

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

© Copyright by
Chenxi Zhou
2022

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

	Page
Abstract	ii
Dedication	iv
Vita	v
List of Figures	x
1. Introduction	1
1.1 An Review of Density Estimation Methods	2
1.1.1 Parametric Approach	2
1.1.2 Nonparametric Approach	3
1.1.2.1 Nonparametric Maximum Likelihood Density Estima- tion	4
1.1.2.1.1 Penalized Maximum Likelihood Density Estimation .	5
1.1.2.1.2 Shape-constrained Maximum Likelihood Density Es- timation	8
1.1.2.2 Nonparametric Score Matching Density Estimation .	12
1.2 Nonparametric Density Estimation in Kernel Exponential Families	14
1.3 Organization of the Remaining Dissertation	16
2. Kernel Exponential Family and Density Estimation Problem in It	18
2.1 Kernel Exponential Families	18
2.1.1 A Review of Finite-dimensional Exponential Family	18
2.1.2 Kernel Exponential Family	19
2.1.3 Properties of \mathcal{Q}_{ker}	20
2.1.3.1 Characterization of \mathcal{F} for Bounded Kernels	20
2.1.3.2 Convexity of A	21
2.1.3.3 Differential Properties of A	22
2.1.4 Connection to Finite-dimensional Exponential Families . . .	25

2.1.5	Assumptions on \mathcal{H} and k and Their Implications	30
2.2	Nonparametric Density Estimation in \mathcal{Q}_{ker}	31
2.2.1	Density Estimation in \mathcal{Q}_{ker} using \hat{L}_{NLL}	32
2.2.2	Density estimation in \mathcal{Q}_{ker} using \hat{L}_{SM}	35
2.3	Proofs	39
2.3.1	Proof of Proposition 2.1	39
2.3.2	Proof of Proposition 2.2	39
2.3.3	Proof of Lemma 2.1	40
2.3.4	Proof of Proposition 2.3	41
2.3.5	Proof of Proposition 2.5	43
3.	Early Stopping Score Matching Density Estimator	45
3.1	An Overview	45
3.2	Early Stopping SM Density Estimator	47
3.2.1	Computation of $\hat{f}^{(t)}$	51
3.2.2	Numerical Examples of Early Stopping SM Density Estimators	54
3.2.3	When to Terminate the Algorithm	55
3.3	Theoretical Properties of Early Stopping SM Density Estimator	57
3.3.1	Limiting SM Density Estimator as $t \rightarrow \infty$	59
3.3.1.1	Decomposition of $\hat{f}^{(t)}$	59
3.3.1.2	Numerical Illustration of Theorem 3.5	62
3.3.2	Rate of Convergence	62
3.3.2.1	An Upper Bound on the Approximation Error	65
3.3.2.2	An Upper Bound on the Sample Error	66
3.3.2.3	Upper Bounds on the Distances between p_0 and $q_{\hat{f}^{(t^*(n))}}$	66
3.3.2.4	Discussion on (B7)	67
3.4	Comparison to Penalized SM Density Estimator	69
3.4.1	Early Stopping SM Density Estimator as the Solution of a Penalized SM Loss Functional	69
3.4.2	Behavior When $\rho \rightarrow 0^+$	69
3.4.3	Comparison through Eigen-decomposition	71
3.4.4	Comparison of Convergence Rates	72
3.4.5	Numerical Examples	73
3.5	Auxiliary Results and Proofs	75
3.5.1	Proof of Theorem 3.2	75
3.5.2	Proof of Theorem 3.3	76
3.5.3	Proof of Theorem 3.4	76
3.5.4	Proofs of Results in Section 3.3.1	81
3.5.5	Proof of Theorem 3.6	84
3.5.6	Proof of Theorem 3.7	85
3.5.7	Proof of Theorem 3.8	87
3.5.8	Proof of Corollary 3.1	92

3.5.9	Proof of Results in Section 3.4	94
4.	Comparison of Regularized ML and SM Density Estimators in \mathcal{Q}_{ker} . . .	98
4.1	Penalized ML Density Estimator	99
4.1.1	Failure of the Representer Theorem	99
4.1.2	Construction of a Finite-dimensional Approximating Space .	101
4.1.3	Computation of the Minimizer of the Penalized NLL Loss Functional	103
4.1.3.1	Batch Monte Carlo Approximation of $\nabla \tilde{A}(\beta)$	105
4.1.3.2	Gradient Descent Algorithm to Minimize $\tilde{J}_{\text{NLL},\lambda}$. . .	105
4.1.4	Numerical Illustration	105
4.2	Regularized SM Density Estimators with $f \in \tilde{\mathcal{H}}$	110
4.2.1	Penalized SM Density Estimator with $f \in \tilde{\mathcal{H}}$	111
4.2.2	Early Stopping SM Density Estimator with $f \in \tilde{\mathcal{H}}$	111
4.3	Comparison of Regularized ML and SM Density Estimators	113
4.4	Discussion on the Presence of a Spike in SM Density Estimates . .	116
4.5	Proofs	118
4.5.1	Proof of Proposition 4.1	118
4.5.2	Details about Example 4.1	118
4.5.3	Proof of Proposition 4.2	122
4.5.4	Proof of Proposition 4.3	123
4.5.5	Proof of Proposition 4.4	124
5.	Influence Function of a (Log-)Density Function and Its Properties	126
5.1	Influence Function and Its Applications in Statistics	126
5.2	Extension of the Influence Function in Density Estimation Problem	131
5.3	Influence Function of (Log-)Density Projection in a Finite-dimensional Exponential Family	135
5.4	Influence Function of (Log-)Density Projection in a Kernel Exponential Family	141
5.5	Proofs	144
5.5.1	Proof of Proposition 5.1	144
5.5.2	Proof of Proposition 5.2	144
5.5.3	Proof of Theorem 5.1	145
5.5.4	Proof of Theorem 5.2	146
6.	Numerical Studies of the Sensitivities of ML and SM (Log-)Density Estimators in \mathcal{Q}_{ker}	150
6.1	Comparison of the Sensitivities of Penalized ML and SM Density Estimators	150
6.1.1	Computation of the Sample Influence Function	151

