文章编号:1671-8860(2012)03-0298-04　　文献标志码:A

半参数模型中影响正则化参数的因素分析

陶肖静[1,2]　朱建军[1,2]　田玉淼[1,2]

(1　中南大学地球科学与信息物理学院,长沙市麓山南路 932 号,410083)
(2　湖南省普通高等学校精密工程测量与形变灾害监测重点实验室,长沙市麓山南路 932 号,410083)

摘　要:通过模拟算例,比较了 L 曲线法、GCV 法(广义交叉核实法)和虚拟观测法确定出的正则化参数,并对影响正则化参数的因素进行了分析,得出结论:正则化参数的确定与信噪比密切相关,当信噪比增大时,各种方法确定的正则化参数变化趋势不同,不同的情况确定正则化参数的适用方法也会有所差异。

关键词:正则化参数;影响因素;L 曲线法;GCV 法;虚拟观测法

中图法分类号:P207.2

目前,半参数模型估计方法[1-3]中,补偿最小二乘法应用最广泛,其关键为正则化矩阵和正则化参数的确定。确定正则化参数的方法[4-7]较多,但各种方法确定的正则化参数并不相同,且没有统一的评价指标。那么,影响正则化参数的主要因素有哪些?它们对正则化参数产生了怎样的影响?本文通过模拟算例分析了影响正则化参数的主要因素。

1　半参数回归模型

一般情况下,半参数回归模型可写为:

$$\underset{n\times 1}{\boldsymbol{L}}=\underset{n\times u}{\boldsymbol{B}}\ \underset{u\times 1}{\boldsymbol{X}}+\underset{n\times 1}{\boldsymbol{S}}+\underset{n\times 1}{\boldsymbol{\Delta}},\boldsymbol{\Delta}\sim N(0,\sigma_0^2\boldsymbol{P}_L^{-1})\quad(1)$$

式中,$\boldsymbol{L}$ 为观测向量;$\boldsymbol{B}$ 为系数阵;$\boldsymbol{X}$ 为参数向量;$\boldsymbol{S}$ 为变化规律复杂的非参数向量;$\boldsymbol{\Delta}$ 为偶然误差;$\sigma_0^2\boldsymbol{P}_L^{-1}$ 为观测误差方差。

按补偿最小二乘估计准则[8]解算式(1):

$$\Phi=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{V}+\alpha\boldsymbol{S}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{S}=\min\qquad(2)$$

其中,$\boldsymbol{R}$ 为正则化矩阵,描述了非参数的光滑性;α 称为平滑参数,调节拟合部分 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{V}$ 和光滑部分 $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{R}\boldsymbol{S}$ 的平衡。解算半参数模型的关键是选择合适的 $\boldsymbol{R}$ 和 α。

选择正则化参数 α 的主要方法有:①L 曲线法[9];② 交叉核实或广义交叉核实法[10];③ 虚拟观测值法[6]。

2　影响正则化参数的因素分析

在补偿最小二乘准则中,正则化参数起到了平衡拟合与平滑部分的作用,因此,可以认为正则化参数与观测噪声和非参数有关,即与信噪比有关。另外,确定正则化参数的方法是从不同的角度来推算的,L 曲线法主要靠曲线拟合来顾及补偿最小二乘准则中的拟合和光滑部分,因此,观测值的个数不能过少;GCV 法是在所有预测点的均方误差最小准则下求得的最佳值;基于 Helmert 方差分量的虚拟观测法是把正则化参数看成两类观测的权来确定。因此,正则化参数与方法本身也有关系。

假设有线性模型 $\boldsymbol{Y}=\boldsymbol{B}\boldsymbol{X}$,并取 $\boldsymbol{X}=[2\quad 3]^{\mathrm{T}}$。$\boldsymbol{B}=(b_{i,j})$ 为 200×2 阶矩阵,$b_{i,1}=7/3\sin(12i/5)$,$b_{i,2}=7/3(\sin(12i/5))^2$,假设系统误差为 $\boldsymbol{S}=[s_1\quad s_2\quad\cdots\quad s_{200}]^{\mathrm{T}}$,$s_i=10(\cos^3(t_i)+((t_i+300)/100)^2)$,$t_i=2(i-1)\pi/100$,$i=1,2,\cdots,200$。观测值的真值为 $\widetilde{\boldsymbol{L}}=\boldsymbol{B}\boldsymbol{X}+\boldsymbol{S}$。观测误差 $\boldsymbol{\Delta}$ 由 200 个服从正态分布的随机数组成列向量,于是观测值为 $\boldsymbol{L}=\boldsymbol{B}\boldsymbol{X}+\boldsymbol{S}+\boldsymbol{\Delta}$,观测值权阵 $\boldsymbol{P}$ 为单位阵。正则化矩阵 $\boldsymbol{R}$ 按照时间序列法选取,本文采用了一阶平滑的方法。

