05 Density Estimation

1 Intro & motivation

Here we take a slight detour from regression estimation to consider the problem of estimating the whole shape of a distribution via its density. As with regression, density estimation is a crucial goal in machine learning and nonparametric statistics in its own right, or it can be used for downstream tasks, such as clustering, anomaly detection, classification, causal inference, and more.

Specifically, here we observe an iid sample $(Y_1, ..., Y_n) \sim \mathbb{P}$ and the goal is to estimate the density function $p(y)$ satisfying $\mathbb{P}(Y \in \mathcal{B}) = \int_{\mathcal{B}} p(y)\, dy$ for any measurable $\mathcal{B}$. As with regression, there are numerous possible measures of risk one could use for an estimator $\widehat{p}(y)$, including pointwise MSE or mean absolute error

$$\mathbb{E}\Big[\{\widehat{p}(y) - p(y)\}^2\Big] \quad \text{or} \quad \mathbb{E}|\widehat{p}(y) - p(y)|$$

expected L_2 error (or integrated MSE)

$$\mathbb{E}\|\widehat{p} - p\|^2 = \mathbb{E}\int \Big\{\widehat{p}(y) - p(y)\Big\}^2\, dy$$

expected L_∞ error $\mathbb{E}\|\widehat{p} - p\|_\infty$ and so on.

2 Kernels

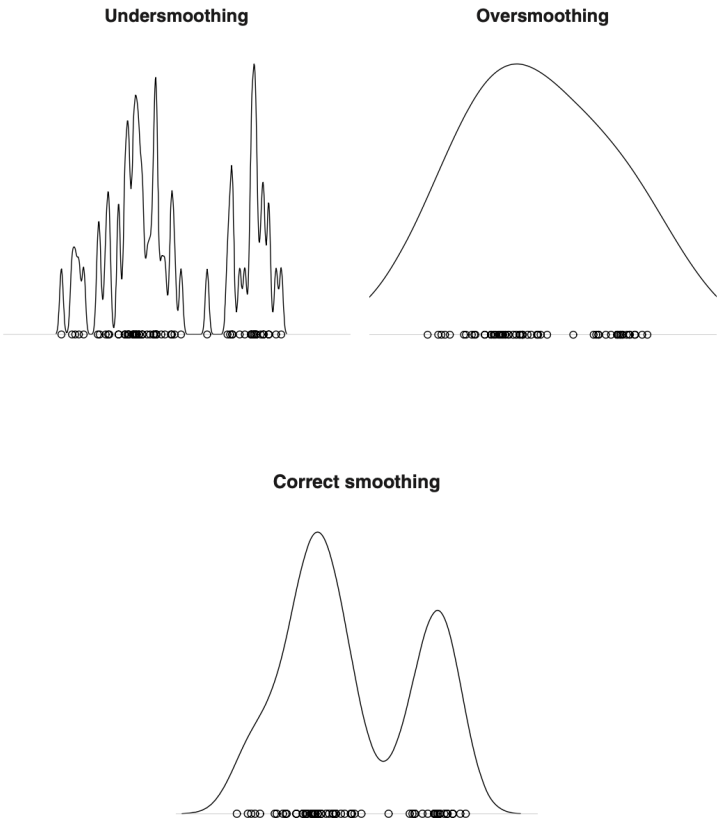
The classic kernel density estimator is given by

$$\widehat{p}(y) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right).$$

This is like a smooth, running average version of a histogram. In the case of the boxcar kernel it can also be viewed an empirical estimate of the approximate derivative, i.e.,

$$\widehat{p}(y) = \frac{\mathbb{P}_n(Y \le y + h/2) - \mathbb{P}_n(Y \le y - h/2)}{h}.$$

As with kernel and local polynomial estimators, the bandwidth parameter h controls the bias/variance trade-off. This is illustrated in the figure below.



