Chapter 1: Introduction

Density estimation is a classical and fundamental problem in statistics. With independent and identically distributed (i.i.d) data drawn from an unknown probability density function (pdf) p_0 over a domain $\mathcal{X} \subseteq \mathbb{R}^d$, the *density estimation problem* seeks to reconstruct p_0 using these data (Silverman, 1986).

Density estimation has found a wide range of statistical applications. On one hand, density estimates can be used in exploratory data analysis. More specifically, they can be utilized as an informal investigation of the properties of a given dataset and provide descriptive features such as multimodality, skewness, and tail behavior (Silverman, 1986; Izenman, 2009). On the other hand, density estimates can be used as an intermediate step to perform further statistical analysis, such as classification (Ripley, 1996) and clustering (Fukunaga and Hostetler, 1975). Due to the wide applications, there have been abundant works contributing to this problem.

We will provide a review of different approaches to the density estimation problem and discuss their advantages and disadvantages in Section 1.1. We then turn to the focus of this dissertation, the density estimation problem in an exponential family induced by a reproducing kernel Hilbert space (RKHS), and introduce the problems we focus on in this dissertation in Section 1.2. The organization of the rest of this dissertation will be given in Section 1.3.

1.1 A Review of Density Estimation Methods

Existing density estimation methods can be broadly categorized into the parametric approach (Subsection 1.1.1) and the nonparametric approach (Subsection 1.1.2). The parametric approach imposes a strong assumption that p_0 belongs to a parametric family known up to a few parameters, whereas the nonparametric approach abandons such a restrictive constraint and makes much fewer assumptions about p_0, which allows data to speak for themselves and offers more flexibility.

In the remaining part of this section, we provide details of these two different approaches and discuss their advantages and disadvantages.

1.1.1 Parametric Approach

For the parametric approach, we assume p_0 belongs to a parametric family

$$\mathcal{Q}_{\text{para}} := \Big\{ q_\theta : \mathcal{X} \to [0, \infty) \;\Big|\; \theta \in \Theta \Big\},$$

where $\theta := (\theta_1, \cdots, \theta_m)^\top$ is an m-dimensional parameter and $\Theta \subseteq \mathbb{R}^m$ is the parameter space; in other words, we assume there exists $\theta_0 \in \Theta$ such that $p_0 = q_{\theta_0}$. The density estimation problem then reduces to a parameter estimation problem. Once we obtain an estimator of θ_0, say $\hat{\theta}$, the resulting density estimator is $q_{\hat{\theta}}$.

Many methods of estimating θ_0 are available, for example, the method of moments and the method of maximum likelihood. The latter method, first proposed by Fisher (1922), considers to maximize the likelihood function

$$\prod_{i=1}^{n} q_\theta(X_i), \qquad \text{subject to } \theta \in \Theta, \tag{1.1}$$

or, equivalently, maximize the (averaged) log-likelihood function

$$\frac{1}{n} \sum_{i=1}^{n} \log q_\theta(X_i), \qquad \text{subject to } \theta \in \Theta. \tag{1.2}$$

The maximum likelihood estimator of θ_0, denoted by $\hat{\theta}_{\mathrm{ML}}$, is any point in Θ maximizing (1.1) or (1.2).

Since the functional form of pdfs in $\mathcal{Q}_{\mathrm{para}}$ is known, $\hat{\theta}_{\mathrm{ML}}$ is generally very easy to compute, where $\hat{\theta}_{\mathrm{ML}}$ either has an analytic form (e.g., when $\mathcal{Q}_{\mathrm{para}}$ is the family of Gaussian pdfs with unknown mean and covariance matrix) or can be obtained by solving an optimization problem of dimensionality at most m.

Statistical properties of these estimators have been well understood. For example, under certain regularity conditions, $\hat{\theta}_{\mathrm{ML}}$ can be shown to be consistent, asymptotically efficient, and asymptotically normally distributed (Casella and Berger, 2002). Some of these favorable statistical properties can be extended to density estimators under additional assumptions.

