Statistics and Application 统计学与应用, 2026, 15(6), 142-147
Published Online June 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.156139

Hans汉斯

高维复杂数据下自适应非参数统计方法优化及应用探究

郭钦尧

金谷成(厦门)私募基金管理有限公司，福建 厦门

收稿日期：2026年5月25日；录用日期：2026年6月17日；发布日期：2026年6月29日

摘　要

本文针对高维复杂数据下传统非参数统计方法出现的维度灾难和异质性适配问题，提出一个自适应非参数统计方法的优化理论框架。该框架从局部带宽动态调整、特征空间自适应权重分配、多尺度自适应调整三个方面对传统方法进行优化。在理论层面，本文对所提估计量的渐近无偏性、方差收敛特性及Oracle不等式等渐近统计性质进行了分析，并探讨了该方法在弱相关数据、各向异性结构及低维内在结构等不同数据结构下的拓展性理论保证。该框架无需数据分布或函数光滑性等先验知识，即可自动适应高维数据的异质性特征并规避维度灾难，旨在为高维复杂数据下的非参数统计分析提供统一的理论基础。

关键词

高维复杂数据，自适应非参数估计，维度灾难，Oracle不等式，内在维度

Optimization and Application Exploration of Adaptive Nonparametric Statistical Methods for High-Dimensional Complex Data

Qinyao Guo

Jingucheng (Xiamen) Private Equity Fund Management Co., Ltd., Xiamen Fujian

Received: May 25, 2026; accepted: June 17, 2026; published: June 29, 2026

Abstract

This paper addresses the curse of dimensionality and heterogeneity adaptation issues encountered by traditional nonparametric statistical methods when applied to high-dimensional complex data.

文章引用: 郭钦尧. 高维复杂数据下自适应非参数统计方法优化及应用探究[J]. 统计学与应用, 2026, 15(6): 142-147.
DOI: 10.12677/sa.2026.156139

An optimized theoretical framework for adaptive nonparametric statistical methods is proposed, which improves upon conventional approaches from three perspectives: dynamic local bandwidth adjustment, adaptive weight allocation in the feature space, and multi-scale adaptive tuning. Theoretically, this paper analyzes the asymptotic statistical properties of the resulting estimators, including asymptotic unbiasedness, variance convergence characteristics, and oracle inequalities, and investigates the theoretical guarantees for the framework's extensibility under various data structures such as weakly dependent data, anisotropic structures, and low intrinsic dimensional structures. Without requiring prior knowledge such as data distribution or function smoothness, the framework can automatically adapt to the heterogeneity of high-dimensional data and circumvent the curse of dimensionality, aiming to provide a unified theoretical foundation for nonparametric statistical analysis of high-dimensional complex data.

Keywords

High-Dimensional Complex Data, Adaptive Nonparametric Estimation, Curse of Dimensionality, Oracle Inequality, Intrinsic Dimension

1. 引言

随着观测技术的发展，高维复杂数据在各个领域统计分析中越来越流行。传统的非参数统计方法处理这类数据的时候，会碰到维度灾难和异质性匹配的理论难题，收敛速度随着观测维度变大而迅速衰减，全局统一估计参数不能适应数据局部异质性[1]。自适应非参数统计方法依靠数据驱动的参数调节机制，给解决这一难题赋予了可行的理论途径，可以自动应对数据局部特性，避开维度灾难的影响。本文针对高维复杂数据下自适应非参数统计方法，建立完整的优化理论框架，从局部带宽调整、特征权重分配、多尺度适配三个方面优化方法，对优化后的估计量的渐近统计性质和方法在不同数据结构下进行拓展性的理论保证，给高维复杂数据下的非参数统计分析提供统一的理论基础。

2. 自适应非参数估计的基础理论框架

2.1. 非参数核估计的基本形式与维度困境

非参数核估计是 nonparametric 统计分析基础方法之一，核心是对邻域样本加权平均实现未知函数局部逼近。非参数回归问题中，经典 Nadaraya-Watson 核估计基本形式有特定表示，其中涉及对协变量边际密度的核密度估计 K(·)、核函数 h、带宽参数 d 及协变量观测维度。在独立同分布样本假设下，核函数与带宽满足正则条件时，估计量有良好渐近性质，偏差、方差有对应表达式，整体均方误差收敛速率为特定形式，β 该速率表达式体现维度灾难问题，即观测维度 d 增大时收敛速率迅速下降 β。而且，传统核估计依赖全局固定带宽参数，隐含回归函数光滑性和协变量分布均匀的强假设，但实际高维复杂数据场景中该假设几乎不成立，高维数据异质性强，不同区域函数光滑性和数据分布密度差异大，全局固定带宽无法适配不同区域特征，会导致估计性能大幅下降。

