UNIVERSITI PUTRA MALAYSIA

ROBUST KERNEL DENSITY FUNCTION ESTIMATION

KOUROSH DADKHAH

IPM 2010 7

ROBUST KERNEL DENSITY FUNCTION ESTIMATION

KOUROSH DADKHAH

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2010

ROBUST KERNEL DENSITY FUNCTION ESTIMATION

By

KOUROSH DADKHAH

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia
In Fulfilment of Requirements for Degree of Doctor of Philosophy

December 2010

Dedicated

To

My parents and my wife

Abstract of the thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy.

ROBUST KERNEL DENSITY FUNCTION ESTIMATION

By

KOUROSH DADKHAH

December 2010

Chairperson: Associate Professor Habshah Midi, PhD

Faculty: Institute for Mathematical Research

The classical kernel density estimation technique is the commonly used method to estimate the density function. It is now evident that the accuracy of such density function estimation technique is easily affected by outliers. To remedy this problem, Kim and Scott (2008) proposed an Iteratively Re-weighted Least Squares (IRWLS) algorithm for Robust Kernel Density Estimation (RKDE). However, the weakness of IRWLS based estimator is that its computation time is very long. The shortcoming of such RKDE has inspired us to propose new non-iterative and unsupervised based approaches which are faster, more accurate and more flexible. The proposed estimators are based on our newly developed Robust Kernel Weight Function (RKWF) and Robust Density Weight Function (RDWF). The basic idea of RKWF based method is to first define a function which measures the outlying distance of observation. The resultant distances are manipulated to obtain the robust weights. The statement of Chandola et al. (2009) that the normal (clean) data appear

in high probability area of stochastic model, while the outliers appear in low probability area of stochastic model, has motivated us to develop RDWF. Based on this notion, we employ the pilot (preliminary) estimate of density function as initial similarity (or distance) measure of observations with the neighbours. The modified similarity measures produce the robust weights to estimate density function robustly. Subsequently, the robust weights are incorporated in the kernel function to formulate the robust density function estimation. An extensive simulation study has been carried out to assess the performance of the RKWF-based estimator and RDWF-based estimator. The RKDE based on RKWF and RDWF perform as good as the classical Kernel Density Estimator (KDE) in outlier free data sets. Nonetheless, their performances are faster, more accurate and more reliable than the IRWLS approach for contaminated data sets.

The classical kernel density function estimation approach is widely used in various formula and methods. Unfortunately, many researchers are not aware that the KDE is easily affected by outliers. We have proposed the RKDE which is more efficient and consumes less time. Our work on RKDE or corresponding robust weights has motivated us to develop alternative location and scale estimators. A modification is made to the classical location and scale estimator by incorporating the robust weight and RKDE. To evaluate the efficiency of the proposed method, comprehensive contaminated models are designed and simulated. The accuracy of the proposed new method was compared with the location and scale estimators based on M,

Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) estimator. The simulation study demonstrates that, on the whole, the accuracy of the proposed method is better than the competitor methods.

The research also develops two new approaches for outlier and potential outlier detection in unimodal and multimodal distributions. The distance of observations from the center of data set is incorporated in the formulation of the first outlier detection method in unimodal distribution. The second method attempts to define an approach that is useable not only for unimodal distribution but also for multimodal distribution. This approach incorporates robust weights, whereby, high weights and low weights are assigned to normal (clean) and outlying observations, respectively. In this thesis, we also illustrate that the sensitivity of RKDE depends on the setting of the tuning constants of the employed loss function. The results of the study indicate that the proposed methods are capable of labelling normal observation and potential outliers in a data set. Additionally, they are able to assign anomaly scores to normal and outlying observations.

Finally this thesis also addresses the estimation of Mutual Information (MI) for mixture distribution which prone to create two distant groups in the data. The formulation of MI involves estimation of density function. Mutual information estimate for bivariate random variables involves the bivariate density estimation.

The bivariate density estimation employs the estimate of covariance matrix. The sensitivity of covariance matrix to the presence of outliers has motivated us to substitute it with robust estimate derived from MCD and MVE. The efficiency of the modified mutual information estimate is evaluated based on its accuracy. To do this evaluation, the mixtures of bivariate normal distribution with different percentage of contribution are simulated. Simulation results show that the new formulation of MI increases the accuracy of mutual information estimation.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah.