Although density estimators from this parametric approach have computational advantages and possess many nice statistical properties, this approach relies on the rigid assumption $p_0 \in \mathcal{Q}_{\mathrm{para}}$, which is hard or even impossible to verify in practice and is "entirely a matter for the practical statistician" (Fisher, 1922). If $p_0 \notin \mathcal{Q}_{\mathrm{para}}$, the resulting density estimator from $\mathcal{Q}_{\mathrm{para}}$ can be misleading and can lead to serious misspecification issues, which has been discussed by Huber (1967) and White (1982). Therefore, Fisher (1922) suggested to perform *a posteriori* test to examine the adequacy of the parametric assumption and the potential existence of misspecification.

1.1.2 Nonparametric Approach

If a parametric model is not postulated, we pursue the nonparametric approach and make as few assumptions as possible about p_0. This is the focus of this dissertation. In particular, we focus on nonparametric methods via minimizing a loss functional plus a penalty functional over a class of pdfs,

$$\underset{q \in \mathcal{Q}}{\operatorname{minimize}} \Big\{ \widehat{L}(q) + \lambda P(q) \Big\}, \tag{1.3}$$

where $\mathcal{Q}$ is a pre-specified class of pdfs over $\mathcal{X}$, $\widehat{L} : \mathcal{Q} \to \mathbb{R}$ is a loss functional, $P : \mathcal{Q} \to [0, \infty)$ is a penalty functional, and $\lambda > 0$ is the penalty parameter. In (1.3), $\widehat{L}$ depends on data and measures the goodness-of-fit of q to data, and a smaller value of $\widehat{L}(q)$ means the better the fit of q to data; and P is typically independent of data and measures the smoothness or size of q, and a larger value of $P(q)$ implies q is less smooth or more complex. Hence, the objective functional in (1.3) represents two conflicting goals: we demand q to have a good fit to data, but we also require it to contain less variations and be not too complex. The penalty parameter λ controls the tradeoff between these two conflicting goals.

We will focus on two different choices of $\widehat{L}$ in this dissertation, the negative log-likelihood loss functional (Subsection 1.1.2.1) and the score matching loss functional (Subsection 1.1.2.2).

1.1.2.1 Nonparametric Maximum Likelihood Density Estimation

Throughout this subsection, we let $\widehat{L}$ in (1.3) be the (averaged) negative log-likelihood (NLL) loss functional

$$\widehat{L}_{\mathrm{NLL}}(q) := -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i). \tag{1.4}$$

We call the density estimator via minimizing $\widehat{L}_{\mathrm{NLL}}$ the maximum likelihood (ML) density estimator, as minimizing the NLL loss functional is equivalent to maximizing the log-likelihood functional.

Minimizing $\widehat{L}_{\mathrm{NLL}}$ can be viewed as minimizing a sample version of the *Kullback-Leibler divergence* (KL-divergence)

$$\mathrm{KL}(p\|q) := \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x, \tag{1.5}$$

where p and q are two pdfs over $\mathcal{X}$ and satisfy $\mathrm{KL}(p\|q) < \infty$. To see this, assuming $\mathrm{KL}(p_0\|q) < \infty$ for all $q \in \mathcal{Q}$, with $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} p_0$, we can approximate $\mathrm{KL}(p_0\|q)$ by

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\big(\log p_0(X_i) - \log q(X_i)\big) &= \frac{1}{n}\sum_{i=1}^{n}\log p_0(X_i) - \frac{1}{n}\sum_{i=1}^{n}\log q(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log p_0(X_i) + \widehat{L}_{\mathrm{NLL}}(q). \end{aligned}$$

Notice that $\frac{1}{n}\sum_{i=1}^{n}\log p_0(X_i)$ is independent of q. The desired conclusion follows.

If we let $\mathcal{Q}$ be the class of all pdfs over $\mathcal{X}$ and $P(q) = 0$ for all $q \in \mathcal{Q}$, $\widehat{L}_{\mathrm{NLL}}$ is unbounded from below. To see this, suppose $\mathcal{X} = \mathbb{R}$ and let

$$q_{\sigma^2}(x) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x - X_j)^2\right), \qquad \text{for all } x \in \mathbb{R},$$

where $\sigma^2 > 0$. It is easy to verify that q_{σ^2} is a valid pdf over $\mathbb{R}$. Then, by shrinking $\sigma^2 \to 0$, we have $\widehat{L}_{\mathrm{NLL}}(q_{\sigma^2}) \to -\infty$.

Therefore, in order to obtain a sensible density estimator using $\widehat{L}_{\mathrm{NLL}}$, we need to impose certain constraints on $\mathcal{Q}$ and/or choose a nonzero P. Several proposals have been made in the literature.

