Density Estimation in Kernel Exponential Families

Methods and Their Sensitivities

Chenxi Zhou

Dissertation Committee

Vincent Q. Vu (advisor), Yoonkyung Lee, Sebastian A. Kurtek

Department of Statistics
The Ohio State University

August 17, 2022

Acknowledgements

I would like to thank

- Dr. Vu, for advising me in the past few years,
- Dr. Lee and Dr. Kurtek, for serving on the dissertation committee, and
- Dr. Conejo, for being the Graduate Faculty Representative.

This dissertation is supported in part by the National Science Foundation under grants DMS-1513621 and DMS-1916446.

Outline

Outline

Introduction to density estimation problem

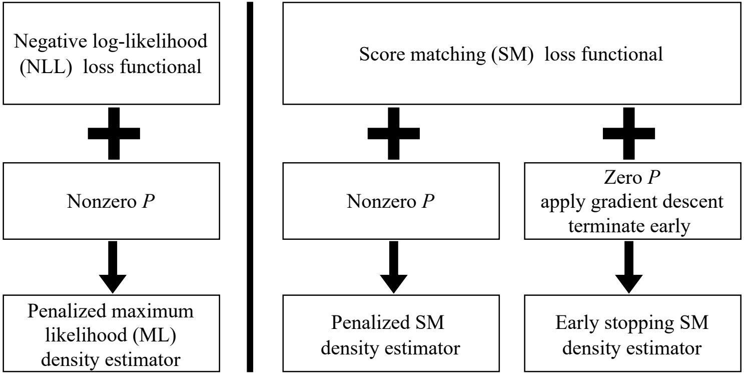
- Density estimation: Let $X_1, \cdots, X_n$ be i.i.d samples from an unknown pdf $p_0 : \mathcal{X} \to [0, \infty)$, where $\mathcal{X} \subseteq \mathbb{R}^d$. Estimate p_0 using these samples.
- Applications: exploratory data analysis, classification, clustering, ...
- Two categories of approaches:
  - parametric approach
  - nonparametric approach: minimizing

$$\widehat{L}(q) + \lambda P(q), \qquad \text{subject to } q \in \mathcal{Q},$$

  $\mathcal{Q}$ is a pre-specified class of pdfs on $\mathcal{X}$, $\widehat{L} : \mathcal{Q} \to \mathbb{R}$ is a loss functional, $P : \mathcal{Q} \to [0, \infty)$ is a penalty functional, and $\lambda \geq 0$ is a penalty parameter.

Our choices of $\mathcal{Q}$, $\widehat{L}$ and P

- Our choice of $\mathcal{Q}$: kernel exponential family — an exponential family induced by a reproducing kernel Hilbert space (RKHS)
- Our choices of $\widehat{L}$ and P:



Review of finite-dimensional exponential family

An m-dimensional exponential family contains all pdfs

$$\tilde{q}_\theta(x) := \mu(x) \exp\big(\langle \theta, \varphi(x) \rangle - B(\theta)\big) \text{ for all } x \in \mathcal{X}, \qquad \theta \in \Theta, \tag{1}$$

- $\mu : \mathcal{X} \to (0, \infty)$ is the *base density*,
- $\theta \in \Theta$ is the *natural parameter*,
- $\varphi : \mathcal{X} \to \mathbb{R}^m$ is the *canonical statistic*,
- $B(\theta) := \log\big(\int_{\mathcal{X}} \mu(x) \exp(\langle \theta, \varphi(x) \rangle) \mathrm{d}x\big)$ is the *log-partition function*, and
- $\Theta := \{\theta \in \mathbb{R}^m \mid B(\theta) < \infty\}$ is the *natural parameter space*.

Observations:

- φ maps to an m-dimensional space, which can be limited in some applications;
- $\tilde{q}_\theta$ depends on φ only through its inner product with θ.

Kernel exponential family $\mathcal{Q}_{\text{ker}}$

Kernel exponential family (Canu and Smola, 2006), $\mathcal{Q}_{\text{ker}}$, contains all pdfs

$$q_f(x) := \mu(x) \exp\big(\underbrace{\langle f, k(x,\cdot)\rangle_{\mathcal{H}}}_{=f(x)} - A(f)\big) \text{ for all } x \in \mathcal{X}, \qquad f \in \mathcal{F}, \tag{2}$$

- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the kernel associated with the RKHS $\mathcal{H}$,
- $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is the inner product in $\mathcal{H}$,
- $A(f) := \log\big(\int_{\mathcal{X}} \mu(x) \exp(f(x)) \mathrm{d}x\big)$ is the *log-partition functional*, and
- $\mathcal{F} := \{f \in \mathcal{H} \mid A(f) < \infty\}$ is the *natural parameter space*.

Assumptions on μ, $\mathcal{H}$, and k

- μ is a continuously differentiable pdf.
- $\mathcal{H}$ does *not* contain constant functions, which ensures identifiability, i.e., $q_{f_1} = q_{f_2}$ if and only if $f_1 = f_2$.
- $\mathcal{H}$ is infinite-dimensional.
- k is
  - twice continuously differentiable, which ensures all quantities shown later to be well-defined, and
  - bounded, i.e., $\sup_{x \in \mathcal{X}} \sqrt{k(x,x)} < \infty$, which implies $\mathcal{F} = \mathcal{H}$ and makes all optimization problems considered later unconstrained.

Minimizing the NLL loss functional

(Averaged) NLL loss functional

$$\widehat{L}_{\mathrm{NLL}}(q) := -\frac{1}{n}\sum_{i=1}^{n} \log q(X_i) \tag{3}$$

With $q = q_f \in \mathcal{Q}_{\mathrm{ker}}$, we can write $\widehat{L}_{\mathrm{NLL}}(q)$ as

$$\widehat{J}_{\mathrm{NLL}}(f) := A(f) - \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{4}$$

Bad News: Minimizing $\widehat{J}_{\mathrm{NLL}}$ over $\mathcal{H}$ has no solution (Fukumizu, 2009).