Before detailing bias and variance bounds it is useful to mention the concept of a higher-order kernel. A kernel of order $\ell \geq 1$ satisfies

$$\int u^k K(u)\ du = 0$$

for $k = 1, ..., \ell$. Section 1.2.2 of Tsybakov [2009] details how one can construct bounded kernels of arbitrary order. For example the function

$$K(u) = \sum_{j=0}^{\ell} b_j(0) b_j(u) \mathbb{1}(|u| \leq 1)$$

is a one-dimensional kernel of order ℓ, when $b_j(u)$ is the j^{th} element of the Legendre polynomials orthonormal on $[-1, 1]$. Note since $u^k = \sum_{j=0}^{k} \alpha_j b_j(u)$ is a polynomial of degree k,

$$\int u^k K(u) = \sum_{m=0}^{k} \alpha_m \sum_{j=0}^{\ell} b_j(0) \int b_m(u) b_j(u) \mathbb{1}(|u| \leq 1)\ du = \sum_{j=0}^{k} \alpha_j b_j(0) = \mathbb{1}(k = 0)$$

Sometimes people complain that, because higher-order kernels can take negative values, the resulting estimator $\widehat{p}(y)$ can potentially take negative values; however one can always replace with the positive part $\widetilde{p}(y) = \max\{0, \widehat{p}(y)\}$ and the risk will never worsen.

The following result gives bounds on the pointwise bias and variance when the underlying density p is Holder smooth, under some conditions.

Theorem 1. *Assume $p(y)$ is s-smooth (with Lipschitz constant L) and:*

1. *a.* $\int K(u)\ du = 1$,
   b. $\int u^j K(u)\ du = 0$ *for* $|j| = 1, ..., \lfloor s \rfloor$,
   c. $\int \|u\|^s |K(u)|\ du \leq B < \infty$,
   d. $\int K(u)^2\ du \leq C < \infty$,
2. $p(y) \leq p_{\max} < \infty$.

Then under 1a–1c

$$|\mathbb{E}\{\widehat{p}(y) - p(y)\}| \leq h^s \left(LB \sum_{|k|=\lfloor s \rfloor} 1/k! \right)$$

and under 1d and 2

$$var\{\widehat{p}(y)\} \leq \frac{p_{\max} C}{nh^d}.$$

Proof. For the bias we have

$$\begin{aligned}
&\frac{1}{h^d} \int K\left(\frac{t-y}{h}\right) p(t)\ dt - p(y) = \int K(u) \Big\{ p(y+hu) - p(y) \Big\}\ du \\
&\quad = \sum_{0<|j|\leq \lfloor s \rfloor - 1} \frac{D^j p(y)}{j!} \int (hu)^j K(u)\ du \\
&\qquad + \sum_{|k|=\lfloor s \rfloor} \frac{\lfloor s \rfloor}{k!} \int \int_0^1 (1-t)^{\lfloor s \rfloor - 1} \Big\{ D^k p(y+thu) - D^k p(y) \Big\}\ dt\ (hu)^k K(u)\ du \\
&\quad \leq C' \int L(h\|u\|)^{s - \lfloor s \rfloor} h^{\lfloor s \rfloor} \|u\|^{\lfloor s \rfloor} K(u)\ du = C' L h^s \int \|u\|^s K(u)\ du
\end{aligned}$$

where the first line follows from Assumption 1a, the second Taylor expansion and 1b, and the third the Holder assumption with $\int_0^1 (1-t)^{\lfloor s \rfloor - 1}\ dt = 1/\lfloor s \rfloor$ and $|u^k| \leq \|u\|^{|k|}$ and $C' = \sum_{|k|=\lfloor s \rfloor} 1/k!$. For the variance

$$\begin{aligned}
\text{var}\{\widehat{p}(y)\} &\leq \frac{1}{nh^{2d}} \int K\left(\frac{t-y}{h}\right)^2 p(t)\ dt \\
&= \frac{1}{nh^d} \int K(u)^2 p(y+hu)\ du \leq \frac{p_{\max} C}{nh^d}
\end{aligned}$$

using Assumptions 1d and 2. □

In the above we need to be a bit careful about whether we require the support of p be unbounded; if not y should be an interior point, otherwise some extra work is needed to obtain small enough bias. Common approaches include boundary corrected kernels or reflection techniques.

