非参数回归

宋歌 2015080086 数52

2018 年 3 月 12 日

1 实验的目的

用近邻加权平均的方法，利用已知的观测值实现模型$y = m(x) + e$的非参数回归模拟。

2 选择的因素

样本量$n = 30, 100, 1000$，误差e分布的标准差$\sigma = 0.1, 1, 2$，不同的模型$m(x) = sin(x), x^2, (log(x))^2, x * sin(1/x)$。

3 详细的研究方法

3.1 产生观测值

对于每一个模型$y = m(x) + e$，产生n个$[0, 1]$上均匀分布的随机数x_i作为自变量，产生n个服从$N(0, \sigma^2)$的随机数e_i作为误差，通过$y = m(x) + e$产生相应的n个y_i。将这些x_i, y_i作为已获得的观测值。

3.2 取定待估计分点并计算距离

将区间$[0, 1]$均匀分割为100份，将这101个分点记为$t_1, t_2, \ldots, t_{101}$。然后对于不同的$n$和$\sigma$，都进行如下操作：

对于每一个分点t_j，都计算出所有x_i到t_j的距离，并用一个数组来记录x_i到t_j的距离。

3.3 取近邻点作平均

对于每一个分点t_j，对距离数组进行排序，取其索引，找到距离t_j最近的五个点x_i，计算对应的五个y_i的平均值，作为在分点t_j 上的估计$\hat{m}(t_j)$。

3.4 评估近邻平均方法

用a来记录在$t_1, t_2, \ldots, t_{101}$这些点上，$\hat{m}(t_j)$与$m(t_j)$的距离平均，以评估该近邻平均方法的好坏。易知，$a$越小，拟合得越好。
同时也可以画出$\hat{m}(t_j)$关于t_j的曲线，与$m(t)$进行比较。

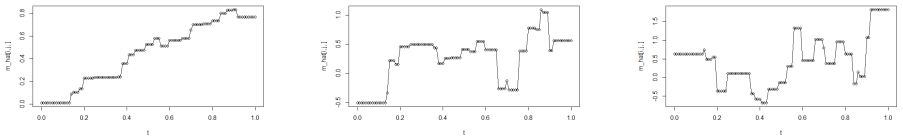
4 结果及讨论

4.1 $m_1(x) = sin(x)$

对于不同的n与σ，a的值如下：

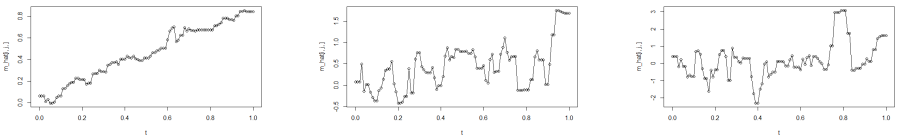
n	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 2$
$n = 30$	0.04206313	0.30743962	0.52042536
$n = 100$	0.03131726	0.34617259	0.78063176
$n = 1000$	0.04072907	0.39862401	0.66379034

在$n = 30, 100, 1000$的时候，$\hat{m}(t_j)$关于t_j的曲线分别如下：



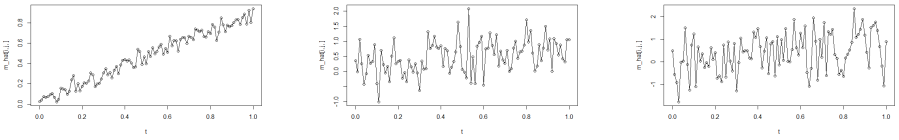
(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

Figure 1: $n = 30$



(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

Figure 2: $n = 100$



(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

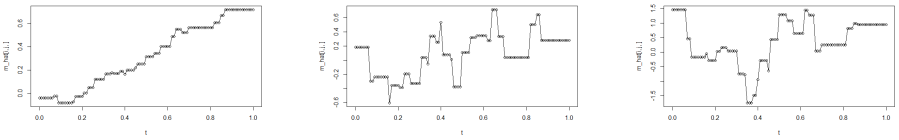
Figure 3: $n = 1000$

4.2 $m_2(x) = x^2$

对于不同的n与σ，a的值如下：

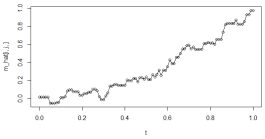
n	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 2$
$n = 30$	0.04726085	0.49771145	0.47433896
$n = 100$	0.03532282	0.32040851	0.64647834
$n = 1000$	0.03197297	0.34728144	0.68696927

