

基于正则惩罚的半参数密度估计方法

谭 欣¹, 晏 梅²

1. 电子科技大学 数学科学学院, 成都 511731

2. 云南师范大学 数学学院, 昆明 650500

摘 要:目的 针对一元随机变量的概率密度函数估计问题, 寻找一个提高非参数估计收敛速度的方法。方法 通过对参数估计和非参数估计组合加权, 提出一个基于正则惩罚的半参数估计方法: 假定一个参数模型得到其对应的参数估计, 用核方法计算得到其非参数估计, 通过带正则惩罚的经验似然损失或积分平方损失得到一个权重, 基于这个权重对得到的参数估计和非参数估计组合, 从而得到所提出的基于正则惩罚的半参数估计。结果 渐近理论性质显示: 所提出的带正则惩罚的半参数密度估计方法结合了参数估计和非参数估计的优点, 不依赖任何模型假设, 在任何情况下都是收敛的, 并且在参数模型假设正确时, 权重偏向参数估计, 此时收敛速度与参数估计一样, 反之则偏向非参数估计, 收敛速度与非参数估计一致。通过数值模拟实验发现: 当数据满足参数模型假设时, 带惩罚的半参数密度估计方法权重偏向参数估计, 与理论结果一致; 当数据不满足参数模型假设时, 所提出的半参数估计权重偏向非参数估计, 这同样与理论结果一致。最后将该方法应用于重庆市降水数据中, 研究了其月降水量的分布。结论 实例分析结果表明: 所提出的基于正则惩罚的半参数估计与非参数估计相比更光滑, 与参数估计相比拟合更好, 验证了该方法的合理性。

关键词: 密度函数估计; 参数估计; 非参数估计; 加权; 正则惩罚

中图分类号: O212.1 **文献标识码:** A doi:10.16055/j.issn.1672-058X.2025.0002.012

Semi-parametric Density Estimation Method Based on Regular Penalty

TAN Xin¹, YAN Mei²

1. School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 511731, China

2. School of Mathematics, Yunnan Normal University, Kunming 650500, China

Abstract: Objective This paper aimed to find a method to improve the convergence speed of nonparametric estimation for the probability density function of a univariate random variable. **Methods** By combining parameter estimation and nonparametric estimation with weighting, a semiparametric estimation method based on regular penalty was proposed. In this method, a parameter model was assumed to obtain its corresponding parameter estimation. Nonparametric estimation was computed using kernel methods. A weight was then obtained through empirical likelihood loss with a regular penalty or integral square loss. Based on this weight, the proposed semiparametric estimation based on regular penalty was obtained by combining the parameter estimation with nonparametric estimation. **Results** Asymptotic theoretical properties show that the proposed semiparametric estimation method based on regular penalty combines the advantages of parameter estimation and nonparametric estimation. It converges in any case without relying on any model assumption, and when the parameter model assumption is correct, the weight tends toward the parameter estimation, resulting in the same convergence speed as

收稿日期: 2023-08-11 **修回日期:** 2023-11-27 **文章编号:** 1672-058X(2025)02-0086-09

基金项目: 云南省科技计划基础研究项目(202201AU070051); 云南师范大学博士科研启动项目(2020ZB014)。

作者简介: 谭欣(1994—), 女, 重庆梁平区人, 博士研究生, 从事非参数统计研究。

通讯作者: 晏梅(1988—), 女, 云南曲靖人, 博士, 讲师, 从事非参数统计、高维数据分析研究。myanlx920@hotmail.com。

引用格式: 谭欣, 晏梅. 基于正则惩罚的半参数密度估计方法[J]. 重庆工商大学学报(自然科学版), 2025, 42(2): 86-94.

TAN Xin, YAN Mei. Semi-parametric density estimation method based on regular penalty[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(2): 86-94.

the parameter estimation; otherwise, it tends toward the nonparametric estimation, with consistent convergence speed as nonparametric estimation. Numerical simulation experiments show that when the data satisfy the parameter model assumption, the weight of the penalized semi-parametric density estimation method tends toward the parameter estimation, consistent with the theoretical results; when the data do not satisfy the parameter model assumption, the weight of the proposed semi-parametric estimation tends toward the nonparametric estimation, also consistent with the theoretical results. Finally, this method was applied to precipitation data in Chongqing to study the distribution of monthly precipitation. **Conclusion** The results of the case study show that the proposed semiparametric density estimation method based on the regular penalty is smoother compared with nonparametric estimation, and fits better compared with parameter estimation, thus validating the rationality of this method.

Keywords: density function estimation; parameter estimation; nonparametric estimation; weighting; regular penalty

1 引言

统计分析的一个重要问题是密度函数的估计,它对于了解数据结构,解决实际中的许多问题都有非常重要的帮助。本文的主要目的是估计一个一元随机变量 X 的概率密度函数 $f_0(x)$ 。经典的参数方法:先假设变量服从带有未知参数的某个分布,然后根据极大似然、矩估计等方法估计其参数,从而得到它的具体分布,更多的可见文献[1]。当假设的参数模型成立时,其估计的收敛速度为 n^{-1} ,然而模型假设不成立时,这个基于模型假设得到的参数估计很可能不会收敛。相比参数估计,非参数估计只依赖数据本身而不依赖于任何模型假设。一个普遍使用的非参数估计方法就是核密度估计(Kernel Density Estimation, KDE),该方法是由 Silverman 首次提出的,它在生物医学、电力工程、地理科学、气象、金融分析学等各个方面被大量使用[2-10]。核密度估计不依赖任何模型假设,总是收敛的,在开区间上其收敛速度约为 $n^{-4/5}$ [11-13]。

为了进一步提高估计的收敛速度,Olkin 等[14]提出了一个加权的半参数方法:

$$f_{sp}(x) = wf_p(x) + (1-w)f_{np}(x)$$

其中, $w \in [0, 1]$ 是一个未知的权重, f_p 是参数估计, f_{np} 是非参数估计。他们根据观测的样本,用极大似然方法得到这个权重参数的估计 $\hat{w}$ 。文献[14]中定理指出,若某些假设条件成立,那么当参数模型假设成立时,所估计的权重概率满足 $P_r(\hat{w} = 1)$ 随着样本量的增加而趋于 1,反之则 $P_r(\hat{w} = 0)$ 趋于 1。