3 Series

One can also construct density estimators via a global modelling or series approach. Let $b(y) = \{b_1(y), ..., b_k(y)\}^{\mathrm{T}}$ denote a basis, e.g., the polynomial, trigonometric, spline, or other bases discussed previously in a regression context. Then the series or projection density estimator is given by

$$\widehat{p}(y) = b(y)^{\mathrm{T}}\mathbb{P}_n\{b(Y)\} \equiv b(y)^{\mathrm{T}}\widehat{\beta}.$$

As in regression, here the tuning parameter is not a bandwidth but instead the number of features k in the basis.

The following result bounds the L_2 error, analogously to our bound for series estimators in regression.

Theorem 2. *Assume $p(y)$ is s-smooth and:*

1. *$p(y)$ has support $\mathcal{Y}$ and $\sup_y p(y) \leq p_{\max} < \infty$,*
2. *a. the basis b is orthonormal on $\mathcal{Y}$,*
   b. $\sup_{f \in \mathcal{H}(s)} \inf_\theta \|b^{\mathrm{T}}\theta - f\| \leq Ck^{-s/d}$.

Then

$$\mathbb{E}\|\widehat{p} - p\| \leq Ck^{-s/d} + \sqrt{\frac{p_{\max}k}{n}}.$$

Proof. Define β as the population projection

$$\beta = \underset{\theta \in \mathbb{R}^k}{\arg\min} \int \left\{p(y) - b(y)^{\mathrm{T}}\theta\right\}^2 dy = \int b(y)p(y)\ dy$$

and note that $\mathbb{E}(\widehat{\beta}) = \beta$. Now we have the decomposition

$$\|\widehat{p} - p\| \leq \|b^{\mathrm{T}}(\widehat{\beta} - \beta)\| + \|b^{\mathrm{T}}\beta - p\|$$

where the second term is approximation error / bias, and is of order $k^{-s/d}$ by Assumption 2b. The first term equals the square root of

$$\begin{aligned}\int \left\{b(y)^{\mathrm{T}}(\widehat{\beta} - \beta)\right\}^2 dy &= (\widehat{\beta} - \beta)^{\mathrm{T}} \int b(y)b(y)^{\mathrm{T}}\ dy(\widehat{\beta} - \beta) \\ &= \|\widehat{\beta} - \beta\|^2\end{aligned}$$

by the orthonormality, and it has squared expectation

$$\begin{aligned}(\mathbb{E}\|\widehat{\beta} - \beta\|)^2 \leq \mathbb{E}\|\widehat{\beta} - \beta\|^2 &= \sum_{j=1}^k \mathbb{E}\left\{(\widehat{\beta}_j - \beta_j)^2\right\} \\ &= \sum_{j=1}^k \mathrm{var}(\widehat{\beta}_j) \leq \frac{1}{n}\sum_{j=1}^k \int b_j(y)^2 p(y)\ dy \leq \frac{p_{\max}k}{n}\end{aligned}$$

again using the orthonormality from 2a along with Assumption 1. □

4 Overall rates

Note the above pointwise and L_2 rates for density estimation match those for regression based on local polynomials and series estimators. Similar balancing of squared bias and variance leads to the optimal tuning parameter choices

$$h \sim n^{-1/(2s+d)} \quad \text{and} \quad k \sim n^{d/(2s+d)}.$$

This in turn yields the overall rate

$$n^{-1/(2+d/s)}$$

which is minimax optimal for Holder smooth functions in density estimation, as well as regression. (We will come back to derive these kinds of minimax rates later).

5 Conditional density estimation

Often we want to estimate a conditional density $p(y \mid x)$, i.e., go beyond regression to learn about the whole distribution of an outcome Y given features X, not just the mean.

It turns out there is a simple heuristic for converting regression estimators to conditional density estimators: simply regress the kernel-transformed outcome

$$\frac{1}{\ell} K\left(\frac{Y-y}{\ell}\right)$$

on covariates. This works since if $p(y \mid x)$ is α-smooth in y then

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{\ell} K\left(\frac{Y-y}{\ell}\right) \,\middle|\, X = x\right\} &= \int \frac{1}{\ell} K\left(\frac{t-y}{\ell}\right) p(t \mid x)\, du = \int K(u) p(y + \ell u \mid x)\, du \\
&= p(y \mid x) + \sum_{0<j\le\lfloor\alpha\rfloor-1} \frac{\partial_y^j p(y \mid x)}{j!} \int (\ell u)^j K(u)\, du \\
&\quad + \int \left\{\partial_y^{\lfloor\alpha\rfloor} p(y + t\ell u \mid x) - \partial_y^{\lfloor\alpha\rfloor} p(y \mid x)\right\} \frac{(\ell u)^{\lfloor\alpha\rfloor}}{\lfloor\alpha\rfloor!} K(u)\, du
\end{aligned}$$

