UTSSRP Project 2025 (Wong)

Monotone density estimation with Wasserstein projection

June 2025

1 Background

Let $x_1, \ldots, x_n$ be independent samples from an unknown density f on $\mathbb{R}^d$. The problem of *density estimation* is to use the samples to compute an estimate $\hat{f}_n$ of f. A common *nonparametric* density estimator is the *kernel density estimator* given by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right), \tag{1.1}$$

where the *kernel* K is a given density function (such as standard normal), and $h > 0$ is a tunable parameter called the *bandwidth*. A basic discussion can be found in [4, Section 18.4]. In R, the function `density()` performs kernel density estimation, where the bandwidth is chosen automatically based on the sample size.

Sometimes it is known *a priori*, or makes sense to *assume*, that the density f satisfies a *shape constraint*. For example, it may be reasonable to restrict f to the class of *unimodel*, *non-increasing*, or *log-concave* densities. See [2] for an overview of *shape-constrained statistical inference*. Note that the kernel density estimator $\hat{f}$ given by (1.1) may not satisfy the given constraint, so other methodologies may be needed to output an estimate in the chosen class.

For the purposes of this project, we consider the following setting:

- $d = 1$, so $X_1, X_2, \ldots$ are *univariate* random variables.
- $X_i \geq 0$, so the density f is supported on $\mathbb{R}_+ = [0, \infty)$, the non-negative real line.
- Our shape constraint is that f is *non-increasing* on $\mathbb{R}_+$:

$$x \leq y \Rightarrow f(x) \geq f(y).$$

- For technical purposes, we assume additionally that the density f has finite second moment.

Let $\mathcal{F}_0$ be the set of all densities on $\mathbb{R}_+$ satisfying this shape constraint.

Example 1. *It is easy to see that $\mathcal{F}_0$ contains all exponential distributions* $\mathrm{Exp}(\lambda)$ *and uniform distributions* $\mathrm{Unif}(0, \theta)$ *for $\theta > 0$.*

A common approach to shape-constrained density estimation is *maximum likelihood estimation*. In our context this means the following. Let data points $x_1, \ldots, x_n > 0$ be given. For $f \in \mathcal{F}_0$, the *log-likelihood* is defined by

$$\sum_{i=1}^{n} \log f(x_i),$$

where $\log f(x_i) = -\infty$ if $f(x_i) = 0$. We say that $\hat{f} \in \mathcal{F}_0$ is a (nonparametric) *maximum likelihood estimate*[1] with respect to the family $\mathcal{F}_0$ if

$$\hat{f} \in \operatorname*{arg\,max}_{f \in \mathcal{F}_0} \sum_{i=1}^{n} \log f(x_i). \tag{1.2}$$

This estimator was introduced by Grenander in [3] and hence is called the *Grenander estimator*. In fact, Grenander proved that $\hat{f}$ is given by the slopes of the *least concave majorant* of the empirical distribution function of the data. One way to implement the Grenander estimator is to use the function `grenader` in the R package `fdrtool`.

2 Density estimation with Wasserstein projection

In this project, we consider an alternative approach to monotone density estimation using the theory of *optimal transport*. We refer the reader to [8] for a systematic introduction to optimal transport. Our approach is motivated by recent interactions between optimal transport and statistical inference, see e.g. [1, 6].

We begin by defining the 2-*Wasserstein distance* between univariate distributions. If μ is a probability distribution (measure) on $\mathbb{R}$,[2] we let

$$F_\mu(x) = \mu((-\infty, x]), \quad u \in \mathbb{R},$$

be its *distribution function*, and

$$Q_\mu(u) = \inf\{x : F_\mu(x) \geq u\}, \quad u \in [0, 1],$$

be its *(generalized) quantile function* which is left-continuous by construction. When F_μ is strictly increasing, we have $Q_\mu = F_\mu^{-1}$. The most important property is that if $U \sim \mathrm{Unif}(0, 1)$, then $X = Q_\mu(X)$ is distributed as μ.

[1] Following standard convention, an *estimate* is the realized value of the estimator given realized data. Conversely, if we replace $x_1, \ldots, x_n$ by the random variables $X_1, \ldots, X_n$, the same formula defines an *estimator* (which is itself a random element).

[2] Since we will work with distributions on $\mathbb{R}_+$ we may replace $\mathbb{R}$ by $\mathbb{R}_+$.