这个加权方法结合了参数和非参数估计的优点,并且不涉及复杂的优化问题,求解方便。然而,在实际应用中,当用核密度估计方法作为非参估计时,它会出现“过拟合”现象:样本点处的函数值拟合得“太好”,使得即使参数模型假设成立时,非参数估计的经验似然损失函数大部分情况下仍然比参数估计的大,此时所得到的权重 $\hat{w}$ 取值很小,这个情况具体的数值例子可

见第4节的表1。这与理论预期的效果矛盾,也许是这个加权方法在实际中应用很少的原因。

基于这个现象,本文改进了 Olkin 和 Clifford 的方法,提出了一个带正则惩罚的半参数估计方法。该方法通过增加一个正则惩罚来选择更好的权重,使其表现与理论预想一致。此外,本文基于理论推导展现了该方法在样本量趋于无穷时的表现性质,并通过数值模拟进行了验证,实例分析进一步说明了该方法的合理性。

2 方法

设 $\{X_i\}_{i=1}^n$ 是 X 的一组独立同分布样本,首先,考虑其参数估计。假设随机变量 X 服从某个已知的参数分布 $f(x; \theta)$, 其中参数 θ 未知需要估计。本文通过极大似然估计其参数 θ , 即

$$\hat{\theta} = \arg \min_{\theta} \prod_{i=1}^n f(X_i; \theta)$$

从而得到其概率密度函数估计 $f_p(x) = f(x; \hat{\theta})$; 然后,再考虑其非参估计。本文使用常用的核密度估计作为非参估计部分:

$$f_{np}(x) = \frac{1}{nh} K\left(\frac{X_i - x}{h}\right)$$

其中, $K(\cdot)$ 是核函数, h 是带宽。为了防止非参估计部分的“过拟合”现象,本文对其增加一个正则惩罚,提出了两种估计权重的损失函数,一个是应用经验似然损失函数,即

$$L_{MLEP}(w) = -\frac{1}{n} \sum_{i=1}^n \log(wf_p(X_i) + (1-w)f_{np}(X_i)) + \lambda_n(1-w)d_f$$

其中, d_f 是非参估计的自由度, $\lambda_n > 0$ 是正则参数。对于核密度估计来说,根据文献[15]可知 $d_f = K(0)/h$ 。另一个损失函数是根据积分平方误差(Integrated Squared Error, ISE)推导的:

$$L_{\text{ISE}}(f(x)) = \int_{\mathcal{X}} (f(x) - f_0(x))^2 dx = \int_{\mathcal{X}} f^2(x) dx - 2 \int_{\mathcal{X}} f(x)f_0(x) dx + \int_{\mathcal{X}} f_0^2(x) dx$$

其中, $\mathcal{X} \subset \mathbf{R}$ 。注意到右边第三项是常数, 对优化问题无关, 所以只考虑用前两项的值作为损失, 而第二项实际上是求 $E[f(X)]$, 因此可以用本均值 $\sum_{i=1}^n f(X_i)/n$ 来近似它。所以, 得到了本文的第二种损失函数:

$$L_{\text{ISEP}}(w) = \int_{\mathcal{X}} [wf_p(x) + (1-w)f_{np}(x)]^2 dx - \frac{2}{n} \sum_{i=1}^n [wf_p(X_i) + (1-w)f_{np}(X_i)] + \lambda_n(1-w)d_f$$

极小化 $L_{\text{MLEP}}(w)$ 或者 $L_{\text{ISEP}}(w)$ 得到权重的估计 $\hat{w}$ 。代入半参数估计的形式得到本文所提出的估计:

$$\hat{f}(x) = \hat{w}f_p(x) + (1-\hat{w})f_{np}(x)$$

3 理论性质

3.1 符号表示

$a_n \sim b_n$ 表示数列 a_n, b_n 等阶, 即 $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = c, 0 \neq c <$

∞ ; $a_n = o(b_n)$ 表示 $\lim_{n \rightarrow \infty} \frac{a_n}{b_n} = 0$; $a_n = O(b_n)$ 表示 $\lim_{n \rightarrow \infty} |a_n/b_n| \leq c$; $a_n = O_p(b_n)$ 表示 $\lim_{n \rightarrow \infty} P_r(|a_n/b_n| \leq c) \rightarrow 1$, 其中 $0 < c < \infty$; $\|\theta\|$ 表示向量 θ 的欧氏范数。此外, 本文只展现用积分平方误差损失估计的渐近性质, 对于经验似然损失的性质可参考文献[14]的定理及其类似的证明, 在此不再赘述。

3.2 假设条件

条件 1 $X \in \mathcal{X}$ 是紧集, $f_0(x)$ 在定义域 $\mathcal{X}$ 上有界且 $f''_0(x)$ 存在。

条件 2 非参估计使用的核函数 $K(\cdot)$ 关于原点对称的。

条件 3 如果模型假设成立, 即 $f_0(x) = f(x; \theta_0)$, 则 $\int_{\mathcal{X}} \nabla^T f(x; \theta_0) \nabla f(x; \theta_0) dx < \infty$, 其中 $\nabla f(x; \theta) = \frac{\partial f(x; \theta)}{\partial \theta} \Big|_{\theta=\theta_0}$ 。

条件 4 若对于任意的 $\theta \in \Theta, f_0(x) \neq f(x; \theta)$, 那么 $\inf_{\theta \in \Theta} \int_{\mathcal{X}} [f(x; \theta) - f_0(x)]^2 dx \geq \delta_n$, 其中, 当 $n \rightarrow \infty$ 时, $\delta_n \rightarrow 0, n^{\frac{3}{4}} \delta_n \rightarrow \infty$ 。