and so the kernel-transformed outcome satisfies

$$\left|\mathbb{E}\left\{\frac{1}{\ell} K\left(\frac{Y-y}{\ell}\right) \,\middle|\, X = x\right\} - p(y \mid x)\right| \le \ell^\alpha \left(\frac{C}{\lfloor\alpha\rfloor!}\right) \int |u|^\alpha |K(u)|\, du$$

and therefore has conditional expectation within ℓ^α of the desired conditional density.

In what follows we show how one can plug these kernel-transformed outcome variables $K\{(Y-y)/\ell\}/\ell$ into standard regression procedures discussed earlier to obtain optimal conditional density estimators.

5.1 Local polynomial estimator

In particular, here we consider plugging a kernel-transformed outcome into the local polynomial estimator, i.e.,

$$\widehat{p}(y \mid x) = \sum_{i=1}^{n} w_i(x; X^n) \frac{1}{\ell} K^* \left(\frac{Y_i - y}{\ell} \right)$$

with weights

$$w_i(x; X^n) = b(x)^{\mathrm{T}} \mathbb{P}_n \left\{ b(X) \frac{1}{h^d} K \left(\frac{X - x}{h} \right) b(X)^{\mathrm{T}} \right\}^{-1} b(X_i) \frac{1}{nh^d} K \left(\frac{X_i - x}{h} \right)$$

for b a polynomial basis including all interactions up to order $\lfloor s \rfloor$. Note here we have two bandwidth parameters, ℓ and h; the former controls window width around y and the latter around x.

Theorem 3. *Define*

$$\widehat{Q}_{hx} \equiv \left\| \mathbb{P}_n \left\{ b(X) \frac{1}{h^d} K \left(\frac{X - x}{h} \right) b(X)^{\mathrm{T}} \right\} \right\| \quad , \quad \xi_n \equiv \mathbb{P}_n \left(\|X - x\| \leq h \right) / h^d,$$

assume $p(y \mid x)$ is α-smooth in y and β-smooth in x, and:

1. $|K(u)| \leq K_{\max} \mathbb{1}(\|u\| \leq 1)$,
2. $\sup_x \|b(x)\| \leq B$,
3. a. $\int K^*(u)\, du = 1$,
   b. $\int u^j K^*(u)\, du = 0$ *for* $|j| = 1, ..., \lfloor \alpha \rfloor$,
   c. $\int \|u\|^\alpha |K^*(u)|\, du \leq B < \infty$,
   d. $\int K^*(u)^2\, du \leq C' < \infty$,
4. $p(y \mid x) \leq p_{\max} < \infty$.

Then:

$$\mathbb{E} \left\{ |\widehat{p}(y \mid x) - p(y \mid x)| \;\middle|\; X^n \right\} \lesssim \left(\ell^\alpha + h^\beta \right) \lambda_n \xi_n + \sqrt{\frac{\lambda_n^2 \xi_n}{nh^d \ell}}.$$

Proof. First note we can write

$$\begin{aligned}
\mathbb{E}\{\widehat{p}(y \mid x) \mid X^n\} - p(y \mid x) &= \sum_{i=1}^{n} w_i(x; X^n) \int \frac{1}{\ell} K \left(\frac{t - y}{\ell} \right) p(t \mid X_i)\, dt - p(y \mid x) \\
&= \sum_{i=1}^{n} w_i(x; X^n) \int K(u) \Big\{ p(y + \ell u \mid X_i) - p(y \mid X_i) \Big\}\, du \\
&\quad + \sum_{i=1}^{n} w_i(x; X^n) \Big\{ p(y \mid X_i) - p(y \mid x) \Big\}
\end{aligned}$$

The first term is handled as in marginal density estimation, while the second is handled as in standard regression.