1.1.2.1.1 Penalized Maximum Likelihood Density Estimation

In their seminal paper, Good and Gaskins (1971) proposed to use a nonzero P and minimize

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n}\log\gamma^2(X_i) + \left(\lambda_1\int_{\mathcal{X}}(\gamma'(x))^2\mathrm{d}x + \lambda_2\int_{\mathcal{X}}(\gamma''(x))^2\mathrm{d}x\right), \\ \text{subject to } \int_{\mathcal{X}}\gamma^2(x)\mathrm{d}x = 1, \end{aligned} \tag{1.6}$$

where $\gamma^2 = q$, $\lambda_1 > 0$, and $\lambda_2 > 0$. The specific form of the penalty functional in (1.6) is motivated from penalizing both the slope and the curvature to ensure the resulting density estimates are very smooth. If $\hat{\gamma}$ is a solution to (1.6), the resulting density estimator is $\hat{q} = \hat{\gamma}^2$.

There are at least two advantages of working with γ rather than the pdf q in (1.6). First, the density estimator by this approach is automatically nonnegative over $\mathcal{X}$, since the density estimator is $\hat{q} = \hat{\gamma}^2$.

Furthermore, due to the constraint $\int_{\mathcal{X}} \gamma^2(x)\mathrm{d}x = 1$, γ must belong to $L^2(\mathcal{X})$, the class of square-integrable functions over $\mathcal{X}$, which is a Hilbert space. This provides computational convenience that one can exploit to compute the minimizer of (1.6). Supposing $\{\varphi_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L^2(\mathcal{X})$, we can then approximate any $\gamma \in L^2(\mathcal{X})$ by $\gamma \approx \sum_{j=1}^{m} c_j \varphi_j$, for a large m. As a result, the minimization problem (1.6) over $L^2(\mathcal{X})$, an infinite-dimensional class of functions, reduces to a computationally tractable problem over $\mathbb{R}^m$, as one only needs to determine the values of $c_1, \cdots, c_m$. The value of m can be determined via cross-validation or through an iterative fashion as Good and Gaskins (1971) did.

When solving (1.6), one has to deal with the constraint $\int_{\mathcal{X}} \gamma^2(x)\mathrm{d}x = 1$, which can be difficult in practice. In order to remedy this, Leonard (1978) introduced the logistic transformation of the density function, proposed to parametrize q as $q(x) = \exp(f(x)) / \int_{\mathcal{X}} \exp(f(t))\mathrm{d}t$ for all $x \in \mathcal{X}$, and considered to minimize

$$-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \log\left(\int_{\mathcal{X}} \exp(f(x))\mathrm{d}x\right) + \frac{\lambda}{2}\widetilde{P}(f), \qquad \text{subject to } f \in \mathcal{H}, \tag{1.7}$$

where $\mathcal{H}$ is a pre-specified class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and $\widetilde{P}(f) := P(q)$. If the solution to (1.7) is $\hat{f}$, the resulting density estimator is $\exp(\hat{f}(x)) / \int_{\mathcal{X}} \exp(\hat{f}(t))\mathrm{d}t$, for all $x \in \mathcal{X}$, which is nonnegative over $\mathcal{X}$ and integrates to 1.

The main disadvantage of the formulation (1.7) is that it may *not* have a unique solution, since, if $\hat{f}$ is a solution, then $\hat{f} + c$ is a solution as well, for any $c \in \mathbb{R}$. To remedy this, Silverman (1982) proposed to minimize

$$-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \int_{\mathcal{X}} \exp(f(x))\mathrm{d}x + \frac{\lambda}{2}\widetilde{P}(f), \qquad \text{subject to } f \in \mathcal{H}. \tag{1.8}$$

Assuming $\widetilde{P}$ depends only on the square of derivatives of f, Silverman (1982) proved the minimizer of (1.8) exists and is unique under very mild conditions, and showed, if $\hat{f}$ is the solution to (1.8), $\exp(\hat{f})$ is automatically the density estimator. Moreover, Silverman (1982) established the consistency and the asymptotic convergence rate of his estimator under various function norms. However, no practical algorithms were proposed to compute his density estimator.

