文章编号：1000-8608(2010)06-1024-04

基于局域波分解的非参数概率密度估计

胡红英*[1,2]， 殷福亮[1]

（1. 大连理工大学 电子信息与电气工程学部，辽宁 大连 116024；
2. 大连民族学院 机电信息工程学院，辽宁 大连 116600）

摘要： 局域波分解在提取信号趋势方面具有异乎寻常的效果. 根据局域波缓变趋势提取算法，在小波概率密度估计思路的基础上，结合密度估计的直方图法，建立了局域波概率密度估计新方法. 此方法能有效去除样本数据直方图中的高频成分，获得低频趋势，即概率密度. 在混合高斯概率密度估计中的应用表明，对于无断点的密度函数，其具有计算简单、精度较高的优点.

关键词： 密度估计；局域波分解；经验模式分解；小波密度估计
中图分类号： TP181 **文献标志码：** A

0 引 言

概率密度估计是机器学习、模式识别、信号处理、特征提取和计算机视觉研究的关键之一，是数据挖掘的基础. 概率密度估计的实质问题，就是要通过从总体中抽得的样本去估计概率密度函数. 概率密度估计方法分为参数法和非参数法. 参数法估计需要事先知道概率密度函数的参数形式，但现实世界中数据的概率密度函数多种多样，很多情况下都无法事先知道，所以非参数估计就成了不可缺少的估计方法. 现有非参数密度估计方法有直方图法、Parzen 窗法[1]、小波法[2]、最大似然法[3]、SVM 法[4]、核方法[5]等. 其中，直方图法是最简单的密度估计方法，但估计结果精度较低，而且和 Parzen 窗法一样对训练数据量的大小较敏感. 而其他的方法虽然能在一定程度上提高估计精度，但计算较复杂.

局域波分解方法是近几年由 Huang 的经验模式分解(EMD)发展起来的一种非平稳信号处理方法，它有着和小波分析非常相似的特征[6~8]，如多分辨分析特性、滤波特性、完备性等. 但局域波也具有自己独特的特性，尤其是在提取信号的趋势时具有异乎寻常的效果. 因此本文根据小波密度估计方法，给出局域波密度估计方法的步骤.

1 局域波分解方法及其特点

局域波分解的方法很多，其中最典型的就是经验模式分解(EMD)方法，其实质是把信号 $s(t)$ 分解成多个基本模式分量和一个趋势项：

$$s(t)=\sum_{i=1}^{n}c_i(t)+r_n(t);\ t\in[0,T] \quad (1)$$

其中 $c_1(t),c_2(t),\cdots,c_n(t)$ 是 n 个基本模式分量，$r_n(t)$ 是趋势项.

每个基本模式分量都是一个单一模式分量，各基本模式分量的频率随着分解级数的增大而降低，而趋势项 $r_n(t)$ 则是频率最低的成分，是信号中无波动的趋势.

局域波分解实质上是一种完全由数据自身驱动的自适应滤波器库. Flandrin 等用此方法对分形高斯噪声进行分解，分解结果可等效成用常 Q 的带通滤波器库对信号进行滤波[9]. 此现象与小波分解类似，但小波分解是截止频率固定的滤波器库，而局域波分解的截止频率却是自适应的.

图 1 是一个局域波分解的实例. 其中，图 1(a)是 2 个正弦和 1 个二次曲线线性叠加的信号，两正弦的幅值和频率分别为 3、0.2 和 5、0.1. 图 1(b)、(c)是此信号经局域波分解后得到的 2 个基本模式分量，分别对应着原信号中的 2

收稿日期： 2008-10-12； **修回日期：** 2010-09-13.
作者简介： 胡红英*(1973-)，女，博士，副教授，E-mail：y_y_who@163.com；殷福亮(1962-)，男，教授，博士生导师.

个正弦；图 1(d)是分解后的趋势项，对应着原信号中的二次曲线. 可见，局域波分解不但能精确地分解出信号中的模式分量，还能精确地分解出信号中的微小趋势.