2.2. 自适应机制的核心逻辑与适配性基础

面对传统非参数方法的局限性，自适应机制的核心就是打破全局参数的限制，使估计器根据数据本

身的局部特征，自动调节自身的结构以及参数，从而达到在各个局部区域内最优偏差和方差之和的目的。与传统固定参数方法不同的是，自适应非参数方法不需要事先知道数据分布、函数的光滑性以及维度结构的先验信息，全部由数据本身决定，从样本中学习局部特征来调整估计参数。从上文可以看出，在函数光滑性较高的地方，选用更大的带宽，引入更多的邻域样本来降低方差并且不增加偏差；而当函数的光滑性较低，例如存在突变或者边缘的地方时，估计器则选择较小的带宽，缩减邻域范围，从而减小偏差，虽然会加大方差，但是总体均方误差依然能达到最优。高维特征空间中，自适应可以自动找到局部的有效维度结构来抛弃无关的特征维度，把局部的有效维度降到数据的内在维度上，从而避免出现维度灾难的情况。自适应机制本质上就是把全局统一估计变为局部自适应估计，使估计器根据实际情况改变自己的行为，以达到提高高维复杂数据估计性能的目的。它不需要先验信息的自适应能力能够匹配高维复杂的数据特征，在高维场景下往往无法事先知道相关的因素，而自适应方法可以凭借自身的数据学习过程来自行决定调整。

3. 高维场景下自适应非参数方法的优化机制

3.1. 局部自适应带宽的动态优化准则

局部自适应带宽选择是自适应非参数方法优化核心，目标是为每个估计 x 点选专属局部带宽 h(x)以适配局部特征。传统全局带宽选择法(如交叉验证)最小化全局预测误差，在异质数据下无法适配局部特征，而局部自适应带宽优化准则基于局部偏差与方差权衡，以数据驱动为各点选最优带宽。基于 Lepski 方法的动态优化准则是具理论保障的方法之一，其核心是选满足偏差约束的最粗带宽，即找最大带宽使不同带宽下估计量差异不超估计标准差水平，保证引入偏差可被方差覆盖，不影响估计精度。该准则由数据驱动，无需提前知函数光滑性就能自动为各点选最优带宽。在光滑区域，函数变化小，不同带宽下估计量差异小，可选大带宽利用邻域样本降方差；在不光滑区域，估计量差异大，只能选小带宽保证偏差不大。同时，该准则能适配数据分布密度差异，数据密集区样本足、标准差小，可选小带宽精准捕捉局部特征；数据稀疏区样本不足、标准差大，会选大带宽借用邻域样本降方差，避免估计失效。这种动态带宽调整让估计器适配函数与数据分布异质性，在高维复杂数据不同区域达最优性能。

3.2. 高维特征空间的自适应权重分配策略

在高维特征空间中，不同特征维度对未知函数影响差异显著，部分是有效维度，部分是冗余噪声维度，且不同维度函数光滑性也有差异，即各向异性结构。传统核估计在所有维度用相同带宽，假设所有维度重要性与光滑性一致，无法适配高维数据各向异性特征，导致估计性能下降。自适应权重分配策略为不同特征维度分配不同带宽参数，用对角带宽矩阵实现差异化调整。对于函数光滑性高或冗余的维度，估计器分配更大带宽，扩大邻域范围，忽略维度差异，降低估计方差；对于函数光滑性低或有效的维度，分配更小带宽，精准捕捉函数变化，保证估计偏差不大[2]。这种自适应权重分配是自动局部降维过程，估计器自动识别不重要维度，调大带宽消除其对估计的影响，将局部有效维度降至真正水平，不受高维度限制。该过程数据驱动，无需提前做变量选择或降维预处理[3]，估计器在估计中自动完成有效维度识别与冗余维度过滤，降低预处理复杂度，避免引入误差，使估计更稳健精准。

3.3. 异质光滑性下的多尺度自适应调整

除局部带宽与权重调整外，多尺度自适应调整是处理高度异质光滑性的重要优化方向，核心是多尺度分解，在不同分辨率上分离信号与噪声[4]，适配光滑与非光滑区域特征。基于小波变换的自适应估计方法，通过对小波系数阈值处理分离信号与噪声，小系数对应噪声被收缩过滤，大系数对应信号被保留，

能在光滑区域强平滑降噪、非光滑区域保留局部细节。该方法计算复杂度 O(n)低，可自动适配函数异质光滑性，同时处理光滑与突变区域特征，实现全局最优估计，适合高维大样本数据。