Remedy: Impose certain kind of regularization to obtain a solution.

Minimizing the penalized NLL loss functional

Gu and Qiu (1993) proposed to minimize

$$\widehat{J}_{\mathrm{NLL}}(f) + \lambda \widetilde{P}(f), \qquad \text{subject to } f \in \mathcal{H}. \tag{5}$$

where $\widetilde{P} : \mathcal{H} \to [0, \infty)$ is $\widetilde{P}(f) := P(q_f)$ for all $f \in \mathcal{H}$.

- Established the existence and the uniqueness of the minimizer in $\mathcal{H}$;
- This minimizer in $\mathcal{H}$ is not computable. Proposed to minimize over

$$\left\{ f \,\middle|\, f := \sum_{i=1}^{n} \alpha_i k(X_i, \cdot), \alpha_1, \cdots, \alpha_n \in \mathbb{R} \right\}.$$

Gu (1993) proposed an iterative algorithm to compute the minimizer. Main difficulty: Need to work with A and its derivatives, which involve integration over a possibly high-dimensional space.

Minimizing the SM loss functional

$$\widehat{L}_{\mathrm{SM}}(q) := \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}\left(\frac{1}{2}\big(\partial_u \log q(X_i)\big)^2 + \partial_u^2 \log q(X_i)\right), \tag{6}$$

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

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

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

$\widehat{L}_{\mathrm{SM}}(q)$ with $q \in \mathcal{Q}_{\mathrm{ker}}$

With $q = q_f \in \mathcal{Q}_{\mathrm{ker}}$, $\widehat{L}_{\mathrm{SM}}(q)$ becomes

$$\widehat{J}_{\mathrm{SM}}(f) := \frac{1}{2}\langle f, \widehat{C} f\rangle_{\mathcal{H}} - \langle f, \hat{z}\rangle_{\mathcal{H}}, \tag{7}$$

where $\widehat{C} : \mathcal{H} \to \mathcal{H}$ is

$$\widehat{C} := \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d} \partial_u k(X_i, \cdot\,) \otimes \partial_u k(X_i, \cdot\,),$$

with $\widehat{C} f = \frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d} \partial_u f(X_i)\partial_u k(X_i, \cdot\,)$ for all $f \in \mathcal{H}$, and

$$\hat{z} := -\frac{1}{n}\sum_{i=1}^{n}\sum_{u=1}^{d}(\partial_u \log\mu(X_i)\partial_u k(X_i, \cdot\,) + \partial_u^2 k(X_i, \cdot\,)) \in \mathcal{H}.$$

With fixed $x \in \mathcal{X}$, $\partial_u^s k(x, y) := \frac{\partial^s}{\partial w_u^s} k(x, y)\Big|_{w=x}$ for all $y \in \mathcal{X}$, for $s = 1, 2$.

Minimizing $\widehat{J}_{\mathrm{SM}}$ over $\mathcal{H}$ has no solution

Good news:

- $\widehat{J}_{\mathrm{SM}}$ does not involve A.
- Minimizing $\widehat{J}_{\mathrm{SM}}$ over $\mathcal{H}$ is a convex problem, as $\widehat{C}$ is self-adjoint positive (semi-)definite.

Bad news: Minimizing $\widehat{J}_{\mathrm{SM}}$ over $\mathcal{H}$ has no solution.

Remedy: Impose certain kind of regularization.

Penalized SM density estimator

Sriperumbudur et al. (2017) proposed to minimize

$$\widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2, \qquad \text{subject to } f \in \mathcal{H}, \tag{8}$$

where $\rho > 0$ is the penalty parameter. This is the Tikhonov regularization.

The minimizer of (8) exists and is unique. Using a general representer theorem,

$$\hat{f}_{\mathrm{SM}}^{(\rho)} := \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2 \right\} = \sum_{i=1}^{n} \sum_{u=1}^{d} \alpha_{(i-1)d+u}^{(\rho)} \partial_u k(X_i, \cdot\,) + \frac{1}{\rho}\hat{z},$$

where $(\alpha_1^{(\rho)}, \cdots, \alpha_{nd}^{(\rho)})^\top \in \mathbb{R}^{nd}$ can be obtained by solving a linear system.

Penalized SM density estimator — comments

- No need to work with A.
- Sriperumbudur et al. (2017) empirically showed penalized SM density estimator outperforms kernel density estimator, especially when d is large.
- No comparison with the (penalized) ML density estimator was conducted by Sriperumbudur et al. (2017).

Outline

Early stopping regularization

- *Early stopping* is a form of regularization based on choosing when to terminate an iterative optimization algorithm.
- Often referred to as *implicit* regularization, in contrast to the penalized approach by explicitly adding a penalty term.

Early stopping SM density estimator

Apply gradient descent algorithm with constant step size to minimizing $\widehat{J}_{\mathrm{SM}}$.

Starting with $\hat{f}^{(0)}_{\mathrm{SM}} = 0 \in \mathcal{H}$, gradient descent iterates are

$$\hat{f}^{(t+1)}_{\mathrm{SM}} = \sum_{i=1}^{n}\sum_{u=1}^{d} \alpha^{(t+1)}_{(i-1)d+u} \partial_u k(X_i, \cdot\,) + (t+1)\tau\hat{z}, \qquad \text{for all } t = 0, 1, 2, \cdots,$$

where $\tau > 0$ is step size, and $(\alpha^{(t+1)}_1, \cdots, \alpha^{(t+1)}_{nd})^\top \in \mathbb{R}^{nd}$ can be obtained by multiplication and addition of certain matrices.

Numerical example

Goals: To illustrate

- early stopping SM density estimator, and
- its similarity with the penalized SM density estimator.

Data: `waiting` variable in the Old Faithful Geyser dataset, which records 299 time intervals (measured in minutes) between the starts of successive eruptions of the Old Faithful Geyser in Yellowstone National Park August 1st – 15th, 1985.