在$n = 30, 100, 1000$的时候，$\hat{m}(t_j)$关于t_j的曲线分别如下：

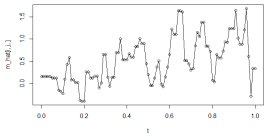


(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

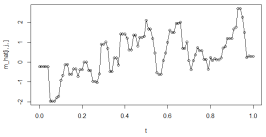
Figure 4: $n = 30$



(a) $\sigma = 0.1$

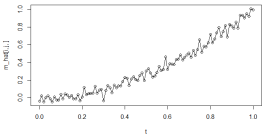


(b) $\sigma = 1$

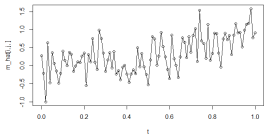


(c) $\sigma = 2$

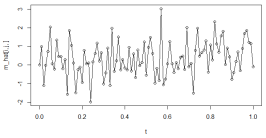
Figure 5: $n = 100$



(a) $\sigma = 0.1$



(b) $\sigma = 1$



(c) $\sigma = 2$

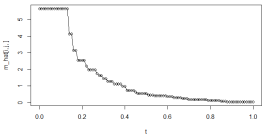
Figure 6: $n = 1000$

4.3 $m_3(x) = (log(x))^2$

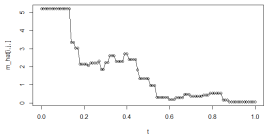
对于不同的n与σ，a的值如下：

n	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 2$
$n = 30$	0.5907299	0.8118271	1.1483001
$n = 100$	0.2363602	0.5331043	0.8445203
$n = 1000$	0.03709858	0.35667904	0.75697030

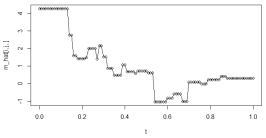
在$n = 30, 100, 1000$的时候，$\hat{m}(t_j)$关于t_j的曲线分别如下：



(a) $\sigma = 0.1$

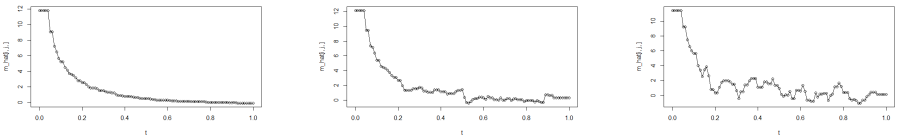


(b) $\sigma = 1$



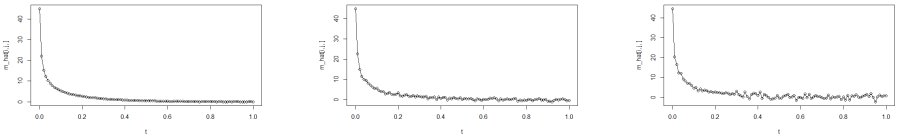
(c) $\sigma = 2$

Figure 7: $n = 30$



(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

Figure 8: $n = 100$



(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

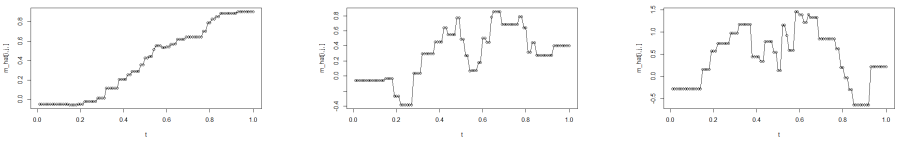
Figure 9: $n = 1000$

4.4 $m_4(x) = x * sin(1/x)$

对于不同的n与σ，a的值如下：

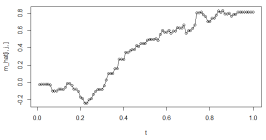
n	$\sigma = 0.1$	$\sigma = 1$	$\sigma = 2$
$n = 30$	0.05162169	0.35776433	0.94005096
$n = 100$	0.03803578	0.35310880	0.73050019
$n = 1000$	0.03326149	0.31989997	0.79323807

在$n = 30, 100, 1000$的时候，$\hat{m}(t_j)$关于t_j的曲线分别如下：

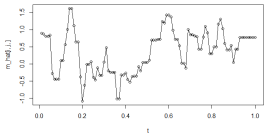


(a) $\sigma = 0.1$ (b) $\sigma = 1$ (c) $\sigma = 2$

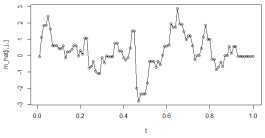
Figure 10: $n = 30$



(a) $\sigma = 0.1$

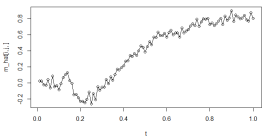


(b) $\sigma = 1$

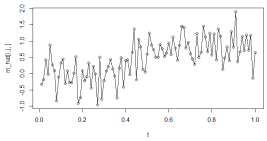


(c) $\sigma = 2$

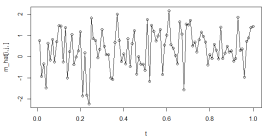
Figure 11: $n = 100$



(a) $\sigma = 0.1$



(b) $\sigma = 1$



(c) $\sigma = 2$

Figure 12: $n = 1000$

由以上结果可知，当n固定时，随着误差服从的正态分布的标准差σ的增大，拟合越来越不好；而当σ固定时，随着样本量n的增大，拟合越来越好。