可得出 β 的真值。

在这里, 由式(7) ~ (10), 通过利用 Fisher scoring 得到的迭代公式如下

$$\bar{X}^{new} = \bar{X} - \left(\sum_{i=1}^n \Phi_i''(b_i \bar{X} + \hat{s}(t_i)) \tilde{b}_i^2 \right)^{-1} \cdot \left(\sum_{i=1}^n \Phi_i'(b_i \bar{X} + \hat{s}(t_i)) \tilde{b}_i \right) \quad (12)$$

$$\hat{s}^{new}(t_j) = \hat{s}(t_j) - \frac{\sum_{i=1}^n K_h(t_i - t_j) \Phi_i'(b_i \bar{X} + \hat{s}(t_j))}{\sum_{i=1}^n K_h(t_i - t_j) \Phi_i''(b_i \bar{X} + \hat{s}(t_j))} \quad (13)$$

其中

$$\tilde{b}_j = b_j + \hat{s}'(t_j) = \frac{\sum_{i=1}^n K_h(t_i - t_j) \Phi_i''(b_i \bar{X} + \hat{s}(t_j)) b_i}{\sum_{i=1}^n K_h(t_i - t_j) \Phi_i''(b_i \bar{X} + \hat{s}(t_j))} \quad (14)$$

为了计算的简便, 我们采用如下记号

$$A_{ij} = K_h(t_i - t_j) \Phi_i''(b_i \bar{X} + \hat{s}(t_j)) / \left[\sum_{i=1}^n K_h(t_i - t_j) \Phi_i''(b_i \bar{X} + \hat{s}(t_j)) \right] \quad (15)$$

$$W = \text{diag}\{ \Phi_1'', \Phi_2'', \dots, \Phi_n'' \}$$

$$v = (\Phi_1', \Phi_2', \dots, \Phi_n')^T$$

则式(12) 用矩阵形式可表示为

$$\bar{X}^{new} = (\tilde{B}^T W \tilde{B})^{-1} \tilde{B}^T W \tilde{Z} \quad (16)$$

其中, $\tilde{B} = (I - A)B$, $\tilde{Z} = \tilde{B}X - W^{-1}v$, I 为单位阵。

为了得出 X 和 S 的具体形式, 当 $p \neq 1$ 时, 计算下列形式的导数(若 $p = 1$, 可用一个接近 1 的数计算导数):

$$\begin{aligned} & \frac{\partial}{\partial \mu} \left[\sum_{i=1}^n \tilde{\omega}_i^p |L_i - \mu|^p \right] = \\ & - p \sum_{i=1}^n \tilde{\omega}_i^p |L_i - \mu|^{p-1} \text{sign}(L_i - \mu) = \\ & - p \sum_{i=1}^n \tilde{\omega}_i^p |L_i - \mu|^{p-2} \cdot (L_i - \mu) \quad (17) \end{aligned}$$

则

$$\Phi_i' = - \frac{p}{\sigma_0^p} \left[\Gamma \left(\frac{3}{p} \right) / \Gamma \left(\frac{1}{p} \right) \right]^{\frac{p}{2}} \cdot \tilde{\omega}_i^p \cdot |L_i - \mu|^{p-2} \cdot (L_i - \mu) \quad (18)$$

$$\Phi_i'' = - \frac{p(p-1)}{\sigma_0^p} \cdot \left[\Gamma \left(\frac{3}{p} \right) / \Gamma \left(\frac{1}{p} \right) \right]^{\frac{p}{2}} \cdot \tilde{\omega}_i^p \cdot |L_i - \mu|^{p-2} \quad (19)$$