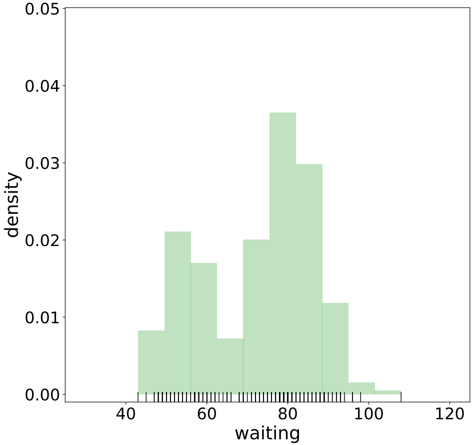


Figure: Histogram of `waiting` data with the bin width selected by the Freedman–Diaconis rule.

Numerical example (continued)

- $\mathcal{X} = (0, \infty)$;
- The base density μ is pdf of Gamma distribution.

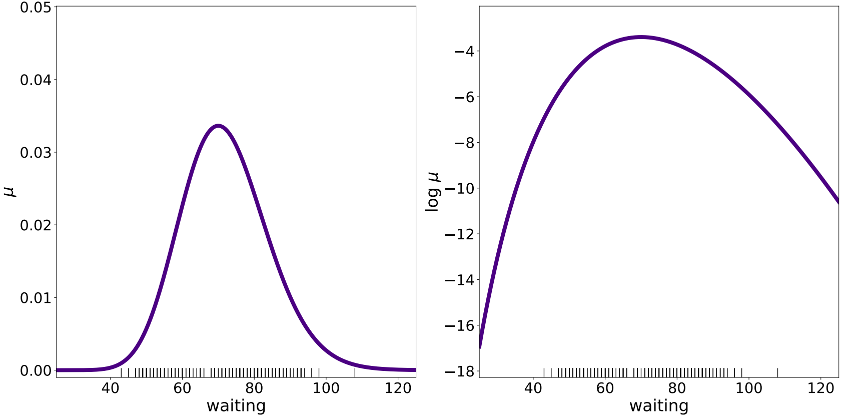


Figure: Left panel: μ; right panel: $\log \mu$.

Numerical example (continued)

The RKHS $\mathcal{H}$ is the one generated by

$$k(s,t) = \exp\left(-\frac{(s-t)^2}{2\sigma^2}\right),$$

for all $s, t \in \mathcal{X}$, with $\sigma = 5$.

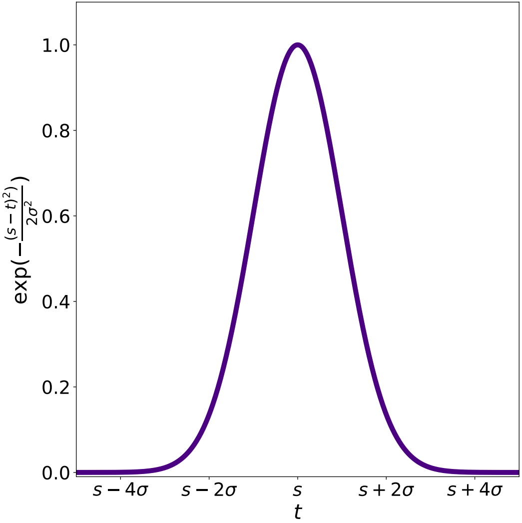


Figure: Gaussian kernel function.

Penalized and early stopping SM density estimators are very similar

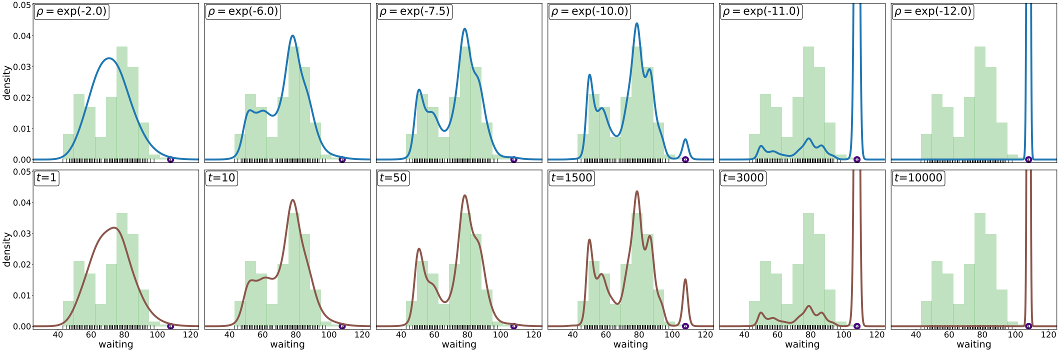


Figure: Penalized (first row) and early stopping (second row) SM density estimates of `waiting` data. Purple circle indicates the location of the isolated observation 108.

Outline

Goal

To compare penalized ML and regularized SM density estimators, and understand their similarities and differences.

Penalized ML density estimator

Choose $\widetilde{P}(f) = \frac{1}{2}\|f\|_{\mathcal{H}}^2$ and minimize

$$\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \qquad \text{subject to } f \in \mathcal{H}. \tag{9}$$

Good news: This minimization problem has a unique minimizer in $\mathcal{H}$.

Bad news: The representer theorem fails. This unique minimizer in the infinite-dimensional $\mathcal{H}$ is not computable.

Remedy: Use a finite-dimensional subspace

$$\widetilde{\mathcal{H}} := \left\{ f \;\middle|\; f := \sum_{j=1}^{m} \beta_j k(w_j, \cdot), \beta_1, \cdots, \beta_m \in \mathbb{R} \right\} \tag{10}$$

to approximate this minimizer, where $w_1, \cdots, w_m \in \mathcal{X}$ are pre-specified.