2.1　信噪比和确定方法对正则化参数的影响

为分析正则化参数与信噪比的关系,本文进行了两步计算:首先,模拟不同的随机噪声和系统

收稿日期:2011-12-15。

项目来源:国家自然科学基金资助项目(40974007);湖南省教育厅基金资助项目(09K006)。

误差,但保持两者之比相同(即信噪比相同);其次,固定随机噪声,而系统误差逐渐增大(即信噪比逐渐增大)。

上述计算中,分别采用本文提到的 3 种方法来确定 α,并对模型解算结果进行对比分析,以说明各种方法本身对正则化参数确定的影响。计算结果见表 1～3 和图 1、2,其中,k 表示系统误差的系数,RMS 表示均方根误差。

从表 1 可以看出,当信噪比相同时,相同方法确定出的正则化参数保持不变。由计算结果可知,信噪比确定时,最优方法是确定的。

表 2 中显示,当信噪比逐渐增大时,用不同方法确定出的正则化参数也随之改变,因此,正则化参数的确定与信噪比密切相关。随机噪声固定不变,系统误差系数相同的情况下(即信噪比相同)3 种方法确定出的正则化参数并不相同。随着系统误差 S 的量级不断增大,用各种方法确定的 α 的变化趋势有如下差异:L 曲线法确定出的 α 增大,其他两种方法确定出的 α 减小。在系统误差量级的变化量相同时,α 的变化量不相同。实验结果表明,对于固定的随机误差都有以上规律。

表 1　取不同的随机噪声和系统误差量级,保持两者之比相等,3 种方法解算的结果

Tab. 1　Results of Three Methods in Condition of the Same Ratio with Different Magnitudes of Systematic Errors and Random Errors

		$N(0,0.5)$, $k=2$	$N(0,1)$, $k=4$	$N(0,2)$, $k=8$
L 曲线法	α	0.806 0	0.806 0	0.806 0
	$\hat{\sigma}_0^2$	0.137 8	0.551 4	2.205 5
	RMS_X	0.033 7	0.067 5	0.134 9
	RMS_S	0.261 7	0.523 4	1.046 7
GCV 法	α	6.143 0	6.143 0	6.143 0
	$\hat{\sigma}_0^2$	0.174 0	0.696 1	2.784 4
	RMS_X	0.032 9	0.065 7	0.131 4
	RMS_S	0.177 2	0.354 4	0.708 8
虚拟观测法	α	2.775 9	2.775 9	2.775 9
	$\hat{\sigma}_0^2$	0.174 9	0.699 6	2.798 4
	RMS_X	0.033 1	0.066 2	0.132 4
	RMS_S	0.197 3	0.394 6	0.789 1

表 2　取 $\Delta\sim N(0,1)$,并保持随机噪声的量级不变,k 的取值为由 2 到 20(步长为 2),3 种方法解算的结果

Tab. 2　Results of Three Methods in Condition of k Ranges from 2 to 20 (Step 2) with Constant Magnitudes of Random Errors