注 1 条件 1 和条件 2 在非参数统计研究中是常见的, 例如文献[12-13]。条件 3 说明了参数模型一阶

导的可积性, 许多常见的分布都满足这个假设条件。条件 4 是为了限制参数估计在参数模型假设不成立时, 即使它是收敛的, 但其收敛速度也不能太快高于 $n^{-3/4}$, 这个条件是很自然的。

3.3 渐近性质

定理 1 若条件 1 至条件 3 成立, 则当 $\lambda_n \sim n^{-1}, h = h_{opt}$ 时, 若 $\mathcal{X} = \mathbf{R}, h_{opt} \propto n^{-1/5}$; 若 $\mathcal{X} = [a, b], h_{opt} \propto n^{-1/4}$, $\lim_{n \rightarrow \infty} P_r(\hat{w} = 1) = 1$, 此时 $L_{\text{ISE}}(\hat{f}(x)) = O_p(n^{-1})$ 。

证明 首先考虑参数估计的收敛速度。

$$L_{\text{ISE}}(f_p(x)) = \int_{\mathcal{X}} [f(x; \hat{\theta}) - f(x; \theta_0)]^2 dx = \int_{\mathcal{X}} [\nabla^T f(x; \theta_0)(\hat{\theta} - \theta_0) + o(\|\hat{\theta} - \theta_0\|^2)]^2 dx = (\hat{\theta} - \theta_0)^T \left[\int_{\mathcal{X}} \nabla^T f(x; \theta_0) \nabla f(x; \theta_0) dx \right] (\hat{\theta} - \theta_0) + o(\|\hat{\theta} - \theta_0\|^2) \leq C \|\hat{\theta} - \theta_0\|^2 = O_p(n^{-1})$$

其中, $C > 0$ 是一个常数。第二个等式是根据 Taylor 展开得到的, 不等式是根据条件 3 得到的, 最后一个等式是根据参数估计的渐近正态性得到的, 更详细的内容可见文献[1]。

另一方面, 考虑核密度估计的收敛速度。由于对于定义域有界的核密度估计存在边界效应 (Boundary Correction), 具体可见文献[16], 所以, 本文分两种情况讨论核密度估计的收敛速度。

当 $\mathcal{X} = \mathbf{R}$ 时, $K(\cdot)$ 在 $\mathbf{R}$ 上对称, 此时对于任意的 $x \in \mathcal{X}$, 有

$$E[f_{np}(x)] = E\left[\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)\right] = E\left[\frac{1}{h} K\left(\frac{X - x}{h}\right)\right] = \frac{1}{h} \int_{\mathbf{R}} K\left(\frac{u - x}{h}\right) f_0(u) du = \int_{\mathbf{R}} K(v) f_0(x + vh) dv = \int_{\mathbf{R}} K(v) [f_0(x) + f'_0(x)vh] dv + \int_{\mathbf{R}} K(v) [f''_0(x)(vh)^2/2 + o(h^2)] dv = f_0(x) + \frac{f''_0(x)\kappa_2}{2}h^2 + o(h^2)$$

其中, $\kappa_l = \int_{\mathbf{R}} u^l K(u) du$, 并且不失一般性地假设 $\kappa_0 = 1$ 。

$$E\left[K^2\left(\frac{X - x}{h}\right)\right] = \int_{\mathbf{R}} K^2\left(\frac{u - x}{h}\right) f_0(u) du = h \int_{\mathbf{R}} K^2(v) f_0(x + vh) dv = v_0 h f_0(x) + o(h^2)$$

其中, $v_l = \int_{\mathbf{R}} u^l K^2(u) du$ 。所以

$$V_{ar}[f_{np}(x)] = V_{ar}\left[\frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)\right] =$$

$$\begin{aligned} \frac{1}{nh^2} V_{ar} \left[K \left(\frac{X-x}{h} \right) \right] &= \\ \frac{1}{nh^2} \left\{ E \left[K^2 \left(\frac{X-x}{h} \right) \right] - E^2 \left[K \left(\frac{X-x}{h} \right) \right] \right\} &= \\ \frac{1}{nh^2} \{ v_0 h f_0(x) + o(h^2) - h^2 E^2[f_{np}(x)] \} &= \\ \frac{f_0(x) v_0}{nh} + o\left(\frac{1}{nh}\right) \end{aligned}$$

因此

$$\begin{aligned} E[L_{ISE}(f_{np}(x))] &= E \left[\int_{\mathbf{R}} (f_{np}(x) - f_0(x))^2 dx \right] = \\ \int_{\mathbf{R}} E[f_{np}(x) - f_0(x)]^2 dx &= \\ \int_{\mathbf{R}} V_{ar}[f_{np}(x)] + E^2[f_{np}(x) - f_0(x)] dx &= \\ \int_{\mathbf{R}} \left\{ \frac{f_0(x) v_0}{nh} + \left[\frac{f''_0(x) \kappa_2}{2} h^2 \right]^2 \right\} dx &+ o\left(\frac{1}{nh} + h^4\right) = \\ \frac{v_0}{nh} + \frac{\kappa_2^2 \left\{ \int_{\mathbf{R}} [f''_0(x)]^2 dx \right\}}{4} h^4 &+ o\left(\frac{1}{nh} + h^4\right) \end{aligned}$$

故当 $h = h_{opt} = \kappa_2^{-2/5} v_0^{1/5} \{ \int_{\mathbf{R}} [f''_0(x)]^2 dx \}^{-1/5} n^{-1/5}$ 时, $L_{ISE}(f_{np}(x)) = O_p(n^{-4/5})$ 。

当 $\mathcal{B} \in [a, b]$ 时, 设 $K(\cdot)$ 在 $[-1, 1]$ 上对称。此时, 对于定义域内部的点 $x \in [a+h, b-h]$, $f_{np}(x)$ 的表现与前面一样。而对于边界点 $x \in [a, a+h]$, 不失一般性, 令 $x = a+rh$, $r \in [0, 1]$, 有

$$\begin{aligned} E[f_{np}(x)] &= \frac{1}{h} E \left[K \left(\frac{X-x}{h} \right) \right] = \\ \frac{1}{h} \int_a^b K \left(\frac{u-x}{h} \right) f_0(u) du &= \\ \int_{-r}^1 K(v) [f_0(x) + f'_0(x)hv] dv &+ \\ \int_{-r}^1 K(v) [f''_0(x)(hv)^2/2 + o(h^2)] dv &= \\ (1 - \kappa_0(r))f_0(x) + h\kappa_1(r)f'_0(x) &+ \\ h^2(\kappa_2 - \kappa_2(r))f''_0(x)/2 + o(h^2) \end{aligned}$$