Penalty functionals we have reviewed so far all depend on the squared L^2 norm of the derivatives of the root-density or a transformation of the log-density. This is computationally convenient as the corresponding objective functional is differentiable so that first- and second-order iterative optimization algorithms can be applied to compute the respective minimizers and the resulting density estimators. Such penalty functionals, nonetheless, allow no jumps or piecewise linear bends in density estimates and often lead to over-smoothed density estimates (Sardy and Tseng, 2010). Motivated by these observations, Koenker and Mizera (2007) considered to minimize (1.8) and chose $\widetilde{P}$ to be the total variation of f',

$$\widetilde{P}(f) = \sup \sum_{i=1}^{m} |f'(u_i) - f'(u_{i-1})|, \tag{1.9}$$

where the supremum is taken over all partitions of $u_1 < u_2 < \cdots < u_m$ in $\mathcal{X} \subset \mathbb{R}$. To facilitate the computation, Koenker and Mizera (2007) assumed f is a piecewise linear function supported on $[X_{(1)}, X_{(n)}]$ with knots at the order statistics $X_{(1)}, \cdots, X_{(n)}$ and proposed to use the interior point method to compute the minimizer. From their simulation studies, Koenker and Mizera (2007) found that the penalty functional (1.9) works particularly well when p_0 is not smooth or contains sharp peaks. However, theoretical properties of their density estimator, such as consistency and the rate of convergence, remain unexplored.

1.1.2.1.2 Shape-constrained Maximum Likelihood Density Estimation

Even though density estimators by minimizing $\widehat{L}_{\text{NLL}}$ plus a penalty functional have very nice statistical properties, the quality of density estimates depends heavily on the choice of the penalty parameter $\lambda > 0$, which is a nontrivial task in general.

A different direction of estimating p_0 via minimizing $\widehat{L}_{\text{NLL}}$ is to impose certain qualitative properties on p_0, such as monotonicity, unimodality, or log-concavity. This shape-constrained approach is attractive as it requires no choice of the penalty parameter and is fully automatic (Cule, Samworth, and Stewart, 2010).

Shape-constrained density estimation originated from Grenander (1956), who studied the problem of estimating a non-increasing pdf over $[0, \infty)$ via minimizing L_{NLL}. It turns out that the solution, called the *Grenander estimator*, exists and is the left derivative of the *least concave majorant* of the empirical distribution function $\widehat{F}_n$, where the *least concave majorant* of a function F on $[0, +\infty)$ is

$$\inf\left\{ G \;\middle|\; G \text{ is concave over } [0, \infty), \text{ and } G(x) \geq F(x) \text{ for all } x \geq 0 \right\}.$$

Various statistical properties of the Grenander estimator, such as pointwise consistency and pointwise asymptotic distribution, have been investigated by Rao (1969), Groeneboom (1984), and Birge (1989).

The Grenander estimator can be extended to the case where p_0 is unimodal with the known mode. Suppose p_0 is unimodal with the known mode m_0, and is non-decreasing on $(-\infty, m_0]$ and is non-increasing on $[m_0, +\infty)$. With the aid of the

Grenander estimator, a natural estimator of p_0 is the derivative of the empirical distribution function obtained by the union of the greatest convex minorant of $\widehat{F}_n$ over $(-\infty, m_0]$ and the least concave majorant over $[m_0, +\infty)$ (Birgé, 1997). Theoretical properties of the Grenander estimator can be carried over to the unimodal density estimator.

When the mode m_0 is unknown, which is typically the case, however, the minimizer of $\widehat{L}_{\text{NLL}}$ over the class of all unimodal pdfs over $\mathbb{R}$ does *not* exist, since one can put the infinite density value at one of the observations (Birgé, 1997). Wegman (1970a), Wegman (1970b) and Birgé (1997) have proposed different unimodal density estimators when the mode is unknown and studied statistical properties of their respective estimators.

Despite their easiness of implementation and nice statistical properties, the monotone and unimodal density estimators discussed so far are hard to be generalized to the multivariate setting. A different shape constraint that has drawn a lot of attention in the past two decades and is easy to be generalized to the multivariate setting is the log-concavity. A function $p : \mathcal{X} \to [0, +\infty)$ is said to be *log-concave* if $\log p$ is a concave function on $\mathcal{X}$ with $\log 0 = -\infty$.