k	L 曲线法				GCV 法				虚拟观测法			
	α	RMS_X	RMS_S	RMS_S/k	α	RMS_X	RMS_S	RMS_S/k	α	RMS_X	RMS_S	RMS_S/k
2	0.772 0	0.067 4	0.528 0	0.264 0	14.274 0	0.064 9	0.305 5	0.152 8	9.008 2	0.065 1	0.313 7	0.156 8
4	0.806 0	0.067 5	0.523 4	0.130 8	6.143 0	0.065 7	0.354 4	0.088 6	2.775 9	0.066 2	0.394 6	0.098 7
6	0.865 0	0.067 5	0.515 8	0.086 0	4.015 0	0.066 3	0.387 1	0.064 5	1.393 1	0.067 1	0.462 1	0.077 0
8	0.949 0	0.067 6	0.506 4	0.063 3	3.008 0	0.066 7	0.412 2	0.051 5	0.824 6	0.067 7	0.523 0	0.065 4
10	1.059 0	0.067 6	0.496 4	0.049 6	2.409 0	0.067 1	0.432 9	0.043 3	0.521 6	0.068 2	0.581 9	0.058 2
12	1.193 0	0.067 7	0.487 6	0.0406	2.009 0	0.067 4	0.450 7	0.037 6	0.337 2	0.068 6	0.640 8	0.053 4
14	1.348 0	0.067 7	0.481 6	0.0344	1.720 0	0.067 6	0.466 5	0.033 3	0.214 7	0.069 0	0.701 3	0.050 1
16	1.517 0	0.067 9	0.480 2	0.030 0	1.500 0	0.067 9	0.480 8	0.030 1	0.128 4	0.069 3	0.764 1	0.047 8
18	1.695 0	0.068 0	0.484 7	0.026 9	1.327 0	0.068 1	0.493 9	0.027 4	0.065 0	0.069 6	0.829 5	0.046 1
20	1.874 0	0.068 2	0.495 7	0.024 8	1.187 0	0.068 3	0.506 1	0.025 3	0.016 6	0.069 9	0.8980	0.044 9

2.2　验证方法本身的影响

为了验证 3 种不同方法的效果,在本文算例的计算中给出了 α 从 1 取到 20 时的解算结果,对比分析得到 α 最优值,从而对确定 α 的最佳方法进行验证。本文还给出了解算过程的图形表示,以更直观地说明问题。

表 3　当 $k=10$,$\Delta\sim N(0,1)$ 时,取不同的 α 的解算结果

Tab. 3　Results of Different α with $k=10$, $\Delta\sim N(0,1)$

α	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	RMS_X	α	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	RMS_X
1	2.018 7	3.093 8	0.067 6	11	2.021 4	3.092 9	0.067 4
2	2.019 1	3.093 0	0.067 2	12	2.021 6	3.093 0	0.067 5
3	2.019 4	3.092 7	0.067 0	13	2.021 8	3.093 1	0.067 6
4	2.019 7	3.092 6	0.066 9	14	2.022 0	3.093 2	0.067 7
5	2.020 0	3.092 5	0.067 0	15	2.022 2	3.093 3	0.067 8
6	2.020 3	3.092 6	0.067 0	16	2.022 4	3.093 4	0.067 9
7	2.020 5	3.092 6	0.067 1	17	2.022 6	3.093 5	0.068 0
8	2.020 8	3.092 7	0.067 2	18	2.022 7	3.093 6	0.068 1
9	2.021 0	3.092 8	0.067 3	19	2.022 9	3.093 7	0.068 2
10	2.021 2	3.092 8	0.067 3	20	2.023 0	3.093 8	0.068 3

表3为固定一组噪声误差和系统误差之后解的情况。表3中 α 的最优值为4，由表2可得GCV法确定的 α 为最优，两者反映的信息一致，验证了确定 α 的最优方法的正确性。

图1、2中，模拟残差为算例中模拟的随机噪声，残差为解得参数之后得到的残差；实线表示系统误差的真值，虚线表示系统误差的估值。可以看出，正则化参数的确定与随机噪声和系统误差同时相关。由图1、2可以看出，固定一组随机噪声 $\boldsymbol{\Delta}\sim N(0,1)$，当系统误差的系数分别取2和20时，用3种方法解算半参数模型结果的拟合和平滑效果有明显不同。当系数取2时，3种方法得到的解算结果拟合效果都较差，同时，L曲线法得到的平滑效果也较差，总体效果并不理想；但系数取20时，3种方法得出的结果都在拟合和平滑效果之间取得了相对较好的平衡，此时用虚拟观测法确定正则化参数之后，残差取得了较小的值，解算结果优于其他方法。综合表2和图1、2可知，当信噪比逐渐增大时，虽然系统误差的均方根误差也逐渐增大，但与系统误差系数的比值逐渐减小，因此，系统误差的相对误差逐渐减小，与图中反映的信息（拟合和平滑效果越来越好）相吻合，因此，评价指标与相对精度有关。