其中, $\kappa_l(r) = \int_r^1 u^l K(u) du$ 。

$$\begin{aligned} V_{ar}[f_{np}(x)] &= \\ \frac{1}{nh^2} \left\{ \int_a^b K^2 \left(\frac{u-x}{h} \right) f_0(u) du \right\} - \frac{E^2[f_{np}(x)]}{n} &= \\ \frac{1}{nh} \left\{ \int_{-r}^1 K^2(v) [f_0(x) + o(1)] du \right\} - & \\ \frac{[1 - \kappa_0(r)]f''_0(x) + o(h)}{n} &= \end{aligned}$$

$$\frac{(v_0 - v_0(r))f_0(x)}{nh} + o\left(\frac{1}{nh}\right)$$

其中, $v_l(r) = \int_r^1 u^l K^2(u) du$ 。此时, 由于

$$\begin{aligned} E[f_{np}(x) - f_0(x)]^2 &= \\ V_{ar}[f_{np}(x)] + E^2[f_{np}(x) - f_0(x)] &= \\ [\kappa_0(r)]^2 f_0^2(x) + O\left(\frac{1}{nh} + h\right) &\rightarrow 0 \end{aligned}$$

此时核密度估计不是相合的。为了保证估计的相合性, 可以考虑对非参估计进行修正: $\tilde{f}_{np}(x) = f_{np}(x)/1 - \kappa_0(r)$ 。对于 $x \in (b-h, h]$, 令 $x = b-rh$, $r \in [0, 1]$, 可以得到类似的结果:

$$\begin{aligned} E[f_{np}(x)] &= (1 - \kappa_0(r))f_0(x) - h\kappa_1(r)f'_0(x) + \\ h^2(\kappa_2 - \kappa_2(r))f''_0(x)/2 + o(h^2) & \\ V_{ar}[f_{np}(x)] &= \frac{(v_0 - v_0(r))f_0(x)}{nh} + o\left(\frac{1}{nh}\right) \end{aligned}$$

结果显示: 对于边界上的核密度估计有一个系数为 $1 - \kappa_0(r)$ 的偏差, 因此可以对此系数进行修正。为了统一整个区间的形式, 当 $r \geq 1$ 时, 令 $\kappa_0(r) = 0$, 则在 $x \in [a, b]$ 上的核密度估计修正为

$$\tilde{f}_{np}(x) = \frac{f_{np}(x)}{1 - \kappa_0(r)}$$

此时 $\tilde{f}_{np}(x)$ 是相合的, 有

$$\begin{aligned} E[L_{ISE}(\tilde{f}_{np}(x))] &= \int_a^b E[\tilde{f}_{np}(x) - f_0(x)]^2 dx = \\ \int_a^b \left\{ \frac{V_{ar}[\tilde{f}_{np}(x)]}{[1 - \kappa_0(r)]^2} + E^2[\tilde{f}_{np}(x) - f_0(x)] \right\} dx &= \\ \frac{\left\{ \int_a^{a+h} + \int_{a+h}^{b-h} + \int_{b-h}^b \right\} V_{ar}[\tilde{f}_{np}(x)] dx}{[1 - \kappa_0(r)]^2} &+ \\ \left\{ \int_a^{a+h} + \int_{a+h}^{b-h} + \int_{b-h}^b \right\} E^2[\tilde{f}_{np}(x) - f_0(x)] dx &= \\ \frac{[v_0 - v_0(r)]}{nh[1 - \kappa_0(r)]^2} + \frac{h^2 \kappa_1^2(r) C_1(h)}{[1 - \kappa_0(r)]^2} &+ \\ \frac{h^3 \kappa_1(r) [\kappa_2 - \kappa_2(r)] C_2(h)}{[1 - \kappa_0(r)]^2} + o\left(\frac{1}{nh} + h^3\right) \end{aligned}$$

其中, 对于 $r \geq 1$, $v_0(r) = 0$, 有

$$\begin{aligned} C_1(h) &= \left\{ \int_a^{a+h} + \int_{b-h}^b \right\} [f'_0(x)]^2 dx \\ C_2(h) &= \left\{ \int_a^{a+h} - \int_{b-h}^b \right\} f'_0(x) f''_0(x) dx \end{aligned}$$

根据条件 1, 不失一般性地可设 $C_1(h) = O(h)$,

$C_2(h) = O(h)$ 。所以, 当 $h = h_{opt} \propto n^{-1/4}$ 时, $L_{ISE}(\tilde{f}_{np}(x)) =$

$O_p(n^{-3/4})$ 。又由于

$$L_{\text{ISEP}}(0) = L_{\text{ISE}}(f_p(x)) + \lambda_n d_f - C_{f_0}$$

$$L_{\text{ISEP}}(1) = L_{\text{ISE}}(f_p(x)) - C_{f_0}$$

其中, $C_{f_0} = \int_{\mathcal{R}} [f_0(x)]^2 dx$ 。所以, 当参数模型假设成立, 且正则惩罚参数满足 $\lambda_n \sim n^{-1}$ 时, $\lim_{n \rightarrow \infty} P_r \{L_{\text{ISEP}}(1) < L_{\text{ISEP}}(0)\} = 1$, 也就是 $\lim_{n \rightarrow \infty} P_r(\hat{w} = 1) = 1$, $L_{\text{ISE}}(\hat{f}(x)) = L_{\text{ISE}}(f_p(x)) = O_p(n^{-1})$ 。