Indeed, many standard families of parametric density functions are log-concave, including all Gaussian pdfs with positive definite covariance matrix, all gamma pdfs $\Gamma(\alpha, \beta)$ with shape parameter $\alpha \geq 1$, all beta pdfs $\text{Beta}(\alpha, \beta)$ with $\alpha \geq 1$ and $\beta \geq 1$, and so on. A comprehensive list of log-concave parametric families together with their applications in economics can be found in Bagnoli and Bergstrom (2005).

Log-concave density functions have many useful properties. The log-concavity implies that the density is automatically unimodal and has convex level sets. The convolution of a log-concave density function with any unimodal density function is again unimodal (Ibragimov, 1956). The convolution of two log-concave pdfs is again log-concave, implying that if random variables X and Y have log-concave densities and are independent, then their sum $X + Y$ also has a log-concave density. Furthermore, random vectors with a log-concave density function have moment generating functions that are finite in a neighborhood of the origin and, thus, have moments of all orders (Samworth, 2018). In addition, if $X = (X_1^\top, X_2^\top)^\top \in \mathbb{R}^d$ has a log-concave density, the marginal densities of X_1 and X_2 are log-concave and the conditional density of X_1 given $X_2 = x_2$ is also log-concave for each x_2. Last but not the least, if X is a d-dimensional random vector with a log-concave pdf and A is a fixed $m \times d$ matrix of rank m, then the random vector AX has a log-concave density on $\mathbb{R}^m$. From these properties, the class of log-concave density functions share many similarities with the class of Gaussian density functions, and can be viewed as an infinite-dimensional generalization of the latter (Cule, Samworth, and Stewart, 2010; Samworth, 2018).

Log-concave density estimation over $\mathbb{R}^d$ via minimizing $\widehat{L}_{\text{NLL}}$ amounts to choosing $\mathcal{Q}$ to be the class of all log-concave pdfs over $\mathbb{R}^d$, which is equivalent to minimizing

$$\begin{aligned} &-\frac{1}{n}\sum_{i=1}^{n} f(X_i) + \int_{\mathcal{X}} \exp(f(x))\mathrm{d}x, \\ &\qquad\qquad\text{subject to } f : \mathcal{X} \to [-\infty, \infty) \text{ is concave.} \end{aligned} \tag{1.10}$$

It turns out that the solution to (1.10) exists and is unique, and admits a finite-dimensional representation. If $d = 1$, the solution is a piecewise linear function with knots at the order statistics $X_{(1)}, \cdots, X_{(n)}$, and is $-\infty$ over $\mathbb{R}\backslash[X_{(1)}, X_{(n)}]$ (Walther, 2002; Pal, Woodroofe, and Meyer, 2007; Dümbgen and Rufibach, 2009); if $d > 1$, the solution is characterized as a "tent function" supported on the convex hull of the data (Cule, Samworth, and Stewart, 2010). Here, for a fixed $y = (y_1, \cdots, y_n) \in \mathbb{R}^n$, a *tent function* is a function $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ with the property that $\bar{h}_y$ is the pointwise least concave function satisfying $\bar{h}_y(X_i) \geq y_i$ for all $i = 1, \cdots, n$. If $\hat{f}_{\text{lc}}$ is the minimizer of (1.10), the ML log-concave density estimator is $\hat{q}_{\text{lc}}(x) := \exp(\hat{f}_{\text{lc}}(x))$ for all $x \in \mathcal{X}$.

Various algorithms have been proposed to compute $\hat{f}_{\text{lc}}$. In the case of $d = 1$, Walther (2002), Pal, Woodroofe, and Meyer (2007), and Rufibach (2007) proposed to use the iterative convex minorant algorithm or a variant of it, and Dümbgen, Huesler, and Rufibach (2007) proposed an active set algorithm that turns out to be very efficient and was implemented in the R package logcondens (Dümbgen and Rufibach, 2010). For the cases of $d > 1$, two main difficulties exist. With the solution characterized as the "tent function", the corresponding objective function is non-differentiable. Cule, Samworth, and Stewart (2010) adopted Shor's r-algorithm, a subgradient method for minimizing non-differentiable convex functions over the Euclidean spaces, to handle the non-differentiability issue. The other difficulty is the computation of the integral appearing in (1.10) and its subgradient. Cule, Samworth, and Stewart (2010) proposed to triangulate the convex hull of data and compute the integral over each simplex in the triangulation. The time of computing $\hat{f}_{\text{lc}}$ unfortunately increases quickly with the sample size n and the dimensionality d and can be intolerably long for large n and d. As is reported in Table 1 by Cule, Samworth, and Stewart (2010), it takes about 224 minutes to compute a 4-dimensional log-concave density estimate using a sample of size 2000. In recent years, more efficient algorithms to compute $\hat{f}_{\text{lc}}$ for $d > 1$ cases have been proposed by Axelrod et al. (2019), Rathke and Schnörr (2019), and Chen, Mazumder, and Samworth (2021).