4. 优化后估计量的渐近统计性质

4.1. 估计量的渐近无偏性与偏差分解

优化后的自适应非参数估计量偏差结构与传统固定带宽估计量有本质差异，能根据局部特征自动调整偏差大小，实现全局偏差最优。传统固定带宽估计量偏差全局统一，光滑区域偏差远小于最优水平，不光滑区域则远大于最优水平，无法局部适配。而自适应估计量局部偏差根据局部光滑性与带宽动态调整，光滑区域带宽大，偏差稍大但因函数光滑仍处较低水平；不光滑区域带宽小，偏差被控制在很小水平，避免破坏估计精度。从偏差分解看，自适应估计量偏差分两部分：逼近偏差由局部带宽与函数光滑性决定，估计偏差由局部样本量与方差决定。自适应优化机制自动调整带宽使两部分偏差和最小，实现局部最优偏差水平。渐近情况下，样本量趋于无穷大时，局部带宽随样本量增加自动缩小，保证局部偏差趋于 0，估计量渐近无偏，此性质只需局部光滑性条件，对高维复杂数据至关重要，因其能适配异质结构，保证估计一致性。

4.2. 渐近方差的收敛特性与维度适配

自适应估计量的方差收敛特性是规避维度灾难的核心保障。与传统固定带宽估计量不同，其方差收敛速度不依赖观测维度，而依赖数据局部内在维度。传统固定带宽核估计方差观测维度影响，观测维度增大时方差迅速增大、收敛速度下降，这是维度灾难来源。自适应估计量方差收敛速度取决于局部内在维度 d_{int}，D 即便观测维度 D 大，只要局部内在维度 d_{int} 小，方差收敛速度就能保持低维数据水平，不受观测维度影响。这是因自适应权重分配机制会调整冗余维度带宽，降低有效维度至内在维度，使方差计算只考虑内在维度。渐近情况下，样本量增加使局部样本量增加，方差趋于 0，收敛速度只与内在维度有关，从根本上解决高维方差爆炸问题[5]。同时，自适应带宽调整优化方差水平，数据稀疏区域大带宽引入更多样本降低方差，数据密集区域小带宽虽减少样本量，但因样本量本身大，方差仍能保持低水平。这种维度适配的方差收敛特性，让自适应估计量在高维观测数据下保持快速收敛，规避维度灾难，适用于高维复杂数据。

4.3. 高维下的 Oracle 不等式与最优收敛速率

Oracle 不等式是衡量自适应估计量性能的核心理论工具，它描述了自适应估计量的风险，与拥有完全先验信息的 Oracle 估计量的风险之间的差距，从而证明自适应方法能够几乎达到理想的最优性能。Oracle 估计量是指那些提前知道真实函数的光滑性、维度结构等先验信息，从而能够选择最优参数的估计量，它是估计性能的理论上界，因为它拥有我们无法获得的先验信息。而自适应估计量，虽然没有这些先验信息，但是它的风险，却能够非常接近 Oracle 估计量的风险，两者之间的差距仅仅是一个对数项。具体来说，自适应估计量的风险，其中 R_{oracle} 是 Oracle 估计量的风险，这意味着自适应估计量的性能，最多比理想的 Oracle 估计量差一个对数因子，而在渐近的情况下，这个对数因子是可以忽略的，因此自适应估计量几乎能够达到 Oracle 的最优性能。例如在小波收缩的自适应估计中，即使没有任何先验信息，自适应估计量的风险也不会超过 Oracle 估计量的两倍对数项，这已经是理论上能够达到的最优的保证了，没有任何其他的估计量能够得到更好的保证。基于 Oracle 不等式，我们可以进一步得到自适应估计量的收敛速率，它能够在一个非常广泛的函数类中，比如 Besov 空间、Nikol’skii 空间，都达到 minimax 的收敛速率[6]，也就是不管函数属于哪个子函数类，不管函数的光滑性是多少，不管是各向同性还是各向异

性的结构，自适应估计量都能够自动的达到这个函数类下的最优收敛速率，这就是所谓的自适应 minimax 性质。这意味着自适应估计量无需提前获知函数光滑性、函数类或维度结构的先验信息，即可自动适应未知特征，达到对应场景下的最优收敛速率，完美匹配了高维场景下先验信息缺失的特点，保证了方法的通用性。

5. 高维复杂数据下方法的拓展性理论分析

5.1. 弱相关数据下的稳健性理论保障

自适应非参数估计量在弱相关数据下仍然具有较好的稳健性，可以保持原来的渐近性质。由于弱相关数据的相关性大多在短程内快速衰减，在局部邻域内样本间的相关性很弱，全局相关性也不会给局部估计带来很大影响。理论上可以证明，在弱相关样本中，估计量的偏差方差收敛速率、Oracle 不等式和自适应 minimax 性质都成立，只需要改变常数项，不会影响收敛速率，所以该方法可以适用于各种相关数据，不需要进行额外的预处理。