将式(19) 代入式(13)、(15) 得

$$A_{ij} = \frac{K_h(t_i - t_j) \tilde{\omega}_i^p |L_i - b_i \bar{X} - \hat{s}(t_j)|^{p-2}}{\sum_{k=1}^n K_h(t_k - t_j) \tilde{\omega}_k^p |L_k - b_k \bar{X} - \hat{s}(t_j)|^{p-2}} \quad (20)$$

$$\hat{s}^{new}(t_j) = \hat{s}(t_j) - \frac{\sum_{i=1}^n K_h(t_i - t_j) \tilde{\omega}_i^p |L_i - b_i \bar{X} - \hat{s}(t_j)|^{p-2} (L_i - b_i \bar{X} - \hat{s}(t_j))}{(p-1) \sum_{i=1}^n K_h(t_i - t_j) \tilde{\omega}_i^p |L_i - b_i \bar{X} - \hat{s}(t_j)|^{p-2}} \quad (21)$$

注: 当 $p = 2$ 时, 一元 p - 范分布的密度函数为

$$f(L) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[- \frac{(L - \mu)^2}{2\sigma^2} \right]$$

这就是正态分布的密度函数。此时

$$\begin{aligned} A_{ij} &= \frac{K_h(t_i - t_j) p_i}{\sum_{k=1}^n K_h(t_k - t_j) p_k} \\ s_j^{new} &= \frac{\sum_{i=1}^n K_h(t_i - t_j) p_i (L_i - b_i \bar{X})}{\sum_{i=1}^n K_h(t_i - t_j) p_i} \end{aligned}$$

则有

$$\bar{X}^{new} = (\tilde{B}^T \tilde{B})^{-1} \tilde{B}^T \tilde{L} \quad (22)$$

$$S^{new} = A(L - B\bar{X}^{new}) \quad (23)$$

其中, $\tilde{L} = (I - A)L$, $\tilde{B} = (I - A)B$, I 为单位阵。式(22)、(23) 即为通常情况下的半参数模型^[3]。

直接解式(16)、(21) 很不方便, 也很不实际, 实际工作中, 可采用 backfitting 方法求解方程组, backfitting 是一个迭代求解的方法, 它在两方程之间交替迭代, 直至收敛为止。现将一元 p - 范半参数模型的极大似然估值的求解步骤总结如下:

1. 确定先验单位方差因子 σ_0^2 , 观测值的先验方差, 并计算 $\tilde{\omega}_i$;
2. 确定未知参数的迭代初始值, X_0 取其最小二乘估计值。
3. 由式(15) 计算系数矩阵 A , W , v ;
4. 计算式(16)、(21), 得到参数估值 \bar{X}^{new}

$\hat{\delta}^{new}$;

- 5. 检查 ΔX , Δs 是否小于迭代阈值, 若 ΔX 或 Δs 大于迭代阈值, 重复 3、4 步;
- 6. 由式(8)计算单位权方差因子的估值 σ_0^2 。

2 算例分析

这里构造两个模拟的平差问题, 来验证本文采用的方法。

算例 1: 有线性模型 $Y = BX$, $X = 1$, $B = (b_i)_{100 \times 1}$, $b_i = -2 + i/50$, 模型误差为 $S = (s_i)_{100 \times 1}$, 其中 $s_i = \sin(t_i\pi/100)$, $t_i = i, i = 1, 2, \dots, 100$, 观测误差 Δ 为一个来自拉普拉斯分布的子样($\mu = 0, \sigma = 1$), 由文献[1]知, 拉普拉斯分布是一元 p - 范分布当 $p = 1$ 时的特殊情况(由于在这里 p 不能等于 1, 不妨选取 $p = 1.0005$)。具体数据参见文献[1], 则观测向量 $L = Y + S + \Delta$ 。真实值与观测值的关系如图 1(图中横坐标表示时间, 纵坐标表示对应的值), 图中虚线为观测向量 L , 实线为 L 的真实值 $Y + S$ 。在这里不妨取权阵 P 为单位阵。利用最小二乘平差得 $X = (B^T PB)^{-1} B^T PL = 0.4093$ 。核函数取常用的 Epanechnikov 核 $K(t) = 0.75(1 - t^2)(|t| \leq 1)$, 带宽 h 不妨取为 0.225, 利用式(16)、(21) 计算, 得到 $X = 1.0002$ 。