Theoretical properties of log-concave density estimators have also been investigated recently. When $d = 1$, under the assumption that p_0 is log-concave, Pal, Woodroofe, and Meyer (2007) proved $\hat{q}_{\text{lc}}$ is consistent for p_0 under the Hellinger distance, and Doss and Wellner (2016) established the corresponding convergence rate.

Dümbgen and Rufibach (2009) established the uniform consistency of $\hat{q}_{\mathrm{lc}}$ and the corresponding convergence rate. Furthermore, Balabdaoui, Rufibach, and Wellner (2009) derived the pointwise limiting distribution of $\hat{q}_{\mathrm{lc}}$. For the multivariate case, if p_0 is log-concave over $\mathbb{R}^d$, then $\hat{q}_{\mathrm{lc}}$ has been shown to be consistent under the exponentially weighted total variation distance and the Hellinger distance (Cule, Samworth, and Stewart, 2010; Dümbgen, Samworth, and Schuhmacher, 2011; Kim and Samworth, 2016). Additional details about log-concave density estimation via minimizing $\widehat{L}_{\mathrm{NLL}}$ can be found in Samworth (2018), a comprehensive survey of the recent progress in this topic.

Based on our discussion so far, the main advantage of the log-concave density estimation via minimizing $\widehat{L}_{\mathrm{NLL}}$ over the penalized approach is that one does *not* need to select the penalty parameter. It does suffer from some disadvantages such as the heavy computational burden discussed earlier. In addition, density estimates have some unsatisfactory qualitative features: they typically contain kinks and are only supported over the convex hull of data, and all boundary points of the convex hull of data are discontinuous points of $\hat{q}_{\mathrm{lc}}$. These unsatisfactory features may lead to serious problems in statistical applications such as classification. Smooth log-concave density estimator has been proposed by Chen and Samworth (2013).

1.1.2.2 Nonparametric Score Matching Density Estimation

Hyvärinen (2005) proposed a different loss functional, called the *score matching (SM) loss functional*, that can be used in the density estimation problem and is given by

$$\widehat{L}_{\mathrm{SM}}(q) := \frac{1}{n} \sum_{i=1}^{n} \sum_{u=1}^{d} \left(\frac{1}{2} \big(\partial_u \log q(X_i) \big)^2 + \partial_u^2 \log q(X_i) \right), \tag{1.11}$$

where $q : \mathcal{X} \to (0, \infty)$ is assumed to be a twice continuously differentiable pdf,

$$\partial_u \log q(x) := \frac{\partial}{\partial w_u} \log q(w)\bigg|_{w=x}, \qquad \text{and} \qquad \partial_u^2 \log q(x) := \frac{\partial^2}{\partial w_u^2} \log q(w)\bigg|_{w=x},$$

for all $u = 1, \cdots, d$, where $w := (w_1, \cdots, w_d)^\top \in \mathcal{X}$.

The SM loss functional originates from the Hyvärinen divergence (H-diveregnce) (Hyvärinen, 2005)

$$\mathrm{H}(p_0 \| q) := \frac{1}{2} \int_{\mathcal{X}} p_0(x) \|\nabla \log p_0(x) - \nabla \log q(x)\|_2^2 \mathrm{d}x, \tag{1.12}$$

where we assume p_0 is continuously differentiable over $\mathcal{X}$ and satisfies

$$\int_{\mathcal{X}} p_0(x) \|\nabla \log p_0(x)\|_2^2 \mathrm{d}x < \infty, \qquad \text{and} \qquad \int_{\mathcal{X}} p_0(x) \|\nabla \log q(x)\|_2^2 \mathrm{d}x < \infty.$$

Using the integration by parts and assuming $p_0(x) \partial_u \log q(x) \to 0$ as x approaches to the boundary of $\mathcal{X}$ for all $u = 1, \cdots, d$, we have

$$\begin{aligned} \mathrm{H}(p_0 \| q) = & \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} \bigg(\frac{1}{2} (\partial_u \log q(x))^2 + \partial_u^2 \log q(x) \bigg) \mathrm{d}x \\ & + \frac{1}{2} \int_{\mathcal{X}} p_0(x) \sum_{u=1}^{d} (\partial_u \log p_0(x))^2 \mathrm{d}x. \end{aligned} \tag{1.13}$$

Note that the last term depends on p_0 only and is independent of q. With $X_1, \cdots, X_n \overset{\text{i.i.d}}{\sim} p_0$, the SM loss functional (2.12) is simply the sample version of (1.13) with the last term omitted.

