Abstract

This dissertation is concerned with the nonparametric density estimation problem in infinite-dimensional kernel exponential families, which amounts to minimizing convex loss functionals in an infinite-dimensional reproducing kernel Hilbert space. 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 both theoretically and empirically investigate various aspects 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 adaptive algorithm to compute the penalized maximum likelihood (ML) density estimator that is obtained by minimizing the penalized NLL loss functional. We empirically compare the regularized SM and ML density estimators and find out that when there is a small amount of regularization, the regularized SM density estimators contain a bump or become a spike at the isolated observation, but the regularized ML density estimator does not. Moreover, if we

remove the isolated observation, the bump or the spike in regularized SM density estimators disappears when the regularization is small. We attempt to explain why this happens.

The presence of a bump or a spike at the isolated observation in SM density estimators motivate us to study the sensitivity of density estimators to the presence of an additional observation. We extend the definition of the influence function and allow its input to be function-valued statistical functionals. We study various properties of this extended influence function in finite-dimensional and kernel exponential families, and empirically demonstrate that regularized SM density estimators in a kernel exponential family is more sensitive to the presence of an additional observation than the regularized ML density estimator when the amount of regularization is small.