Computation of penalized ML density estimator

With $f = \sum_{j=1}^{m} \beta_j k(w_j, \cdot\,) \in \widetilde{\mathcal{H}}$, we can write $\widehat{J}_{\mathrm{NLL}}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2$ as

$$\widetilde{J}_{\mathrm{NLL},\lambda}(\boldsymbol{\beta}) := \widetilde{A}(\boldsymbol{\beta}) - \boldsymbol{\beta}^\top \left(\frac{1}{n}\mathbf{K}_1 \mathbf{1}_n\right) + \frac{\lambda}{2}\boldsymbol{\beta}^\top \mathbf{K}_2 \boldsymbol{\beta}, \tag{11}$$

- $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_m)^\top \in \mathbb{R}^m$,
- $\widetilde{A}(\boldsymbol{\beta}) := A\big(\sum_{j=1}^{m} \beta_j k(w_j, \cdot\,)\big)$,
- the (j, i)-entry of $\mathbf{K}_1 \in \mathbb{R}^{m\times n}$ is $k(w_j, X_i)$,
- the (j, j')-entry of $\mathbf{K}_2 \in \mathbb{R}^{m\times m}$ is $k(w_j, w_{j'})$,
- $\mathbf{1}_n := (1, \cdots, 1)^\top \in \mathbb{R}^n$.

Use gradient descent algorithm to compute the minimizer of $\widetilde{J}_{\mathrm{NLL},\lambda}$ over $\mathbb{R}^m$.
Approximate $\nabla \widetilde{A}(\boldsymbol{\beta})$ using the Monte Carlo method.

Computation of regularized SM density estimators

For comparability purpose, also compute regularized SM density estimators in $\widetilde{\mathcal{H}}$.

- Penalized SM density estimator: with a fixed $\rho > 0$, $\arg\min_{f \in \widetilde{\mathcal{H}}} \left\{ \widehat{J}_{\mathrm{SM}}(f) + \frac{\rho}{2} \| f \|_{\mathcal{H}}^2 \right\}$ can be obtained by solving a linear system;
- Early stopping SM density estimator: with $\tilde{f}_{\mathrm{SM}}^{(0)} = 0$, gradient descent iterates are

$$\tilde{f}_{\mathrm{SM}}^{(t)} := \sum_{j=1}^{m} \tilde{\beta}_j^{(t)} k(w_j, \cdot), \qquad \text{for all } t = 0, 1, 2, \cdots,$$

where $(\tilde{\beta}_1^{(t)}, \cdots, \tilde{\beta}_m^{(t)})^\top \in \mathbb{R}^m$ can be obtained by matrix addition and multiplication.

Numerical example

Still use `waiting` data to empirically compare penalized ML and regularized SM density estimators.

Choose $\widetilde{\mathcal{H}}$ to be

$$\left\{ f \;\middle|\; f := \sum_{j=1}^{m} \beta_j k(w_j, \cdot\,), \beta_1, \cdots, \beta_m \in \mathbb{R} \right\}, \tag{12}$$

where $m = 201$, and $(w_1, \cdots, w_{201}) = (1, \cdots, 201)$.

No bump/spike in penalized ML density estimates

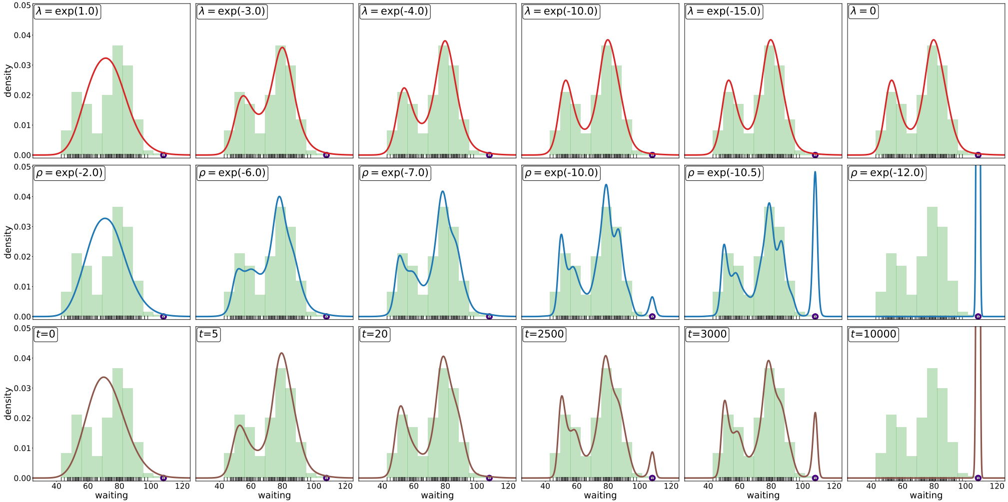


Figure: Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of `waiting` data. Purple circle indicates the location of isolated observation 108.

No bump/spike in regularized SM density estimates if 108 is removed

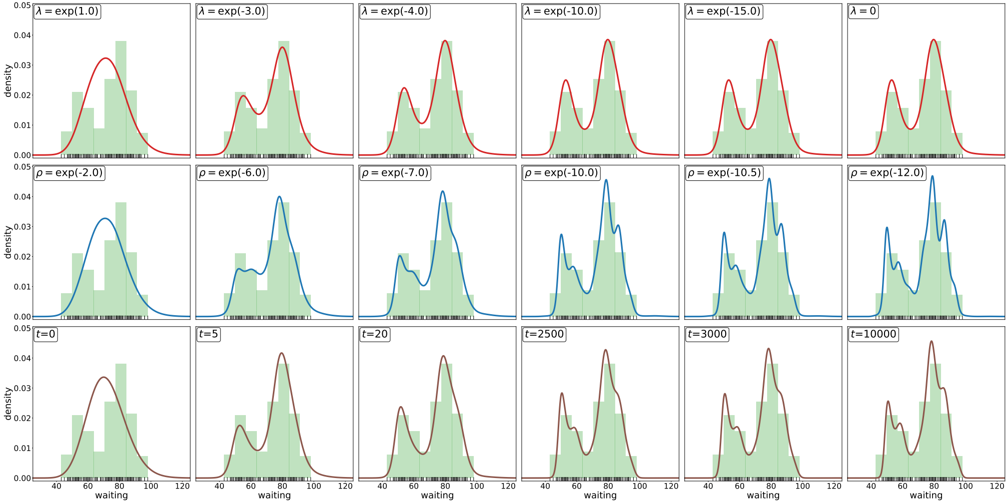


Figure: Penalized ML (first row), penalized SM (second row), and early stopping SM (third row) density estimates of **waiting** data with isolated observation 108 removed.

