On Nonparametric Density Estimation in Kernel Exponential Families and the Sensitivity of Density Estimators

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Chenxi Zhou, B.S.

Graduate Program in Department of Statistics

The Ohio State University

2022

Dissertation Committee:

Vincent Q. Vu, Advisor

Yoonkyung Lee

Sebastian A. Kurtek

Abstract

This dissertation is concerned with the nonparametric density estimation problem in a kernel exponential family, which is an exponential family induced by a reproducing kernel Hilbert space (RKHS). The corresponding density estimation problem can be formulated as a convex minimization problem over a RKHS or a subset of it. The loss functionals we focus on are the negative log-likelihood (NLL) loss functional and the score matching (SM) loss functional.

We propose a new density estimator called the early stopping SM density estimator, which is obtained by applying the gradient descent algorithm to minimizing the SM loss functional and terminating the algorithm early. We investigate various statistical properties of this density estimator. We also compare this early stopping SM density estimator with the penalized SM density estimator that has been studied in the literature and address their similarities and differences.

In addition, we propose an algorithm to compute the penalized maximum likelihood (ML) density estimator that is obtained by minimizing the penalized NLL loss functional. We empirically compare the penalized and early stopping SM density estimators with the penalized ML density estimators and find out that when there is a small amount of regularization (corresponding to small values of the penalty parameter or large values of the number of iterations), the regularized SM density estimates contain a bump or become a spike at the isolated observation, but the penalized ML

density estimates do not. Moreover, if we remove the isolated observation, the resulting regularized SM density estimates do not contain a bump or a spike when the regularization is small. We attempt to explain why this happens.

Observations above motivate us to study the sensitivities of different density estimators to the presence of an additional observation. We extend the definition of the influence function by allowing its input to be function-valued statistical functionals. We study various properties of this extended influence function of ML and SM (log-)density projections in finite-dimensional and kernel exponential families, and empirically demonstrate that regularized SM density estimators in a kernel exponential family are more sensitive to the presence of an additional observation than the penalized ML density estimator when the amount of regularization is small.

To my family

Vita

May 2015	B.S. Mathematics and Economics
August 2015 — May 2019	Graduate Teaching and Research Associate, Department of Statistics, The Ohio State University
August 2019 — present	Graduate Research Associate, Nationwide Center for Advanced Customer Insights, Fisher College of Business, The Ohio State University

Fields of Study

Major Field: Statistics

Table of Contents

List of Figures

Figure **Page**