证毕。

定理 2 如果条件 1、条件 2 以及条件 4 成立, 当 $\lambda_n = O(\delta_n n^{-1/4})$, $h = h_{\text{opt}}$ 时, 若 $\mathcal{R} = \mathbf{R}$, $h_{\text{opt}} \propto n^{-1/5}$; 若 $\mathcal{R} = [a, b]$, $h_{\text{opt}} \propto n^{-1/4}$; $\lim_{n \rightarrow \infty} P_r(\hat{w} = 0) = 1$, 并且此时

$$L_{\text{ISE}}(\hat{f}(x)) = \begin{cases} O_p(n^{-4/5}) & \mathcal{R} = \mathbf{R} \\ O_p(n^{-3/4}) & \mathcal{R} = [a, b] \end{cases}$$

证明 根据定理 1 的证明过程可知: 当 $\mathcal{R} = \mathbf{R}$ 时, $G_{\text{ISEP}}(0) = O_p(n^{-4/5}) + \lambda_n d_f - C_{f_0}$; 当 $\mathcal{R} = [a, b]$, $L_{\text{ISEP}}(0) = O_p(n^{-3/4}) + \lambda_n d_f - C_{f_0}$ 。又根据条件 4 可得 $L_{\text{ISEP}}(1) = L_{\text{ISE}}(f_p(x)) \geq \delta_n$, 其中 $n^{3/4} \delta_n \rightarrow \infty$ 。故当 $\lambda_n = O(\delta_n n^{-1/4})$ 时, $\lim_{n \rightarrow \infty} P_r(\hat{w} = 0) = 1$ 。

证毕。

4 数值模拟实验

4.1 正则参数 λ_n 以及带宽 h 的选择

在实际应用时, 涉及两个参数的选择, 即正则参数 λ_n 以及带宽 h 。正则参数可以按照传统的交叉验证法结合格点搜索进行选择, 但在本文的数值模拟和实例分析中为了计算简便直接设定 $\lambda_n = \ln n/n$ 。至于带宽 h , 本文采用 Sheather-Jones 方法来选择, 见文献[17]。根据有限的数值例子表明: 这样选择的正则参数和带宽所得到的估计具有不错的表现。

4.2 数值实验

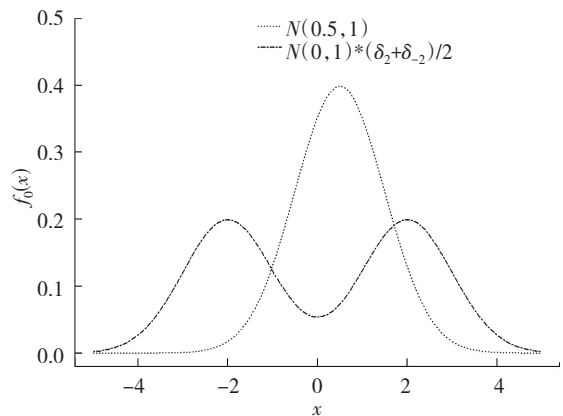
为了验证该方法的估计效果, 本文用 R 软件设计了 4 个例子进行比较, 包括核密度估计(KDE)、参数估计(MLE)、Olkin 和 Clifford 的不带惩罚的半参数似然方法(MLE.sp)以及文章中所提出的带惩罚的半参数估计方法(MLEP.sp 和 ISEP.sp)。关于参数估计部分, 为了计算简单, 本小节只将正态分布、t-分布以及指数分布 3 种常见的分布作为备选的参数分布假设, 非参数估计直接调用 R 软件中自带的函数 `density()` 进行估计, 设置参数 `bw = "SJ"` 即可调用 Sheather-Jones 选择带宽的方法。此外, 为了比较不同密度估计方法的性能, 使用了两种损失函数: 积分平方误差(ISE)和

Hellinger 距离(HD)来比较估计的效果。

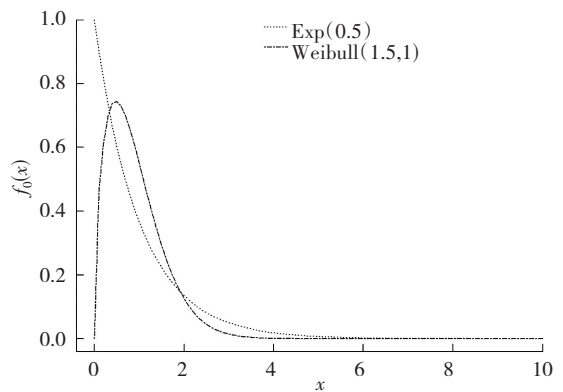
$$G_{\text{ISE}}(f(x), f_0(x)) = \int_{\mathcal{R}} \{f(x) - f_0(x)\}^2 dx$$

$$G_{\text{HD}}(f(x), f_0(x)) = \int_{\mathcal{R}} \{\sqrt{f(x)} - \sqrt{f_0(x)}\}^2 dx$$

考虑以下 4 种分布: (1) 正态分布 $N(0.5, 1)$; (2) 指数分布 $\text{Exp}(1)$; (3) 混合正态分布 $N(0, 1) * (\delta_{-2} + \delta_2)/2$; (4) 韦布尔分布 $\text{Weibull}(1, 1.5)$, 其中 δ_a 是在 a 点退化的分布函数, $*$ 表示卷积。可以看出, 前两种分布在参数模型的假设中, 另外两种不属于参数模型假设, 其中分布(3)也同样出现在文献[18]中。首先, 绘制了这 4 个分布的概率密度函数曲线, 如图 1 所示。



(a) 正态分布与混合正态分布



(b) 指数分布与韦布尔分布

图 1 概率密度函数曲线

Fig. 1 The curves of the probability density function

对于每种分布, 考虑样本量 $n = 100, n = 200$ 两种情形, 每种重复计算 500 次, 取其均值(G_{MISE} 和 G_{MHD})和标准差展现。

对于前两种分布, 其结果如表 1 和表 2 所示。它们均满足参数模型假设。从表中可以看出: 由于假设的成立使得参数估计具有最好的表现, 它们的 G_{MISE} 和 G_{MHD} 在所有情况下都是最小的, 非参估计的误差较大; 然而, 此时不带惩罚的 MLE.sp 选择的权重平均值都较