Explanation of spike in penalized SM density estimates when ρ is tiny

The penalized SM density estimator is obtained by minimizing $\widehat{L}_{\mathrm{SM}}(q_f) + \frac{\rho}{2}\|f\|_{\mathcal{H}}^2$, where, with $d = 1$,

$$\widehat{L}_{\mathrm{SM}}(q_f) = \underbrace{\frac{1}{n}\sum_{i \neq i^*}\left(\frac{1}{2}\big((\log q_f)'(X_i)\big)^2 + (\log q_f)''(X_i)\right)}_{=:(\mathrm{I})} + \underbrace{\frac{1}{n}\left(\frac{1}{2}\big((\log q_f)'(X_{i^*})\big)^2 + (\log q_f)''(X_{i^*})\right)}_{=:(\mathrm{II})}, \tag{13}$$

and X_{i^*} denotes the isolated observation.

When ρ is tiny, we are effectively minimizing $\widehat{L}_{\mathrm{SM}}$ part.

Explanation of spike in penalized SM density estimates when ρ is tiny

Notice

1. $\log q_f$ is a linear combination of $\log \mu$ and Gaussian kernel functions,
2. Gaussian kernel functions are local basis functions, and
3. a spike is essentially a local maximum.

Then, putting a spike in $\log q_f$ at X_{i^*} has the effects of

- forcing $((\log q_f)'(X_{i^*}))^2 \approx 0$ (due to 1 and 3),
- reducing the value of $(\log q_f)''(X_{i^*})$ (due to 1 – 3) and that of (II) a lot,
- not affecting (I) much (due to 2), and
- reducing the value of $\widehat{L}_{\mathrm{SM}}$ a lot.

Outline

Goals

1. To develop a set of tools to understand the sensitivity of density estimators, and
2. To understand the sensitivities of penalized ML and SM density estimators[1] in $\mathcal{Q}_{\mathrm{ker}}$ to the presence of an isolated observation.

[1] We drop early stopping SM density estimator as it is very similar to penalized SM density estimator.

Using influence function in density estimation problem

Influence function (Hampel, 1968) was traditionally defined for real- and vector-valued statistical functionals.

The object of main interest in density estimation problem is a pdf.

Need to extend the definition of influence function to allow function-valued statistical functionals.

Influence functions of log-density function evaluated at x

Let T be a map from the collection of distribution functions over $\mathcal{X}$ to the class of log-density functions over $\mathcal{X}$.

The *influence function of* $T(F)$ *evaluated at* $x \in \mathcal{X}$ is

$$\mathrm{IF}_x(T, F, y) := \lim_{\varepsilon \to 0^+} \frac{1}{\varepsilon} \Big(T((1-\varepsilon)F + \varepsilon\delta_y)(x) - T(F)(x) \Big), \tag{14}$$

where δ_y is the point mass 1 at $y \in \mathcal{X}$.

$\mathrm{IF}_x(T, F, y)$

- is the directional derivative of T at F in direction δ_y evaluated at x, and
- measures the effect of an infinitesimal amount of contamination at y on $T(F)$ evaluated at x.

Example: normal location model

Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Q}$ contain all pdfs

$$\tilde{q}_\theta(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right) \text{ for all } x \in \mathcal{X},$$

where $\theta \in \mathbb{R}$.

Define $T(F) = \log q_{\theta(F)}$, where

$$q_{\theta(F)} := \arg\max_{q_\theta \in \mathcal{Q}} \left\{ \int_{\mathcal{X}} \log q_\theta(x) \mathrm{d}F(x) \right\}.$$

Assume $m_0 := \mathbb{E}_F[X]$ exists. For all $x \in \mathcal{X}$,

$$\mathrm{IF}_x(T, F, y) = (y - m_0)(x - m_0).$$

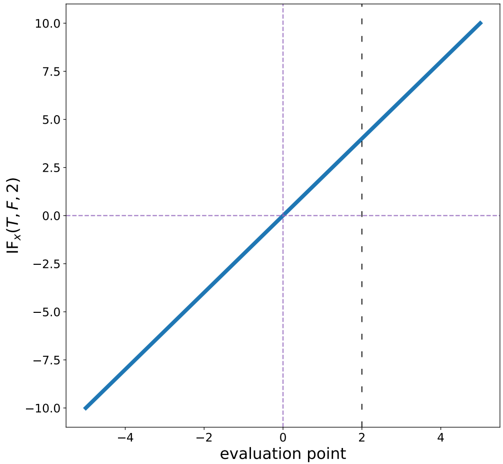


Figure: $\mathrm{IF}_x(T, F, y)$ evaluated at different points with $y = 2$ and $m_0 = 0$. The black dashed vertical line indicates the location of y.

Overall influence

Fix y and F, and view $\mathrm{IF}_x(T,F,y)$ as a function of x. $\mathrm{IF}_x(T,F,y)$ varies with x.

Define the *overall influence* of y on $T(F)$ to be

$$\sup_{x\in\mathcal{X}} \left|\mathrm{IF}_x(T,F,y)\right|,$$

which describes the maximal possible effect of y on $T(F)$.