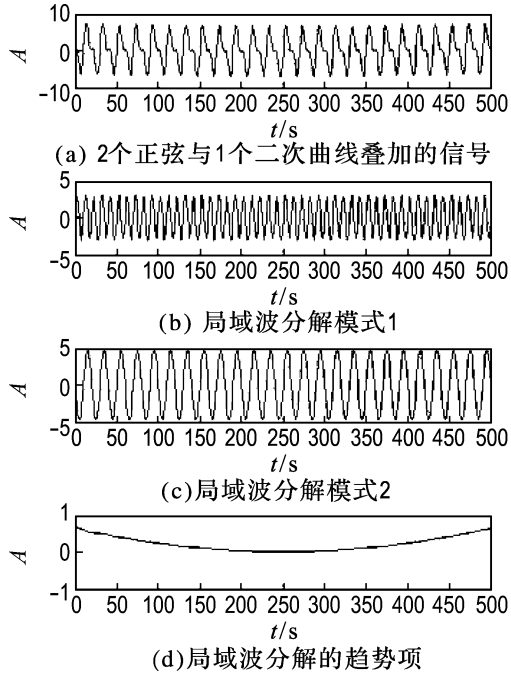


图 1　局域波分解实例

Fig. 1　An example of local wave decomposition

2　局域波缓变趋势提取及非参数密度估计

2.1　局域波缓变趋势提取方法

局域波分解的趋势项和小波分解的近似信号相类似，但小波分解的近似信号是小波 n 级分解后频率在$[0,2^{-n}]$的信号，可能包含低频波动，也可能不包含任何波动. 如果要用小波分解获得局域波分解的趋势项，分解级数可能很大，相应的计算量也会很大. 由此可见，在提取信号中的趋势，去除局部波动、平滑信号的能力上，局域波比小波更强.

然而，在工程应用中，有价值的信息往往不仅仅是信号的趋势项，还包括后几个基本模式分量叠加后的低频成分，叫做信号的缓变趋势. 信号的缓变趋势去除了信号中的高频噪声，能精确反映信号的变化过程和衍化趋势，对分析信号具有重要的意义.

文献[10]提到用 EMD 方法提取信号中快变波动(高频成分)的思路. 下面基于这种思路，给出相应的提取缓变趋势(低频成分)的具体方法.

首先对信号进行局域波分解(结果见式(1))，定义前 d 个基本模式分量的标准均值为

$$meanimf(d)=\frac{\sum_{t=0}^{T}\sum_{i=1}^{d}c_i(t)/N}{\sum_{t=0}^{T}s(t)/N} \tag{2}$$

其中 $d=1,2,\cdots,n$.

理论上讲，每个模式分量的均值和标准均值都为零，但实际上是一个非零小数. 而且，如果信号中存在非零趋势，那么在分解过程中，非零趋势就会对序数较大的模式分量产生影响，使得标准均值 $meanimf(d)$ 在 $d\geqslant D$ 时快速偏离零点. 因此，可以认为从第 D 个基本模式分量到最后一个基本模式分量再加上趋势项就是信号缓变趋势 $f(t)$ 的估计：

$$\tilde{f}(t)=\sum_{j=D}^{n}c_j(t)+r_n(t) \tag{3}$$

其中 n 是信号 $s(t)$ 经局域波分解得到的基本模式分量的个数.

图 2 是用此方法提取缓变趋势的例子. 图 2(b)是 4 000 点的某系统输出信号，经局域波分解后，得到 10 个基本模式分量和一个趋势项. 图 2(a)是前 d 个基本模式分量的标准均值 $meanimf(d)$ 随 d 变化的曲线. 可以看出，当 $d=7$ 时，$meanimf(d)$ 的值快速偏离了零点，因此取 $D=7$. 用第 7 到第 10 个基本模式分量和趋势项进行重构(见式(3))，得到信号的缓变趋势(见图 2(c)). 缓变趋势保留了信号中的较大波动，精确地展现了信号的衍化过程.

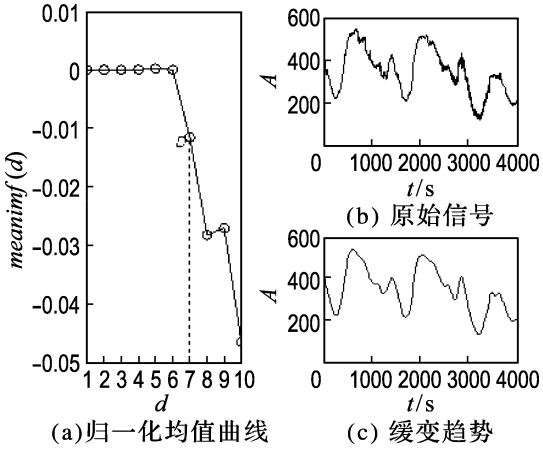


图 2　信号及其缓变趋势

Fig. 2　Signal and its slow-varying trend

2.2　基于局域波分解的非参数概率密度估计

由以上的分析可知，局域波方法能精确地去

除信号中的高频成分，获得信号的缓变趋势. 而小波密度估计的思想正是通过小波降噪，去除样本数据直方图中的高频成分，获得低频成分作为样本数据的概率密度估计. 因此，对照小波密度估计方法和步骤[2]，得到局域波密度估计步骤：

步骤 1 计算样本数据 X 的直方图，将数据转化为(X_b, Y_b). 即根据采样数据 X 的大小，均匀划分出 m 个柱条，$X_b(i)$ 是第 i 个柱条的中心坐标，$Y_b(i)$ 是样本数据 X 落在第 i 个柱条里的频数.