小,尤其对于正态分布来说,这与理论预期相矛盾;此时,本文所提出的带惩罚的 MLEP. sp 和 ISEP. sp 的预测误差虽然不是最小的,但是选择的权重是按照理论预期的设想靠近 1 的。

表 1 正态分布的数据结果
Table 1 Results of normal distributed data

Method	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$
	$n = 100$			$n = 200$		
KDE	7.106(5.755)	13.721(8.598)	—	3.945(2.927)	7.769(3.948)	—
MLE	2.527(2.518)	5.100(4.975)	—	1.183(1.117)	2.397(2.212)	—
MLE. sp	6.314(5.750)	12.158(8.512)	0.171(0.238)	3.577(2.926)	7.026(4.030)	0.112(0.171)
MLEP. sp	2.533(2.511)	5.133(4.973)	0.992(0.031)	1.183(1.117)	2.404(2.210)	0.995(0.021)
ISEP. sp	2.537(2.521)	5.114(4.975)	0.999(0.183)	1.189(1.122)	2.401(2.212)	0.999(0.015)

表 2 指数分布的数据结果
Table 2 Results of exponential distribution data

Method	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$
	$n = 100$			$n = 200$		
KDE	54.116(12.769)	99.574(13.734)	—	40.404(8.198)	76.153(8.464)	—
MLE	2.345(3.467)	2.311(3.316)	—	1.153(1.649)	1.140(1.598)	—
MLE. sp	16.303(16.691)	36.272(25.485)	0.567(0.287)	8.451(1.149)	22.111(16.403)	0.648(0.243)
MLEP. sp	2.365(3.477)	2.617(3.580)	0.995(0.018)	1.163(1.655)	1.286(1.734)	0.997(0.011)
ISEP. sp	2.404(0.506)	2.564(3.659)	0.995(0.032)	1.153(1.633)	1.238(1.806)	0.998(0.017)

对于后面两种分布,其结果如表 3 和表 4 所示。它们都倾向于选择更小的权重,这和定理 2 的预想一样。此外,当分布服从韦布尔分布时,某段区间内的概率密度函数与参数模型重合,故增加了权重的半参方法在样本量较大时的估计效果比非参估计要好。

表 3 混合正态分布的数据结果
Table 3 Results of mix-normal distribution data

Method	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$
	$n = 100$			$n = 200$		
KDE	0.617(0.345)	1.812(0.917)	—	0.350(0.177)	1.021(0.455)	—
MLE	3.741(0.282)	9.995(0.418)	—	3.693 0(0.190)	9.803(0.212)	—
MLE. sp	0.617(0.345)	1.812(0.917)	0.000(0.000)	0.350(0.177)	1.021(0.455)	0.000(0.000)
MLEP. sp	0.618(0.346)	1.812(0.917)	0.000(0.003)	0.352(0.178)	1.026(0.460)	0.003(0.011)
ISEP. sp	0.617(0.366)	1.840(0.972)	0.007(0.027)	0.350(0.177)	1.021(0.456)	0.000(0.000)

表 4 韦布尔分布的数据结果
Table 4 Results of Weibull distribution data

Method	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$	$G_{MISE}(\times 10^{-3})$	$G_{MHD}(\times 10^{-3})$	$\hat{w}$
	$n = 100$			$n = 200$		
KDE	3.992(2.026)	4.818(1.126)	—	2.530(1.160)	3.530(0.702)	—
MLE	17.580(6.064)	8.143(1.559)	—	17.471(5.380)	7.774(1.463)	—
MLE. sp	4.028(2.112)	4.669(1.005)	0.035(0.082)	2.514(1.147)	3.451(0.633)	0.021(0.048)
MLEP. sp	5.977(3.827)	4.944(1.372)	0.321(0.218)	2.828(1.540)	3.348(0.710)	0.114(0.126)

5 实际案例与结果分析

5.1 数据集

这一节将前面提到的 4 个估计方法应用到重庆市降水数据中。原始数据集从 <https://www.ncei.noaa.gov/> 上下载。该数据集记录了重庆市从 1951 年 1 月 1 日至 2023 年 8 月 3 日约 26 509 天的日平均气温、最高气温、最低气温、降水量、降雪量 5 个指标。本文主要研究其降水量的分布。由于每日降水量中许多值为零,且按照每天为单位进行考虑样本量太大,所以先将其转化为每月的降水量,总计约 871 个月的样本,绘制其时间序列曲线,如图 2 所示。

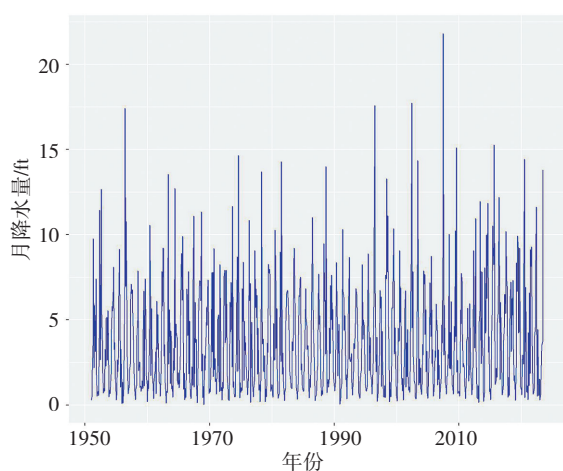


图 2 重庆市 1951—2023 年月降水量

Fig. 2 Monthly precipitation in Chongqing (1951–2023)

首先,对数据进行简单的统计分析,得出它们基本的统计特征,如表 5 所示。

表 5 重庆市 1951-01—2023-08 降水量数据统计描述

Table 5 Statistical description of precipitation data in Chongqing from January 1951 to August 2023