Normal location model example:

$$\sup_{x\in\mathcal{X}} \left|\mathrm{IF}_x(T,F,y)\right| = \begin{cases} 0, & \text{if } y = m_0 \\ \infty, & \text{otherwise} \end{cases}$$

Sample influence function

Define the *sample influence function of* $T(F_n)$ *evaluated at* $x \in \mathcal{X}$ to be

$$\mathrm{SIF}_{x,\varepsilon}(T, F_n, y) := \frac{1}{\varepsilon}\Big(T((1-\varepsilon)F_n + \varepsilon\delta_y)(x) - T(F_n)(x)\Big), \tag{15}$$

where $\varepsilon > 0$ and F_n is the empirical distribution function of $X_1, \cdots, X_n$.

The corresponding *overall influence* is

$$\sup_{x \in \mathcal{X}} |\mathrm{SIF}_{x,\varepsilon}(T, F, y)|.$$

A special sample influence function

With $\varepsilon = \varepsilon_0 := \frac{1}{n+1}$,

$$(1 - \varepsilon_0)F_n + \varepsilon_0 \delta_y = F_{n+1},$$

where F_{n+1} is the empirical distribution function of $X_1, \cdots, X_n$ and y.

Then,

$$\mathrm{SIF}_{x,\varepsilon_0}(T, F_n, y) = (n+1)\Big(T(F_{n+1})(x) - T(F_n)(x)\Big) \tag{16}$$

describes how the value of $T(F_n)$ at x is affected by an additional observation y.

Notation

In order to compare the sensitivities of penalized ML and SM density estimators in $\mathcal{Q}_{\mathrm{ker}}$, define

$$T_\lambda(F) = \log q_{f^{(\lambda)}_{\mathrm{ML},F}}, \qquad \text{and} \qquad S_\rho(F) = \log q_{f^{(\rho)}_{\mathrm{SM},F}},$$

where

$$f^{(\lambda)}_{\mathrm{ML},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{ A(f) - \int_{\mathcal{X}} f(x)\mathrm{d}F(x) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}} \right\},$$
$$f^{(\rho)}_{\mathrm{SM},F} := \operatorname*{arg\,min}_{f\in\mathcal{F}} \left\{ \frac{1}{2}\langle f, C_F f\rangle_{\mathcal{H}} - \langle f, z_F\rangle_{\mathcal{H}} + \frac{\rho}{2}\|f\|^2_{\mathcal{H}} \right\},$$

and $C_F := \sum_{u=1}^d \int_{\mathcal{X}} \partial_u k(x,\,\cdot\,) \otimes \partial_u k(x,\,\cdot\,)\mathrm{d}F(x)$, and
$z_F := -\sum_{u=1}^d \int_{\mathcal{X}} \left(\partial_u^2 k(x,\,\cdot\,) + \partial_u \log\mu(x)\partial_u k(x,\,\cdot\,)\right)\mathrm{d}F(x)$.

Focus on the sample influence function

Expressions of $\mathrm{IF}_x(T_\lambda, F, y)$ and $\mathrm{IF}_x(S_\rho, F, y)$ exist but are hard to work with.

We focus on the sample influence function with $\varepsilon = \varepsilon_0 := \frac{1}{n+1}$ and compare numerically.

Setup:

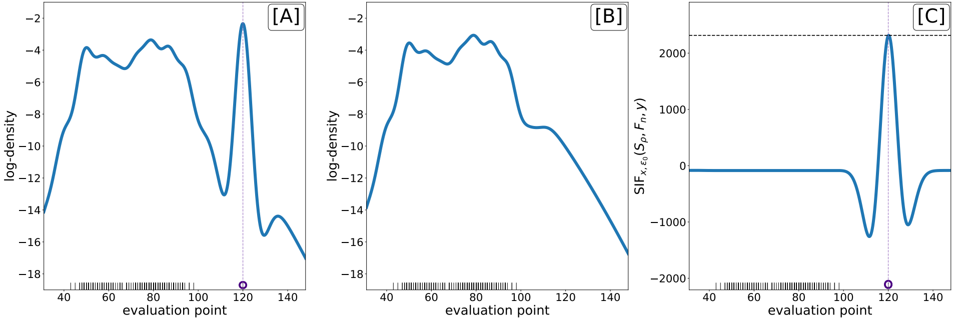
- Use `waiting` data with the isolated observation 108 removed.
- Use the same $\mathcal{X}$, μ, k, and $\widetilde{\mathcal{H}}$ as before.
- Choose $y = 20, 22, 24, \cdots, 180$.
- Choose
  - $\lambda = 0, e^{-15}, e^{-14.5}, \cdots, e^{0.5}, e^{1}$, and
  - $\rho = e^{-12}, e^{-11.5}, \cdots, e^{0}$.

Example

Fix $y = 120$ and $\rho = e^{-11}$. [A] and [B] show $S_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)$ and $S_\rho(F_n)$, respectively, and [C] shows

$$\mathrm{SIF}_{x,\varepsilon_0}(S_\rho, F_n, y) \propto S_\rho((1-\varepsilon_0)F_n + \varepsilon_0\delta_y)(x) - S_\rho(F_n)(x).$$

Overall influence of $y = 120$ is ≈ 2315.48, achieved roughly at $x = 120$.



Different y locations give different overall influences

Still fix $\rho = e^{-11}$ and vary y:

- if y is < 40 or > 100 (in low-density region), it has a larger overall influence, and
- if y is between 40 and 100 (in high-density region), it has a smaller overall influence.

Need to look at different y values.

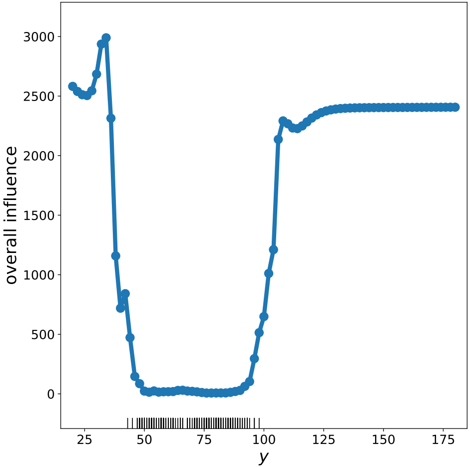


Figure: Overall influence of y on S_ρ vs. y, with $\rho = e^{-11}$.