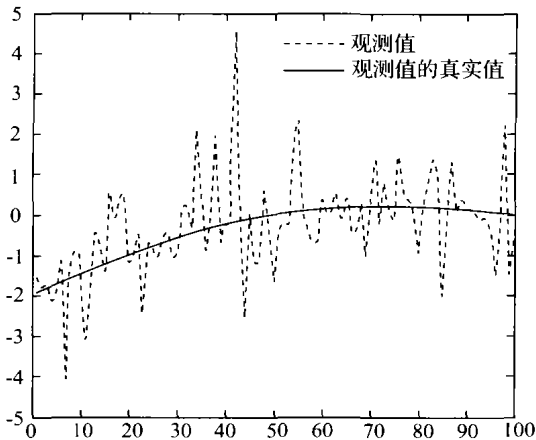


图 1 观测值与真实值的关系($p = 1$)

Fig. 1 The Observations and their true values($p = 1$)

算例 2: 当 $p = 2$ 时, p - 范分布即为常见的正态分布, 则观测向量 $L = Y + S + \Delta$ 。 Y, S 同算例 1, 观测误差 $\Delta \sim N(0, 1)$ 。在这里不妨取权阵 P 为单位阵, 利用最小二乘平差得 $X = (B^T PB)^{-1} B^T PL = 0.9677$ 。真实值与观测值的关系如图 2(图中横坐标表示时间, 纵坐标表示对

应的值), 图中虚线为观测向量 L , 实线为 L 的真实值 $Y + \Delta$ 。从图中还可以看出, 由于加入了误差向量 S , 从而使观测值 X 与它的真实值的误差较大。利用半参数模型进行计算。核函数同上, 对于式(22)、(23) 应用参数求解的迭代步骤 1~6, 对于不同的带宽 h 得到 X 的估值如下表:

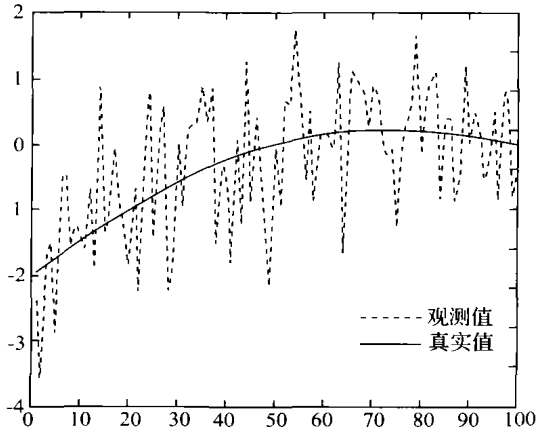


图 2 观测值与真实值的关系($p = 2$)

Fig. 2 The Observations and their true values($p = 2$)

表 1 带宽参数 h 与半参数估值 p - 范 X_p

Tab. 1 Bandwidth parameter and estimation of parametric component

带宽参数 h	0.1	0.2	0.3	0.4
参数估值 X_p	1.0569	1.0824	1.2550	1.0101
带宽参数 h	0.5	0.6	0.7	0.8
参数估值 X_p	0.9427	0.9574	0.9795	0.9865
带宽参数 h	0.9	1		
参数估值 X_p	0.9903	0.9926		

从表中可以看出, 当 $h = 1$ 时, 半参数平差模型参数 X 的估值 X 与真值相当接近, 半参数模型的拟合效果得到大大提高。

3 结 论

通过以上分析可以看出, 将误差分布的一般模式——一元 p 范分布, 扩充到半参数 p 范分布模型理论中, 通过选取适当的核函数和带宽参数, 能够有效地减少模型中参数估值的误差, 同时还可以从有噪声的观察向量中分离出非参数分量。