Definition 2.1. *We let $\mathcal{P}_2(\mathbb{R})$ be the set of probability distributions on $\mathbb{R}$ with finite second moment, i.e., $\int_{\mathbb{R}} x^2 \mathrm{d}\mu(x) < \infty$.*

Proposition 2.2. *A probability distribution μ on $\mathbb{R}$ is an element of $\mathcal{P}_2(\mathbb{R})$ if and only if $Q_\mu \in L^2([0,1])$, i.e.,*

$$\|Q_\mu\|_{L^2([0,1])}^2 := \int_0^1 Q_\mu(u)^2 \mathrm{d}u < \infty.$$

Proof. We simply note that

$$\begin{aligned}\int_{\mathbb{R}} x^2 \mathrm{d}\mu(x) &= \mathbb{E}_{X\sim\mu}[X^2] = \mathbb{E}_{U\sim\mathrm{Unif}(0,1)}[Q_\mu(U)^2] \\ &= \int_0^1 Q_\mu(u)^2 \mathrm{d}u = \|Q_\mu\|_{L^2([0,1])}^2.\end{aligned}$$

So μ has finite second moment if and only if $\|Q_\mu\|_{L^2([0,1])} < \infty$. □

Definition 2.3 (2-Wasserstein distance on $\mathcal{P}_2(\mathbb{R})$)**.** *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$, we define the 2-Wasserstein distance $\mathcal{W}_2(\mu,\nu)$ by*

$$\mathcal{W}_2(\mu,\nu) = \|Q_\mu - Q_\nu\|_{L^2([0,1])}. \tag{2.1}$$

That is, $\mathcal{W}_2(\mu,\nu)$ is simply the L^2-distance between the quantile functions of μ and ν. It can be shown that this is equivalent to the (original) definition

$$\mathcal{W}_2(\mu,\nu) = \left(\inf\left\{\mathbb{E}\left[(X-Y)^2\right] : X\sim\mu, Y\sim\nu\right\}\right)^{1/2}, \tag{2.2}$$

where the infimum is over all joint distributions (i.e., *couplings*) for (X,Y) such that $X\sim\mu$ and $Y\sim\nu$. The 2-Wasserstein distance defines a metric (distance) on $\mathcal{P}_2(\mathbb{R})$. The representation (2.1) shows that the metric space $(\mathcal{P}_2(\mathbb{R}), \mathcal{W}_2)$ is *isometric* to the set of L^2 quantile functions equipped with the L^2 distance. This is a tremendous simplification which will play a crucial role in our study.

We consider monotoncity of the density in terms of the quantile function. To fix ideas, we begin with an example.

Example 2. *Let $\mu \in \mathrm{Exp}(1)$. Its density is $f_\mu(x) = e^{-x}$, which is non-increasing. The cdf is $F_\mu(x) = 1 - e^{-x}$. The quantile function is*

$$Q_\mu(u) = -\log(1-u), \quad u \in [0,1].$$

Note that $Q_\mu(0) = \lim_{u\to 0^+} Q_\mu(u) = 0$ and Q_μ is convex on $[0,1]$.

Proposition 2.4. *Suppose f is a non-increasing probability density on $\mathbb{R}_+$. Let Q be the corresponding quantile function. Then $Q(0) = \lim_{u\to 0^+} Q(u) = 0$ and Q is convex on $[0,1]$.*

Proof. Exercise. Also try to formulate a converse. That is, characterize μ if Q_μ satisfies the two properties. □

We introduce a set $\mathcal{F} \subset \mathcal{P}_2(\mathbb{R})$ that captures the shape constraint. It is slightly different from $\mathcal{F}_0$.

Definition 2.5. *Let $\mathcal{F}$ be the set of probability distribution μ on $\mathbb{R}_+$ whose quantile fucntion Q_μ satisfies the following properties:*

(i) *Q_μ is non-increasing (this always holds and is stated for completeness).*

(ii) *$Q_\mu(0) = \lim_{u \to 0^+} Q_\mu(u) = 0$.*

(iii) *Q_μ is convex.*

Proposition 2.6. *Let $\mathcal{Q} = \{Q_\mu : \mu \in \mathcal{F}\}$ be the set of quantile functions of distributions in $\mathcal{F}$, regarded as a subset of $L^2([0,1])$. Then:*

(i) *$\mathcal{Q}$ is convex.*

(ii) *$\mathcal{Q}$ is closed in L^2: if $Q_n \in \mathcal{Q}$, $n \geq 1$, is a sequence such that $\|Q_n - Q\|_{L^2([0,1])} \to 0$ for some $Q \in L^2([0,1])$, then $Q \in \mathcal{Q}$.*[3]