Different ρ values give different overall influences

For a fixed $y = 120$, the smaller the penalty parameter value is, the larger the overall influence of y is.

Need to look at different penalty parameter values.

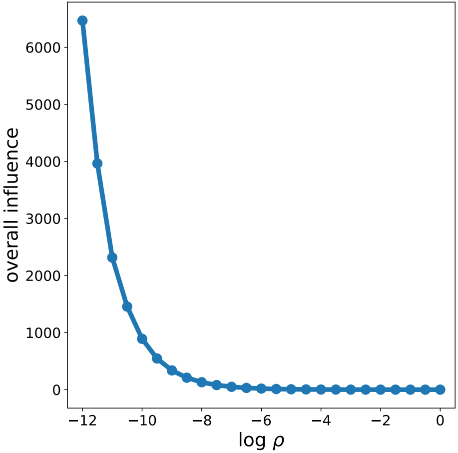


Figure: Overall influence of $y = 120$ on S_ρ vs. penalty parameter.

Plotting penalty parameter on one axis is a bad idea

To show effects of different y and different penalty parameter values on overall influence, we can produce a heat map similar to the right.

Recall our goal is to compare sensitivities of penalized ML and SM density estimators.

Plotting penalty parameter on one axis is not conducive to comparison, as λ and ρ are on different scales.

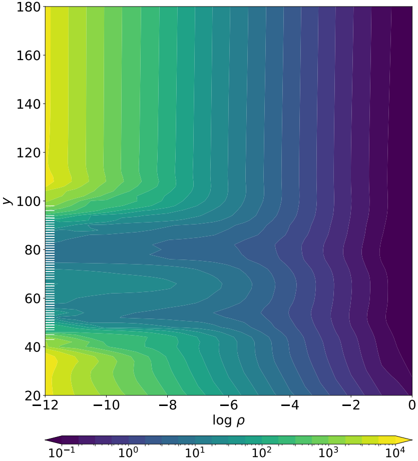


Figure: Heat map of overall influence on S_ρ vs. y and ρ. White rugs indicate locations of `waiting` data.

Penalized SM density estimator is more sensitive to y

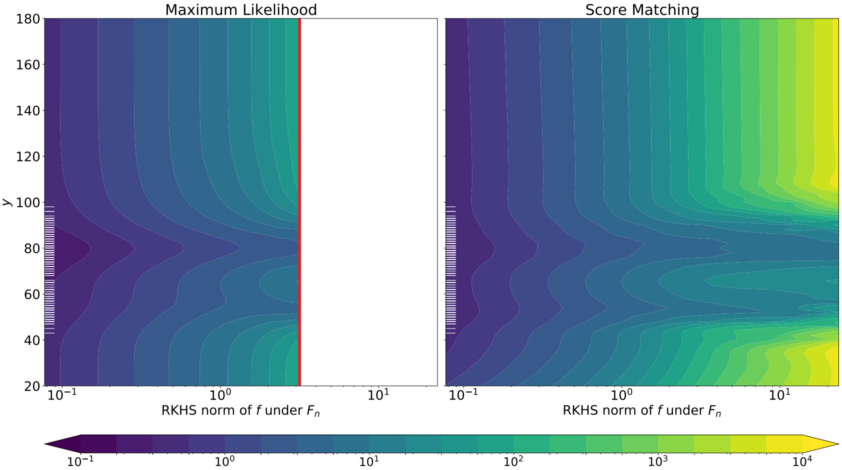


Figure: Heat maps of overall influence of y on T_λ and S_ρ vs. y and RKHS norm of the natural parameter under F_n. Red vertical line in left panel indicates the case $\lambda = 0$.

Penalized SM density estimator is more sensitive to y

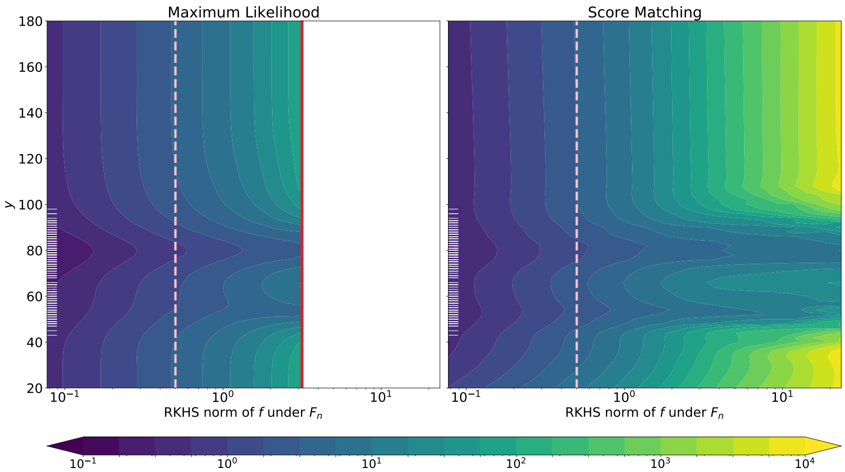


Figure: Heat maps of overall influence of y on T_λ and S_ρ vs. y and RKHS norm of the natural parameter under F_n. Red vertical line in left panel indicates the case $\lambda = 0$.