6.1.2	Comparison of the Sample Influence Functions of Log-density and Density Estimators	153
6.1.3	Comparison of the Sensitivities	159
6.2	The Sensitivity of K -fold Cross-validated Penalized SM Density Estimator	163
6.3	Which One to Use: Penalized ML or Regularized SM Density Estimators?	164
7.	Summary and Future Directions	166
7.1	Summary	166
7.2	Future Directions	167
	Appendices	170
A.	Math Background	170
A.1	Fréchet Differentiability and Derivative	170
A.2	Bochner Integral	173
A.3	Partial Derivative of a Kernel Function	175
A.4	Some Theories on Bounded Linear Operators	176

List of Figures

Figure		Page
1.1	Penalized SM density estimates of the <code>waiting</code> data with (first row) and without (second row) the isolated observation 108 (indicated by the purple circle). Histogram of the <code>waiting</code> variable with the bin width selected by the Freedman-Diaconis rule (Freedman and Diaconis, 1981) is shown in green.	15
3.1	Left panel shows μ and right panel shows $\log \mu$. The rug plot indicates the location of data.	54
3.2	Early stopping SM density estimates for different values of number of iterations labeled at the upper left corner. Histogram of data using the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.	55
3.3	Plots of \hat{z} (left panel), \hat{z}_1 (middle panel) and \hat{z}_2 (right panel). The rug plot indicates the location of data.	62
3.4	Density value at 108 against the number of iterations.	63
3.5	Density value at 108 against $\log \rho$	71
3.6	The penalized (first row) and early stopping (second row) SM density estimates with various choices of ρ and t , respectively, shown at the upper left corner. Histogram of data using the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.	74

4.1	Left panel: the minimum of $\tilde{\mathcal{J}}_{\text{NLL},\lambda}$ against the gap between two adjacent points at which kernel functions are centered in different choices of finite-dimensional approximating subspace. Different opacity indicates different values of λ , and the more opaque indicates the smaller λ value. Right panel: the minimum of $\tilde{\mathcal{J}}_{\text{NLL},\lambda}$ against different values of $\log \lambda$	109
4.2	Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of the <code>waiting</code> variable. Histogram of data using the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.	114
4.3	Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of the <code>waiting</code> variable with isolated observation 108 removed. Histogram of data using the Freedman-Diaconis rule is shown in green. The rug plot indicates the location of data and the purple circle indicates the location of the observation 108.	115
5.1	$\text{IF}_x(T, F, y)$ (left panel) and $\text{IF}_x(\tilde{T}, F, y)$ (right panel) evaluated at different $x \in \mathcal{X}$ with $\mathbb{E}_F[X] = 0$ and $y = 2$. The black dashed vertical line indicates the location of the contaminant y	134
6.1	Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$	155
6.2	Fix $\rho = e^{-11}$. Panels [A] and [B] show the penalized SM log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$	156

6.3	Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 120$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 120$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 120$	157
6.4	Fix $\lambda = e^{-15}$. Panels [A] and [B] show the penalized ML log-density estimates with and without the additional observation $y = 180$. Panel [C] shows the sample influence function of the log-density estimator. Panels [D] and [E] show the penalized SM density estimates with and without the additional observation $y = 180$. Panel [F] shows the resulting sample influence function of the density estimator. The rug plot indicates the location of data and the purple circle indicates the location of the observation $y = 180$	158
6.5	Left panel shows the overall influence versus different choices of y , where we fix $\rho = e^{-11}$ and the rugs indicate the location of the waiting data. The right panel shows the overall influence against different choices of ρ (shown in log scale), where we fix $y = 120$	160
6.6	Heat map of the overall influence on the penalized SM log-density estimates against y and ρ (shown in log scale). White rugs indicate locations of the waiting data.	161
6.7	Heat maps of the overall influence on penalized ML (left) and SM (right) log-density estimates against y and RKHS norm of the natural parameter under F_n (shown in log scale). Red vertical line in left panel indicates the case $\lambda = 0$. White rugs indicate locations of waiting data.	162
6.8	Overall influence of y on the K -fold cross-validated penalized SM density estimates against the values of y . We choose $K = 3$ (left panel), 5 (middle panel), and 10 (right panel).	164
7.1	ML log-concave density estimate with 100 random samples from the standard normal distribution, where the density estimate is computed using the R package logcondens (Dümbgen and Rufibach, 2010). Histogram with the bin width chosen by the Freedman-Diaconis rule is shown in green.	169