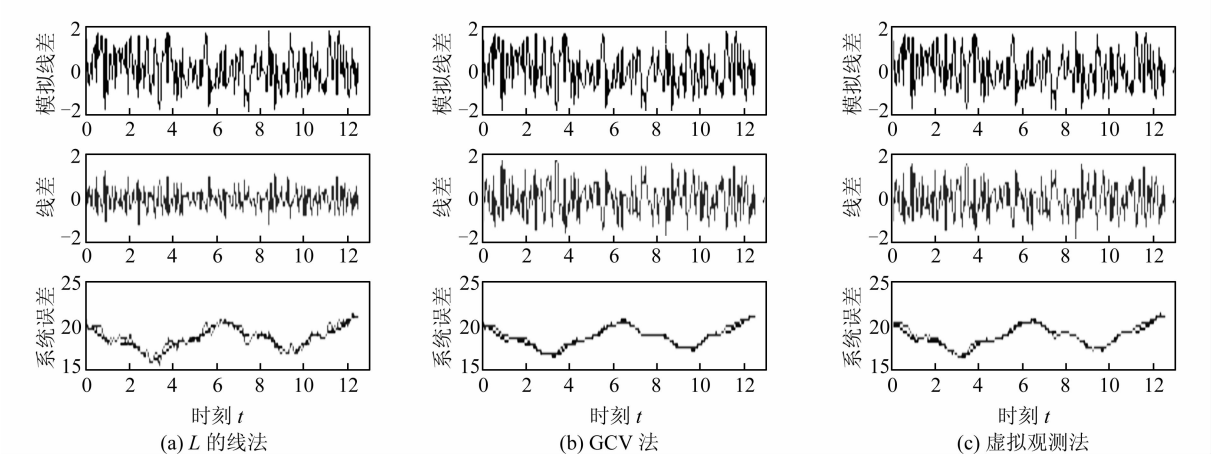

图1 系统误差的系数为2（$k=2$）时，3种方法得到的拟合和光滑效果图

Fig. 1 Diagram of Fitting and Smoothing of Three Methods when Coefficient of Systemic Errors is 2 ($k=2$)

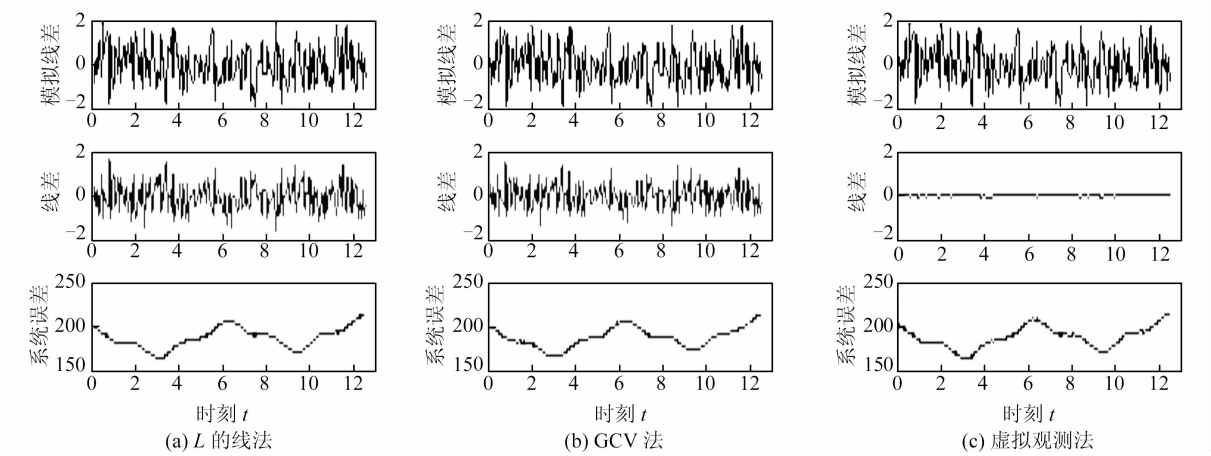

图2 系统误差的系数为20（$k=20$）时，3种方法得到的拟合和光滑效果图

Fig. 2 Diagram of Fitting and Smoothing of Three Methods When Coefficient of Systemic Errors is 20 ($k=20$)