PENGANGGARAN FUNGSI KETUMPATAN KERNEL TEGUH

Oleh

KOUROSH DADKHAH

Desember 2010

Pengerusi: Profesor Madya Habshah Midi, PhD

Fakulti: Institut Penyelidikan Matematik

Teknik penganggaran fungsi ketumpatan kernel klasik adalah kaedah yang sering digunakan untuk menganggar fungsi ketumpatan. Sekarang telah terbukti bahawa ketepatan penganggarn fungsi ketumpatan tersebut mudah dipengaruhi oleh titik terpencil. Untuk mengatasi masalah ini, Kim and Scott (2008) mencadangkan Algoritma Kuasadua Terkecil Berpemberat (IRWLS) bagi Penganggaran Ketumpatan Kernal Teguh (RKDE). Namun, kelemahan penganggar yang berasaskan IRWLS ialah masa komputasi yang sangat panjang. Kelemahan penganggar RKDE tersebut telah memberi inspirasi kepada kami untuk mencadangkan pendekatan baru berasaskan tak-berulang tanpa pengawasan yang lebih cepat, lebih tepat dan lebih fleksibel. Penganggar yang dicadangkan berasaskan Fungsi Berpemberat Kernel (RKWF) dan Fungsi Berpemberat Ketumpatan Teguh (RDWF) baharu yang kami bangunkan. Ide permulaan kaedah berasaskan RKWF ialah pertamanya mendefinasikan suatu fungsi yang mengukur

jarak keterpencilan cerapan. Jarak yang terhasil dimanipulasi untuk mendapatkan pemberat teguh. Pernyataan Chandola et al. (2009) yang merumuskan bahawa data normal (bersih) muncul di kawasan model stokastik dengan kebarangkalian tinggi , manakala titik terpencil muncul di kawasn model stokastik dengan kebarangkalian rendah, telah member motivasi kepada kami untuk membangunkan RDWF. Berdasarkan ide ini, kami menggunakan penganggar pilot (awal) fungsi ketumpatan sebagai ukuran kesamaan awal (atau jarak) cerapan dengan jiran-jiran. Ukuran kesamaan yang diubahsuai menghasilkan pemberat teguh untuk menganggar fungsi ketumpatan secara teguh. Selanjutnya, pemberat teguh digabungkan ke dalam fungsi kernel untuk merumuskan penganggaran fungsi ketumpatan teguh. Kajian simulasi yang meluas telah dijalankan untuk menilai prestasi penganggar yang berasaskan RKWF dan RDWF. Bagi data set yang bebas daripada titik terpencil, prestasi penganggar RKDE yang berasaskan RKWF dan RDWF ada lah sama seperti Penganggar Ketumpatan Kernel Klasik (KDE). Walau bagaimana pun, bagi set data tercemar, prestasi kedua-dua penganggur lebih cepat, lebih tepat dan lebih dipercayai daripada pendekatan IRWLS.

Pendekatan penganggaran fungsi ketumpatan kernel klasik digunakan secara meluas dalam pelbagi formula dan kaedah. Malangnya, ramai penyelidik tidak menyedari bahawa KDE mudah dipengaruhi oleh titik terpencil. Kami telah mencadangkan RKDE yang lebih efisien dan kurang penggunaan masa. Penyelidikan kami keatas RKDE atau pemberat teguh yang berkaitan telah memberi motivasi kepada kami

untuk membangunkan penganggar alternatif bagi lokasi dan skala. Suatu pengubahsuaian telah dibuat keatas penganggar klasik lokasi dan skala dengan menggabungkan pemberat teguh dan RKDE. Untuk menilai kecekapan kaedah yang dicadang, model-model tercemar telah direkabentuk dan disimulasikan secara meluas. Kecekapan kaedah baharu yang dicadangkan ini telah dibandingkan dengan penganggar lokasi dan skala yang berasaskan penganggar M, Penentu Minimum Kovarians (MCD) and Minimum Isipadu Ellipsoid (MVE). Kajian simulasi telah menunjukkan bahawa secara keseluruhan, ketepatan kaedah yang dicadangkan adalah lebih baik daripada kaedah pesaing.