步骤 2 将 Y_b 作为一个信号进行局域波分解.

步骤 3 用式(2) 计算基本模式分量的标准均值，确定标准均值快速偏离零点时的 d 值，令 $D=d$.

步骤 4 用式(3)重构密度函数 f 的一个估计 $\tilde{f}$.

为了验证以上估计方法的性能，可以用仿真数据进行分析.

3 仿真分析

对于单一密度估计，简单的 Parzen 窗就有较好的估计效果，本文提出的局域波方法估计精度亦较高. 下面针对较复杂的混合高斯密度估计进行计算机仿真. 为了突出算法性能，仿真采用较少的样本数据.

由下式给出的混合高斯密度函数随机产生 100 个样本：

$$p(x) = \alpha_1 g(\mu_1, \sigma_1) + \alpha_2 g(\mu_2, \sigma_2) \quad (4)$$

其中 $g(\mu, \sigma)$ 是均值为 μ、方差为 σ 的高斯函数. 式(4)中各参数取值见表 1.

表 1 混合高斯密度参数

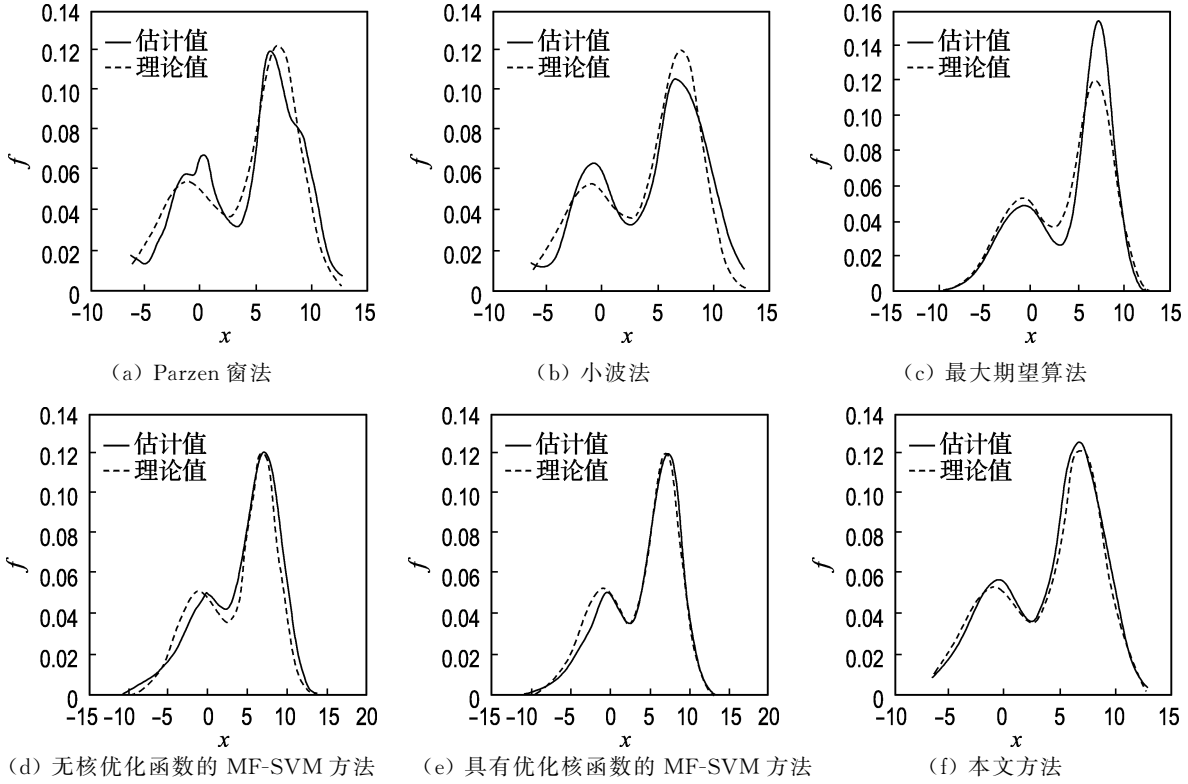
Tab. 1 Parameters of Gaussian mixture model density

参数	数值	参数	数值
μ_1	−1	μ_2	7
σ_1	3	σ_2	2
α_1	0.4	α_2	0.6

分别用 Parzen 窗法、小波法和局域波分解法估计此样本的密度. 计算过程中使用的柱条数均为 $m=100$.

(1)Parzen 窗法 调整 Parzen 窗的宽度，获得较光滑估计曲线，此处窗宽为 1. 结果见图 3(a). 其中短划线是样本 X 的理论密度曲线. 估计密度与理论密度的偏差平方和为 0.009 7.