另外, 半参数模型在处理参数估计问题时, 一方面解决了单纯线性回归模型与非线性回归模型

难以解决的问题,增强了模型的适应性。另一方面克服了非参数方法信息损失过多的问题,能够分离出模型中的非参数分量;其次,它是一套解决实际问题的工具,使得原先在参数情形下使用的工具无力解决的问题找到了新方法,因此,半参数回归模型较参数线性模型有较强的适应性。由于实际工作中经常会遇到某个变量有影响,尤其是当应变量与自变量间函数关系不清楚时,参数模型难以进行拟合处理,而半参数法却能加以有效分析。

参考文献:

[1] YU Zong chou, TAO Ben zao, LIU Da jie. Generalized Surveying Adjustment [M]. Wuhan: WTUSM

Press, 1996. 83-95. (in Chinese)

[2] SUN Hai yan. Theory of P norm Distribution and Application of Surveying Data Processing[M]. Wuhan: WTUSM Press, 1994. (in Chinese)

[3] CHAI Geng xiang, SHI Yun chi, QIAN Zhi jian. Semiparametric Model of Time Series under Fixed Design [J]. Chinese Ann. Math. (Series A), 2001, 22 (2): 163-176. (in Chinese)

[4] MARLENE M. Semiparametric Extensions to Generalized Linear Models[M]. Berlin: Schrift zur Habilitation im Fach Statistik, 2000. 35-55.

[5] NELDER J A, WEDDERBUM R W. Generalized Linear Models[J]. J Roy Statis Soc A, 1972, 135, 370-184.

中国测绘学会 2004 年优秀地图作品奖获奖名单

[本刊讯] 根据《国家科学技术奖励条例》和《中国测绘学会科学技术奖励办法》,中国测绘学会组成 2004 年“优秀地图作品评审委员会”,进行了 2004 年优秀地图作品奖的评选工作,通过对推荐项目严格评审,经向社会公示广泛征求意见和中国测绘学会科学技术奖励委员会审核批准,《中国西部地区生态环境现状遥感图集》等 6 项作品获 2004 年优秀地图作品一等奖;《长江防洪地图集》等 11 项作品获 2004 年优秀地图作品二等奖。

	地图作品名称	编制单位	出版单位	申报单位
一 等 奖	中国西部地区生态环境现状 遥感图集	中国测绘科学研究院	科学出版社	中国测绘科学研究院
	长江经济带可持续发展图集	中科院成都山地灾害与环境 研究所	科学出版社	科学出版社
	中华人民共和国政区标准地 名图集	国家民政部/总参测绘局	星球地图出版社	星球地图出版社
	中国自然灾害系统地图集	北京师范大学资源学院	科学出版社	北京师范大学资源学院
	中国地质图集	中国地质科学院地质研究所	地质出版社	地质出版社 中国地质科学院地质研究所
	世界分国地图	中国地图出版社	中国地图出版社	中国地图出版社
二 等 奖	长江防洪地图集	长江水利委员会	科学出版社	长江水利委员会航测信息工 程院
	土地利用动态遥感监测图集	中国测绘科学研究院	科学出版社	中国测绘科学研究院
	广东省地图集	广东省地图出版社	广东省地图出版社	广东省地图出版社
	中国道路网地图集	广东省地图出版社	广东省地图出版社	广东省地图出版社
	上海市影像地图集	上海市测绘院	上海科学技术出版社	上海市测绘院
	(新课标北师大)历史地 图册	中国地图出版社	中国地图出版社	中国地图出版社
	兰州城区影像地图集	甘肃省地图院	甘肃省人民出版社	甘肃省地图院
	中国交通图集	成都地图出版社	成都地图出版社	成都地图出版社
	少儿中国·世界地图	成都地图出版社	成都地图出版社	成都地图出版社
	南极洲全图	武汉大学中国南极测绘研究 中心	中国地图出版社	中国南极测绘研究中心
	宁波旅游交通图	宁波市测绘设计研究院	西安地图出版社	宁波市测绘设计研究院