The first term equals

$$\begin{aligned}
&\sum_{i=1}^n w_i(x;X^n)\Bigg[\sum_{0<j\le\lfloor\alpha\rfloor-1}\frac{\partial_y^j p(y\mid X_i)}{j!}\int(\ell u)^j K(u)\,du \\
&\qquad\qquad+\int\Big\{\partial_y^{\lfloor\alpha\rfloor}p(y+t\ell u\mid X_i)-\partial_y^{\lfloor\alpha\rfloor}p(y\mid X_i)\Big\}\frac{(\ell u)^{\lfloor\alpha\rfloor}}{\lfloor\alpha\rfloor!}K(u)\,du\Bigg] \\
&=\sum_{i=1}^n w_i(x;X^n)\int\Big\{\partial_y^{\lfloor\alpha\rfloor}p(y+t\ell u\mid X_i)-\partial_y^{\lfloor\alpha\rfloor}p(y\mid X_i)\Big\}\frac{(\ell u)^{\lfloor\alpha\rfloor}}{\lfloor\alpha\rfloor!}K(u)\,du \\
&\le\sum_{i=1}^n |w_i(x;X^n)|\int L(\ell u)^{\alpha-\lfloor\alpha\rfloor}\frac{(\ell u)^{\lfloor\alpha\rfloor}}{\lfloor\alpha\rfloor!}|K(u)|\,du \\
&\le\ell^\alpha\Big(L\lambda_n\xi_n\Big)\int|u|^\alpha|K(u)|\,du/\lfloor\alpha\rfloor!
\end{aligned}$$

The second term satisfies

$$\sum_{i=1}^n w_i(x;X^n)\Big\{p(y\mid X_i)-p(y\mid x)\Big\}\lesssim h^\beta\lambda_n\xi_n$$

For the variance we have

$$\begin{aligned}
\operatorname{var}\left\{\sum_{i=1}^n w_i(x;X^n)\frac{1}{\ell}K\left(\frac{Y_i-y}{\ell}\right)\,\Big|\,X^n\right\}&\le\sum_{i=1}^n w_i(x;X^n)^2\frac{1}{\ell^{2d}}\int K\left(\frac{t-y}{h}\right)^2 p(t\mid X_i)\,dt \\
&\le\frac{p_{\max}}{\ell}\left(\int K(u)^2du\right)\max_i|w_i(x;X^n)|\sum_{i=1}^n|w_i(x;X^n)| \\
&\lesssim\frac{p_{\max}}{\ell}\left(\int K(u)^2du\right)\frac{\lambda_n^2\xi_n}{nh^d}
\end{aligned}$$

which yields the result. □

Balancing $h^{2\beta}$ with $1/nh^d\ell$ gives

$$h^{2\beta}\sim\frac{1}{nh^d\ell}\implies h^{2\beta+d}\sim\frac{1}{n\ell}\implies h\sim\left(\frac{1}{n\ell}\right)^{1/(2\beta+d)}$$

Now balancing $h^{2\beta}$ with $\ell^{2\alpha}$ gives

$$h^{2\beta}\sim\left(\frac{1}{n\ell}\right)^{2\beta/(2\beta+d)}\sim\ell^{2\alpha}\implies\ell^{(2\alpha(2\beta+d)+2\beta)/(2\beta+d)}\sim n^{-2\beta/(2\beta+d)}\implies\ell\sim n^{-2\beta/(2\alpha(2\beta+d)+2\beta)}$$

which yields

$$\ell^\alpha\sim n^{-2\alpha\beta/(2\alpha(2\beta+d)+2\beta)}\sim n^{-1/(2+1/\alpha+\beta/d)}.$$

Therefore the rate of convergence for conditional density estimation is

$$n^{\frac{-1}{2+1/\alpha+d/\beta}}.$$

For example, if $\alpha\to\infty$ then this rate tends to that of estimating a β-smooth d-dimensional regression. Otherwise the rate is slower due to having to estimate the conditional density of Y, beyond the mean.

Efromovich [2007] shows that the rate $n^{-1/(2+1/\alpha+d/\beta)}$ is actually minimax optimal for a conditional density that is α-smooth in the outcome and β-smooth in the features.