Proof. Since (i) is immediate, we will only prove (ii). Since Q_n is non-decreasing and convex with $Q_n(0) = \lim_{u \to 0^+} Q_n(u) = 0$ for all n, we may extend Q_n to a non-decreasing convex function on $\mathbb{R}$ by letting $Q_n(u) = 0$ for $u \leq 0$. Since $Q_n \to Q$ in $L^2([0,1])$, there exists a subsequence $Q_{n'}$ along which $Q_{n'} \to Q$ almost everywhere on $[0,1]$. By [7, Theorem 10.8], Q_n converges pointwise on $\mathbb{R}$ to a convex function $\tilde{Q}$ which is a.e. equal to Q on $[0,1]$. Clearly, $\tilde{Q}$ is non-decreasing. Also, we have $\tilde{Q}(0) = \lim_{n' \to 0} Q_{n'}(0) = 0$. Furthermore, since $\tilde{Q}$ is continuous at 0 by [7, Theorem 10.1], we have $\tilde{Q}(0) = \lim_{u \to 0^+} \tilde{Q}(u) = 0$. □

Remark 2.7 (Discussion)**.**

(i) *In fact, we may show that $\mathcal{F} = \overline{\mathcal{F}_0}^{\mathcal{W}_2}$ is the closure of $\mathcal{F}_0$ in the Wasserstein space $(\mathcal{P}_2(\mathbb{R}), \mathcal{W}_2)$.*

(ii) *Proposition 2.6 states that $\mathcal{F}$ may be regarded as a closed convex subset of the Wasserstein space under the isometry* (2.1)*. Convexity corresponds to geodesic convexity in the sense of McCann's displacement interpolation [5].*

Theorem 2.8 (Wasserstein projection)**.** *For any $\mu \in \mathcal{P}_2(\mathbb{R}_+)$, there exists unique $\mu^* \in \mathcal{F}$ such that*

$$\mu^* = \arg\min_{\nu \in \mathcal{F}} \mathcal{W}_2(\nu, \mu). \tag{2.3}$$

We call $\mu^* = \operatorname{proj}_{\mathcal{F}} \mu$ the Wasserstein projection of μ onto $\mathcal{F}$.

[3] More precisely, there exists a quantile function $\tilde{Q} \in \mathcal{Q}$ such that $Q = \tilde{Q}$ almost everywhere. This technicality is needed since two functions are identified in $L^2([0,1])$ if they agree almost everywhere.

Proof. From the isometry (2.1) under which a distribution ν is identified with its quantile function Q_ν, the problem $\min_{\nu \in \mathcal{F}} \mathcal{W}_2(\nu, \mu)$ is equivalent to the L^2 projection problem

$$\min_{Q \in \mathcal{Q}} \|Q - Q_\mu\|^2_{L^2([0,1])}. \tag{2.4}$$

Since $\mathcal{Q} \subset L^2([0,1])$ is a closed and convex subset by Proposition 2.6, by the Hilbert projection theorem there exists a unique solution $Q^* \in \mathcal{Q}$. We may let μ^* be the (unique) distribution whose quantile function is Q^*. □

We define a new monotone density estimator based on the Wasserstein projection.

Theorem 2.9 (Wasserstein density estimator)**.** *Given data points $x_1, \ldots, x_n \geq 0$, consider the empirical distribution $\mu_n = \sum_{i=1}^n \frac{1}{n}\delta_{x_i}$. The Wasserstein density estimate $\hat{\mu}_n$ with respect to $\mathcal{F}$ is defined by the Wasserstein projection onto $\mathcal{F}$:*

$$\hat{\mu}_n = \operatorname{proj}_{\mathcal{F}} \mu_n. \tag{2.5}$$

Remark 2.10. *While other shape constraints may be considered, for concreteness and simplicity we focus on $\mathcal{F}$.*

3 Research directions

(i) Implementation (possibly after a suitable discretization) of the Wassersten density estimator (2.3). From (2.4), it is an L^2 projection problem which is *convex*.

(ii) Numerical experiments using simulated data.

(iii) (May be difficult.) Theoretical properties of the Wasserstein density estimator, esp. in comparison with those of the Grenander estimator. For example, is it true that $\hat{\mu}_n$ always has a piecewise constant density? Also, can we prove that the density of $\hat{\mu}_n$ converges to that of μ? If so, at what rate? What happens if $\mu \notin \mathcal{F}$, i.e., the model is misspecified?

References

[1] Shun-ichi Amari and Takeru Matsuda. Wasserstein statistics in one-dimensional location scale models. *Annals of the Institute of Statistical Mathematics*, 74(1):33–47, 2022.

[2] Lutz Dümbgen. Shape-constrained statistical inference. *Annual Review of Statistics and Its Application*, 11, 2024.

[3] Ulf Grenander. On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956.

[4] Robert W Keener. *Theoretical Statistics.* Springer, 2010.

[5] Robert J McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[6] Victor M Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer, 2020.

[7] R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[8] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkäuser, 2015.