5.2. 各向异性结构下的自适应适配能力

高维复杂数据常常具备明显的各向异性结构，即不同特征维度上的函数变化情况存在着明显差别，有的维度上的函数变化快，光滑性低，有的维度上的函数变化慢，光滑性高，这种各向异性的结构，是传统的各向同性非参数方法所不能应对的，由于这些方法都假定所有的维度都有相同的光滑性，这样就会使得估计的性能大幅度降低。自适应非参数估计量可以自动地适应各个向量的异质性结构，不管各个向量的光滑性有多大差别，都可以自动地调节参数来达到最好的性能。因此自适应的带宽矩阵可以给每个维度分配不同的带宽，在光滑性低的维度上给它分配小的带宽来准确地识别函数的变化，控制偏差，在光滑性高的维度上给它分配大的带宽用更多的样本去降低方差。理论上证明，在各向异性的 Nikol’skii 空间或者 Besov 空间中，自适应估计量仍然可以达到 minmax 收敛速率，不管各个方向上的光滑性参数有多大差别，都能够自动适应，不需要提前知道这些差异的存在。由于各向异性的适配性，自适应估计量可以处理具有大量特征的高维数据，不论不同特征之间的差异有多大，都能自动调节，得到最好的估计。

5.3. 内在维度感知下的维度灾难规避机制

高维数据的一个核心特征是，虽然观测维度很高，但是数据本身往往分布在一个低维的流形上，也就是数据的内在维度远低于观测维度，比如图像数据、文本数据、基因数据等，都是典型的例子，它们的观测维度可能有上千甚至上万，但是内在维度却只有几个或者几十个[7]。传统的非参数方法，无法识别这种内在维度的结构，仍然按照观测维度来计算收敛速度，因此会受到维度灾难的影响，收敛速度非常慢。而自适应非参数估计量，能够自动地感知局部的内在维度，并且根据内在维度来调整自身的参数，从而完全规避维度灾难的影响。具体来说，自适应的权重分配机制，会自动地识别出那些流形之外的冗余维度，将这些维度的带宽调整到足够大，从而忽略这些维度的差异，将局部的有效维度降低到内在维度的水平，因此估计的收敛速度，是根据内在维度来计算的，而不是观测维度。也就是说，不管观测维度 D 有多高，只要内在维度 d 很低，收敛速度就会保持在低维数据的水平。而且，这个内在维度的感知过程，是完全自动的，不需要我们提前估计内在维度，也不需要做流形学习的预处理，估计器在估计的过程中，就自动地完成了内在维度的识别，以及参数的调整。这就从根本上解决了高维非参数估计中的维度灾难问题，不管观测维度有多高，只要数据具有低维的内在结构，自适应估计量就能够自动的利用这个结构，达到和低维数据一样的估计性能，这也是它能够适用于超高维复杂数据的核心原因。

6. 方法的理论应用边界与拓展方向

自适应非参数方法的优化框架有很强的扩展性，可以被推广到许多非参数统计问题中，给高维复杂数据下的统计分析提供统一的理论基础。目前的方法还存在着一定的理论边界，对于强相关数据和重尾分布数据的适应性还有待提高，相关的优化方向是未来理论研究的重点，可以拓宽方法的应用范围。

参考文献

[1] 黄菊红. 基于非参数统计的分类方法研究及应用[D]: [硕士学位论文]. 长沙: 湖南师范大学, 2016.

[2] 夏亚峰, 何佳. 高维数据下广义线性模型自适应桥惩罚估计的变量选择[J]. 甘肃科学学报, 2022, 34(1): 7-15.

[3] 郭婧璇, 田茂再. 基于充分降维的半参数不可忽略无响应光滑分位回归[J]. 系统科学与数学, 2024, 44(2): 471-507.

[4] 朱洪俊. 非平稳信号自适应滤波的小波模型与滤波方法[J]. 机械工程学报, 2006(8): 201-204.

[5] 李敏. 高维变系数模型的误差方差估计[D]: [硕士学位论文]. 重庆: 重庆大学, 2017.

[6] 谈凯. 稀疏切片逆回归: 最优收敛速度及其自适应估计[D]: [硕士学位论文]. 上海: 华东师范大学, 2018.

[7] 刘金灵. 多响应充分降维方法的改进[D]: [硕士学位论文]. 昆明: 云南财经大学, 2021.