最小值	最大值	平均值	1/4 分位数	中位数	3/4 分位数
0.030	21.790	3.688	1.160	2.850	5.515

5.2 数据分析及其结果

为了对其密度函数进行估计,首先绘制其直方图,如图 3 所示。

从图 3 中可以看出:降水量服从具有偏态、厚尾以及在 $[0, +\infty)$ 上取值等特征的分布。因此,关于备选的参数估计,本节不像第 4 节数值模拟实验中那样考虑在整个实数域上取值的正态分布和 t 分布,而是考虑只在非负区间上取值的指数分布、Gamma 分布和韦布尔分布这 3 种作为参数估计待选的分布,至于非参数估计,仍然使用核密度估计。

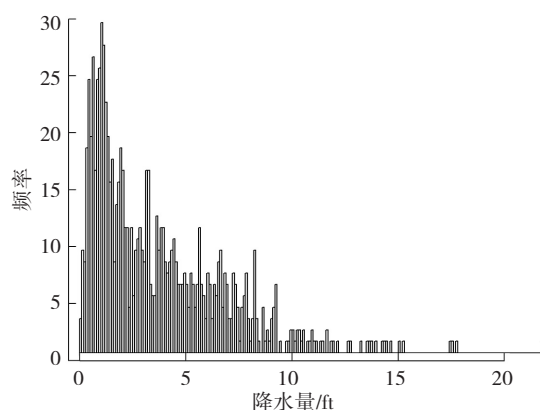


图 3 降水量直方图

Fig. 3 Histogram of precipitation

通过 R 软件进行计算,在 3 个备选的参数模型中,参数估计选择 Gamma 分布模型,应用前面介绍的 3 个半参数加权方法各自估计的权重如表 6 所示。

表 6 半参数估计所得到的权重

Table 6 Weights estimated by the semiparametric estimation

Method	MLE. sp	MLEP. sp	ISEP. sp
$\hat{w}$	0.000	0.296	0.743

表 6 可以看出:Olkin 和 Clifford 的不带惩罚的半参数方法仍然和数值模拟结果一样,选择了一个基本为 0 的权重,即基本上选择了非参数估计;而本文提出的两个基于不同损失的增加了惩罚的半参数估计都选择了大一点的权重,其中,基于经验似然损失的半参数估计所得到的权重约为 0.296,基于积分平方误差损失的半参数估计所得到的权重约为 0.743,远大于经验似然损失得到的权重。

为了直观地比较各个方法所得到的估计拟合的效果,本文将各个方法所估计的曲线与直方图放一起,如图 4 所示。

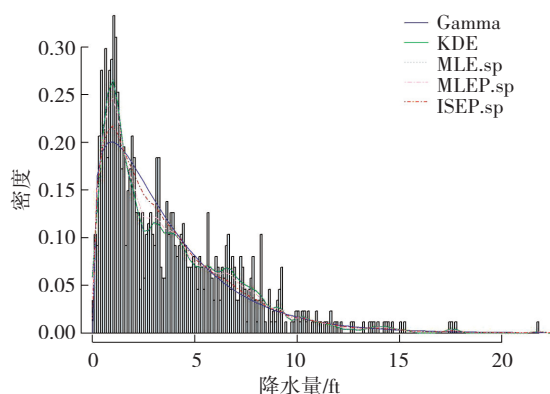


图 4 概率密度函数拟合曲线

Fig. 4 Fitting curve of the probability density function

从图4可以看出:参数估计(蓝色的Gamma分布曲线)比较光滑,和直方图比较一致,也和文献[19-20]提出的最具有代表性的Gamma分布函数的结果一致。但对于数据拟合来说,核密度估计(绿色实线)拟合得最好,其形状走势最接近直方图的走势,但是它最不光滑。不带惩罚的半参数估计(紫色虚线)基本与核密度估计重合,而文章所提出的带惩罚的半参估计(粉色和红色虚线)与非参估计相比更光滑,与参数估计相比拟合效果更好。其中,基于积分平方损失的半参数估计方法所得到的概率密度函数曲线最为光滑,且接近直方图的走势。

6 结 论

数值实验结果显示:所提出的半参数方法结合了参数和非参方法的优点,使得概率密度函数估计在参数模型假设成立时收敛速度与参数估计一致,比非参数核密度估计有效,同时在参数模型假设不成立时仍然收敛,速度与非参数核密度估计一致。进一步扩展应用该方法,可以考虑从以下方面进行优化:

参数估计部分只考虑了正态分布、t分布和指数分布等3个分布,对于不同的数据存在不同的分布,比如在第5节应用该方法到降水数据时,结合降水数据分布的特点,考虑了Gamma分布;工程数据中如果经常出现韦布尔分布,可以考虑在参数模型假设中加入此分布。

非参估计部分只考虑了核密度估计,可以考虑应用其他改进的非参估计方法,比如文献[21]的局部参数化核密度估计、文献[22]的局部似然估计方法、文献[23-24]的自适应核密度估计方法等。

文章中的权重是一个取值为0到1之间的常数。然而,实际中所要估计的概率密度函数可能会出现某一段或者某些段与参数分布重合的情况。因此,可以考虑寻找一个可变的权重来进一步改进估计效果。

文章只考虑了一元连续性随机变量的密度函数估计问题,实际中多元连续型甚至多元混合型随机变量并不少见。近年来,关于核密度估计应用到此类复杂数据中的方法已有不少,例如文献[25]考虑了连续与离散混合的多元核密度估计,后续可以考虑将这些方法应用到非参数估计部分,从而使本文的半参数方法扩展到多元复杂的数据中。

参考文献(References):