(2)小波法 把样本点的统计结果 Y_b(局域波估计步骤 1)作为一个信号，用小波法进行密度估计. 为了获得最好的效果，尝试各种小波基和分解层数. 最后确定采用bior3.3小波，2层分解，软阈



(a) Parzen 窗法 (b) 小波法 (c) 最大期望算法

(d) 无核优化函数的 MF-SVM 方法 (e) 具有优化核函数的 MF-SVM 方法 (f) 本文方法

图 3 不同方法的混合高斯密度估计结果

Fig. 3 The estimations of Gaussian mixture model density with different methods

值规则进行估计，结果见图 3(b). 估计曲线出现了和理论曲线相似的形状，但并不吻合，误差较大. 估计密度与理论密度的偏差平方和为 0.011 7.

(3)局域波法　根据 2.2 节给出的局域波密度估计步骤，对此样本数据进行估计，结果见图 3(f). 估计结果与理论曲线基本吻合，偏差平方和为 0.002 7.

图 3 中的(c)、(d)、(e)是根据文献[11]介绍的方法获得的估计结果，分别使用了最大期望算法、无核优化函数的 MF-SVM 方法和具有优化核函数的 MF-SVM 方法.

从图 3 可以看出，在样本较少的情况下，局域波估计的密度更接近样本数据的理论值，比图3(a)～(d)的估计结果精度高. 虽然本文方法的结果与图 3(e)的结果基本相当，但计算方法比图 3(e)采用的优化核函数的 MF-SVM 方法简单很多.

4　结　论

实践证明，当密度函数具有断点或其一阶导数有断点时，采用小波进行密度估计是一个好的解决方法. 但当密度函数曲线较光滑时，本文的局域波法效果较突出. 本文的结果较其他估计方法精度高，方法简单，且计算量小. 当然，由于局域波分解时，边界问题还没有完全解决，有可能影响估计曲线的两端形状，产生些许误差. 随着局域波分解方法的不断完善，边界问题解决后，局域波法的估计精度将会更高.

参考文献：

[1] 吴喜之，王兆军. 非参数统计方法[M]. 北京：高等教育出版社，1996

[2] 飞思科技产品研发中心. Matlab6.5 辅助小波分析与应用[M]. 北京：电子工业出版社，2003

[3] BAHL L R, BROWN P F, DE SOUZA P V, *et al*. A new algorithm for the estimation of hidden Markov model parameters [C] //**IEEE International Conference on Acoustics, Speech and Signal Processing**. New York: IEEE, 1988: 493-496

[4] 张　炤，张　素，章琛曦，等. 基于支持向量机的概率密度估计方法[J]. 系统仿真学报，2005，**17**(10)：2355-2357

[5] HU Yu-suo, CHEN Hua, LOU Jian-guang, *et al*. Distributed density estimation using non-parametric statistics [C] // **The 27th International Conference on Distributed Computing Systems**. Toronto: IEEE, 2007: 25-27

[6] 马孝江，余　伯，张志新. 一种新的时频分析方法——局域波法[J]. 振动工程学报，2000，**13**(9)：24-29

[7] HUANG N E, SHEN Z, LONG S R, *et al*. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis [J]. **Proceeding of the Royal Society London A**, 1998, **454**: 903-995

[8] 胡红英. 局域波分解方法、特征剖析及工程应用研究[D]. 大连：大连理工大学，2006

[9] FLANDRIN P, RILLING G, GONCALVÉS P. Empirical mode decomposition as a filter bank [J]. **IEEE Signal Processing Letters**, 2004, **11**(2): 112-114

[10] FLANDRIN P, GONCALVES P, RILLING G. Detrending and denoising with empirical mode decompositions [C]//**Proceedings of the European Conference on Signal Processing**. Vienna: EURASIP, 2004: 1581-1584

[11] MOHAMED R M, EL-BAZ A, FARAG A A. Probability density estimation using advanced support vector machines and the expectation maximization algorithm [J]. **International Journal of Signal Processing**, 2005, **1**: 185-188

Nonparametric probability density estimation based on local wave decomposition

HU　Hong-ying[*1,2]，　YIN　Fu-liang[1]

(1. Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China;
2. Electromechanical Engineering College, Dalian Nationalities University, Dalian 116600, China)

Abstract: Local wave decomposition is good at accurate trend extracting. Therefore, based on the algorithm of slow-varying trend extracting of local wave method, as well as wavelet probability density estimation method and histogram method, a new probability density estimation method is presented. The proposed method can get rid of high frequency components of histogram and obtain the low frequency trend, i. e. probability density. The application of this method to Gaussian mixture model density estimation proves the advantages of the approach for non-breakpoint density function estimation. And it is easier in computation and more accurate.

Key words: density estimation; local wave decomposition; empirical mode decomposition; wavelet density estimation