Penalized SM density estimator is more sensitive to y

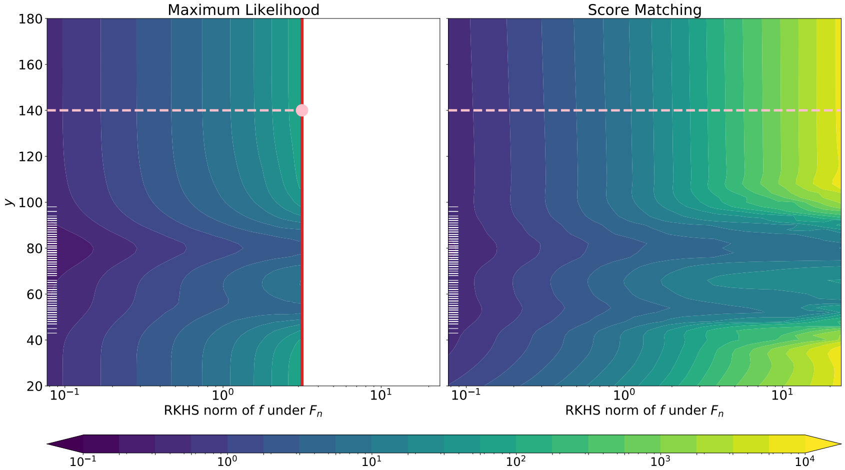


Figure: Heat maps of overall influence of y on T_λ and S_ρ vs. y and RKHS norm of the natural parameter under F_n. Red vertical line in left panel indicates the case $\lambda = 0$.

Summary

Regularized SM density estimators

- are easy to compute (matrix calculation), but
- are very sensitive to the presence of an isolated observation, especially when there is a small amount of regularization.

Penalized ML density estimator

- is hard to compute (mainly because we need to work with A and its derivatives), but
- is not sensitive to the presence of an isolated observation, even when no penalty is imposed.

Use regularized SM density estimators
with an appropriate amount of regularization!

Outline

Sensitivity of regularized SM density estimators in higher dimensions

All numerical examples provided here are restricted to $d = 1$.

Conjecture: Regularized SM density estimators in higher dimensions ($d \geq 2$) are also very sensitive to isolated observations when the regularization is small.

Need numerical examples to confirm this.

Sensitivity issues when generalized SM loss functional is used

Several generalized SM loss functionals have been proposed (Parry, Dawid and Lauritzen, 2012; Yu, Drton and Shojaie, 2020).

Interesting to

- apply these generalized SM loss functionals to density estimation problem in $\mathcal{Q}_{\mathrm{ker}}$, and
- investigate the sensitivity issue of the resulting density estimators.

Questions?

Bibliography

Canu, Stéphane and Alex Smola. “Kernel methods and the exponential family”. *Neurocomputing* 69.7 (Mar. 2006): 714–720. Print.

Fukumizu, Kenji. “Exponential manifold by reproducing kernel Hilbert spaces”. *Algebraic and Geometric methods in statistics* (2009): 291–306. Print.

Gu, Chong. “Smoothing Spline Density Estimation: A Dimensionless Automatic Algorithm”. *J. Am. Stat. Assoc.* 88.422 (June 1993): 495–504. Print.

Gu, Chong and Chunfu Qiu. “Smoothing Spline Density Estimation: Theory”. *Ann. Stat.* 21.1 (1993): 217–234. Print.

Hampel, Frank R. “Contributions to the theory of robust estimation”. University of California, 1968. Print.

Hyvärinen, Aapo. “Estimation of Non-Normalized Statistical Models by Score Matching”. *J. Mach. Learn. Res.* 6.Apr (2005): 695–709. Print.

Parry, Matthew, A Philip Dawid and Steffen Lauritzen. “Proper local scoring rules”. *Ann. Stat.* 40.1 (Feb. 2012): 561–592. Print.

Sriperumbudur, Bharath, et al. “Density Estimation in Infinite Dimensional Exponential Families”. *Journal of Machine Learning Research* 18.57 (2017): 1–59. Web. <http://jmlr.org/papers/v18/16-011.html>.

Yu, Shiqing, Mathias Drton and Ali Shojaie. “Generalized Score Matching for General Domains”. arXiv: 2009.11428 [stat.ME] (Sept. 2020). arXiv: 2009.11428 [stat.ME]. arXiv: 2009.11428 [stat.ME].

Genesis of $\widehat{L}_{\mathrm{SM}}$

$\widehat{L}_{\mathrm{SM}}$ comes from the *Hyvärinen divergence* (Hyvärinen, 2005)

$$\mathrm{H}(p_0 \| q) := \frac{1}{2} \int_{\mathcal{X}} p_0(x) \left\| \nabla \log p_0(x) - \nabla \log q(x) \right\|_2^2 \mathrm{d}x. \tag{17}$$

Under certain regularity conditions, using integration by parts,

$$\mathrm{H}(p_0 \| q) = \sum_{u=1}^{d} \int_{\mathcal{X}} p_0(x) \left[\frac{1}{2} \left(\partial_u \log q(x) \right)^2 + \partial_u^2 \log q(x) \right] \mathrm{d}x + \mathrm{const}. \tag{18}$$

$\widehat{L}_{\mathrm{SM}}$ is the empirical counterpart of $\mathrm{H}(p_0 \| q)$ with const omitted.

$\mathrm{IF}_x(T, F, y)$ and $\mathrm{SIF}_{x,\varepsilon}(T, F, y)$ do not depend on $\mu(x)$

Let G be a distribution function over $\mathcal{X}$.

Suppose $T(G) = \log q_{f_G}$, where $q_{f_G} \in \mathcal{Q}_{\mathrm{ker}}$ for some $f_G \in \mathcal{H}$. Then,

$$\begin{aligned} & T((1-\varepsilon)G + \varepsilon\delta_y)(x) - T(G)(x) \\ = & \left[\cancel{\log \mu(x)} + f_{(1-\varepsilon)G+\varepsilon\delta_y}(x) - A(f_{(1-\varepsilon)G+\varepsilon\delta_y})\right] \\ & \quad - \left[\cancel{\log \mu(x)} + f_G(x) - A(f_G)\right] \\ = & \left[f_{(1-\varepsilon)G+\varepsilon\delta_y}(x) - A(f_{(1-\varepsilon)G+\varepsilon\delta_y})\right] - \left[f_G(x) - A(f_G)\right]. \end{aligned}$$

Hence, $\mathrm{IF}_x(T, F, y)$ and $\mathrm{SIF}_{x,\varepsilon}(T, F, y)$ do not depend on $\mu(x)$, but only on natural parameter part.