- [1] 茆诗松,王静龙,濮晓龙. 高等数理统计[M]. 2版. 北京: 高等教育出版社, 2006.
MAO Shi-song, WANG Jing-long, PU Xiao-long. Advanced mathematical statistics[M]. 2nd ed. Beijing: Higher Education Press, 2006.
- [2] SILVERMAN B W. Kernel density estimation using the fast Fourier transform[J]. Journal of the Royal Statistical Society Series C (Applied Statistics), 1982, 31(1): 93-99.
- [3] 徐询, 谢丽蓉, 叶林, 等. 基于非参数核密度估计的风电场有功功率双层优化模型[J]. 电力系统自动化, 2022, 46(2): 43-55.
XU Xun, XIE Li-rong, YE Lin, et al. Bi-level optimization model of active power for wind farm based on nonparametric kernel density estimation[J]. Automation of Electric Power Systems, 2022, 46(2): 43-55.
- [4] 石慧, 康辉, 赵李志, 等. 基于Copula理论的多部件系统剩余寿命核密度预测方法[J]. 计算机集成制造系统, 2023, 29(1): 212-223.
SHI Hui, KANG Hui, ZHAO Li-zhi, et al. Remaining useful life kernel density prediction method for multi-component system based on Copula theory[J]. Computer Integrated Manufacturing Systems, 2023, 29(1): 212-223.
- [5] 张均胜, 孙晓平, 刘志辉. 自动迭代聚类数据集训练的虚假信息识别方法[J]. 情报学报, 2023, 42(1): 59-73.
ZHANG Jun-sheng, SUN Xiao-ping, LIU Zhi-hui. Misinformation identification method by automatic iterative clustering data set for training[J]. Journal of the China Society for Scientific and Technical Information, 2023, 42(1): 59-73.
- [6] 贺小洋, 肖绯雄, 杨强, 等. 基于核密度估计的构架应力谱外推与疲劳寿命分析[J]. 起重运输机械, 2023(1): 51-56.
HE Xiao-yang, XIAO Fei-xiong, YANG Qiang, et al. Stress spectrum extrapolation and fatigue life analysis of frame based on kernel density estimation[J]. Hoisting and Conveying Machinery, 2023(1): 51-56.
- [7] 李淑艳, 翟友邦, 王小龙, 等. 基于核密度估计的拖拉机传动轴载荷外推方法[J]. 中国农业大学学报, 2021, 26(10): 175-184.
LI Shu-yan, ZHAI You-bang, WANG Xiao-long, et al. An

- extrapolation method of tractor propeller shaft load based on kernel density estimation [J]. Journal of China Agricultural University, 2021, 26(10): 175-184.
- [8] 宋晓东, 颜永逸, 秦顺全, 等. 基于自适应核密度估计的桥梁短时风速预测方法[J]. 桥梁建设, 2020, 50(5): 56-61.
SONG Xiao-dong, YAN Yong-yi, QIN Shun-quan, et al. Short-term wind speed prediction method for bridges based on adaptive kernel density estimation [J]. Bridge Construction, 2020, 50(5): 56-61.
- [9] 温婷, 肖文彬, 亓思楠. 中国数字经济生产效率测度及其时空动态演进[J]. 统计与决策, 2022, 38(23): 5-10.
WEN Ting, XIAO Wen-bin, QI Si-nan. Measurement and spatial-temporal dynamic evolution of production efficiency of China's digital economy [J]. Statistics & Decision, 2022, 38(23): 5-10.
- [10] 马馨悦. 基于核密度估计的互联网金融产品收益率对比分析[J]. 重庆工商大学学报(自然科学版), 2018, 35(4): 37-43.
MA Xin-yue. Comparison and analysis of the yield of Internet financial products based on the kernel density estimation [J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2018, 35(4): 37-43.
- [11] IZENMAN A J. Review papers: Recent developments in nonparametric density estimation [J]. Journal of the American Statistical Association, 1991, 86(413): 205-224.
- [12] WAND M P, JONES M C. Kernel smoothing [M]. London: Chapman & Hall, 1995.
- [13] FAN J, GIJBELS I. Local polynomial modelling and its applications [M]. Boston: MA Springer US, 1994.
- [14] OLKIN I, SPIEGELMAN C H. A semiparametric approach to density estimation [J]. Journal of the American Statistical Association, 1987, 82(399): 858-865.
- [15] LOADER C R. Bandwidth selection: classical or plug-in? [J]. The Annals of Statistics, 1999, 27(2): 380-389.
- [16] JONES M C. Simple boundary correction for kernel density estimation [J]. Statistics and Computing, 1993, 3(3): 135-146.
- [17] SHEATHER S J, JONES M C. A reliable data-based bandwidth selection method for kernel density estimation [J]. Journal of the Royal Statistical Society Series B (Statistical Methodology), 1991, 53(3): 683-690.
- [18] WANG Y, CHEE C S. Density estimation using non-parametric and semi-parametric mixtures [J]. Statistical Modelling, 2012, 12(1): 67-92.
- [19] 田付友, 郑永光, 毛冬艳, 等. 基于 Γ 函数的暖季小时降水概率分布 [J]. 气象, 2014, 40(7): 787-795.
TIAN Fu-you, ZHENG Yong-guang, MAO Dong-yan, et al. Study on probability distribution of warm season hourly rainfall with Γ Distribution [J]. Meteorological Monthly, 2014, 40(7): 787-795.
- [20] 顾学志, 叶磊, 赵钢铁, 等. 中国日降水量的概率分布 [J]. 水利学报, 2021, 52(10): 1248-1262.
GU Xue-zhi, YE Lei, ZHAO Tong-tie-gang, et al. Probability distribution of daily precipitation in China [J]. Journal of Hydraulic Engineering, 2021, 52(10): 1248-1262.
- [21] HJORT N L, JONES M C. Locally parametric nonparametric density estimation [J]. The Annals of Statistics, 1996, 24(4): 475-486.
- [22] LOADER C R. Local likelihood density estimation [J]. The Annals of Statistics, 1996, 24(4): 236-247.
- [23] VAN KERM P. Adaptive kernel density estimation [J]. The Stata Journal: Promoting Communications on Statistics and Stata, 2003, 3(2): 148-156.
- [24] BOTEV Z I, GROTOWSKI J F, KROESE D P. Kernel density estimation via diffusion [J]. The Annals of Statistics, 2010, 38(5): 567-579.
- [25] MCCLOUD N, PARMETER C F. Determining the number of effective parameters in kernel density estimation [J]. Computational Statistics & Data Analysis, 2020, 143: 106843-106853.

责任编辑:李翠薇