Penyelidikan ini juga membangunkan dua pendekatan baharu untuk mengenalpasti titik terpencil dan titik terpencil berpotensi bagi taburan unimodal and multimodal. Jarak cerapan-cerapan dari pusat suatu set data digabungkan dalam perumusan kaedah pertama pengesanan titik terpencil bagi taburan unimodal. Kaedah kedua dicuba dengan mentakrifkan suatu pendekatan yang boleh digunakan bagi kedua-dua taburan unimodal dan multimodal. Pendekatan ini digabungkan dengan pemberat teguh, yang mana, pemberat tinggi dan pemberat rendah diberikan masing-masing kepada cerapan-cerapan normal (bersih) dan cerapan-cerapan terpencil. Dalam tesis ini, kami juga mengilustrasikan bahawa kepekaan RKDE bergantung kepada penetapan pemalar tala bagi fungsi pengaruh yang digunakan. Keputusan kajian menunjukkan bahawa kaedah yang disarankan berupaya untuk melabel cerapan-cerapan normal dan titik terpencil berpotensi dalam suatu data set.

Selain dari itu, mereka juga berupaya untuk mengumpukkan skor anomali kepada cerapan normal dan cerapan terpencil.

Akhirnya, tesis ini juga mengutarakan penganggaran *Mutual Information* (MI) bagi taburan campuran yang mirip ke arah pembentukan dua kumpulan jarak bagi suatu data. Perumusan MI melibatkan penganggarn fungsi ketumpatan. Anggaran MI bagi pembolehubah rawak *bivariate* melibatkan penganggarn ketumpatan *bivariate*. Penganggaran ketumpatan *bivariate* menggunakan anggaran Matrik kovarians. Kepekaan matrik kovarians terhadap titik terpencil telah memberi motivasi kepada kami untuk menggantikannya dengan anggaran teguh yang di dapati daripada penganggar MCD dan MVE. Prestasi anggaran MI yang diubahsuai dinilai berdasarkan kecekapannya. Untuk melakukan penilaian ini, taburan campuran *bivariate normal* dengan beberapa peratusan sumbangan, disimulasikan. Keputusan simulasi menunjukkan bahawa ketepatan penganggaran perumusan baharu MI telah meningkat.

ACKNOWLEDGEMENTS

First and foremost, I would like to give grace to the Almighty God for sparing my life and for seeing me through the completion of this research work. I also wish to express my sincere appreciation and deep sense of gratitude to my supervisor Assoc. Prof. Habshah Midi for her guidance, encouragement and personal concern throughout the course of this research work. I would like to extend my gratitude to my supervisory committee for their guidance and support on this research.

Special thanks for Prof. Dr. A. M. H. Rahmatullah Imon, Statistics Professor from Ball State University, USA for his useful remarks and being my supervisory committee member.

It is also my great pleasure to give a due recognition to my family members for their all the time love, understanding and support in the course of this program and also for their prayers and words of encouragement whenever my enthusiasm waned. Specifically, I want to use this opportunity to express my sincere thankfulness to my father and mother for their constant support for my education over the years. I only hope that I can be as helpful to them in life as they have been to me.

My family deserves special recognition. This study would not have been possible without the encouragement, patience and overwhelming support of the author's wife Shahla Hosseini, during the period of this research that is especially acknowledged.

I certify that a Thesis Examination Committee has met on December 2010 to conduct the final examination of Kourosh Dadkhah on his thesis entitled "Robust Kernel Density Estimation" in accordance with the Universities and University Colleges Act 1971 and Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

Isa Daud, PhD
Associated Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Isthrinayagy A/P S. Krishnarajah, PhD
Lecturer
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mohd Rizam Abu Bakar, PhD
Associated Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Mohammed Nasser, PhD
Professor
University of Rajshani
Bangladesh
(Eternal Examiner)

SHAMSUDDIN SULAIMAN, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 18 January 2011

This thesis was submitted to the Senate of Universiti Putra Malaysia has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of supervisory committee were as follow:

Habshah Midi, PhD
Associate Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Chairperson)

A. M. H. Rahmatullah Imon, PhD
Associate Professor
Mathematical Sciences
Ball State University, USA
(Member)

Mohd Bakri Adam, PhD
Assistant Professor
Institute for Mathematical Research
Universiti Putra Malaysia
(Member)

Nasir Sulaiman, PhD
Associate Professor
Faculty of Science Computer and Information Technology
Universiti Putra Malaysia
(Member)

HASANAH MOHD GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I declare that the thesis is my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

KOUROSH DADKHAH
Date: 8 December 2010

TABLE OF CONTENTS