The main motivation of using $\widehat{L}_{\mathrm{SM}}$ is that one can avoid working with the normalizing constant. Let us return to the logistic transformation of the density function discussed earlier and let $q(x) = \exp(f(x)) / \int_{\mathcal{X}} \exp(f(t)) \mathrm{d}t$ for all $x \in \mathcal{X}$, and suppose we know the functional form of $f : \mathcal{X} \to \mathbb{R}$. It is typical that the normalizing constant $\int_{\mathcal{X}} \exp(f(t)) \mathrm{d}t$ is unknown and is analytically and computationally intractable. Then, since $\partial_u \log q(x) = \partial_u f(x)$ and $\partial_u^2 \log q(x) = \partial_u^2 f(x)$ do *not* depend on the normalizing constant $\int_{\mathcal{X}} \exp(f(t)) \mathrm{d}t$, neither does $\widehat{L}_{\mathrm{SM}}$.

Minimizing $\widehat{L}_{\mathrm{SM}}$ over the class of all pdfs over $\mathcal{X}$ is unbounded below. To see this, assume $\mathcal{X} = \mathbb{R}$ and again consider

$$q_{\sigma^2}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - X_j)^2\right), \qquad \text{for all } x \in \mathbb{R},$$

which is a valid pdf over $\mathbb{R}$. In addition, note that q_{σ^2} is twice continuously differentiable over $\mathbb{R}$. With some algebra, it can be shown that $\widehat{L}_{\mathrm{SM}}(q_{\sigma^2}) \to -\infty$ as $\sigma^2 \to 0^+$. This suggests that, in order to obtain a sensible density function, one has to put certain constraints on $\mathcal{Q}$ or use a nonzero P.

1.2 Nonparametric Density Estimation in Kernel Exponential Families

So far, we have discussed various approaches to density estimation. In the rest of this dissertation, we will focus on the nonparametric approach via minimizing $\widehat{L}_{\mathrm{NLL}}$ and $\widehat{L}_{\mathrm{SM}}$, and restrict $\mathcal{Q}$ to be an exponential family induced by a RKHS, which we call the *kernel exponential family* and will introduce formally in Chapter 2, and discuss various density estimators in it.

More specifically, for $\widehat{L}_{\mathrm{NLL}}$, we choose a nonzero penalty functional P and consider the penalized ML density estimator. For $\widehat{L}_{\mathrm{SM}}$, we consider two kinds of regularized density estimators: the penalized SM density estimator obtained by minimizing the SM loss functional plus a nonzero penalty functional P (Sriperumbudur et al., 2017), and the early stopping SM density estimator obtained by applying the gradient descent algorithm to minimizing the SM loss functional (with a zero P) and terminating the algorithm early to regularize (see Chapter 3).

Let us focus on the penalized SM density estimator for now. The first row in Figure 1.1 shows penalized SM density estimates of the `waiting` variable in the Old Faithful Geyser dataset (Azzalini and Bowman, 1990), with the corresponding penalty