5.2 Series estimator

Theorem 4. *Assume $p(y \mid x)$ is α-smooth in y and β-smooth in x, and:*

1. $\sup_{f \in \mathcal{H}(\beta)} \inf_\theta \|b^{\mathrm{T}}\theta - f\| \lesssim k^{-\beta/d}$,
2. *the eigenvalues of* $Q \equiv \mathbb{E}\{b(X)b(X)^{\mathrm{T}}\}$ *are bounded above and below away from zero,*
3. *a.* $\int K^*(u)\, du = 1$,
   b. $\int u^j K^*(u)\, du = 0$ *for* $|j| = 1, ..., \lfloor \alpha \rfloor$,
   c. $\int \|u\|^\alpha |K^*(u)|\, du \leq B < \infty$,
   d. $\int K^*(u)^2\, du \leq C' < \infty$,
4. $p(y \mid x) \leq p_{\max} < \infty$.

Then as long as $\sup_x \|b(x)\|^2 \log k/n \to 0$ *the estimator* $\widehat{p}(y \mid x)$ *satisfies*

$$\|\widehat{p} - p\| = O_{\mathbb{P}}\left(k^{-\beta/d} + \ell^\alpha + \sqrt{\frac{k}{n\ell}}\right).$$

Proof. Define $\widetilde{p}(y \mid x) = b(x)^{\mathrm{T}}\beta_h$ where

$$\beta_h = \arg\min_\theta \mathbb{E}\left[\left\{\frac{1}{h^d}K\left(\frac{Y-y}{h}\right) - b(X)^{\mathrm{T}}\theta\right\}^2\right] = \Big[\mathbb{E}\{b(X)b(X)^{\mathrm{T}}\}\Big]^{-1}\mathbb{E}\left\{b(X)\frac{1}{h^d}K\left(\frac{Y-y}{h}\right)\right\}$$

as the linear projection of the kernel-transformed outcome, and define $r(y \mid x) = b(x)^{\mathrm{T}}\beta$ where

$$\beta = \arg\min_\theta \mathbb{E}\left[\left\{\frac{1}{h^d}K\left(\frac{Y-y}{h}\right) - b(X)^{\mathrm{T}}\theta\right\}^2\right] = \Big[\mathbb{E}\{b(X)b(X)^{\mathrm{T}}\}\Big]^{-1}\mathbb{E}\Big\{b(X)p(y \mid X)\Big\}$$

as the linear projection of the true conditional density. Now we have the decomposition

$$\|\widehat{p} - p\| \leq \|\widehat{p} - \widetilde{p}\| + \|\widetilde{p} - r\| + \|r - p\|.$$

Intuitively, the first term is estimation error akin to that in our original series analysis, the second term is approximation error from the kernel smoothing of the outcome, and the third term is approximation error from approximating the conditional density as a function of features with a k-dimensional basis (akin to the first approximation error term in our original series analysis). The third term is of order $k^{-\beta/d}$ since $p(y \mid x)$ is β-smooth in x and by the assumed approximation properties of the basis. The second term is of order ℓ^α since $p(y \mid x)$ is α-smooth in y and by the fact that $\sup_x |\mathbb{E}\{K((Y-y)/\ell)/\ell \mid X = x\} - p(y \mid x)| \lesssim \ell^\alpha$. For the first term we can replicate the original series analysis to show that it is of order $\frac{k}{n\ell} + k^{-\beta/d}$. □

Balancing ℓ^α with $k^{-\beta/d}$ gives

$$\ell \sim k^{-\beta/d\alpha}$$

and balancing $k^{-2\beta/d}$ with $k/n\ell$ gives

$$k^{-2\beta/d} \sim \frac{k^{1+\beta/d\alpha}}{n} \implies k^{(d+\beta/\alpha+2\beta)/d} \sim n \implies k \sim n^{d/(d+\beta/\alpha+2\beta)}$$

which implies

$$k^{-\beta/d} \sim n^{-\beta/(d+\beta/\alpha+2\beta)} = n^{-1/(2+1/\alpha+d/\beta)}$$

which again is the minimax optimal rate.

References

S. Efromovich. Conditional density estimation in a regression setting. *The Annals of Statistics*, 35:2504–2535, 2007.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.