本文算例是选取了两个参数的情况，若选取参数真值为 $\boldsymbol{X}=[2\quad 3\quad 5]$，相应的系数矩阵加一列，取 $b_{i,3}=7/3(\sin(12i/5))^3$，其余条件与本文算例相同，按照同样的方法进行了一系列验证，最后得出了与两个参数时相同的结论。

3 结　语

1) 同一方法相同信噪比可得出相同的 α;确定 α 的最优方法是随信噪比的变化而变化的,也就是说,假如信噪比取定某一个值后,用 L 曲线法确定的 α 最优,那么,当信噪比变化时,L 曲线法就不一定最优了。

2) 对同一个模型来说,在方法相同的情况下,不同信噪比得出的 α 有差异;在选定了方法后,正则化参数是随信噪比的变化而变化的,但对应的变化趋势却不相同,L 曲线法得到的 α 随信噪比的增大而增大,其他两种方法相反。本文通过大量实例分析了两个和三个参数时的情况,得出的结论是一致的,此规律有一定的普适性。

3) 由于实际问题中非参数部分具有一定的复杂性,信噪比的大小势必受到影响,从而影响到正则化参数的确定,因此,在确定正则化参数时,要充分分析各种情况来选择不同的方法。对于不同方法的评价指标,本文初步得出与参数的相对精度有关的结论,具体的影响因素量级上的研究以及对各种方法选取的评价未来将进一步探讨。

参 考 文 献

[1] 丁士俊.测量数据的建模和半参数估计[D].武汉:武汉大学,2005

[2] 胡宏昌.半参数模型的估计方法及其应用[D].武汉:武汉大学,2004

[3] 吴云,孙海燕,马学忠.半参数估计的自然样条函数法[J].武汉大学学报·信息科学版,2004,29(5):398-401

[4] 王振杰,欧吉坤.用 L-曲线法确定半参数模型中的平滑因子[J].武汉大学学报·信息科学版,2004,29(7):651-653

[5] 李功胜,王家军.一种新的正则化方法的正则参数的最优后验选取[J].数学杂志,2002,22(1):103-106

[6] 朱建军,冯光财,戴吾蛟.半参数模型解算的一种虚拟观测法[J].工程勘察,2006(9):54-57

[7] 孙海燕,吴云.半参数回归与模型精化[J].武汉大学学报·信息科学版,2002,27(2):172-174

[8] Green P J, Silverman B W. Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach [M]. London: Chapman and Hall, 1994

[9] Fischer B, Hegland M. Collocation, Filtering and Nonparametric Regression, Part Ⅱ [J]. ZfV, 1999(2):46-52

[10] Golub G H, Heath M, Wahba G. Generalized Ross Validation as a Method for Choosing a Good Ridgeparamete[J]. Technometrics, 1979, 21(2): 215-223

第一作者简介:陶肖静,博士生,主要从事半参数模型方面的数据处理。

E-mail:txjing1008@163.com

Analysis of Factors Influencing Smoothing Parameter in Semiparametric Model

TAO Xiaojing[1,2]　*ZHU Jianjun*[1,2]　*TIAN Yumiao*[1,2]

(1 School of Geosciences and Info-physics, Central South University, 932 South Lushan Road, Changsha 410083, China)

(2 Key Laboratory of Precise Engineering Surveying & Deformation Hazard Monitoring, Universities of Hunan Province, 932 South Lushan Road, Changsha 410083, China)

Abstract: Many methods can be used to determine smoothing parameters in semiparametric model, but smoothing parameters vary with these methods for the same model. After comparative analysis of the L-Curve, GCV(Generalize Cross Validation) and virtual observational approach, the factors influencing smoothing parameter are studied by simulated example. Smoothing parameters have a close relation with signal-to-noise ratio. When the signal-to-noise ratio increases, variation trend of smoothing parameters is different. The best method determining smoothing parameters in different situations is different.

Key words: smoothing parameters; influencing factors; L-Curve; GCV; virtual observation

About the first author: TAO Xiaojing, Ph. D candidate, majors in data processing of semi-parametric model.

E-mail: txjing1008@163.com