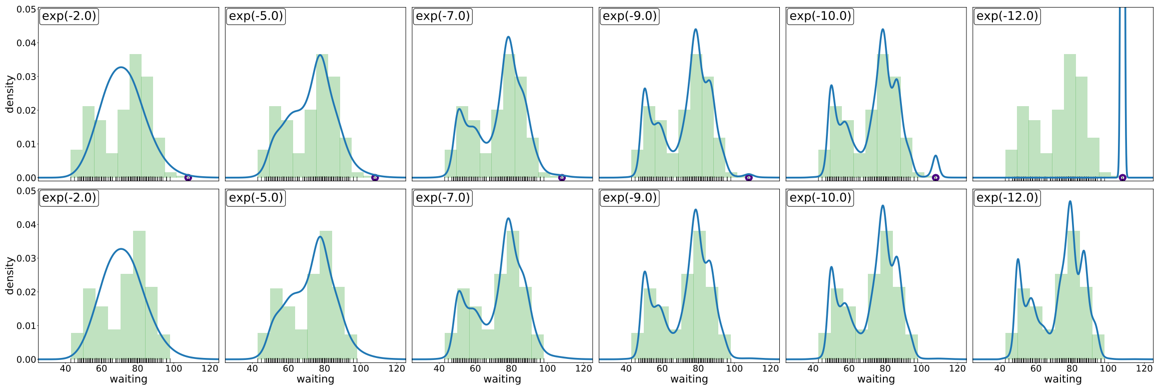

Figure 1.1: Penalized SM density estimates of the `waiting` variable with (first row) and without (second row) the isolated observation 108 (indicated by the purple circle). Histogram of the `waiting` variable with the bin width selected by the Freedman-Diaconis rule (Freedman and Diaconis, 1981) is shown in green.

parameter values listed in the upper left corner. The `waiting` variable records 299 time intervals (measured in minutes) between the starts of successive eruptions of the Old Faithful Geyser in Yellowstone National Park collected continuously from August 1st to August 15th, 1985. Notice, as the value of the penalty parameter becomes smaller, the penalized SM density estimates contain a bump or become a spike at the isolated observation 108. If we remove this isolated observation, as is shown in the second row of Figure 1.1, the resulting density estimates do *not* contain a bump or a spike when the value of the penalty parameter is small. As we will see in Chapter 3, the early stopping SM density estimator is qualitatively very similar to the penalized SM density estimator. When we compare these two kinds of regularized SM density estimators with the penalized ML density estimator in Chapter 4, however, the latter does not contain a bump or a spike, even when there is no penalty.

The observations above motivate us to study the sensitivity of these different density estimators to the presence of an isolated observation. The tool we choose is the influence function (Hampel, 1968), a classic concept from the robust statistics.

Traditionally, the influence function was defined for real- and vector-valued statistical functionals and was used to study the robustness properties of various real- and vector-valued estimators. But the object of primary interest in the density estimation problem is a function. The classic notion of the influence function is not directly applicable. We extend its definition to allow function-valued statistical functionals and to facilitate our understanding of the sensitivity of various density estimators.

1.3 Organization of the Remaining Dissertation

The rest of this dissertation is organized as follows.

In Chapter 2, we will formally introduce the kernel exponential family, discuss its properties, show its connection with the classic finite-dimensional exponential family, and discuss the density estimation problem in it using $\widehat{L}_{\text{NLL}}$ and $\widehat{L}_{\text{SM}}$ found in the literature.

In Chapter 3, we will focus on the early stopping SM density estimator and discuss its theoretical properties. We also compare it with the penalized SM density estimator and address their similarities and differences.

In Chapter 4, we will numerically compare two kinds of regularized SM density estimators with the penalized ML density estimator. We will demonstrate that the representer theorem, a classic theorem that characterizes the minimizer of a penalized convex loss functional over a possibly infinite-dimensional RKHS, cannot be used to characterize the minimizer of the penalized NLL loss functional. Instead, we discuss how to find a finite-dimensional subspace in $\mathcal{H}$ to approximate the minimizer of the penalize NLL loss functional, and propose an algorithm to compute the minimizer in such a subspace. In order to ensure the comparability, we also minimize the (penalized) SM loss functional in this finite-dimensional subspace and discuss how to achieve this. Furthermore, we will explain why the regularized SM density estimates contain a bump or a spike at the isolated observation when there is very small amount of regularization.

In Chapter 5, we will discuss our approach of extending the classic notion of the influence function to the studies of the sensitivity of density estimators. We will derive the influence functions of ML and SM (log-)density projections (to be defined) in both finite-dimensional and kernel exponential families. In Chapter 6, we compare the sensitivities of penalized ML and SM log-density estimators in the kernel exponential family through numerical examples and show that the penalized SM log-density estimator is more sensitive to the presence of an isolated observation than the penalized ML log-density estimator. Since we can use regularized SM or penalized ML density estimators, we will discuss which density estimator should be used and how it should be used in practice.

Finally, in Chapter 7, we will summarize this dissertation and discuss possible future directions based on the current work.